Quantification et caractérisation de l'incertitude de segmentation d'images médicales pardes réseaux profonds

Benjamin Lambert

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : ISCE - Ingénierie pour la Santé la Cognition et l'Environnement
Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement
Unité de recherche : Grenoble Institut des Neurosciences

Quantification et caractérisation de l'incertitude de segmentation d'images médicales par des réseaux profonds

Quantifying and understanding uncertainty in deep-learning-based medical image segmentation

Présentée par :

Benjamin LAMBERT

Direction de thèse :

Michel DOJAT DIRECTEUR DE RECHERCHE, INSERM DELEGATION AUVERGNE-RHONE-ALPES	Directeur de thèse

Rapporteurs :

MARIA ZULUAGA
ASSOCIATE PROFESSOR, EURECOM
NINON BURGOS
CHARGEE DE RECHERCHE HDR, CNRS DELEGATION PARIS CENTRE

Thèse soutenue publiquement le **16 mai 2024**, devant le jury composé de :

OLIVIER FRANÇOIS, PROFESSEUR DES UNIVERSITES, GRENOBLE INP	Président
MICHEL DOJAT, DIRECTEUR DE RECHERCHE, INSERM DELEGATION AUVERGNE-RHONE-ALPES	Directeur de thèse
MARIA ZULUAGA, ASSOCIATE PROFESSOR, EURECOM	Rapporteure
NINON BURGOS, CHARGEE DE RECHERCHE HDR, CNRS DELEGATION PARIS CENTRE	Rapporteure
PIERRICK COUPE, DIRECTEUR DE RECHERCHE, CNRS DELEGATION AQUITAINE	Examinateur
JULIEN MAIRAL, DIRECTEUR DE RECHERCHE, CENTRE INRIA UNIVERSITE GRENOBLE ALPES	Examinateur
CHRISTIAN BAUMGARTNER, DOCTEUR EN SCIENCES, Eberhard Karls Universität Tübingen	Examinateur

Invités :

SENAN DOYLE
DOCTEUR EN SCIENCES HDR, PIXYL
FRANCESCA GALASSI
MAITRESSE DE CONFERENCES, ECOLE SUPERIEURE D'INGENIEURS DE RENNES (ESIR)

Acknowledgements

I would first like to express my deep gratitude to my advisors Michel and Florence. Thank you for your guidance, kindness, and trust which enabled me to grow as a researcher. It has truly been a pleasure to carry this project by your side. Senan, Julien and Alan, thank you for the trust you have placed in me throughout this thesis. You put me in the best possible conditions to carry out my research and I am proud of what we were able to accomplish together at Pixyl. Emmanuel, Benjamin, and Thomas, thank you for welcoming me to your GIN's team and allowing me to contribute to the exciting research happening there.

A kind thank you to the rapporteurs and jury members who agreed to review this thesis. I look forward to meeting you and discussing our research with you.

I was lucky enough to be part of two wonderful teams. Many thanks to my colleagues from Pixyl and GIN Team 5 for your support and for the good times we had together. A special thanks to my colleagues from team $R\&D$: Pascal, Veronica, Pauline, and Harmonie. Working with you is a breeze, and I deeply value your feedback throughout the years.

To my fellow GIN's non-permanents: you are the best! Thank you all for creating such a special environment that makes the thesis journey so enjoyable, inside and outside the lab. A special thanks to Aurélien for always providing great advice, especially when it came to writing this manuscript.

To my friends in Grenoble, Jules, Amélie, Paul, Olivia, and Katell: I was very lucky to meet you in engineering school and I'm thrilled that we could continue our adventures in Grenoble together. You have been a very important part of my life in Grenoble and I thank you from the bottom of my heart for being there through the good times and the bad.

Raphaël and Léopoldine, I've learned a lot from your strength and wisdom. You inspired me to always give my best even in stormy conditions, and I am truly fortunate to have you both in my life.

Finally, I would like to thank my parents for their unconditional support and for always trying their best to try and understand what my job is all about. Mom, Dad, you always have been an example of commitment and caring. I dedicate this thesis to you!

Contents

List of Abbreviations

AI	Artificial Intelligence
AE	Autoencoder
AUPR	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BDL	Bayesian Deep Learning
BN	Batch Normalization
BNN	Bayesian Neural Network
CNN	Convolutional Neural Network
CP	Conformal Prediction
CT	Computed Tomography
DE	Deep Ensemble
DL	Deep Learning
DS	Domain Shift
ECE	Expected Calibration Error
EDL	Evidential Deep Learning
FLAIR	Fluid Attenuated Inversion Recovery
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
GIN	Graph Isomorphism Network
GNN	Graph Neural Network
HD	Haussdorff Distance
ID	In Distribution
MC	Monte Carlo
MAE	Mean Average Error
MSE	Mean Squared Error
ML	Machine Learning

MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
NN	Neural Network
OOD	Out-Of-Distribution
PI	Predictive Interval
PR	Precision Recall
ROC	Receiver Operating Characteristic
SD	Surface Dice
SRT	Stochastic Regularization Technique
SSIM	Structural Similarity Index Measure
T1ce	T1 weighted MRI with contrast enhancement
T1w	T1 weighted MRI
T2w	T2 weighted MRI
TTA	Test Time Augmentation
TN	True Negative
TP	True Positive
UQ	Uncertainty Quantification
UAD	Unsupervised Anomaly Detection
UEO	Uncertainty Error Overlap
WCP	Weighted Conformal Prediction
WMH	White Matter Hyperintensity

List of Figures

Notation

This section provides a reference describing notations used throughout this thesis. These notations are taken from the Deep Learning Book (Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning* MIT Press, 2016).

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix

Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$\mathcal{G}$	A graph

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
$A_{i,j}$	Element i, j of matrix $\boldsymbol{A}$
$\mathsf{A}_{i,j,k}$	Element (i, j, k) of a 3-D tensor A

Linear Algebra Operations

$\boldsymbol{A}^\top$	Transpose of matrix $\boldsymbol{A}$

Calculus

$\frac{dy}{dx}$	Derivative of y with respect to x
$\nabla_{\boldsymbol{x}} y$	Gradient of y with respect to $\boldsymbol{x}$

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$	A probability distribution over a continuous variable
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathbb{E}(f(x))$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$
$\mathrm{Var}(f(x))$	Variance of $f(x)$ under $P(\mathrm{x})$
$H(\mathrm{x})$	Shannon entropy of the random variable x
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Functions

$f : \mathbb{A} \to \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$.
$\|\|\boldsymbol{x}\|\|_p$	L^p norm of $\boldsymbol{x}$
$\mathbf{1}_{\text{condition}}$	is 1 if the condition is true, 0 otherwise

Datasets and Distributions

p_{data}	The data generating distribution
$\hat{p}_{\text{data}}$	The empirical distribution defined by the training set
$\mathbb{X}$	A set of training examples
$\boldsymbol{x}^{(i)}$	The i-th example (input) from a dataset
$y^{(i)}$ or $\boldsymbol{y}^{(i)}$	The target associated with $\boldsymbol{x}^{(i)}$ for supervised learning

General Introduction

Decision-making in healthcare is inherently uncertain. Pieces of evidence collected to establish a diagnosis can be limited, ambiguous, incomplete, or conflicting. In the presence of a rare disease, clinicians may be uncertain about the exact cause of the symptoms or disagree with other experts' opinions. Moreover, there are usually several plausible outcomes for a given patient that should be taken into account to select the adequate treatment. In radiology, image artifacts or poor resolution can result in uncertain conclusions, for instance regarding the presence or malignancy of a lesion. As an additional challenge, the lack of experience with complex pathologies as well as the accumulated fatigue over the course of a clinical shift can contribute to reduced precision, potentially leading to incomplete or erroneous diagnoses. It is thus commonly acknowledged that uncertainty is intrinsic to clinical practice [2], and medical training generally includes learning how to optimize decision-making based on the natural ambiguity and complexity of clinical scenarios [3].

In recent years, Machine Learning (ML) algorithms have achieved remarkable results in many tasks, including medical image segmentation and classification. In radiology, ML algorithms, mostly based on Deep Learning (DL) approaches, have the potential to considerably assist clinicians by automatizing time-consuming and error-prone tasks, such as the segmentation of the brain into hundreds of regions, or the detection of small subtle lesions in brain Magnetic Resonance Imaging (MRI). As an example, ML tools allow faster quantification of Multiple Sclerosis (MS) disease progression by providing a count of the new lesions in different brain regions, a measure of the total lesions volume, and a precise description of the shape of lesions. This analysis can be used as an additional source of knowledge for the neuroradiologist, allowing for improved clinical decision-making and patient care. Yet, contrary to clinicians who navigate daily with uncertainty, these algorithms generally produce predictions without any information concerning their confidence. As a result, these models are often referred to as "black boxes". This prevents the full adoption of AI algorithms in critical fields such as healthcare, as they tend to make mistakes without confidence estimates, which could have warned the user about model deficiency. These silent errors are particularly dangerous and can lead to erroneous conclusions. Identifying and understanding these failure cases is crucial to maximizing the utility of AI models, as well as their acceptance — in particular by healthcare professionals, and their integration into the medical information flow.

Uncertainty Quantification (UQ) in DL models is challenging. These complex models are composed of millions of parameters that cannot be easily interpreted. Moreover, training strategies involve learning statistical features from the data itself, without any human supervision regarding the choice of features. As a result, the learned decision rules are opaque. In medical applications, however, uncertainty has to be provided to the healthcare professional in an intelligible and useful way to assist decision-making. This requires defining levels of uncertainty that are meaningful to clinicians. In the context of volumetric medical image segmentation, the straightforward option is to quantify voxel uncertainty, which involves assigning a score to each voxel representing the confidence of the model regarding the predicted

label. However, it may not be the most relevant and meaningful option in a clinical context. For example in MS, clinicians may be interested in the confidence of the model at the lesion level to help determine if the identified lesion is not a false positive detection. They may also be interested in confidence intervals associated with high-level metrics derived from the segmentation, such as lesion volumes. Finally, overall quality scores could be envisaged for both the input and the output. For instance, an input quality control measure could warn the user if the input image does not meet established quality standards due to important artifacts, thus potentially impacting the output analysis. Thus, it appears that the limitation of UQ to the voxel level is not satisfying to fully quantify the ambiguity in DL-based medical image analysis.

Challenges and Contributions

This thesis aims to address the current identified limitations by designing uncertainty estimates that are relevant for clinicians. More particularly, we propose to investigate 4 different scales of uncertainty that are useful in automated medical image analysis. They are illustrated in Figure A, and a motivation for each is provided below:

- **Voxel-level.** DL segmentation models operate at the voxel level, and thus the application of standard UQ methodology yields voxel-level uncertainty estimates. In these maps, each voxel in the image volume is associated with a confidence estimate. These maps can be superimposed to the image or segmentation to identify uncertain areas. Interestingly, these voxel-level scores are efficient in identifying misclassified voxels, due for instance to partial volume effect in MRI or to model deficiencies. Many performing methods have been proposed to quantify voxel uncertainty. As a contribution, a benchmark of such methods is proposed with a particular focus on robustness to domain shift, which occurs when the test data originates from a different distribution than the training data. In MRI processing, domain shift primarily originates from variations in the image acquisition protocol or inconsistencies in image quality.
- **Lesion-level.** For pathologies such as MS, the segmentation of lesion voxels usually identifies several dozen individual brain lesions. For these diseases, the attention of the clinician is at the lesion level, and estimating the overall confidence of each unique lesion instance is critical. It will allow the user to directly review the most uncertain ones to validate or reject them if they estimate that the lesion is a false positive finding. This emerging level of UQ is still rarely considered in the literature. This thesis contributes by proposing three different means of computing lesion-wise uncertainty. To this end, we propose three lesion models that can predict the probability that the lesion is a false positive, providing interpretable lesion uncertainty estimates.
- **Subject-level.** Quality Control (QC) is a crucial step of medical image analysis, and uncertainty can play a major role in its automatization. QC can be implemented at two levels: the input image, and the output analysis. Regarding the first case, the rationale is that model confidence can be expected to be low for input images that have poor qualities (*e.g.* important artifacts or poor resolution). Monitoring the model confidence could thus be used to automatically detect poor-quality images. This input QC can be framed as an out-of-distribution detection problem. Alternatively, QC can be performed on the output segmentation to identify model failures at the subject level

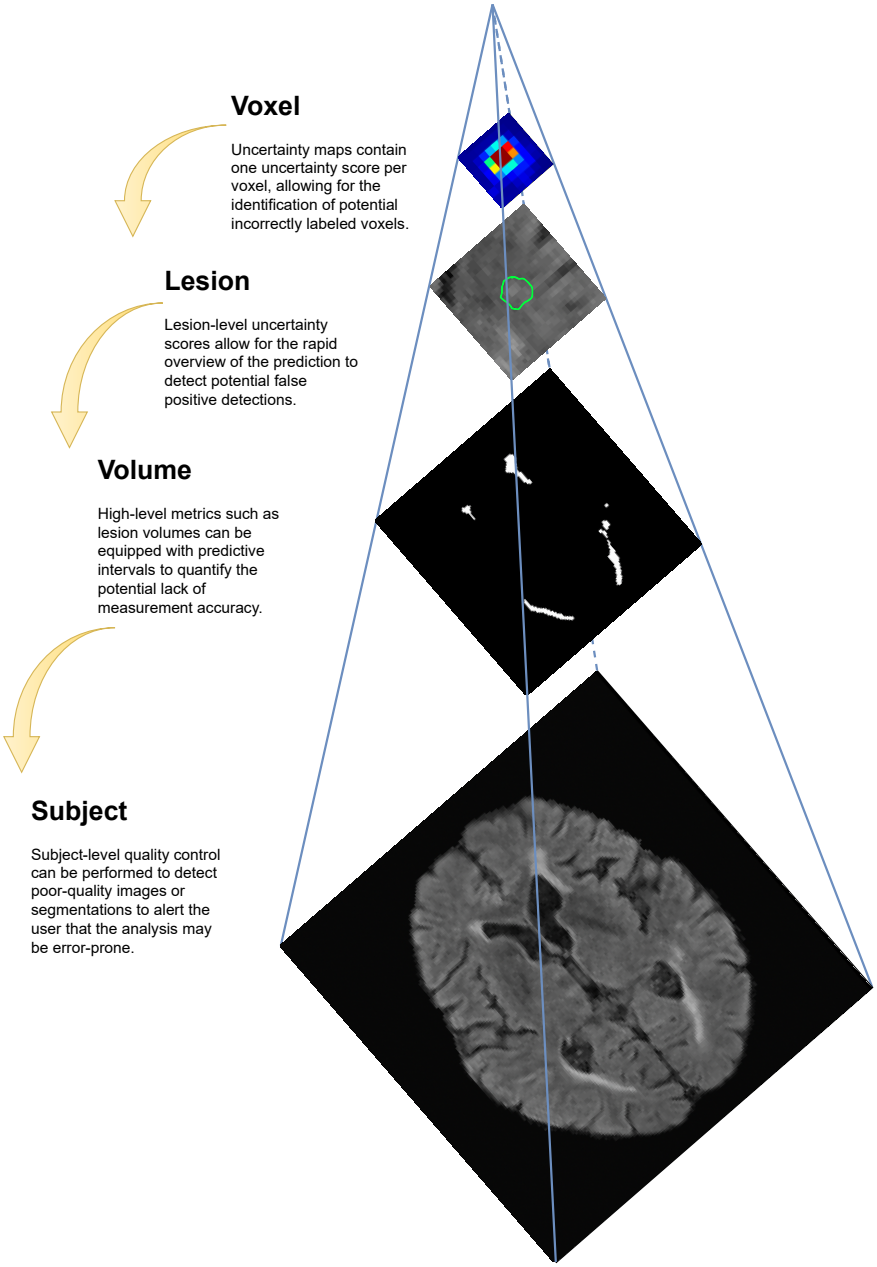


Figure A: Illustration of the different scales of uncertainty useful for automatic medical image analysis. From the lowest to the highest level: voxel, lesion, volume and subject-level estimates can be useful to estimate the confidence of the automated analysis.

Figure B: Pixyl logo. The company, founded in 2015, specializes in developing AI-based software for automated medical image analysis.

(*e.g.* erroneous segmentation). These two types of case-level uncertainties can be used to automatically warn the user in cases where the automated analysis is expected to be poor. However, they are generally tackled in completely different literature. We thus propose here as a contribution a unified framework that combines these two levels for an enriched automated QC procedure.

- **Predictive Intervals on Volumes.** Segmentation is usually the first step to a more advanced analysis pipeline, including the extraction of lesional volumes. Indeed, this information is a powerful biomarker to assess the extent and progression of the disease. However, this volume estimation is also prone to uncertainty, and it can be taken into account by associating predictive intervals with the estimation. To do so, we propose the first investigation of the Conformal Prediction framework for predictive intervals on volumes, using an efficient approach called TriadNet.

Thesis Context

Pixyl [1] (Figure B) is a company founded in 2015 and located in Grenoble, France, that specializes in the development of AI-based software dedicated to the automated analysis of MR images. More particularly, it commercializes Pixyl.Neuro, a software that automatically analyses brain MR images to support rapid detection, early diagnosis, and objective monitoring of neurological disorders. The solution, which is CE-marked (MDR class IIa) and FDA-cleared (class II), is composed of two distinct but complementary modules: Pixyl.Neuro.MS and Pixyl.Neuro.BV.

Pixyl.Neuro.BV, dedicated to brain volumetric analyses, delivers automatic neuroimaging biomarker extraction to assist the diagnosis, prognosis, and follow-up of patients with various neurodegenerative pathologies. It provides brain volume quantification using 3D T1 Gradient Echo MRI, allowing a better understanding of the pattern of atrophy with objective measurements and comparison with normative values. The Pixyl.Neuro.MS module, dedicated to neuroinflammatory disorders, automatically detects, quantifies, and categorizes white-matter hyperintensities on 3D T2-FLAIR MR images. Supporting patient follow-up (longitudinal analysis), the software provides information on individual lesion activity, highlighting even subtle changes between visits. In practice, it provides segmentation of white matter hyperintensities, with lesion load classified by relevant regions (McDonald regions: infratentorial, juxtacortical, periventricular, deep white matter). It may also provide information on the disease activity and change of individual white matter hyperintensities

[1] https://pixyl.ai/

since the previous visit, if available. This can be used to support the diagnosis, prognosis, and follow-up of subjects suffering from neuroinflammatory disorders (in particular, but not exclusively, Multiple Sclerosis). Pixyl is used today by over 100 centers in more than 12 countries across Europe, North America, and Africa.

Although such AI tools are becoming pivotal in clinical decision-making, with benefits for both the clinician and the patient, their deployment in the real clinical routine raises many challenges linked to uncertainty. For example, the deployed model can be confronted with data acquired with unusual imaging protocols, or the brain can present an unusual lesion that is not present in the training dataset. These scenarios could potentially undermine the performance of the AI model at hand. Being able to flag up these possibly problematical cases and warn the user is crucial to guaranteeing the reliability of the software. Importantly, an industrial requirement for computing uncertainty-related measures is speed and computational efficiency. Indeed, Pixyl's automated analyses should be sent to the user no longer than 5 minutes after receiving the input image, so that the clinician can examine the report while the patient is still in the examination room. Thus, we aim in this Ph.D to incorporate the different uncertainty quantification modules seamlessly within a Deep Learning-based analysis pipeline, as illustrated in Figure C.

This Ph.D. is a collaboration between Pixyl and two research teams, namely the Grenoble Institute of Neurosciences (GIN), Team **Functional Neuroimaging and Brain Perfusion** directed by Benjamin Lemasson and Thomas Christen on one side, and INRIA, Team **STATIFY**, directed by Florence Forbes on the other side. Team Lemasson-Christen specializes in the development of innovative MRI scan acquisition techniques and their application in preclinical and clinical neuroscience studies. Moreover, it develops mathematical models used to assist MRI analysis. Team STATIFY specializes in the development of innovative statistical models applied to complex and large-dimensional data. The thesis was funded by a CIFRE convention granted by the National Technology Research Association (ANRT 2020/1555).

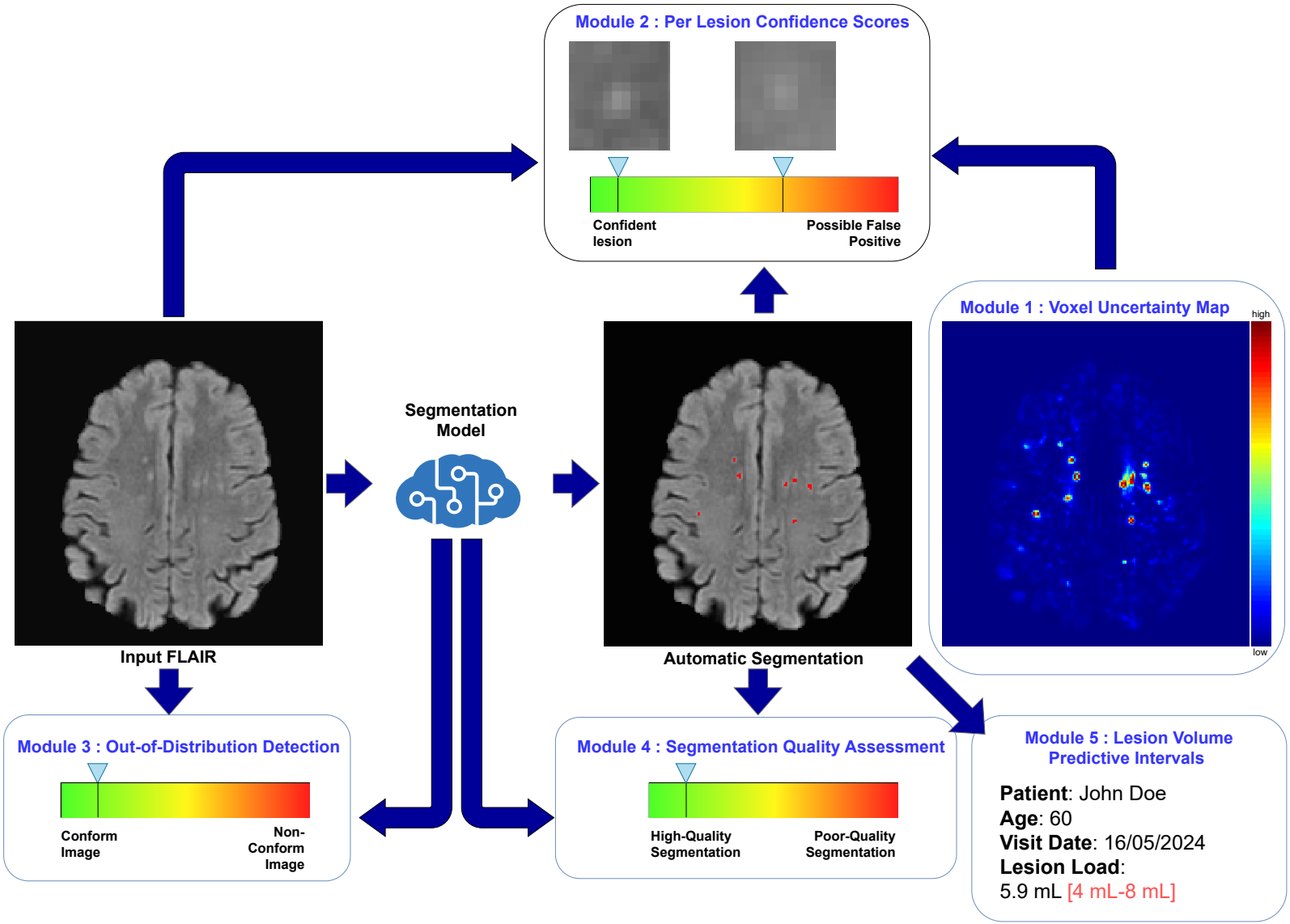


Figure C: Expected thesis framework, incorporating the different uncertainty quantification levels within a Deep Learning analysis pipeline.

Thesis Organization

The present dissertation is organized into 5 chapters.

Chapter I introduces the main concepts of DL networks. This helps to introduce their pitfalls in terms of explainability and uncertainty quantification. Popular architectures and training strategies in the context of medical image analysis are more specifically detailed.

Chapter II introduces the definitions and concepts of UQ in DL. The different types of uncertainty (aleatoric, epistemic) inherent to medical image analysis are introduced. Then, an in-depth literature review is proposed focusing on the applications of UQ for medical image classification and segmentation. A particular focus is given to ways of quantitatively evaluating the quality of uncertainty, a prerequisite for the development of useful UQ tools. Eventually, a benchmark of the prevailing UQ methods is performed on three brain lesion segmentation tasks to identify a baseline UQ method for the remainder of the manuscript.

Chapter III focuses on the emerging notion of structural uncertainty. Most UQ methods, when applied to 3D medical image segmentation, provide uncertainty estimates at the voxel level. Yet, for lesion segmentation tasks such as MS-lesion detection in brain MRI or lung-nodule detection in chest CT, the interest of the clinician is at the lesion level. For efficient review of the automated results, it is more appropriate to provide one uncertainty score per lesion, instead of one score for each voxel. This chapter introduces our contribution related to structural UQ, based on a model that builds graphs associated with each lesion, and leverages Graph Neural Networks (GNN) to merge voxel uncertainty estimates into lesion uncertainty.

Chapter IV enlarges the scope of uncertainty quantification to the case level. More particularly, as classical UQ approaches fail at this crucial detection task, the emerging technique of latent-space detection is explored in detail on a large and complete benchmark of out-of-distribution images. Second, this promising latent-space detector is combined with an output segmentation quality estimation strategy, resulting in a unified input-output QC protocol.

Chapter V is dedicated to the conformal prediction (CP) framework, which is becoming a prevailing UQ technique but is under-explored for medical-image processing. CP is a mathematical framework that can be employed to equip DL predictions with predictive intervals, guaranteed to contain the ground truth quantity with a user-defined level of confidence. It is thus particularly promising for medical applications where the accountability of algorithms is mandatory. In this thesis, we propose to use CP to compute predictive intervals associated with estimated lesion volumes. As a perspective, we propose a discussion on how a weighted formulation of CP can be used to tackle domain shift issues.

Associated Publications

In some parts, the studies presented in this thesis have given place to publications. The complete list of this published work is provided in Appendix A9. This thesis proposes further experiments and applications with respect to these publications. Additions are indicated at the beginning of each thesis chapter.

Chapter I

Deep Learning for medical image analysis

Deep Learning (DL) is one of the most promising technological innovations of this last decade. Contrary to past ML frameworks that required handcrafted features, DL learns features directly from the raw data itself, alleviating the need for formal specification of task-specific knowledge. It is now well established as the state-of-the-art approach in many challenging tasks, including image classification, segmentation, natural language processing, or regression. This chapter provides an overview of the principles of DL, including neural networks design, loss objectives, and optimization. These core steps are illustrated in Figure I.1.1. Current trends regarding architecture choices and training paradigms in the context of medical-image analysis are also briefly presented. This chapter lays the technical foundations necessary for the methodological developments presented in the rest of the thesis.

Contents

I.1 Neural networks

The building and training of Neural Networks (NNs) is the core of the DL framework. These models can learn a function f^* that maps an input $\boldsymbol{x}$ to a prediction $\hat{\boldsymbol{y}}$ by approximating a set of parameters $\boldsymbol{\theta}$ during training, which are learned to provide the optimal mapping $\hat{\boldsymbol{y}} \rightarrow f(\boldsymbol{x};\boldsymbol{\theta})$. For classification tasks, the input $\boldsymbol{x}$ is an image and the prediction $\hat{\boldsymbol{y}}$ is a label. For segmentation tasks, which are the primary focus of this thesis, the classification is performed for each pixel (or voxel in 3D), and the output is a label image with the same dimensions as the input.

NNs are traditionally composed of several layers with learnable weights, which are the network's high-level building blocks. Each layer i computes a function ${f^{(\mathrm{i})}}_{(i=1,\ldots,n)}$, meaning that the final approximated function f^* can be written as a chain of compound function $f^*(\boldsymbol{x}) = f^{(\mathrm{n})}...f^{(\mathrm{h})}...f^{(1)}(\boldsymbol{x})$. $f^{(1)}$ is the input layer receiving $\boldsymbol{x}$, $f^{(\mathrm{n})}$ the output layer yielding the prediction $\hat{\boldsymbol{y}}$ and ${f^{(\mathrm{h})}}_{(h=2,\ldots,n-1)}$ are referred as hidden layers (Figure I.1.2). In the standard supervised-learning scenario, pairs $\boldsymbol{(x;y)}$ of labeled training data are used to teach the NN the expected behavior of the output layer, given the properties of the input. The behavior of the other layers is not directly conditioned by the data, hence their *hidden* denomination. Interestingly, it has been shown that even a simple feedforward NN with a single hidden layer is a *universal function approximator*, meaning that for a given function, there exists a finite number of neurons for which the network will be able to approximate it with arbitrary accuracy [4].

The primary distinction of DL from other learning models (e.g. Random Forests, Logistic Regression, XGBoost) is that features are not explicitly defined. Instead, a neural network learns by itself to extract meaningful features from raw data. Early layers enable the extraction of low-level features (e.g. edges, corners), while deeper layers can build from these simple notions to build more complicated decision rules [5]. As a counterpart of this automated learning, the model lacks interpretability, as the user does not have direct access to the learned features.

In DL, handcraft feature engineering is thus replaced by sophisticated architecture engineering, aiming to design a sequence of layers that maximizes the performance of an automated task. While a large variety of NN architectures can be found in the literature, they are generally composed of the same building blocks that are described below: fully connected (FC), convolution, and transformer layers, as well as normalization layers (e.g. batch, instance or channel normalization) and activation functions (e.g. ReLU, or sigmoid). Each block has different functionalities and properties and their combination allows to perform the desired task. With the rise of the number of layers grows the depth of the network, giving birth to *deep learning networks*. In the following, the most commonly used NN blocks for medical image analysis are presented.

Fully-connected (FC) layers, also called linear or dense layers, are the building blocks of the Multi-Layer Perceptron (MLP) which has been one of the earliest NN proposed in the literature (1991) [6]. FC layers are composed of a set of neurons, which are the network's lowest-level element, and are suitable for 1D input vectors. In FC, each neuron

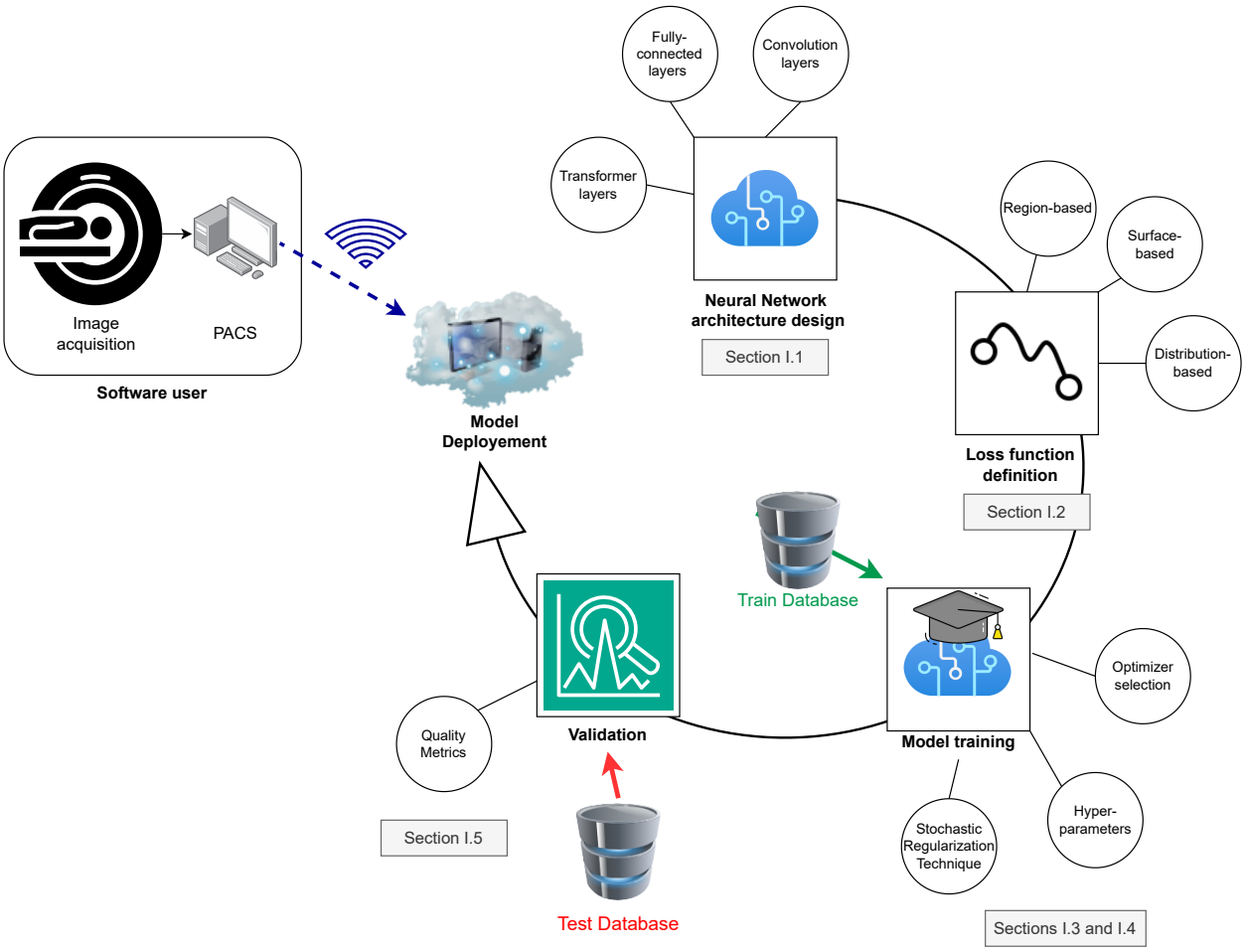


Figure I.1.1: Illustration of the principal steps of the development of a Deep Learning model. At Pixyl, the ultimate step is the deployment of the model in the cloud, accessible by users via the Picture Archiving and Communication System (PACS) of the hospital, where medical images are stored. Each step of the development is further detailed in this chapter.

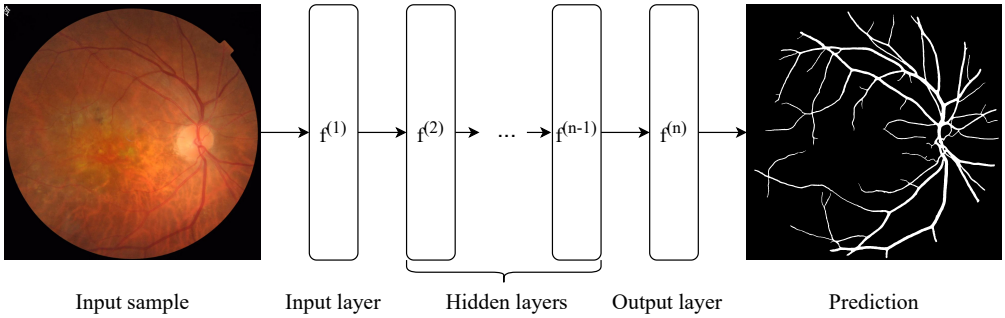


Figure I.1.2: A standard feedforward Neural Network.

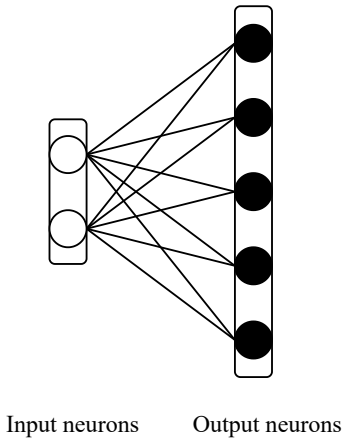


Figure I.1.3: A fully connected layer with 2 input neurons and 5 output neurons.

from the previous layer is connected to each of the neurons of the output layer, hence their *fully-connected* denomination (see Figure I.1.3). Writing M and N the number of input and output neurons, respectively, then the learnable weights matrix of the FC layer is a matrix $\boldsymbol{W} \in \mathbb{R}^{M \times N}$, and the learnable bias term is a vector $\boldsymbol{b} \in \mathbb{R}^{N}$. Bias is used to offset the layer activation towards the positive or negative side. For an input $\boldsymbol{x} \in \mathbb{R}^{M}$, the output of the layer can be expressed as:

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{W}^{\top} + \boldsymbol{b} \tag{I.1.1}$$

The computation is thus akin to a linear transform. Following the computation of y, an activation function h can be applied to get the final neuron output $\boldsymbol{y'} = h(\boldsymbol{y})$. This function is essential to add non-linearity to the networks, the root of their powerful modeling capabilities. Commonly used functions are the Sigmoid or the ReLU activation functions.

FC layers are built for 1D inputs making them poorly suited for image processing. One option would consist in flattening images into a 1D vector, but this process is inefficient for large images, especially 3D medical images. Moreover, it would discard the important spatial

information contained in the image. This motivated the development of parameter-efficient layers suited for images, such as the Convolutional layer, that can be used in combination with FC layers.

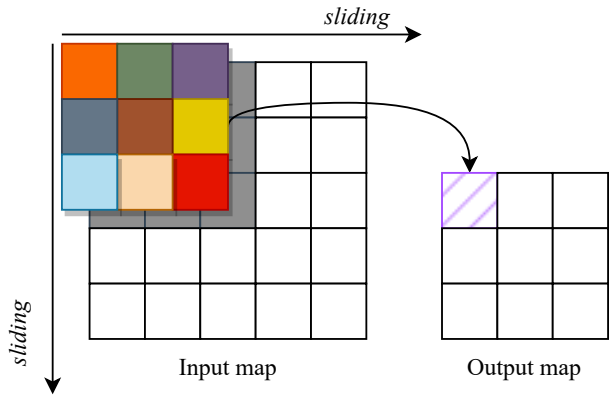


Figure I.1.4: A 2D convolution layer with a kernel size of 3×3 and a stride of 1. Applied to an input image of 5×5 pixels, it produces an output map of shape 3×3.

Convolutional layers The dedicated DL architecture for computer vision is the Convolutional Neural Network (CNN). This kind of model obtained state-of-the-art results in image classification and pattern recognition tasks and has been widely studied since a CNN called AlexNet won the ImageNet challenge in 2012, a famous benchmark for image classification. [7]. As their denomination indicates, CNNs are networks that include at least one convolution operation in their architectures. The convolution ($*$) is an operator that takes two real-valued functions as inputs and can be expressed as:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n) K(m,n) \tag{I.1.2}$$

where I is a 2D image, (i,j) the coordinate of a pixel, and K is a 2D kernel of shape $m \times n$. In most DL library, such as PyTorch [8] or TensorFlow [9], the convolution operation (represented in Figure I.1.4) is actually implemented using the *cross-correlation* function, which is defined as :

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n) \tag{I.1.3}$$

which is essentially the same as the standard convolution operator, but without the flipping of the kernel. A kernel, also referred to as a convolution filter, is an array of parameters that are learned during training, usually defined as a small isotropic window (e.g. 3×3 or 5×5). As the kernel is convolved across the input image, results are stored in an output array called a *feature map*. The map contains the activation of the filter, and can thus be

used to detect the presence of a particular pattern. As a single kernel can only learn a single kind of feature, a convolution layer generally consists of many applications of different kernels in parallel. A convolution layer is thus characterized by a set of hyperparameters, including the size of the kernel and the number of filters. The *stride* of the convolution can also be defined, quantifying the translation of the filter across the image. Ultimately, *padding* can be applied to the borders of the image to allow convolution to be applied to image edges.

After convolution, a *pooling* operation is traditionally performed on the features maps to scale them down. Essentially, it replaces the value of a feature map at a certain position with a summary statistic of neighboring values. A common choice is using the max operation [10] (Max Pooling) which effectively highlights the most prominent features in a predefined square region (e.g. 3×3 squares). The resulting reduction of the feature map dimensionality allows the preservation of important structural elements, while fine details not relevant to the task at hand are discarded. By doing so, pooling helps increase the robustness to translations. Indeed, if an object included in the input image is translated by a small amount, the values of most pooled outputs will not be modified [1]. Pooling also helps accelerate computation, thus reducing training time. A CNN can therefore be obtained by chaining convolution and pooling operations.

For classification and regression tasks, a FC layer is generally placed at the hand of the CNN, after a series of convolutions, to produce the final class scores. For segmentation, FC layers are generally not used in order to preserve the spatial information required to perform the task, and models are thus generally fully convolutional. For medical image classification, popular convolutional architectures include Residual and Dense Convolutional Neural Networks (CNNs) [11] and EfficientNets [12]. For medical image segmentation, popular choices include U-Net [13] and its variants, such as Residual U-Net [14], V-Net [15], Attention U-Net [16] or Dynamic U-Net [17]. While very successful for image processing, convolutions have some pitfalls. Indeed, the size of the kernel is limited, and thus CNNs struggle to model long-range dependencies within images [18]. This has motivated the exploration of convolution-free blocks, such as the vision transformer.

Vision Transformer Transformer models were originally proposed for sequence-to-sequence tasks in natural language processing (NLP) [19]. The concept was further adapted to computer vision tasks through the development of Vision Transformer blocks (ViTs, Figure I.1.5). In ViT, the input image is divided into a series of non-overlapping patches of fixed size. They are flattened and linearly projected into a lower-dimensional embedding sequence. Positional encoding of the patches is added to embeddings to maintain the spatial information that has been discarded when the input image has been divided into patches. ViT then exploits the self-attention mechanism to consider dependencies between different patches of the image when making predictions for a specific region. This allows the modeling of long-range dependencies between distant patches. The final element is a MLP composed of FC and normalization layers in order to produce the output of the ViT.

Popular transformer-based models in medical image processing include the U-NETR [20], Swin U-NETR [18] or Trans U-Net [21] models. All these models combine convolutions and transformer layers. However, transformer-based models are generally greedy in terms of

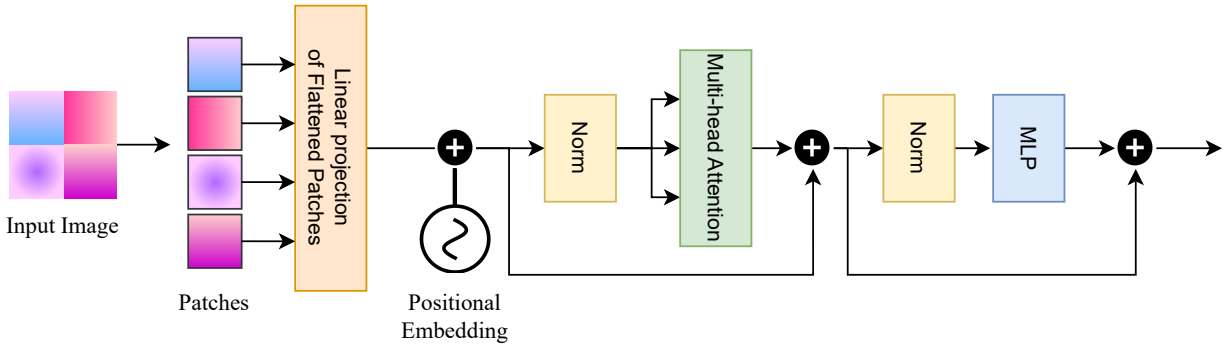


Figure I.1.5: A Vision Transformer block. The input image is first divided into patches, that undergo a linear projection into a lower-dimension embedding space. Positional embedding is added to maintain spatial information. Then, multi-head attention is applied to model the dependencies between patches. Finally, a Multi-Layer Perceptron (MLP) produces the final layer output. Note that skip connections are used to facilitate gradient flow during training.

parameters and computation and require extensive data points to be trained [22]. This is a challenge for 3D medical image processing as datasets are limited, and images take up a large amount of memory. As a result, training transformer-based models on 3D medical images represent a challenge in terms of computational cost.

Normalization layers are a crucial element of deep NNs, aiming at facilitating their training. They are generally applied to the hidden layers in the network. The most known approach, Batch Normalization (BN), was proposed in 2015 by Ioffe et al. [23]. The main objective was to tackle interval covariate shifts that can occur during the training of NN, leading to instabilities. More precisely, during optimization, the gradients for a given layer are computed assuming that the other layers are fixed. However, in practice we optimize all layers in parallel [1]. As a result, the input distribution for the current layer is constantly changing as the parameters of previous layers are updated, so that the current layer must constantly readapt to the shifts. BN proposes to alleviate this optimization issue by fixing the means and variances of the normalized layer.

Let's write $\boldsymbol{H}$ the input of the layer we wish to normalize for a given batch of inputs. Specifically, the i-th row of $\boldsymbol{H}$ contains the activation of the i-th element in the batch. Batch normalization operates by normalizing the input across the batch during training. The output of the batch normalization layer is expressed as:

$$\mathrm{BN}(\boldsymbol{H}) = \gamma \boldsymbol{H}' + \beta \tag{I.1.4}$$

$$\text{with } \boldsymbol{H}' = \frac{\boldsymbol{H} - \mu}{\sigma} \tag{I.1.5}$$

where μ and σ are the vectors containing the mean of each unit (neuron for FC layers, kernel

for convolutional layers) over the current batch. γ and β are learnable parameters named the scale and the shift. γ allows the model to learn the optimal scaling factor for each feature in the normalized layer, while β helps to control the spread of the activations. By normalizing the mean and variance of the layer, BN helps stabilizing learning and provides a protection against internal covariate shift. One drawback of BN is that its robustness depends on batch size, with larger batches being preferable [24]. For high-dimensional medical image processing, however, batch size is limited by hardware. Other types of normalization layers have therefore been proposed, such as instance [25], layer [26], or group normalization [27], which perform normalization for each sample in the batch independently, without taking into account the other elements in the batch, which is more suitable for limited batch sizes.

Once a NN has been defined using the previously introduced blocks, training can be performed to teach the model how to perform a given task. A pivotal step is the selection of a training objective, used to quantify the performance of the NN towards the target goal. In the following, the main principles of training objectives are presented.

I.2 Training objectives

In this section, the main concepts of training objectives are presented. They are introduced for 2D image classification for notation simplicity. Segmentation tasks are then directly derived from the presented framework as pixel-wise classification tasks.

For supervised classification, the goal is to build a discriminative model f that maps images $\boldsymbol{x} \in \mathbb{R}^{H \times W}$ into labels $\hat{\boldsymbol{y}} \in \{0, ..., K-1\}$ where K corresponds to a pre-defined number of classes. To build the model, a training dataset $\mathbb{D}_{train}$ is constructed, composed of pairs of images and ground truth labels $\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{N}$, considered independently drawn from the true, but unknown, data generating distribution p_{data}. In reality, only the empirical training distribution defined as $\hat{p}_{\text{data}}$ is available.

For a given image, the segmentation model outputs a categorical (or multinouilli) probability distribution, defined by the model's parameters $\hat{\theta}$. It corresponds to a distribution over the discrete label y that can have K different states (e.g. class):

$$P(\mathrm{y}|\boldsymbol{x}; \hat{\theta}) = \mathbf{Cat}(y; \pi) \tag{I.2.1}$$

It is parameterized by a vector $\pi \in [0, 1]^K$ where π_i indicates the probability of the i-th class, such as $\sum_{i=0}^{K-1} \pi_i = 1$ with $\pi_i \geq 0$. Typically, π is predicted by the DL model f by applying a softmax function to the raw model's logit predictions $z \in [-\infty, +\infty]^K$:

$$\pi = \frac{e^z}{\sum_j e^{z_j}} \text{ with } z = f(\boldsymbol{x}|\hat{\theta}) \tag{I.2.2}$$

Training requires the definition of a loss function $J(\theta)$, used to quantify the performance of

the network considering the task at hand. The baseline choice to train classification models is the **Negative Log-Likelihood** (NLL) loss. The optimization problem can then be defined as:

$$\hat{\theta} = \arg\min_{\theta} \Big\{ -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}\{\boldsymbol{y}^{(i)} = k\} \log P(\hat{\boldsymbol{y}} = k | \boldsymbol{x}^{(i)}; \theta) \Big\} \tag{I.2.3}$$
$$\hat{\theta} = \arg\min_{\theta} \Big\{ \mathbb{E}_{\hat{p}_{\text{data}}} [\mathcal{L}^{NLL}(\boldsymbol{y}, \boldsymbol{x}, \theta)] \Big\}$$

where $\mathbf{1}$ is the indicator function. The expectation is taken over the empirical training distribution $\hat{p}_{\text{data}}$, hence the denomination of **Empirical Risk Minimization**. Several derivates of the NLL loss function have been proposed for classification and segmentation tasks, including the Focal and Top-K cross-entropy losses. They are referred to as **Distribution** losses, as they aim at reducing the differences between the predicted and target probability distributions.

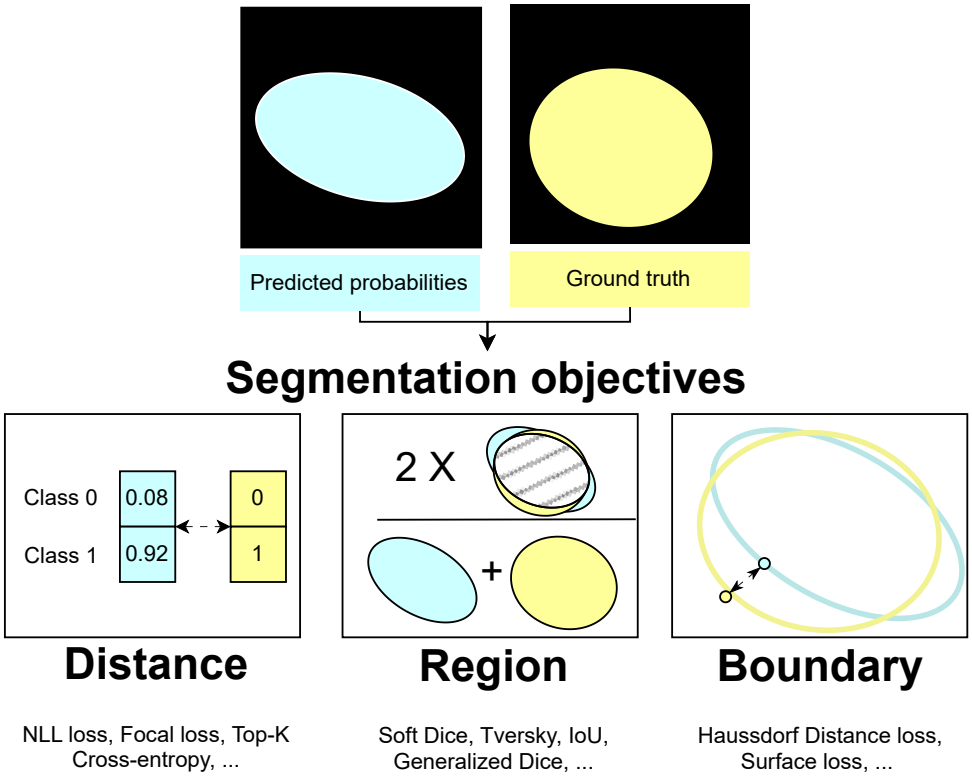


Figure I.2.1: Different categories of segmentation loss functions.

For segmentation tasks, defined as **per-pixel** classification, specific losses derived from popular segmentation quality metrics can be applied. The different existing categories are illustrated in Figure I.2.1. For instance, the Soft Dice loss [15] is predominantly used for

medical-image segmentation. Contrary to the NLL, the soft Dice is a **Region**-based loss [28] that aims to maximize the overlap between the predicted and ground truth masks. For a binary segmentation task, let p_{0i} the predicted probability that the pixel indexed by i belongs to class 0, and p_{1i} the probability that it belongs to class 1. Similarly, the ground truth label g_{0i} is 1 if i belongs to class to 0, else 0 (same for g_{1i}). Then the Soft Dice is expressed as:

$$\mathcal{L}^{Dice} = \frac{2 \times \sum_{i=1}^{N} p_{1i} g_{1i}}{\sum_{i=1}^{N} p_{1i} g_{1i} + \beta \sum_{i=1}^{N} p_{0i} g_{1i} + \alpha \sum_{i=1}^{N} p_{1i} g_{0i}} \tag{I.2.4}$$

with α and β the weights of False Positives (FP) and Negatives (FN) in the loss computation, respectively. The baseline Soft Dice loss is obtained with $\alpha = \beta = 0.5$, while alternative weightings are explored in the Tversky variant [29]. Another popular region-based loss is the Generalized Dice, designed for imbalanced segmentation problems [30].

However, region metrics including the Dice score are known to be biased toward large volumes [31]. For small targets (e.g. new MS lesions), region losses may thus produce unstable results [32]. This motivated the development of **Boundary**-based losses, inspired by the Haussdorff Distance (HD) [33] or the Mean Squared Error between the boundaries [34, 32]. They seek to minimize the distance between the contours of the predicted and ground-truth segmentations. Oversegmentation or undersegmentation will be penalized, which enforces a precise delineation of the target object.

Note that in practice, there is no obligation to use a single loss function to perform training. For medical image segmentation task, a popular choice is to combine the Dice loss and the cross-entropy. This is for instance the default setting of the nn U-Net framework [17]. Now that the loss function has been defined, the goal of training is to tune the model parameters so that the loss is minimized. This is akin to an optimization process that is presented in the following.

I.3 Optimization

In ML, optimization refers to the process of finding the optimal set of model parameters (here, weights and biases) to minimize the empirical error, assessed using a loss function. As previously presented, the loss function quantifies how well the model performs on the given task. In DL, the loss function is optimized using a gradient-descent approach based on the back-propagation algorithm [35]. The first step consists of the forward propagation process, during which input $\boldsymbol{x}$ is passed to the NN, flows through the hidden layers, and finally yields a prediction $\hat{\boldsymbol{y}}$. An error value is then computed by comparing $\hat{\boldsymbol{y}}$ to the ground truth $\boldsymbol{y}$ using the selected loss function $\mathcal{L}(\boldsymbol{y}, \boldsymbol{x}, \theta)$. The gradient of this scalar cost with respect to the parameters of the network $\Delta_\theta \mathcal{L}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ is computed during a step called **back propagation**. It consists of the iterative computation of the gradient using the Chain-Rule of calculus, stating that the gradient can be computed by multiplying derivatives of each function composing the network. [1]. Let x be the input of the network. Each layer, represented by functions f, g, h, operates a transformation of the input, yielding intermediate values v and

w. Output z of the network is thus obtained by compounding functions: $z = h(g(f(x)))$ and the gradient of z regarding the input x can be written as:

$$\begin{aligned} \frac{\partial z}{\partial x} &= \frac{\partial z}{\partial w}\frac{\partial w}{\partial v}\frac{\partial v}{\partial x} \\ &= h'(w)g'(v)f'(x) \end{aligned} \tag{I.3.1}$$

This principle is applied to backpropagate the error within the NN. By going back up the layers and chaining the derivatives, it is therefore possible to estimate $\Delta_\theta \mathcal{L}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ regarding the parameters of the network. Following the backpropagation step, a gradient-descent algorithm is performed, during which the model weights are updated accordingly to the gradient computed as follows:

$$\theta = \theta - \eta.\Delta_\theta \mathcal{L}(\boldsymbol{y}, \boldsymbol{x}, \theta) \tag{I.3.2}$$

where η is an hyperparameter called the **learning rate**. Several optimization algorithms called optimizers were developed to perform this gradient descent. Popular optimizers are **Adagrad** [36] and **Adam** [37], which allow refinement of the descent direction according to previous gradients.

Computing the error on the entire training dataset at each gradient descent step is too costly when dealing with large datasets. A convenient way to circumvent this challenge is to use a **mini-batch** approach [38]. More precisely, the error is computed over a reduced set of training samples, called a batch. Gradient computation and parameter updates are then carried out for each batch of data. This process continues until all the training samples have been passed through the model, this cycle being called an epoch. Mini-batch training is a stochastic process, as the batches are randomly sampled from the training dataset. This approach has several advantages, including limiting the computational cost of the algorithm, as the training dataset is decomposed into smaller batches. Yet the selection of the batch size is crucial. Indeed, large batch sizes allow for a more accurate estimation of the gradient. However, the memory usage scales up with the batch size, so small batch sizes (one or two samples) are generally used when dealing with high-dimensional 3D images [15], which has the disadvantage of making the gradient descent more noisy. This training procedure ultimately leads to the decrease of the empirical error through an iterative process. Training of a DL network is traditionally performed over several hundred epochs.

The goal of optimization is to minimize the error on the training dataset, with the assumption that the model will generalize on fresh test data. However, complications may arise, which have led to the development of methods to improve generalization. They are presented in the following.

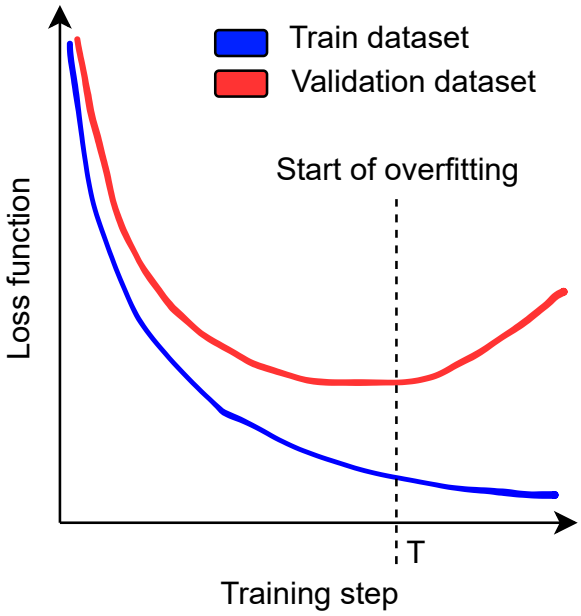


Figure I.4.1: Illustration of the monitoring of overfitting. The loss is monitored on the training dataset as well as a set-aside validation that is not used to perform gradient descent. At the training step T, the validation loss starts to increase while the training loss steadily decreases, which is an indicator of overfitting.

I.4 Generalization and stochastic regularization techniques

The starting assumption of DL (and more broadly ML) is that a model trained using $\hat{p}_{\text{data}}$ will be able to generalize to new unseen test data points drawn from the same underlying data distribution. The main obstacle to generalization is overfitting, a deficiency of the network for which the cost attained a small value on the training set, but is significantly larger when processing new data. This mismatch is often referred to as the *generalization gap* in the literature [1]. It implies that the model has failed to learn the global trend of the data, and rather simply has learned a function that fits all the training points. Overfitting is a common threat in medical image analysis: the size of the training dataset is often limited, hence there is a risk that the model simply memorizes training data instead of learning generalizable rules [39]. The NN is more prone to overfitting when it has a high complexity (high number of parameters), which is generally desired to model complex decision rules.

In practice, overfitting can be detected by monitoring the value of the loss function on two datasets: the training dataset, and a set-aside validation dataset that is not used to perform gradient descent (see Figure I.4.1). Overfitting occurs when the training loss function steadily decreases while the validation loss increases.

Several tools were designed throughout the history of DL to improve the generalization of trained models, including *Stochastic Regularization Techniques* (SRTs) [40]. These techniques are designed to regularize the model (i.e. reduce the generalization gap) by injecting random

noise during training. The most widely known SRT technique is Dropout [41]. The concept is to randomly drop (i.e. set to 0) the layer outputs within the network, with a given probability (e.g. 20%). By doing so, the complexity of the model is impaired during training, preventing memorization. Interestingly, using Dropout in a network allows the simulation of a virtually infinite ensemble from a single network, as various network configurations are obtained by stochastically dropping out activations. This property is at the root of the Monte Carlo (MC) Dropout technique, a very popular UQ method that will be introduced later in this thesis (Section II.2.6). Finally, data augmentation plays a pivotal role in generalization. This technique consists of artificially increasing the size of the training dataset using image augmentation such as rotation, flipping, contrast enhancement, or noise injection. From a single image, an infinite number of variants can be potentially obtained by randomizing the parameters of these transforms. Hence, data augmentation can be seen as a form of implicit SRT [42]. Moreover, data augmentation is efficient in making the model robust to specific types of noises and artifacts that are likely to be encountered during inference [43]. Another recent lead of research is the use of generative AI to create synthetic images that can be used to complement the training dataset. Important advancements in this direction were achieved by diffusion models [44, 45, 46, 47].

After training is completed, the last important step of the model development is validating the model on test data. This procedure is briefly detailed in the following for segmentation models, which are the main focus of this thesis.

I.5 Evaluation of segmentation deep learning models

Once the model has been trained, a crucial step is the validation of the DL model to make sure that it achieves satisfying performance on unseen test data. For medical applications, careful evaluation is required to ensure that the model is clinically useful. The gold standard is to rely on a test database of images manually annotated by human experts. It allows computing overlap measures between predicted and reference segmentations, with the Dice score being a predominant option in medical image processing. However this metric may not always reflect the biomedical need, and other relevant metrics can be used to assess the quality of a segmentation based on the task's specifications. For settings where the exact delineation of the target object is important (e.g. organ delineation), distance-based metrics such as the Haussdorf distance [48] can be employed. For tasks involving the detecting of lesions, lesion-wise detection metrics may be more relevant [31].

Finally, in industrial applications, other key metrics are monitored, not directly related to the predictive performance. It includes the model memory consumption and inference time. Particular attention is given to designing high-performance models that are compatible with real-time clinical use, meaning that their time and memory consumption should be reasonable.

Conventional DL model development stops at this validation stage. Nevertheless, it should be noted that we have not yet turned our interest to quantifying uncertainty. An immediate idea would be to look at the probability associated with a prediction. This concept is presented in

the following section.

I.6 Probability calibration

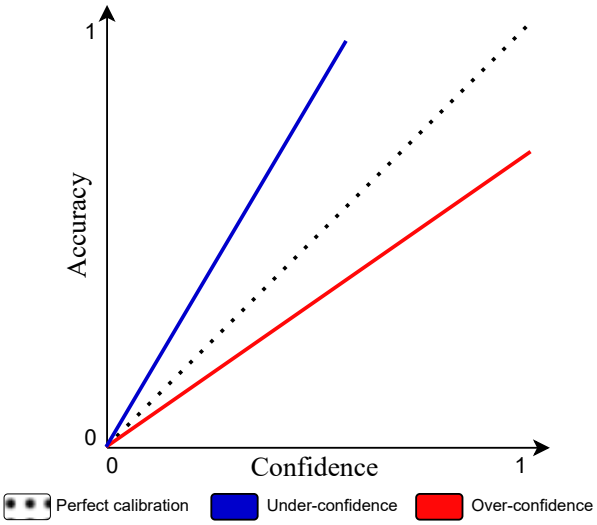


Figure I.6.1: Example of a reliability diagram showing perfect calibration (black dashed line), under (blue), and over-confidence (red).

By combining the tools introduced in this chapter, discriminative NN models can be trained and applied to unseen test data points. As presented in Section I.2, the NN produces a categorical probability distribution over the possible class labels. An immediate and intuitive notion of uncertainty can be derived from this output: a high probability is expected to indicate a confident prediction, while an uncertain choice should be associated with the probability of random guessing (e.g. $p = 0.50$ for binary classification tasks). Ideally, *correct* predictions should be made with high confidence, while erroneous ones should be associated with low confidence. This alignment between confidence and accuracy is called the calibration of the model [49].

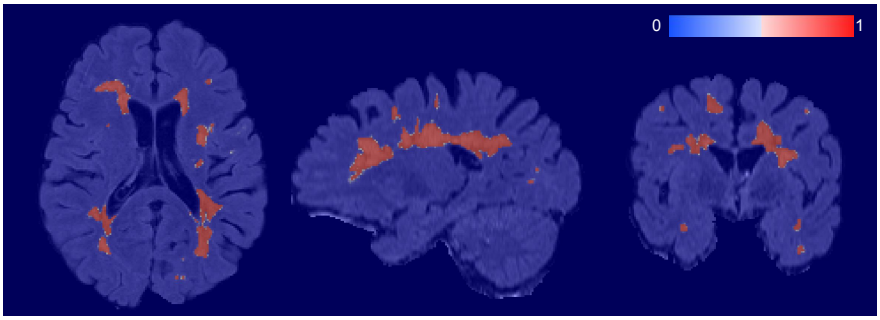
For a perfectly calibrated model, the predicted probability perfectly reflects the true probability of an event, e.g. when considering all predictions made with a probability of 0.80, the model is correct 80% of times. More formally, writing Y and $\hat{y}$ the ground truth and predicted label classes and $\hat{P}$ the associated probability, a perfectly calibrated model respects:

$$P(\hat{Y} = Y|\hat{P} = p) = p, \forall p \in [0, 1] \tag{I.6.1}$$

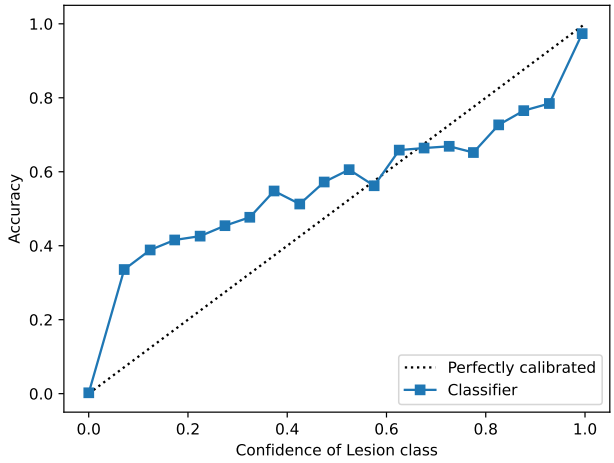
Of course, this ideal calibration property is rarely achieved, especially for modern NN [49]. A convenient way to measure the actual calibration of the model is through reliability diagrams (also called calibration plots). Such a diagram is illustrated in Figure I.6.1. The main idea is to plot the model accuracy as a function of its confidence (e.g. defined here as the probability

of the predicted class). This is performed by binning predictions according to the associated predicted confidence, and computing the accuracy for each bin. Following Equation I.6.1, the calibration plot of a perfectly calibrated model should correspond to the identity function (Figure I.6.1, black dashed line). Any deviation to this identity function corresponds to a *calibration gap* (either under or overconfidence, represented by the blue and red lines in Figure I.6.1).

It is important to stress that many UQ methods that will be presented in the next section are building on predicted probabilities to compute uncertainty scores, including the entropy, variance, or mutual information (MI) of the categorical probability distribution. However, severe miscalibration is usual in medical-image segmentation models, notably due to inappropriate choice of loss functions (e.g. using the Dice loss for image segmentation) [50]. An example of such a phenomenon is presented in Figure I.6.2: a model is trained to segment MS lesions using the Dice loss. A test image is then overlayed with the predicted lesion class probability map. It appears that the large majority of voxels is associated with a probability extremely close to 1. The associated reliability diagram is presented below. The first step before exploiting probabilities for UQ is thus to make sure they are properly calibrated. Calibration techniques are further introduced later in this thesis (Section II.6.3).



a



b

Figure I.6.2: Illustration of a miscalibrated model trained using the Dice loss. (a) Input FLAIR brain MRI overlayed with the predicted lesion class probability map. (b) Associated reliability diagram showing both under-confidence for probabilities below $p = 0.5$, and over-confidence for probabilities above $p = 0.5$.

I.7 Chapter conclusion

In this introductory chapter, the main concepts of DL models have been presented, including NN building and loss function optimization. We also introduced the main challenges associated with deep NN, including the generalization issue and poor calibration. In the following chapter, the tools that have been proposed to enhance NN with proper uncertainty estimates are presented.

Chapter II

Uncertainty for Deep Learning-based medical image analysis: definitions, motivations and literature review

In the previous chapter, the DL paradigm was introduced. While it has revolutionized medical image processing, it has yet to be accepted and used by clinicians due to the black-box effect of NN. Indeed, NN do not have explicit decision rules, as the implicit features of NN that are learned during training are generally unintelligible to the user, and there is a lack of reliable confidence estimates associated with their predictions [49]. Moreover, it has been shown that DL models can be overconfident about their predictions on outlier data [51], which suggests a global lack of robustness of these predictive models. Due to these limitations, detecting failures or inconsistencies produced by DL models is complex, raising concerns regarding the reliability and safety of these algorithms in clinical-routine use [52]. To tackle this important issue, Uncertainty Quantification (UQ) methods [53] have been developed to quantify the predictive uncertainty of a given DL model and it has emerged, from a clinical point of view, as one of the expected properties of any deployed AI algorithm [54]. As a result, the medical-imaging community is becoming increasingly interested in incorporating UQ into image-processing pipelines in order to highlight model failures or weaknesses. In this chapter, we propose a literature overview of the proposed UQ tools integrated into medical-image processing pipelines.

Contents

II.1 Sources of uncertainty in medical images

Identifying the different sources of uncertainty that can arise in supervised ML classification problems is crucial for their proper quantification. Predictive uncertainty, meaning the uncertainty associated with the prediction of a DL model, is generally divided into two parts: epistemic (or model) and aleatoric (or data) uncertainty [55]. For convenience, these different sources of uncertainty are first illustrated on a simple 1D regression problem (Figure II.1.1). In this scenario, the red line corresponds to the function to be learned using the training samples (black dots). However, the learning is complicated by the presence of both epistemic and aleatoric uncertainties.

Epistemic uncertainty describes the lack of knowledge of the model concerning the current input being processed [55, 56]. It is considered to be reducible, meaning that it can be limited by using additional data. In practice, epistemic uncertainty is expected to be high when the model is confronted with samples that are unusual or different from those observed during the training stage [57]. In the regression example, it can be observed that data points are not uniformly distributed over the possible values of x. Instead, points are concentrated in the intervals $[-10, -5] \cup [5, 10]$. The intermediate interval $[-5, 5]$ corresponds to a *moderate* epistemic uncertainty region, as data points in this area are very scarce. The intervals $[-\infty, -10] \cup [10, \infty]$ correspond to *high* epistemic uncertainty regions, as training samples are absent. At inference, test samples belonging to these intervals would be considered as **out-of-distribution** (OOD) data points, and it is expected that the model's prediction would be uncertain and suboptimal. In medical-image analysis, such situations are frequent, as there may be significant variation between training and test images, for example, if they were acquired at different hospitals or using different machines [58]. Additionally, unexpected patterns can be encountered in test images, such as diseases not encountered during training and artifacts. In ML, epistemic uncertainty is generally modeled using distributions over the model's parameters [57].

Aleatoric uncertainty describes intrinsic noise and random effects within the data [55]. It is not intrinsic to the model, but rather a property of the underlying distribution of the data. In the 1D regression example, the green section ($x \in [-10, -5]$) corresponds to a low aleatoric uncertainty region: measurements are noiseless and provide an accurate approximation of the true function. On the contrary, the red section ($x \in [5, 10]$) corresponds to a high aleatoric uncertainty setting: the measurements are much more noisy, and there is a significant variability in the outcome for similar values of x. Aleatoric uncertainty can be further split into two categories: homoscedastic uncertainty which is identical for each sample of the dataset, and heteroscedastic uncertainty which depends on the query input. Finally, aleatoric uncertainty is challenging to reduce. The only way to mitigate it would be to change the data acquisition strategy, for instance by increasing the quality of the sensors. In ML, aleatoric uncertainty is generally modeled by placing a distribution over the model's outputs [57].

Label uncertainty : In the context of medical-image analysis, aleatoric uncertainty can be observed not only in the input data (low signal-to-noise ratio, artifacts, partial volume

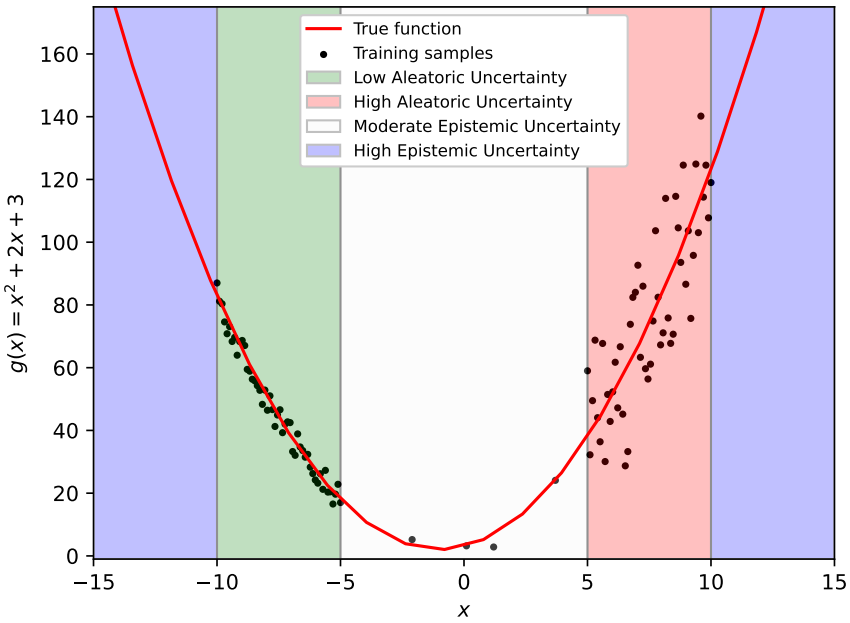


Figure II.1.1: Visualization of uncertainty sources on a simple 1D regression task. Blue areas correspond to out-of-distribution regions where no training samples are available.

effect), but also in the ground truths. It has been observed that inter-rater variability in the context of ground truth annotations of medical images is important [59, 60]. We refer to this type of uncertainty as label uncertainty. Two real-world examples of this inter-rater disagreement are provided in Figure II.1.3, for prostate MRI segmentation and new MS lesions segmentation in longitudinal brain MRI. In these examples, the experts do not agree about the exact delineation of the organ (prostate segmentation) or the presence of a brain lesion (new MS lesions segmentation). This has a direct impact on the model's overall uncertainty as the same object of interest (e.g. a brain tumor) may have significantly different ground truth delineations depending on the rater. The noisiness of ground truths for medical-image segmentation is thus a serious issue that can heavily impair the performance of the trained model [61].

While the regression example is convenient for introducing the concepts of aleatoric and epistemic uncertainties, this thesis deals with the processing of high-dimensional medical images, for which uncertainty sources manifest in specific ways. In Figure II.1.2, several synthetic images are presented, each expressing one particular uncertainty setting that can be encountered in medical applications. The first setting, **No uncertainty**, corresponds to an ideal scenario where there are no uncertainty sources: the target object is easily distinguishable from the background, and the ground truth perfectly matches the target object. **Aleatoric uncertainty** is illustrated in two settings, respectively, label and image uncertainties. **Label uncertainty** depicts the case where the ground truth is noisy and contains annotation errors. More precisely, the delineation under-segments the object, which would increase the model uncertainty about the expected boundary of the object. **Image uncertainty** represents uncertainty contained within the image or ground truth. More

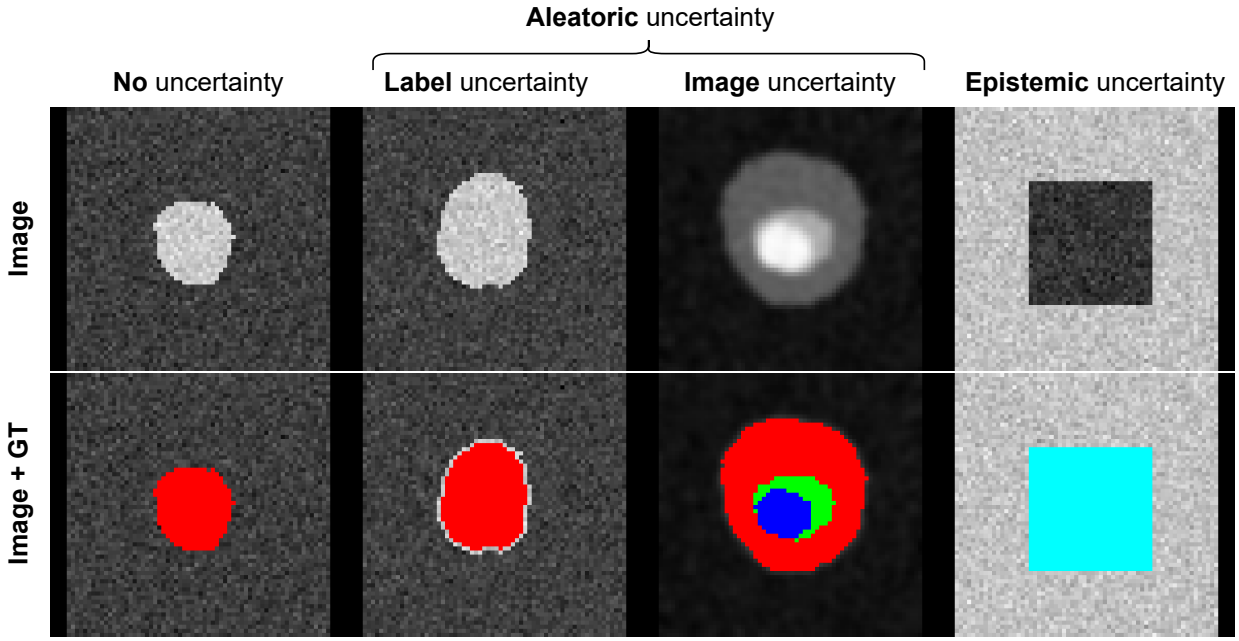
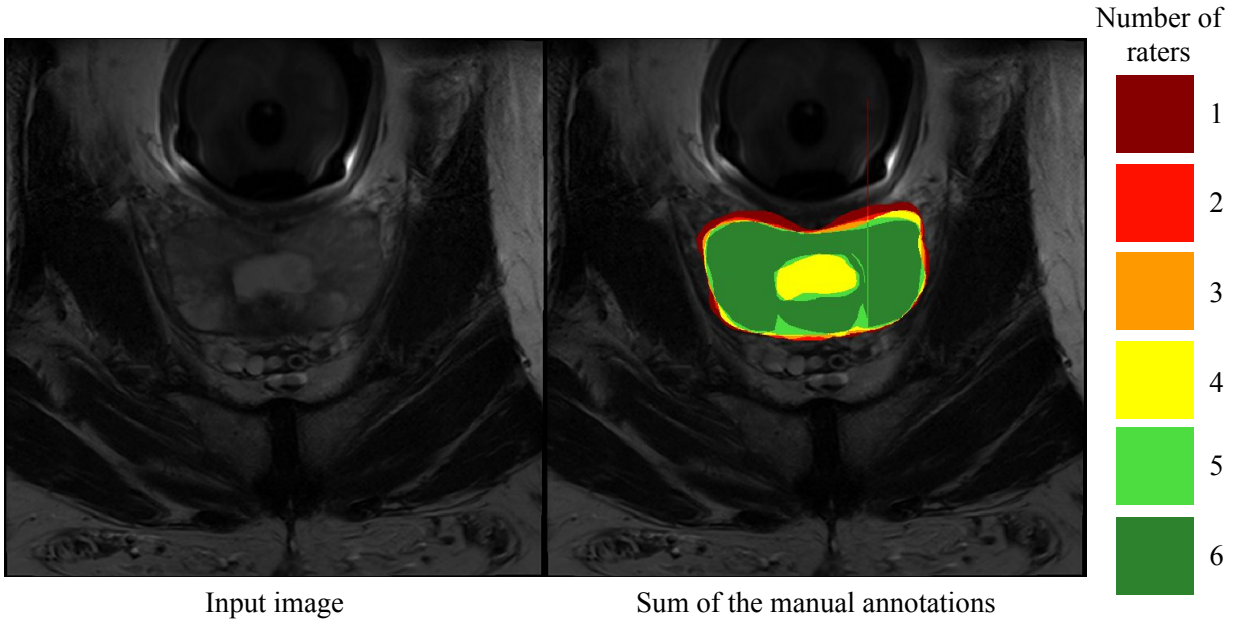


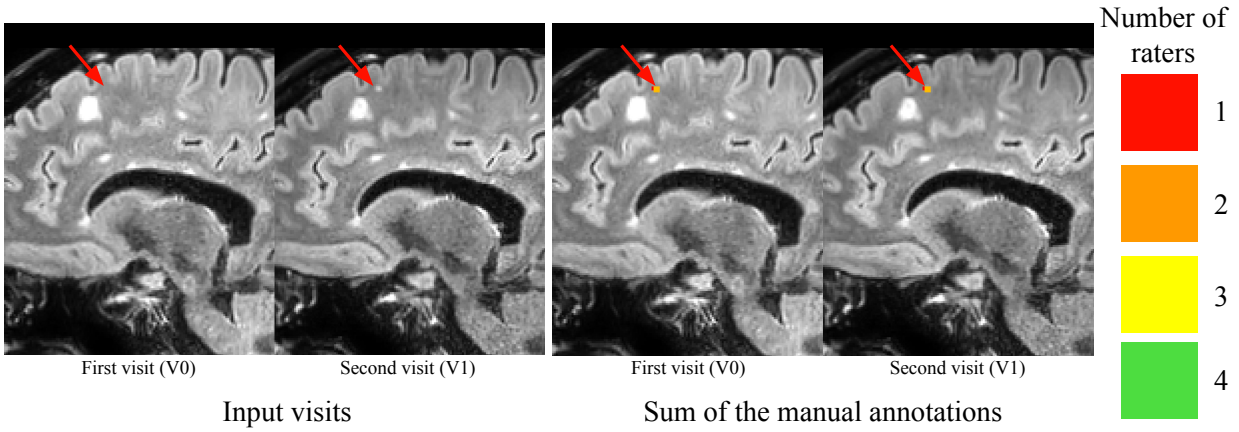
Figure II.1.2: Types of uncertainties illustrated on synthetic 3D medical images.

particularly, 3 different classes are nested inside each other (blue, green, and red labels). Partial volume, simulated using Gaussian blur, complicates the precise delineation of the boundaries between classes. Such partial volume effect is extremely common on MRI [62], where a single voxel can contain a combination of several tissues, for instance, healthy white matter *and* MS lesion. Thus, it is expected that a segmentation model would hesitate about both classes. Finally, the last example depicts **Epistemic uncertainty**. As compared to the first image which contains a hyper-intense spheroid and a hypo-intense background, the image contains a hypo-intense cube and a hyper-intense background. A model trained on the former type of images would therefore lack knowledge about the expected output for the second image.

These two examples (1D regression and synthetic images) allow us to intuitively distinguish between the 2 main sources of uncertainty in supervised medical-image classification, namely data (which encompass image and label-induced uncertainties) and model. In real-world applications, these phenomena are often intertwined. A flourishing literature has been proposed to quantify one particular type of uncertainty or both, in the context of DL-based medical image analysis. In the following section, an in-depth review of this literature is presented.



a



b

Figure II.1.3: Illustration of label uncertainty in prostate MRI segmentation (a, QUBIQ dataset [a]) and in new MS lesions segmentation in longitudinal brain MRI (b, MSSEG-2 dataset [63]). Images (left) are superimposed with the sum of the annotations (right) made by distinct annotators. For prostate segmentation (a), label uncertainty can be observed around the boundaries of the prostate. For new MS lesions segmentation (b), only 2 out of 4 raters segmented the pointed lesion as new, although the lesion appears clearly in V1 but not in V0.

[a] https://qubiq.grand-challenge.org/

II.2 Review of UQ techniques applied to medical-image analysis

II.2.1 Additional contributions to the paper "Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis"

The literature review presented in this section is the subject of an article published in Artificial Intelligence in Medicine [64]. To complement this study, two benchmarks are added in this thesis. While one focuses on the comparison of calibration techniques, the second compares the most popular UQ paradigm on brain lesion segmentation tasks.

II.2.2 Overview

To examine existing methods proposed to quantify uncertainty in medical-image analysis, the Google Scholar and PubMed search engines were employed. The goal was to identify studies that implemented UQ methods applied to medical-image classification or segmentation. The search was restricted to the following period: January 2015 to October 2023 (included). The following keywords were employed: **Deep Learning**, **Uncertainty**, **MRI**, **CT**, **PET**, **X-RAY**, **Ultrasound** and **Medical Image**. Matching articles were included if 1) they presented DL approaches for medical-image classification or segmentation, and 2) they proposed an uncertainty quantification of their algorithms. Non-peer-reviewed studies were excluded, with exceptions for papers exceeding 30 citations. Non-English publications were also removed, as well as review articles and animal studies. This ultimately narrowed down the number of papers from 241 to 218. A total of 338 UQ methods were identified in this pool of papers, implemented either as principal contributions or as comparison methods. To identify trends in this collection of papers, they have been clustered according to 1) the method used for uncertainty estimation and 2) the type of uncertainty that is considered, namely epistemic, aleatoric of both (see Figure II.2.1). In the following section, each UQ framework is introduced.

II.2.3 Softmax uncertainty

As mentioned in the previous chapter (Section I.6), a segmentation NN produces categorical probability distributions over the possible class labels, which can be used as an immediate and intuitive uncertainty estimate (see Figure II.2.2). In practice, this is true if and only if probabilities are calibrated. However, modern NNs tend to be highly miscalibrated, meaning that the produced probabilities are unreliable, and usually over-confident [49]. To transform the raw probabilities into real certainty estimates, various calibration methods have been proposed in the literature. Pioneering work proposed the Temperature Scaling approach [49], which consists of rescaling the logits of the NN by a single scalar value, the *temperature*, which proves to empirically reduce the calibration error without altering the classification result. However, Temperature Scaling also reduces the confidence of correct predictions. More sophisticated approaches were proposed afterward, based on binning approaches [65] or Dirichlet distributions [66]. Finally, while these methods imply a post-hoc calibration

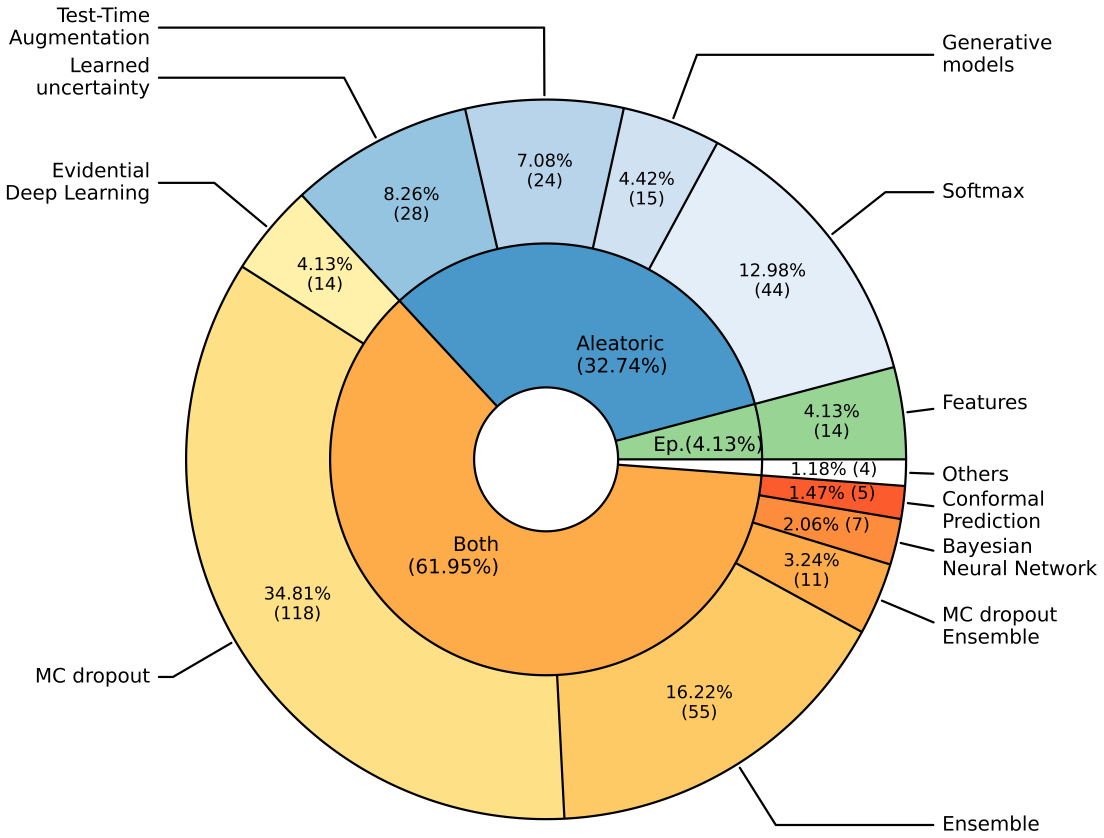


Figure II.2.1: Pie chart of Uncertainty Quantification methods in the 218 selected papers. Percentages (and numbers) of selected papers for each class of methods are indicated in the outer ring. The inner ring classifies methods according to the type of uncertainty modeled: aleatoric, epistemic (Ep.), or both.

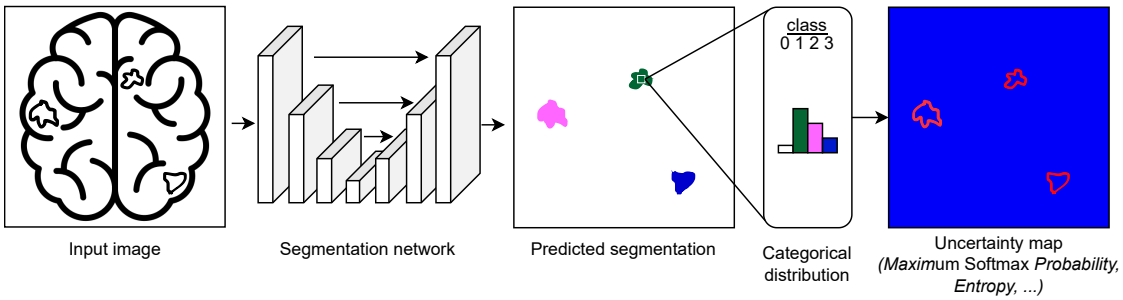


Figure II.2.2: Illustration of the Softmax uncertainty paradigm. A standard segmentation network produces a categorical probability distribution for each voxel, from which different uncertainty scores can be derived such as the Maximum Softmax Probability or the entropy.

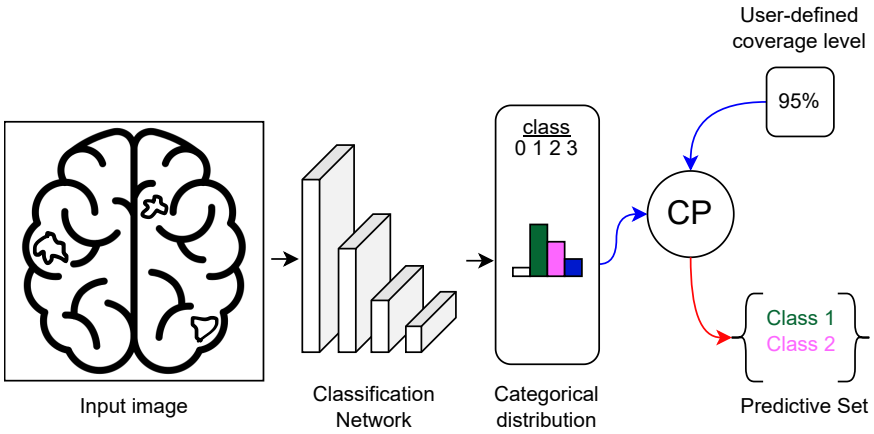


Figure II.2.3: Illustration of the Conformal Prediction (CP) paradigm for image classification. In the usual setting, the most probable class is predicted for a given image. Instead, CP operates a post-processing of the predicted probabilities to generate a set of labels conditioned to contain the true label with a user-defined confidence level, such as 95%.

once training is completed, another field of work enforces calibration through the learning objective, to obtain *ad-hoc* calibrated models. Recent examples of these loss functions that promote calibration are the Margin loss [67] and the Dice++ loss[50].

Due to its simplicity, the utilization of Softmax probabilities as uncertainty estimates was naturally explored for medical-image processing applications, often serving as a simple baseline for comparison to more sophisticated approaches. As an illustrative example, [68] and [69] leveraged the entropy of (uncalibrated) Softmax probability vectors for brain tumor segmentation in MRI. Alternatively, [70] used the Maximum Softmax Probability (MSP) uncertainty estimator, corresponding to the highest probability class for each voxel, for skin lesion segmentation in RGB images. A similar score is used for out-of-distribution (OOD) detection experiments in the context of chest X-ray pathology classification [71] and COVID-19 lesions segmentation in CT scans [72], respectively. Calibration was explored in Carneiro et al. [73], where authors employ Temperature Scaling to recalibrate the predicted probabilities of a polyp classification model. Finally, [67] and [74] proposed incorporating calibration terms in the training objective of their NN in the context of segmentation and classification of medical images, respectively, to obtain well-calibrated predicted probabilities.

It is important to note that UQ based on Softmax probabilities only considers the distribution over the model's outputs and not the model's weights. Thus, this type of deterministic uncertainty estimate only considers aleatoric uncertainty [55, 57].

II.2.4 Conformal prediction

Conformal Prediction (CP, Figure II.2.3) is a statistical approach for uncertainty quantification that has been attracting a lot of attention lately in the ML community. When applied to image classification, CP operates a post-processing of the raw softmax probabilities. Hence,

it can be seen as an extension of the previously presented Softmax uncertainty approach.

While its fundamental concepts are not new [75], CP has been extensively revisited in DL pipelines thanks to its several appealing properties: it makes no assumption about the black-box predictor nor the distribution of the data, and it provides provable statistical guarantees. The core concept of CP is to transform the point-wise prediction of a model into a *predictive set*. In the classification setting, these predictive sets correspond to a list of probable class labels, while for regression tasks, they correspond to predictive intervals (PIs) associated with the regressed value [76]. These sets are constructed so that the ground truth label is guaranteed to be included with a user-defined confidence level, such as 90% or 95%. This corresponds to the desired *coverage* level. To achieve this result, CP performs a post-processing of the raw predictions of the model (class probabilities for classification, or predicted scores for regression) and is usually fit using a set-aside labeled calibration dataset that comes from the same distribution as the test dataset. This procedure is called *split CP*. It is important to note that in contrast to other UQ methods that aim at complementing a prediction with an uncertainty estimate, CP instead starts by defining a target level of uncertainty, and then adapts the prediction accordingly. CP has found many applications in natural image classification [77], regression tasks [78], or drug discovery [79]. Applications to medical images are emerging: CP is employed in the AmnioML framework [80] to provide PIs associated with Amniotic Fluid volume prediction. As opposed to volume prediction, [81] focuses on computing PIs for counting tasks, applied to cell and brain lesions counting. Finally, CP is also investigated in two recent studies focusing on medical-image classification [82, 83].

While extremely promising for medical applications, as it provides statistical guarantees to the user concerning the error of the deployed model, CP suffers from 2 major limitations that may hinder its usage in the field. First, it is based on the assumption that calibration and test data are exchangeable, meaning that they come from the same distribution. However, it is well known that domain shifts are extremely common in medical-imaging applications, which hinder the effectiveness of the conformal procedure [84]. Second, the calibration dataset must be large enough to perform the split CP procedure. The current guideline is to use 1000 calibration samples [76]. In medical applications, data is often scarce, hence obtaining high-confidence PIs using split CP may not always be feasible.

II.2.5 Bayesian deep learning

Bayesian modeling is a very convenient paradigm for dealing with uncertainty and is thus the preferred theoretical approach to uncertainty in Deep Learning [40]. In Bayesian Deep Learning (BDL, Figure II.2.4), each weight of the neural network is replaced by a distribution, rather than having a single fixed value [85]. To achieve this, a prior distribution $p(\theta)$ (usually Gaussian) is first initialized over the neural network parameters θ. It follows that each weight is represented by a mean and a variance (thus effectively doubling the number of parameters of the model). Then, during training, the model learns the posterior distribution $p(\theta|X, Y)$, given the dataset and the prior distribution, which accounts for the less and more likely parameters given the observed data. Using Bayes theorem, the posterior is expressed as:

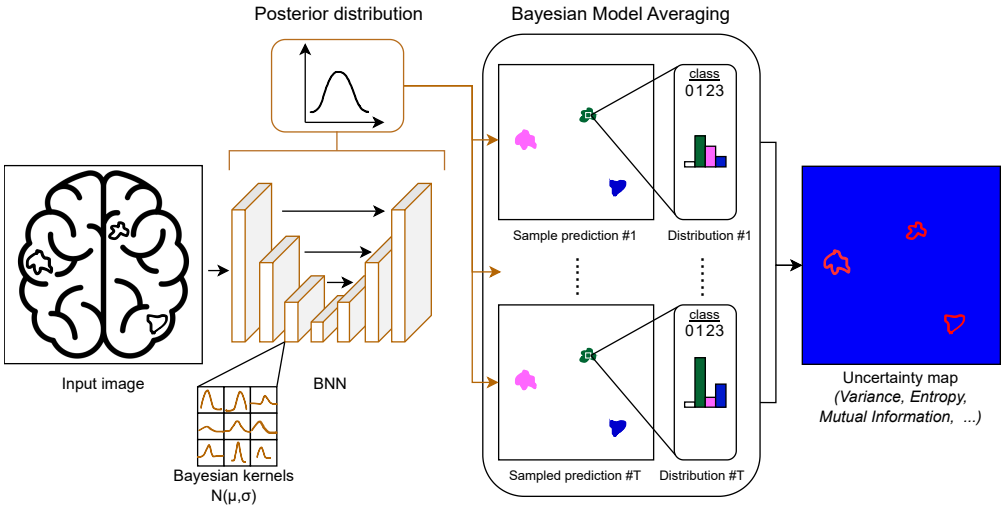


Figure II.2.4: Illustration of the Bayesian Deep Learning paradigm. In a Bayesian network, weights are represented by distributions in place of the usual deterministic weights. The distributions are represented by a set of two parameters: the mean and the variance. To perform inference, Bayesian Model Averaging is employed, following which the weights are sampled from the posterior distribution.

$$p(\theta|X,Y) = \frac{p(Y|X,\theta)p(\theta)}{p(Y|X)} \tag{II.2.1}$$

In this equation, $p(Y|X,\theta)$ is called the likelihood distribution of the model and is responsible for generating outputs y based on a given query input x and parameters θ. Note that in the context of image classification or segmentation, the likelihood corresponds to the softmax function. The normalizer of equation II.2.5 is called the model evidence and can be written as:

$$p(Y|X) = \int p(Y|X,\theta)p(\theta)d\theta \tag{II.2.2}$$

A trained Bayesian Neural Network (BNN) is akin to a virtually infinite ensemble of neural networks, where each instance has its weights drawn from the learned posterior distribution. During inference, the distribution is marginalized by repeatedly sampling weights from the shared distribution and averaging the predictions. This process is called Bayesian Model Averaging [86]. The inference process for a given query input x^* can be written as:

$$p(y^*|x^*,X,Y) = \int p(y^*|x^*,\theta)p(\theta|X,Y)d\theta \tag{II.2.3}$$

Computing this integral requires marginalizing over all model parameters. Although it is possible for very simple neural networks, modern neural networks are over-parameterized, which makes the exact computation of the posterior intractable [40, 87]. To account for this issue, a branch of work has focused on approximating the true posterior using Variational Inference (VI). VI proposes to approximate the posterior using a variational distribution $q(\theta|w)$ [88]. The parameters w of the variational distribution are learned during training to be as close as possible to the exact posterior. This is achieved by minimizing the variational free-energy cost function, usually referred to as the expected lower bound (ELBO) [85]:

$$F(D, w) = KL[q(\theta|w) \parallel p(\theta)] - \mathbb{E}_{q(\theta|w)}[log(P|\theta)] \tag{II.2.4}$$

Minimization of this loss is achieved using Stochastic Gradient Descent (SGD), as in standard neural networks. This training paradigm is called Bayes by Backprop (BBB) [85]. VI allows to address Bayesian Inference as a classical optimization problem. Once training is completed, various uncertainty estimates can be obtained, such as the entropy of the predictive distribution, its variance, or its mutual information. BDL places a distribution on the model's weights, hence it is rooted in epistemic uncertainty quantification. However, BDL applied to classification and segmentation tasks also produces a categorical probability distribution, so it can be easily coupled with the Softmax probabilities framework previously introduced (Section II.2.3) to also quantify aleatoric uncertainty [89].

Applications of BDL to medical-image processing are so far scarce. Studies initially focused on applying Bayesian convolutions associated with VI approaches. We found applications for 2D medical-image classification [90, 91], knee abnormality detection [92], lung and nasal endoscopy CT segmentation [93] and brain tumor segmentation [94]. However, this approach requires extensive changes in the model architecture and training paradigm [85, 89], associated with an increase in the computational cost of both training and inference. This has motivated recent studies on scalable BDL solutions. For example, in Adams et al. [95], authors evaluate Rank-1 Bayesian networks [96] as well as latent posterior BNN on a task of organ segmentation in 3D CT. These two approaches were recently proposed as scalable alternatives to the standard BDL framework.

II.2.6 Monte Carlo dropout methods

In Gal et al. [97], authors demonstrated that a NN trained with the dropout technique (introduced in Section I.4) can efficiently approximate Bayesian inference without the associated prohibitive computational cost. Based on this principle, Monte Carlo Dropout (MC dropout, illustrated in Figure II.2.5) proposes to train a model with dropout and keep it activated during inference. For a given query input, multiple forward passes are then performed. Each time, a different dropout mask is randomly sampled (generally following a Bernoulli distribution), producing different predictions. Following this process, a predictive distribution is obtained, similar to BNN. As for BDL, MC dropout was initially proposed to tackle epistemic uncertainty, although it still produces a categorical probability distribution from which aleatoric uncertainty estimates can be computed [98]. MC dropout allows the approximation of a BNN in any network trained with dropout, it thus rapidly gained popu-

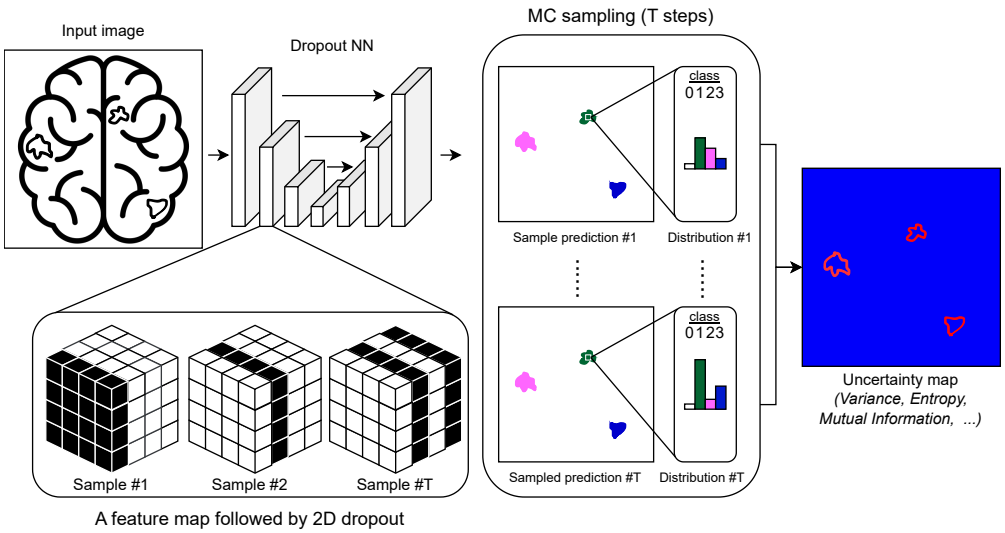


Figure II.2.5: Illustration of the Monte Carlo Dropout framework in a 2D CNN. Blocks colored in black indicate dropped-out channels. The key concept is to keep dropout activated at test time and repeat the inference process multiple times. It generates a set of Monte Carlo samples from which usual uncertainty scores can be derived (e.g. entropy, variance).

larity, and applications in the medical-imaging field are numerous. Implementation of this framework varies little. Studies that stand out have studied in further detail the importance of the dropout layer's position and type, rate, and the number of drawn MC dropout samples. Jungo et al. [99, 69] studied the importance of the positioning of the dropout layer within a convolutional network for brain tumor segmentation in MRI. However, experiments did not allow to draw clear conclusions. [100] evaluated the impact of the number of MC samples at inference on the segmentation accuracy of the photoreceptor layer in OCT scans, and found no improvement after 20 samples. Similar work was carried out in Camarasa et al. [101], where the impact of the dropout rate and type (Bernouilli or Gaussian dropout) was assessed on a task of cardiac MRI segmentation. While the dropout type had little impact on the segmentation performance, they found that the choice of the dropout probability was critical, as the performance of their model significantly decreased for a dropout probability superior to $p = 0.50$. In Ghoshal et al. [102], authors propose to quantify both aleatoric and epistemic uncertainty using MC dropout in a task of nuclei segmentation in microscopy images and found that increasing the number of MC samples led to a decrease of the measured aleatoric uncertainty.

Other studies have proposed improvements to the standard MC dropout technique. In Jungo et al. [69], authors propose to use concrete dropout, a variant where the dropout rate at each layer is learned as part of the optimization process [103], but found no improvement as compared to the standard Bernouilli dropout. In a similar vein, [104] proposed a novel Spike-and-Slab dropout strategy, allowing to learn during training the dropout probability for each convolutional filter independently, for brain parcellation in T1-weighted MRI. Alternatively, [105] applied a variant of dropout called DropConnect [106], following which weights are

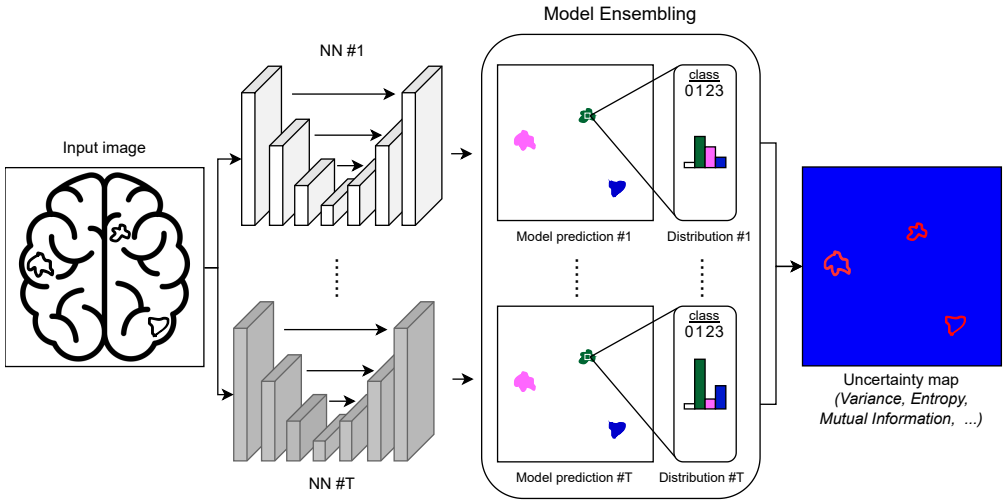


Figure II.2.6: Illustration of the Deep Ensemble framework. In its simplest setting, identical neural networks are trained on the same dataset. Due to the inherent stochasticity of gradient descent, the resulting trained models are different. The variability in their predictions can then be used to compute uncertainty scores.

randomly set to zero instead of activation, and demonstrated its advantages on various segmentation tasks, including organ segmentation from CT scans.

II.2.7 Ensembling methods

Deep Ensemble (DE, illustrated in Figure II.2.6) [107] proposes to quantify uncertainty from a series of sequentially trained NN. As the weights of the neural networks are initialized randomly, the models reach different optimum during training. As a result, they produce diverse predictions for the same query input. As for BDL and MC dropout, uncertainty estimates of both aleatoric and epistemic uncertainties can then be extracted from the ensemble's predictive distribution [108]. A DE does not require any changes to model architecture or training paradigm and is known for boosting predictive performance. Yet, it requires to repeat the training several times, which is cumbersome when the model is complex. Moreover, the aggregation of each individual prediction at inference increases the computational cost of this approach.

Ensembling techniques have been widely studied for UQ in medical imaging, with various studies demonstrating superior predictive performance and uncertainty-quantification quality, as compared to the MC dropout approach [109, 110, 111]. Noticeably, efforts are carried out to develop more efficient ensembling strategies that preserve the performance and uncertainty gain of the standard DE approach, without the prohibitive computational cost. Typical examples include multi-output architectures, that are able to produce a diverse set of predictions in a single pass using different branches or heads. This concept has been applied to both medical-image classification [112] and segmentation [94]. In a similar vein, Layer Ensemble [113] proposes to add a dedicated output to each intermediate layer of a segmentation model to form an ensemble within a single network. The same concept

is adopted in the Early-Exit Ensemble [114] and applied to medical-image classification tasks. Additional illustrative examples include Checkpoint Ensembling [115], which builds an ensemble from different checkpoints saved during the course of a single NN training. Alternatively, the Stochastic Weight Averaging Gaussian (SWAG) framework [116] can be viewed as an efficient way of ensembling. It aims at estimating a Gaussian approximate posterior over the weights of a NN by sampling its weight configurations during training, using a constant learning rate. At inference, it is possible to sample an ensemble of diverse models from this distribution. Two applications of SWAG can be found in the medical-image processing literature, for retinal artery-venous segmentation [117] and chest X-Ray classification [118], respectively. Finally, another research lead corresponds to developing techniques to improve diversity within ensembles, which is known to be a key factor of this technique [119, 120]. Two efforts in this direction are i) Orthogonal Ensembles [121], which optimize the orthogonality between ensemble member's weights to promote variety among predictions, and ii) diversity-promoting ensembles, composed of varied NN architectures explicitly chosen to minimize the correlation between their predictions [122].

Finally, it is worth noticing that some works propose to associate ensemble and MC dropout, forming the so-called Ensemble Monte Carlo (EMC) [123, 124, 90]. This allows to investigate two different types of uncertainty, namely i) uncertainty resulting from the random seed used to perform SGD training, yielding to different optima when sequentially training NNs, and ii) the weight uncertainty within each unique ensemble member assessed using the MC dropout approach.

II.2.8 Learning-based uncertainty quantification

Learned uncertainty (LU) frameworks are built on the idea that aleatoric uncertainty can be learned during training directly from the data itself. The most immediate approach, for segmentation tasks, is to treat the inter-rater variability as a ground truth for uncertainty. Supervised-learning strategies can then be adopted to reproduce the distribution of the raters annotations [125, 126, 127, 128]. However, this approach is limited to datasets where multiple ground truth segmentations are available per image, which is usually not the case. Most LU approaches have thus developed strategies to learn the segmentation and uncertainty conjointly without the need for explicit ground truth labels for uncertainty (see Figure II.2.7). In this direction, the initial proposal is to suppose that the network output logits z can be modeled by a Gaussian distribution parametrized by $\mathcal{N}(z; \rho, \sigma^2)$, where ρ and σ are the outputs of the NN, obtained by providing the model with two separate output branches [57]. High values of σ represent high heteroscedastic aleatoric uncertainty. To train the model to predict both quantities, a sampling-based uncertainty-aware loss is adopted. Illustrative applications of this approach can be found for Multiple Sclerosis lesions [129], tumor [130], and atlas segmentation [131] in brain MRI. This framework was recently extended to skewed Gaussian distributions in order to quantify asymmetric-contour uncertainty [132]. Another lead consists of learning uncertainty estimates directly correlated with the errors of the model, usually for segmentation applications. The starting observation is that errors are available at each training iteration by computing differences between the ground truth labels and the predicted labels. Thus, it is possible to use the computed-error maps as ground truth indicators for uncertainty. In McKinley et al. [133], authors present a modified segmentation

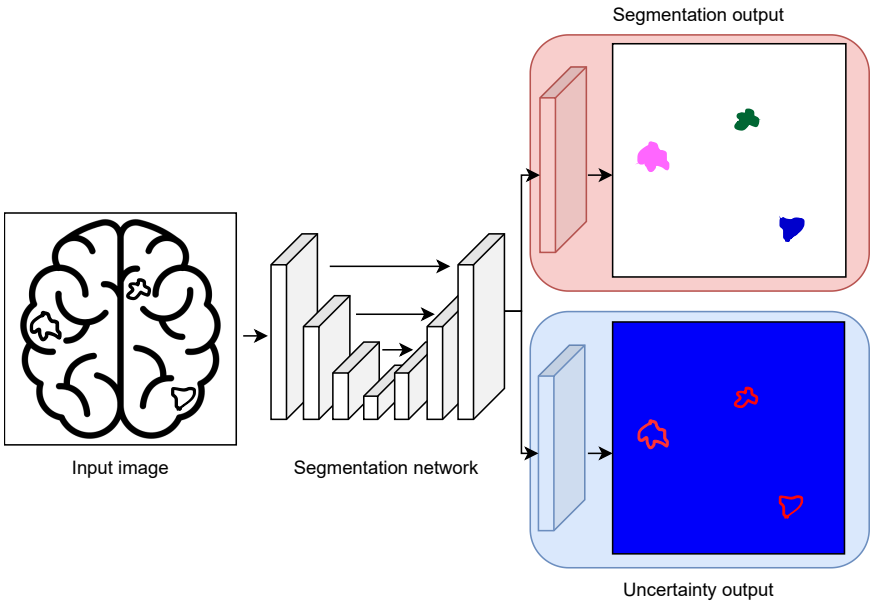


Figure II.2.7: Illustration of the Learned Uncertainty (LU) framework. A segmentation network is equipped with two output heads: one for the usual segmentation output (red), and one for the predicted uncertainty (blue). In practice, the model learns to predict both quantities (segmentation and uncertainty) using a dedicated loss function (e.g. Labelflip and Learned Confidence Estimates losses).

architecture with 2 output convolutions: one for the segmentation probabilities, and one for an uncertainty output called labelflip probability. They also propose the Labelflip loss in order to correlate this predicted uncertainty score with the incorrectly segmented voxels. The loss aims at reducing the weight of voxels at the interface between classes, that are inherently uncertain and inconsistently annotated in the ground truths. Instead, learning is emphasized for voxels that are incorrect but not inherently uncertain. An alternative approach is the Learned Confidence Estimates (LCE) loss proposed in Devries et al. [70]. In this framework, the segmentation model is also equipped with two separate outputs for the segmentation and a confidence estimate, respectively. This latter estimate is used to interpolate between the predicted probability distribution and the target distribution so that low-confidence voxels are pushed toward the correct output. Finally, [68] proposed the uncertainty cross-entropy loss, which is an extension of the standard cross entropy with an extra class corresponding to uncertain cases, for which no other class can be predicted with confidence. The loss can be minimized in two fashions, either by i) predicting the correct class label or ii) predicting the uncertainty class. The same motivation is at the core of the Deep Gambler model [134], recently applied to medical-image classification [135], that converts a $m-$class classification problem into a $m+1$ problem where the extra class represents abstention from answering.

II.2.9 Generative models

Image-conditional generative models have been explored for UQ in medical-image segmentation. The main objective is to generate various plausible and spatially-correlated segmentation

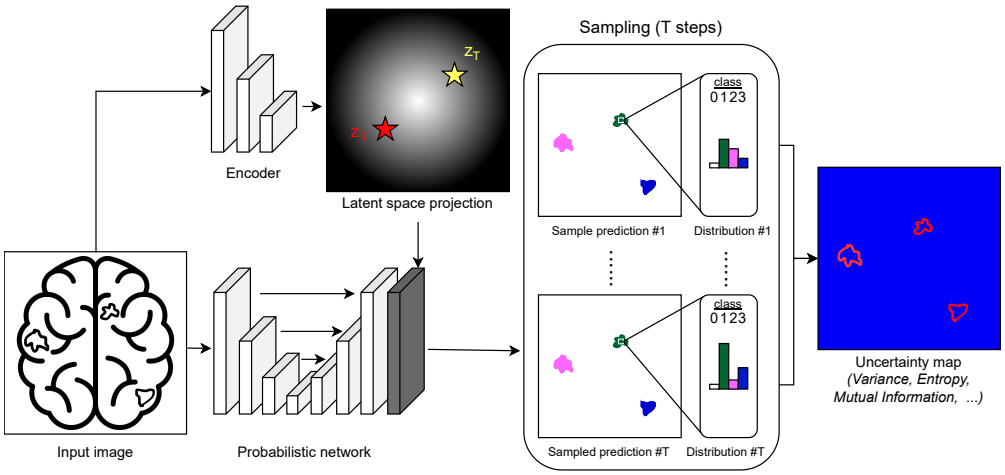


Figure II.2.8: Illustration of the Probabilistic U-Net framework. The input image is projected into a latent space using a dedicated encoder network. This latent space encodes the different plausible segmentations for the given input image. Samples from this latent space $z_1, ..., z_T$ are injected at the output of the segmentation network to produce a set of plausible masks, from which uncertainty estimates can be derived.

masks for a given input image. The first attempt in this direction was achieved with the Probabilistic U-Net [136] which proposed a segmentation architecture based on a Variational Autoencoder (VAE). This allows the encoding of the input image into several multivariate normal latent variables which are then decoded into diverse variations of the same region of interest. This process is illustrated in Figure II.2.8. Several improvements were then proposed to extend the expressivity of this generative model, such as the Hierarchical Probabilistic U-Net [137], the PHISeg [138], the RevPHISeg models [139], or by the insertion of Normalizing Flows [140, 141, 142]. Another interesting variant was proposed in the Stochastic Segmentation Network [143], which places low-rank multivariate normal distributions on the predicted logit space, allowing the sampling of a set of spatially-coherent segmentation for each input image. More recently, diffusion models were applied to this problem [144]. Note, however, that these different approaches are based on the sampling (either sampling several plausible masks at test time for Probabilistic U-Net and variants, or an iterative generative process for diffusion models), hence their computation cost is higher than that of standard segmentation models.

II.2.10 Test-time augmentation

Test-Time Augmentation (TTA, illustrated in Figure II.2.9) [145] was proposed as an UQ method to evaluate aleatoric uncertainty. At test time, multiple variants of the input image are generated using Data Augmentation. This can include spatial transformations (e.g. flipping, rotation) as well as intensity augmentations (e.g. contrast modification, noise injection, or artifacts). The model generates a prediction for each augmented variant of the input image. From this distribution, uncertainty metrics can be extracted such as the median or the variance. The TTA process aims to explore the impact of input-image transformations on the prediction. TTA is particularly interesting as it is completely model-agnostic: it

does not require any particular architecture or training design, and can thus be used with any pretrained or open-source model. Moreover, TTA can mimic the natural variability of medical imaging devices (contrast, noise level), which is thus relevant for medical tasks. Various TTA strategies were experimented for medical-image applications. In its simplest setting, only flipping is applied [72]. Noise and intensity shifts are also commonly added to the augmentation pipeline [146, 147, 68]. [145] gives an example of a more elaborate setting, where 128 variants are generated per input image using both extensive intensity and geometry transformations. In a more original manner, [148] proposed to use in-painting as TTA for uncertainty estimation. A downside of this approach is the increased cost of generating variants of high-dimensional images, which can be time-consuming for instance when using elastic deformations.

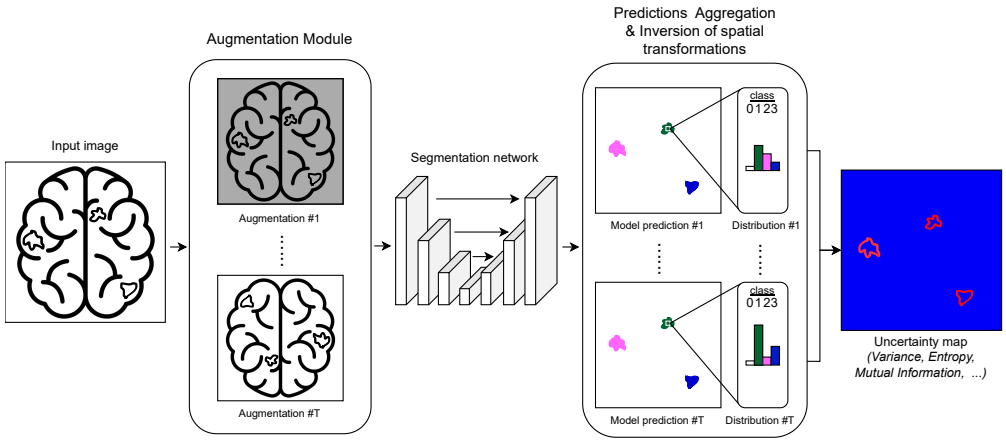


Figure II.2.9: Illustration of the Test Time Augmentation framework. An augmentation module generates multiple variants of the input image, which are each processed by the trained model. This produces a set of heterogeneous predictions from which uncertainty scores can be derived.

II.2.11 Latent-space OOD detection methods

From a practical point of view, epistemic uncertainty is expected to be high for Out-of-distribution (OOD) images, corresponding to images that are far from the training image distribution. Based on this, efficient epistemic-uncertainty techniques have been recently proposed to detect OOD images using the intermediate features of a trained NN [149]. This builds on the hypothesis that the feature maps computed when processing an input image contain information regarding its conformity. These methods, illustrated in Figure II.2.10, are computationally efficient and are increasingly experimented in medical-image processing applications. For instance, the Mahalanobis Distance was investigated to detect outliers in the context of COVID-19 lesions segmentation [72], X-ray classification [150, 71, 151] mammography classification [152] and more recently liver segmentation in T1-w MRI [153]. As alternatives, [154] proposed to study the spectral signature of the intermediate feature map by computing its Singular Value Decomposition in order to detect OOD images, while [68] computed class-wise prototypes from the feature representations in the context of brain tumor segmentation in MRI, allowing to detect train-test mismatches. Finally, we contributed

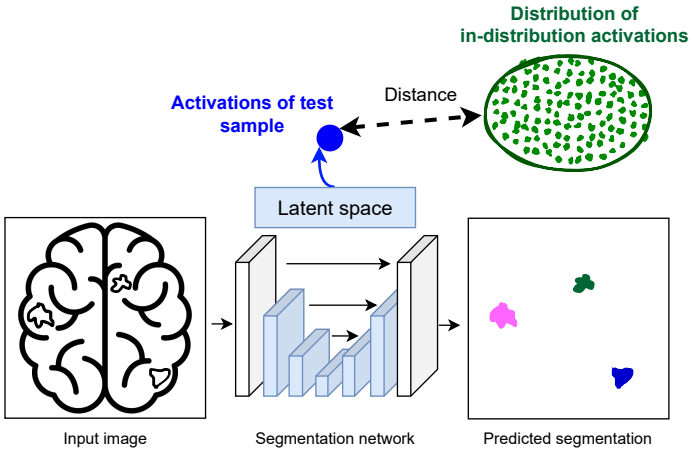


Figure II.2.10: Illustration of latent-space Out-Of-Distribution (OOD) detection. A compressed latent representation of the input image is collected from the intermediate activations of the network. A distance metric is then usually computed to estimate the distance between the test image and the training ones.

to this field by proposing a benchmark of the aforementioned solutions on a task of tumor segmentation in 3D brain MRI, and by analyzing the importance of the number of feature maps used to compute OOD scores [155]. This study is presented in Chapter IV of this thesis.

II.2.12 Evidential deep learning

The Dempster–Shafer Theory of Evidence (DST) is a framework for dealing with both epistemic and aleatoric uncertainty [156]. In a K-class segmentation problem, DST proposes to assign belief masses b_i^k to each possible class for each voxel i, as well as an overall uncertainty mass such that:

$$1 = \sum_{k=1}^{K} b_i^k + u_i \tag{II.2.5}$$

where $b_i^k > 0$ and $u > 0$. Beliefs are computed from *evidence* e_i^k, which is typically obtained by applying a Softplus operator to the raw logits predicted by the NN [157], such that:

$$b_i^k = \frac{e_i^k}{S} \text{ and } u_i = \frac{K}{S} \text{ with } S = \sum_{k=1}^{K}(e_i^k + 1) = \sum_{k=1}^{K} \alpha_i^k \tag{II.2.6}$$

where S is called the Dirichlet strength. When there is no evidence collected guiding to any of the K classes, the beliefs reach their minimal values 0, while the overall uncertainty reaches its maximal value 1. Finally, DST proposed to parametrize a Dirichlet distribution on the model's outputs in place of the categorical distribution, using the parameters $\{\alpha_i^1, \alpha_i^2, ..., \alpha_i^K\}$.

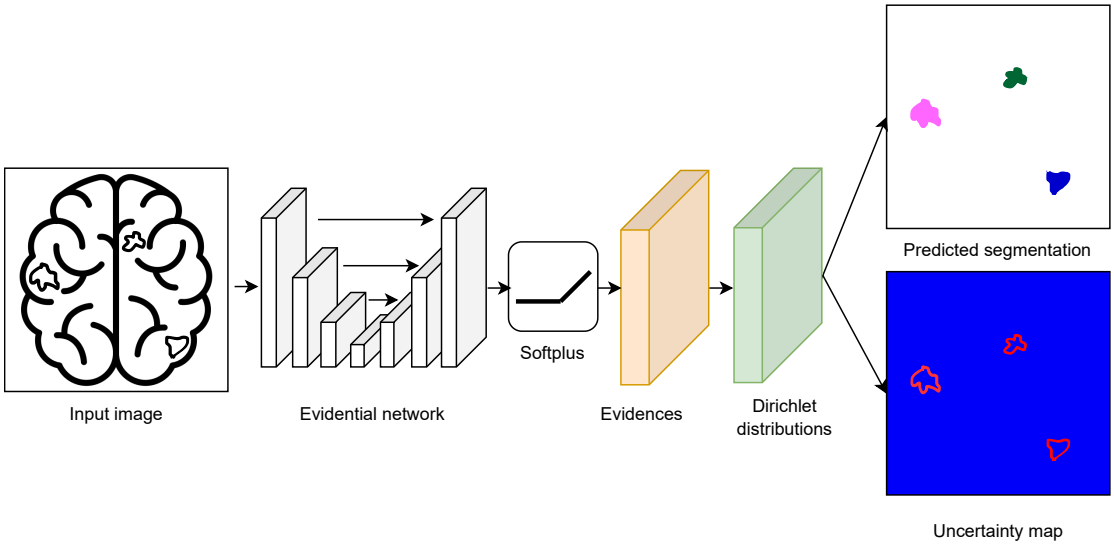


Figure II.2.11: Illustration of the Evidential Deep Learning uncertainty paradigm. A Softplus activation is inserted at the end of the network to generate evidence for each class and for each voxel. They are used to parameterize Dirichlet distributions, from which the segmentation and uncertainty estimates can be derived.

Interestingly, the realization of a Dirichlet distribution is still a distribution. It can be intelligently used to replace the standard categorical probability distribution of a classification NN by a distribution over possible Softmax outputs, thus modeling second-order probabilities and uncertainty [158]. It is thus much more expressive in terms of UQ than the standard Softmax probability framework (Section II.2.3). Different configurations of the Dirichlet distribution in the case of $K = 3$ are illustrated in Figure II.2.12, showing cases where epistemic or aleatoric uncertainties are high. Figure II.2.11 illustrates how EDL can be implemented in a segmentation network in practice.

In EDL, the probability that voxel i belongs to class k is obtained following [158]:

$$p_i^k = \frac{\alpha_i^k}{S} \tag{II.2.7}$$

and training is traditionally performed using the expected cross-entropy loss $\mathcal{L}_{ce}$ on the Dirichlet distribution [157, 158, 159]:

$$\mathcal{L}_{ice} = \mathbb{E}_{Dir(p|\alpha)} = \sum_{k=1}^{K} y_K \log p_k = \sum_{k=1}^{K} y_i(\psi(S) - \psi(\alpha)) \tag{II.2.8}$$

where ψ denotes the digamma function. A KL divergence term is usually added to the loss

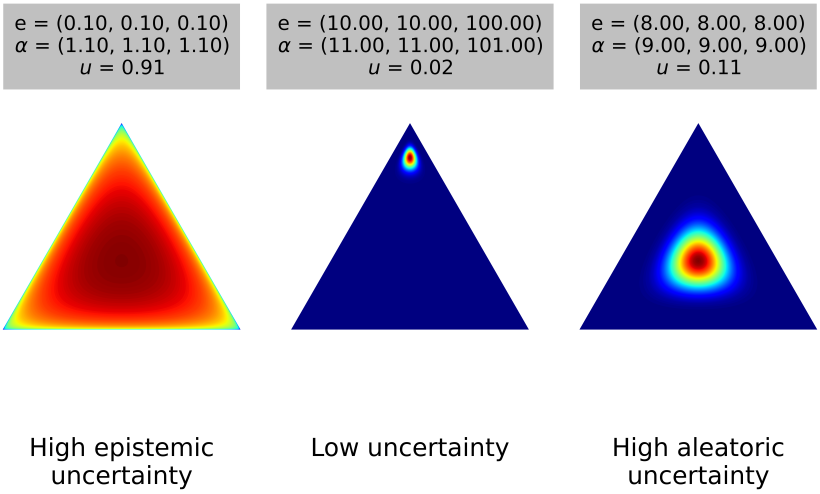


Figure II.2.12: Example of a Dirichlet distribution over categorical class probability distributions for $K = 3$. Left: a setting with high epistemic uncertainty, without collected evidence for any of the three classes. Center: Low uncertainty setting, with collected evidence confidently in favor of the third class. Right: high aleatoric uncertainty, with equivalent evidence for each of the three classes.

to make sure that incorrect predictions lead to less evidence [158, 157]:

$$\mathcal{L}_{KL} = \log \frac{\Gamma(\sum_{k=1}^{K} \tilde{\alpha}_k)}{\Gamma(K) \sum_{k=1}^{K} \Gamma(\tilde{\alpha}^k)} + \sum_{k=1}^{K} (\tilde{\alpha}_k - 1) \left[\psi(\tilde{\alpha}_k) - \psi(\sum_{k=1}^{K} \tilde{\alpha}_k) \right] \tag{II.2.9}$$

Aleatoric-uncertainty estimates can be obtained from the estimated probability, similarly to the Softmax uncertainty framework, while epistemic uncertainty can be evaluated using the u_i, which encompasses the accumulated evidence. DSL applications can be found for both medical-images segmentation [157, 160, 161] and classification [162, 163, 164].

II.2.13 Other UQ methods

Finally, a few UQ methods considered in this literature review do not conform to any of the previously introduced frameworks. In Jensen et al. [165], the authors explore the Monte Carlo Batch Normalization (MCBN) framework, a variant of MC dropout which makes use of the stochasticity of batch normalization layers. In Jungo et al. [69], an auxiliary net is proposed to detect the errors of a segmentation model, and the voxel-wise error probability is used as an uncertainty metric, allowing the decoupling of the uncertainty and segmentation tasks. [166] plug a Gaussian Process at the end of a DL feature extractor for diabetic retinopathy classification. Finally, Wang et al. [167] address contour uncertainty by replacing the binary segmentation masks with a soft alpha matte mask.

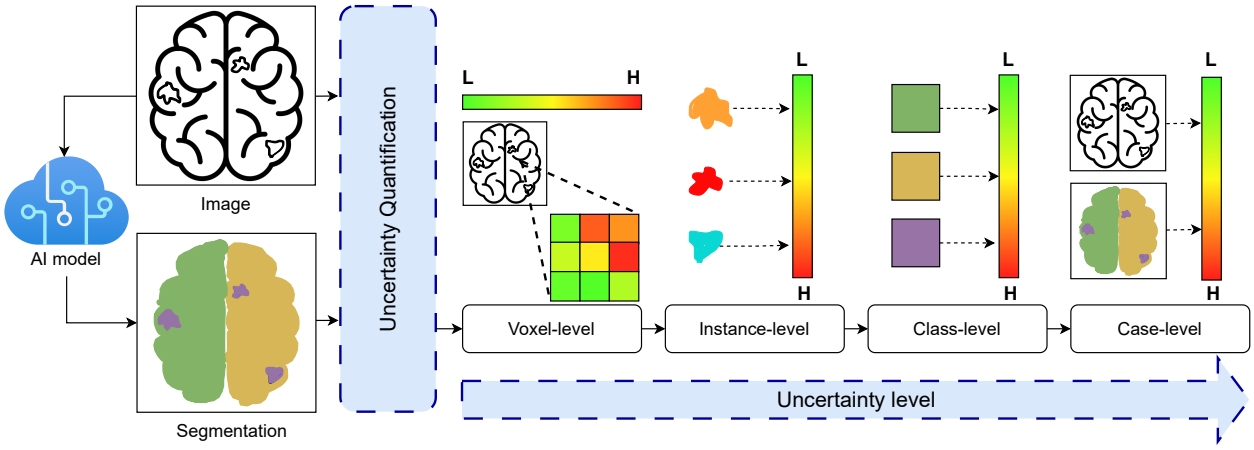


Figure II.3.1: Different levels of uncertainty in medical image illustrated on a 3-classes segmentation task, containing one lesion class (purple) and two anatomical classes (green and yellow).

II.3 From voxel uncertainty to lesion and case-level uncertainties

In the previous section, the most popular approaches for UQ in medical-image analysis have been introduced. In a segmentation setting, all methods except the *feature-based* approach produce voxel-wise uncertainty estimates when applied to 3D medical-image segmentation. While this is convenient for visualization purposes, this is not exactly aligned with medical attention, which is usually located at the structure level (e.g. lesion or anatomical region). Moreover, when processing 3D medical images, the visual inspection of the entire volume to monitor uncertain areas can be time-consuming. A branch of the medical-imaging UQ literature has thus focused on estimating uncertainty estimates at higher levels. These structural uncertainty scores have been mainly explored in two settings: i) the binary segmentation of lesions, for which an uncertainty score is assigned to each identified lesion (Section II.3.1), and ii) case-level QC II.3.2, where the goal is to identify non-conform images (input QC) or poor predictions (output QC). These different scales are illustrated in Figure II.3.1, and the corresponding literature is further presented in the following.

II.3.1 Lesion-level uncertainty estimates

Lesion-level confidence estimation is an emerging UQ application that consists in attributing a single uncertainty score for each identified lesion in a medical image. This is relevant for applications that rely on the detection of multiple lesions, and for which precise counting is required. A pioneering study in this direction is the work of Nair et al. [129] which proposes to fuse voxel-level certainty scores to lesion-level scores, for MS-lesion segmentation. The proposed strategy consists of using the *logsum* operator on the uncertainty of the voxels composing each lesion. One downside is that this metric systematically attributes higher uncertainties to smaller lesions. However, in practice, false positive detections are usually

small lesions, which is why this operator provides useful estimates. Other aggregation operators have further been proposed for MS, such as the average of voxel uncertainties [168]. In a different direction, several plausible masks can be obtained for each unique MS lesion using a DE, and the disagreement between the ensemble members has been proposed as a lesion-level uncertainty score, although not using directly the voxel uncertainty maps [168].

Lesion-level estimates were also explored for liver-tumor lesions [169, 170]. In these two studies, the authors start by computing voxel-uncertainty maps, for example using MC dropout, DE or TTA. Then, they compute radiomics [171] for each lesion, using the uncertainty maps. These features are used to train a classifier (SVM) to predict the status of the lesion: True Positive (TP) or False Positive (FP). While their primary focus is the reduction of FP liver lesions, it seems that the FP probability of the lesion constitutes a viable lesion-level uncertainty score, where higher values are associated with more ambiguous lesions. A similar strategy was proposed by Ozdemir et al. [172] for FP reduction in nodules detected in lung CT. The CT image is first processed by a segmentation model, providing a segmentation and uncertainty map. Bounding boxes are further extracted from the segmentation, centered at each identified nodule. An auxiliary CNN classifier then predicts the probability that the lesion is a False Positive.

II.3.2 Case-level uncertainty estimates

A frequent question when dealing with uncertainty is to wonder if the *overall* prediction can be trusted. To answer this, case-level uncertainty scores can be furnished to the user to provide a general impression regarding the model confidence for a given case. In the context of medical image segmentation, such scores can actually be computed to detect non-conform input (input-level QC) or poor-quality output segmentation (output-level QC).

II.3.2.1 Input Quality Control

Epistemic (i.e. model) uncertainty is expected to be high for images that are significantly different from the ones encountered during training [57]. Thus, monitoring the output uncertainty of predictive models could theoretically be used to detect poor-quality input images. This idea was explored in McClure et al. [104], where authors used MC dropout to estimate the voxel uncertainty of a brain-tumor segmentation model. From these uncertainty maps, image-level scores are derived by averaging voxel scores across the volume. The scores are then used to detect poor-quality scans. A similar process is adopted by Gonzalez et al. [72] for non-conform input detection in chest CT segmentation, using MC dropout and TTA as voxel uncertainty estimators. In line with these works, the previously presented feature-based OOD detection methods (Section II.2.11) propose case-level scores that can be used to perform input-level QC.

II.3.2.2 Output Quality Control

Output-level uncertainty estimates aim at evaluating the overall quality of an automated segmentation. An intuitive solution is to fuse the voxel uncertainty estimates computed by a standard UQ methodology (MC dropout, DE, TTA...) to a case-level score, for example, using the mean [173]. Following the observation that voxel uncertainty is often concentrated

at the boundaries between classes, two studies have proposed to reduce the weight of contour voxels to get more accurate structural uncertainties. This *prior knowledge-based aggregation* was investigated in Jungo et al. [69] and Graham et al. [174].

Instead of relying on voxel uncertainties, a series of studies have proposed to focus on the set of plausible segmentation masks generated by standard UQ methodologies. Pioneering work in this direction is the study carried out by Roy et al. for whole brain segmentation [173]. They use MC dropout to generate a set of segmentation masks for each input image and use the disagreement between the samples as an estimate of structural uncertainty. This follows the intuition that disagreement between the predictions should be higher for poor predictions. In this direction, they propose different proxies: the Coefficient of Variations among the volumes (CoV), the Dice & IoU agreements between the MC samples, and the mean voxel uncertainty in the segmented volume. They further show that these proxy metrics correlate strongly with the true Dice, unknown during inference. This concept of using a set of plausible segmentation masks to compute structural uncertainty metrics was further explored via MC dropout sampling [175, 176] (Section II.2.6), Deep ensemble [177] (Section II.2.7) and TTA [146, 178] (Section II.2.10). Other proxies were proposed, including the Predictive Dice Coefficient [175], the Contour Quality metric [179] or the Doubt score [180], all of which have demonstrated a strong correlation with the true Dice coefficient.

To further improve the output QC procedure, several studies have explored the use of these uncertainty metrics as features to train a ML model to directly infer the prediction quality, in a regression setting. Ghosal et al. [181] and Hann et al. [182] trained linear regression models to predict the Dice directly from uncertainty estimates of MC dropout models, for digital histopathology image segmentation and cardiac-MRI segmentation, respectively. Alternatively, Arega et al. [183] used a Random Forest (RF) either in a binary classification approach (accept/reject poor segmentation) or regression (predict the Dice score) from the outputs of a MC dropout model. These approaches require building a training dataset comprising automated predictions together with their associated quality to allow the training of the auxiliary ML model.

II.4 How to evaluate uncertainty quantification approaches

In the previous sections, the main UQ approaches applied to DL-based medical-image classification and segmentation were presented. To compare them, different evaluations have been proposed in the literature. Evaluating UQ approaches is not straightforward, as there are generally no ground truth uncertainty values. Proxy metrics are thus developed to circumvent this limitation. More precisely, 6 different types of evaluation protocols can be identified in the reviewed papers (see Figure II.4.1). In the following, we present each protocol and identify their use cases.

II.4.1 Qualitative assessment protocol

As computing quantitative metrics for uncertainty is not direct, several works focused on a qualitative assessment of the predicted uncertainty estimates. In this context, a visual

inspection of the cases considered as certain/uncertain is usually performed to verify whether they correspond to cases that a human would consider as such [184, 185, 186]. Alternatively, the relevance of the incorporation of UQ in a medical-image processing pipeline can be assessed via the monitoring of its beneficial impact on a downstream task such as active learning [187], curriculum learning [188, 189, 190], weakly-supervised learning [191], semi-supervised learning [192, 193, 194, 195, 196], cascaded inference tasks [197, 198], segmentation refinement [199], federated learning [200], cross-domain generalization [201], or predictive performance [202].

II.4.2 Calibration metrics

Calibration metrics are designed to evaluate the reliability of predicted probability estimates. In the previous chapter, the reliability diagram was presented as a way to verify the correspondence between predicted probabilities and actual error rates (see Section I.6 and Figure I.6.1). From this graphical representation, the popular Expected Calibration Error (ECE) score can be derived to quantitatively estimate calibration [49, 172, 164, 69, 67, 110]:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \text{acc}(B_m) - \text{conf}(B_m) \Big| \tag{II.4.1}$$

where $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are the average accuracy and average confidence, respectively, in the m-th bin. n designates the total number of test samples (images for classification, voxels for segmentation). The ECE was initially introduced for classification tasks. However, for medical-image segmentation, it has some pitfalls. Indeed, the majority of voxels correspond to the background class, which is generally segmented correctly and with very high confidence. Thus, it overestimates the true calibration of the NN. To circumvent this, recent works on calibration compute the calibration metrics only on the foreground classes [67].

Alternatively, proper scoring rules can be employed to evaluate calibration, corresponding to metrics that are minimized when the model predicts probabilities that are consistent with the true-event probabilities. Popular scoring rules include the Negative Log-Likelihood (NLL) score [109] (introduced in Equation I.2.3) and the Brier score [203], framed as:

$$\text{Brier} = \frac{1}{K} \sum_{k=1}^{K} (p_{i,k} - y_{i,k})^2 \tag{II.4.2}$$

where i is the voxel index and K the number of classes.

II.4.3 Coverage error

As presented in Section II.2.3, Conformal Prediction (CP) is traditionally defined around the notion of *coverage*, following which a fraction of the ground truth labels should be included in the predictive sets (e.g. 95% or 99%). A natural way of evaluating uncertainty under this

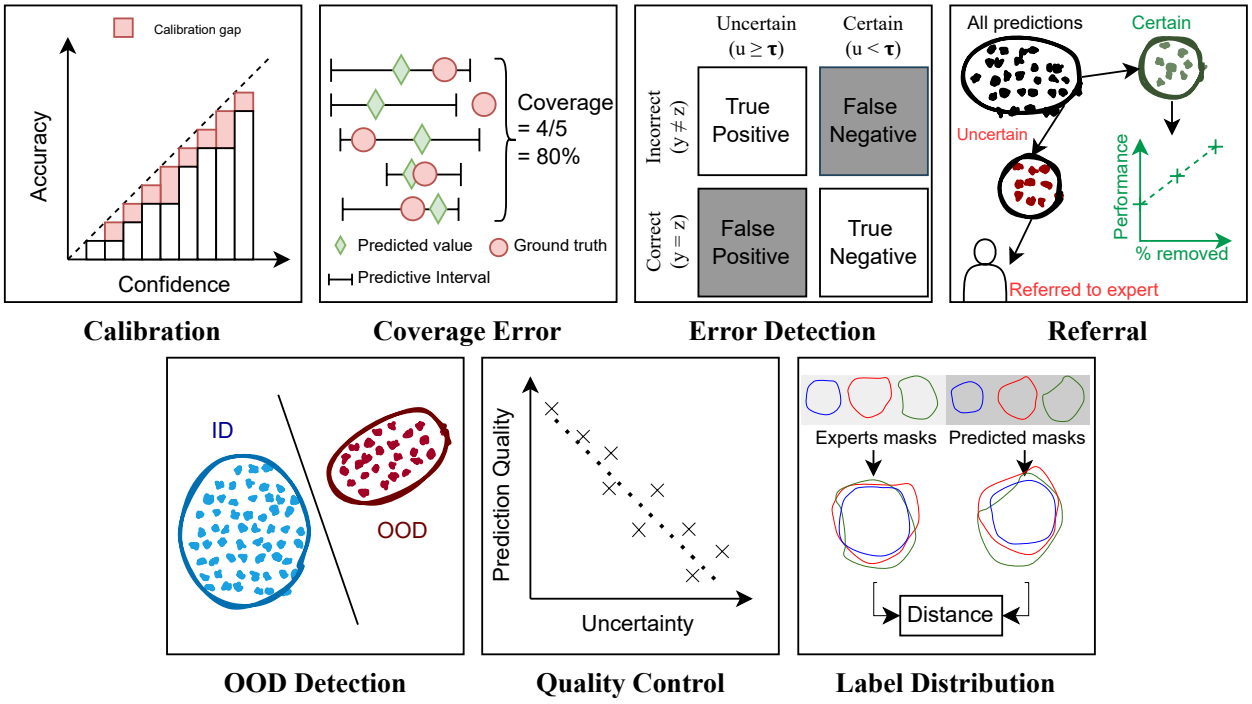


Figure II.4.1: Illustration of the different approaches used to estimate the quality of uncertainty estimates. See text for details.

framework is to compute the distance between the empirical coverage on test data and the user-defined target coverage [76]. If implemented properly, CP is statistically guaranteed to approximate the target coverage. However, this can be achieved with unnecessarily large intervals. Let's take the example of predicting the volume of a brain lesion, that we want to equip with a predictive interval using CP. A perfect coverage of 100% could be achieved by predicting a lower bound of 0 ml, and an upper bound corresponding to the overall intracranial volume. However, these intervals would be useless in practice. Thus, CP methods are also generally evaluated with respect to the average interval width, where narrower values are preferred [81].

II.4.4 Error detection and referral

A direct downstream application of uncertainty in an automated pipeline is the detection of samples for which the prediction is likely to be incorrect. This is crucial to prevent silent errors that could have a dramatic impact, especially in real-world medical-image applications. Error detection is thus commonly used to estimate the quality of uncertainty estimates. In this scenario, the model's predictions are classified into two groups, certain and uncertain samples, by setting a threshold on their associated uncertainty. The result of this classification is then compared to the correctness of each sample, namely correct or incorrect. In that context, a confusion matrix from the uncertainty point of view can be constructed, by distinguishing 4 possible cases, as shown in Figure II.4.1 (*Error Detection* case, where τ is the set threshold). Usual classification metrics (e.g. Accuracy) can then be computed based on the counts of

each case: i) True Positive (TP) cases when the classification is uncertain and the expected label and the prediction differ; ii) False Negative (FN) cases when the classification is certain but the expected label and the prediction differ; iii) True Negative (TN) cases when the classification is certain and the expected label and the prediction are identical; and iv) False Positive (FP) cases when the classification is uncertain but the prediction and the expected label are identical. Illustrative applications of this metric can be found for COVID-19 detection [111], cell colony segmentation [204] and skin-disease assessment [205]. In a similar vein, but specifically for image segmentation, the uncertainty-error overlap was also proposed [69] and further extended using mutual information [206]. Other variants include the use of distance metrics such as the Wasserstein distance [207] or Jensen-Shannon [208] to measure how much the predicted uncertainty correlates with the distribution of the model errors.

Another variant of this framework is the referral mechanism (also sometimes referred to as rejection or filtering) [209]. In this context, predictions of the model are ordered from the most certain to the most uncertain. A fraction of the most uncertain predictions are then rejected (for instance, referred to the expert), and the performance of the model is computed on the remaining predictions. If uncertainty estimates efficiently identify as uncertain the cases that are more likely to be incorrect, then the error rate on the remaining prediction should decrease. Multiple fractions can then be used in this way, producing a curve showing the error rate of the model with respect to the fraction of rejected data. The area under the resulting curve is used as a qualitative score. This referral-based evaluation protocol aims at mimicking a human-in-the-loop process where the model abstains on uncertain predictions, which are eventually redirected to an expert for correction. It essentially highlights the same trends as the previous misclassification detection setting. Such metric was for instance used for MS-lesion segmentation in brain MRI [129], cardiac MRI segmentation [210], brain stroke detection [211] or diabetic-retinopathy detection [212]. Interestingly, such referral metrics were also used in the context of the QU-BraTS 2020 challenge focusing on UQ for brain-tumor segmentation in MRI [213], as well as in the more recent SHIFT 2023 challenge focusing on MS-lesion segmentation [214].

II.4.5 Out-of-Distribution detection protocol

A desired property of uncertainty is to be high in the context of a train-test mismatch, occurring when input images are significantly different from the images seen during training. Similarly to the misclassification-detection setting, the uncertainty estimates can be translated into a binary classifier that aims at distinguishing between in-distribution (ID) and OOD images. Standard classification metrics can further be computed.

Protocols for OOD detection can be characterized based on the type of OOD data used. The most obvious setting corresponds to *far OOD* data, corresponding to samples that share little to no similarity with the training data. For instance, [72] proposes to train a model to segment COVID lesions in chest CT, and further use colon and spleen CT as OOD data. A more realistic shift in the context of medical-image processing is *diagnostic shift*, where a disease unobserved during training is included in the test images. This setting has been explored in Berger et al. [71], where authors train a binary classifier to distinguish between cardiomegaly and pneumothorax in chest X-ray, then use images with fracture as OOD data.

Similarly, in the context of digital pathology detection, [112] and [110] train breast-metastasis detection models and include images with new unseen abnormalities at test-time as OOD data. Combalia et al. [215] train an 8-class classifier on the ISIC 2019 dataset to detect skin disease, which also contains a test set of OOD images belonging to none of the 8 classes for OOD evaluation. *Modality shifts* can also be encountered in medical-image processing tasks, where the imaging acquisition protocol is different between training and testing images. Tardy et al. [152] trained a 2D mammography classification model and used images acquired from a different manufacturer at test time, simulating a common data shift encountered in clinical routine. Calli et al. [150] consider posteroanterior chest X-rays as training images and tested OOD detection on anteroposterior images. Finally, *Transformation shifts* have also been investigated, where transformations are applied to the input images to push them away from the training images distribution. For instance, Gonzalez et al. [72] generate OOD data by applying affine transformations and synthetic artifacts to the input images. This simulates image quality degradation or the presence of artifacts that can complicate analysis.

II.4.6 Quality control

For segmentation tasks, uncertainty is expected to be higher for poorly-segmented images than for well-segmented ones. Based on this desired property, several works studied the correlation between image-wise uncertainty scores (in contrast with the usual pixel-wise estimates) and the segmentation quality, such as the Dice score. In an automated medical-image segmentation pipeline, this process can be used to detect images for which the produced segmentation does not meet quality standards. We refer to this mode of evaluation, specific to segmentation tasks, as QC-based evaluation protocols. In this setting, the goal is generally to maximize Pearson's correlation between the true segmentation quality and the proxy score [173, 146, 177, 216, 175, 179, 113]. However, we found that such QC-based protocols usually only focus on the correlation with the Dice score, which is correlated to the size of the segmented object [31, 217]. As a result, these QC protocols may demonstrate a correlation between the size of the segmented region and the uncertainty level (with smaller regions being more uncertain), rather than the true quality of the segmentation. Studying the correlation with other segmentation metrics is only rarely envisaged, with an exception in Kushibar et al. [113], where authors demonstrate a correlation for both the Dice and Haussdorf metrics.

II.4.7 Label-distribution protocol

Finally, for segmentation tasks where several expert delineations are available per image, label-distribution metrics can be employed. This consists of comparing the predicted distribution of labels P_{out} with the ground truth distribution of the experts P_{gt}, and is thus commonly used paired with Generative UQ models presented in subsection II.2.9. A popular choice of metric is the Generalized Energy Distance [136, 137, 138] between both distributions. Other proposed metrics include the normalized cross-entropy [128], the normalized cross-correlation [138] or the weighted mean of the predictive entropy between the true and predicted uncertainty maps [218].

The distribution of the UQ evaluation protocols in the studied corpus of papers is shown in Figure II.4.2.

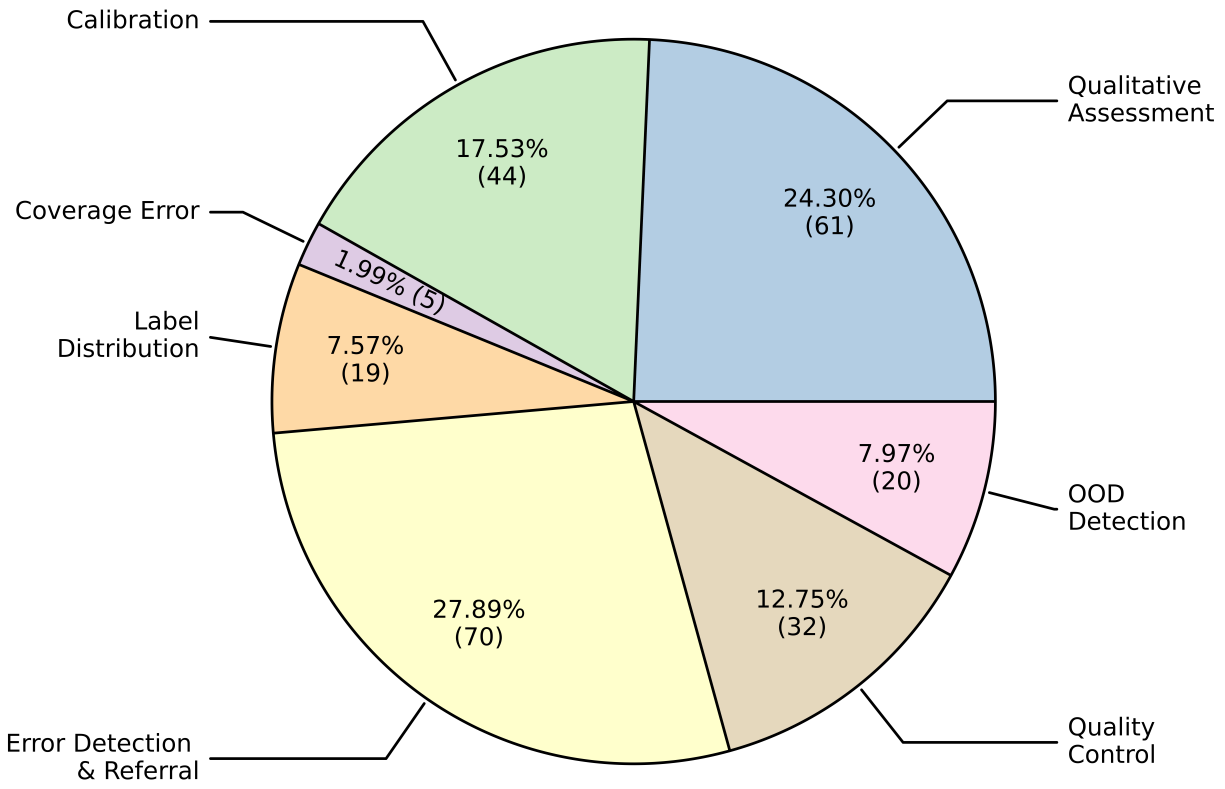


Figure II.4.2: Implemented UQ evaluation protocols in the reviewed papers. The percentage (and number) of the reviewed papers per class is mentioned in the Pie chart.

II.4.8 Distinguishing aleatoric and epistemic uncertainties during evaluation

While the distinction between aleatoric and epistemic uncertainties is possible when defining UQ approaches, this distinction is less clear when dealing with evaluation. Aleatoric and epistemic uncertainty estimates are often compared using the same metrics [68, 69, 157, 206, 219]. Nevertheless, some evaluation scenarios specifically focus on one particular side of uncertainty. For instance, OOD detection is clearly linked to epistemic uncertainty, while label distribution is akin to aleatoric uncertainty. Other metrics including Coverage Error, Error Detection, Referral, and Quality Control do not make any assumption about the source of error, which could come from a noisy data point (aleatoric uncertainty) or the lack of knowledge of the model (epistemic uncertainty), and hence cannot be associated with either of them. Calibration metrics, that consider the output probabilities of the model, could be cast as a way to evaluate aleatoric uncertainty. However, there is an increasing literature on the calibration under domain-shift, which bridges the gap with epistemic uncertainty [220, 221, 222, 223].

II.5 Discussion on the literature review

Now that the most popular UQ methods and evaluation schemes for DL-based medical-image analysis have been reviewed, we can take a step back and analyze the major trends in this field.

First, the large number of studies (228) that have been included in this review proves that the need for UQ is well taken into account by the DL community. This shows that efforts are being made to develop responsible and understandable AI that can be used in real clinical settings, without only focusing on the raw prediction accuracy.

Considering **UQ methods**, it appears that, although Bayesian methodology provides a strong theoretical background for uncertainty, it is scarcely implemented for medical-image analysis (only 2.06% of the papers, see Figure II.2.1). This can be explained by the complex implementation that requires the modification of the NN and training paradigm. Less formal approximations of the Bayesian framework, such as dropout-based methods, are thus generally preferred. Overall, MC dropout method seems to be the most popular approach for UQ in medical-image analysis, representing more than a third of the implemented methods (38.05%), considering both the standard MC dropout methods (34.81%) as well as MC dropout Ensemble models (3.24%), which are an MC-dropout extension. This popularity can be explained by its easy implementation in any NN trained with dropout. Additionally, dropout helps prevent over-fitting during training, which is a common problem in the medical domain, where the training dataset size is limited. However, the performance of MC dropout is highly dependent on the applied dropout rate [101, 224], which can make it impractical to tune. Moreover, it requires multiple inferences for the same input image, increasing the inference time, which may not be compatible with AI applications in clinical routine. Ensembling approaches are also commonly employed for UQ (16.22% of the implemented UQ methods), although less common than MC dropout models. Aggregating the predictions of multiple models is a

well-known trick to improve predictive performance, while also providing quality uncertainty estimates. The drawback is an increased computational cost and time, as it requires multiple training and the aggregation of their predictions at inference. Other popular UQ metrics are Softmax probability (12.98%), which provides intuitive and easy-to-use uncertainty estimates, and Learned Uncertainty methods (8.26%) which propose to learn aleatoric uncertainty from data. Finally, we note (see Figure II.2.1) that most implemented methods simultaneously estimate aleatoric and epistemic uncertainties (61.95%), while a third only evaluated aleatoric uncertainty (32.74%), and only a few simply considered epistemic uncertainty (4.13%). It can also be noted that the majority (67.83%) of the implemented UQ methods are based on a sampling protocol (MC dropout, Deep Ensemble, BNN, MC dropout Ensemble, TTA, and Generative models), aiming at generating multiple plausible predictions for the same query input. Yet, this process may significantly increase the computational burden of UQ, especially when processing large 3D volumes, which may prevent its adoption in an automated pipeline in the medical domain (where latency is a practical concern). Single-step UQ methods such as Softmax (calibrated) probabilities, Evidential Deep Learning, or Features-based methods are thus promising especially for time-critical applications.

UQ evaluation protocols. In the literature, a large variety of evaluation protocols are reported, aiming at assessing the quality of uncertainty estimates. In the context of medical-image segmentation, if multiple manual expert delineations are available for a given input image, the inter-rater variability can be used as ground truth uncertainty, to be compared with the one predicted (representing 7.57% of the implemented evaluation protocols, see Figure II.4.2). However, most of the time, such an uncertainty gold standard is not provided. Thus, the evaluation of UQ usually relies on proxy tasks, such as the detection of errors (27.89%), Quality Control (12.75%), or Out-of-distribution (7.97%). These methods are inspired by concrete applications of uncertainty in a real-world scenario. Yet, although commonly used, UQ evaluation based on error detection is not ideal for ranking methods. Indeed, the set of correct and incorrect predictions is specific to each predictive model. It is then inappropriate to compare them directly [225]. Calibration evaluation metrics are also commonly used (17.53%). The use of such metrics seems particularly interesting because many popular uncertainty estimates, such as variance, entropy, or mutual information, can be directly extracted from probability distributions. Thus, guaranteeing that the probability estimates are well-calibrated seems to be an essential prerequisite to obtaining meaningful uncertainty estimates.

Finally, it must be acknowledged that the effort of the community is promoted by the organization of uncertainty-oriented challenges. The 2020 edition of the BraTS challenge included an uncertainty task (QU-BraTS): participants were expected to provide brain tumor segmentation models that are able to provide voxel-wise uncertainty estimates correlating with segmentation errors [213, 226]. Furthermore, the MICCAI QUBIQ challenge[1], hosted in 2020 and 2021, focused on label uncertainty. Participants were provided with images annotated by several experts, on a variety of medical-image segmentation tasks, and were asked to develop methods able to reproduce the annotation distribution from the different raters. Finally, the SHIFT 2023 challenge[2] contained a task of uncertainty quantification for

[1] https://qubiq21.grand-challenge.org/
[2] https://shifts.grand-challenge.org/

MS-lesion segmentation [214]. The challenge focused on the development of models robust to train-test mismatches and uncertainty was evaluated through an error-detection setting. Our participation in this challenge, which ranked at the second position, is further presented in Appendix (A7).

II.6 Benchmark of voxel-level uncertainty estimates for brain MRI segmentation

In this chapter, the flourishing literature on UQ for medical image analysis has been presented. The 6 most popular approaches are further retained for benchmarking: Softmax, MC dropout, Deep Ensemble, TTA, EDL, and Learned uncertainty. Altogether this pool of methods regroup 83.28% of the reviewed UQ strategies. This benchmark has two main objectives: i) selecting a baseline segmentation loss function that will be used to train the segmentation models in this thesis, that shouldn't hurt the calibration of the models and ii) selecting a voxel-level uncertainty baseline.

II.6.1 Benchmark materials

The benchmark is based on three tasks with different characteristics: segmentation of WMH in brain T2w FLAIR MRI, multi-class tumor segmentation in multi-modal brain MRI, and binary stroke lesion segmentation in T1w brain MRI.

II.6.1.1 MS lesions segmentation in brain T2w FLAIR MRI

Pathology Description Multiple Sclerosis is a neurodegenerative, demyelinating disease causing damage to the nerve fibers in the brain as well as the optic nerves and spinal cord. It affects 2.8 million individuals worldwide in 2020 [227]. The diagnosis relies on clinical symptoms (including balance, speech, or reflex impairments) and is supported by MRI findings. The role of imagery is to highlight lesions within the brain white matter, which are disseminated both in space and time for MS [228]. The preferred MRI sequence is a T2-weighted FLAIR image, in which MS lesions appear as White Matter Hyperintensities (WMH).

Data Description T2w FLAIR MRI of MS patients are collected from several open-source datasets, along with the manual ground truth annotations of WMH: MSSEG-1 [229] (53 images), ISBI 2015 [230] (21 images), MSLUB [231] (30 images) and WMH Challenge (170 images). It results in a total of 274 images that are stratified as presented in Table II.1. A total of 149 images are used to train the models, 21 to perform validation and model checkpointing, 49 for an in-distribution test dataset with images from the same distribution as training images. Note that contrary to the other datasets, ISBI 2015 is a longitudinal dataset and actually contains multiple imaging visits (4 or 5) of a limited pool of 5 patients. To avoid train-test contamination, the 4 visits of patient 1 are used for validation, the 4 visits of patient 5 are kept for ID testing, and the rest of the patients (2, 3, 4) are used to training. We then evaluate two domain-shift (DS) datasets to validate the robustness of the models. MSLUB [231] is a test dataset containing images from a site unseen during training, while

Dataset	Centers	Devices	Train	Val	Test ID	Test DS 1	Test DS 2	Total
MSSEG-1	Rennes Bordeaux Lyon	Siemens Verio 3T GE Discovery 3T Philips Ingenia 3T Siemens Aera 1.5T	23	4	11	0	15	53
ISBI 2015	Best	Philips Tesla 3T	13	4	4	0	0	21
WMH 2017	Utrecht Singapore Amsterdam	Philips Achieva 3T Siemens Trio 3T GE Sigma HDxT 3T Philips Ingenuity 3T GE Sigma HDxT 1.5T	124	13	23	0	10	170
MSLUB	Ljubljana	Siemens Trio 3T	0	0	0	30	0	30
Total			150	20	49	30	25	274

Table II.1: Data sources and stratification for the Multiple Sclerosis experiments. ID: In-distribution, DS: Domain-shift.

the 1.5 Tesla dataset contains images from seen sites (Lyon and Amsterdam) but that were acquired at 1.5 Tesla, in contrast to the training images acquired with 3 Tesla MRI devices. All images are preprocessed uniformly, comprising a resampling to a $1\,\mathrm{mm}^3$ resolution and skull-stripping using HD-BET [232].

II.6.1.2 Tumor segmentation in multi-modal brain MRI

Pathology Description Glioblastoma is the prevalent form of brain tumor, representing 60% of brain tumors in the adult population [233]. It affects 0.59 to 5 per 100 000 persons [234], and it is associated with a poor prognosis: the median survival is 14 to 15 months starting from the diagnosis. Treatment relies greatly on surgery, which can be complemented by radiotherapy and/or chemotherapy [235]. The diagnosis and follow-up of the disease relies on MRI imaging. The standard protocol consists in a multi-parametric MRI acquisition, including T1w, T2w, T1 with contrast enhancement, and FLAIR sequences [236].

Data Description The task is based on the large-scale BraTS 2023 dataset [237, 226] for glioblastoma segmentation from brain MRI, comprising T1w, T2w, FLAIR, and T1 with contrast enhancement (T1ce). This setting allows the exploration of multi-class segmentation: background, necrosis, edematous, and GD-enhancing tumor (GDE). The 2023 edition of the dataset is used, comprising 1133 subjects, stratified in 876 for training, 30 for validation, and 227 for ID testing. The 2023 challenge also included various auxiliary datasets allowing to evaluate the robustness of the model on DS settings. For this experiment, the BraTS Africa [238] dataset ($N = 60$) is used, comprising sub-saharan (SSA) patients imaged with lower quality devices and presenting more advanced stages of the disease. Due to this shift in both image quality and population, the SSA dataset is particularly suitable to test the robustness of the model. All images are provided pre-processed, including resampling to a $1\,\mathrm{mm}^3$ resolution, brain extraction, and registration of the MRI sequences to a common anatomical template.

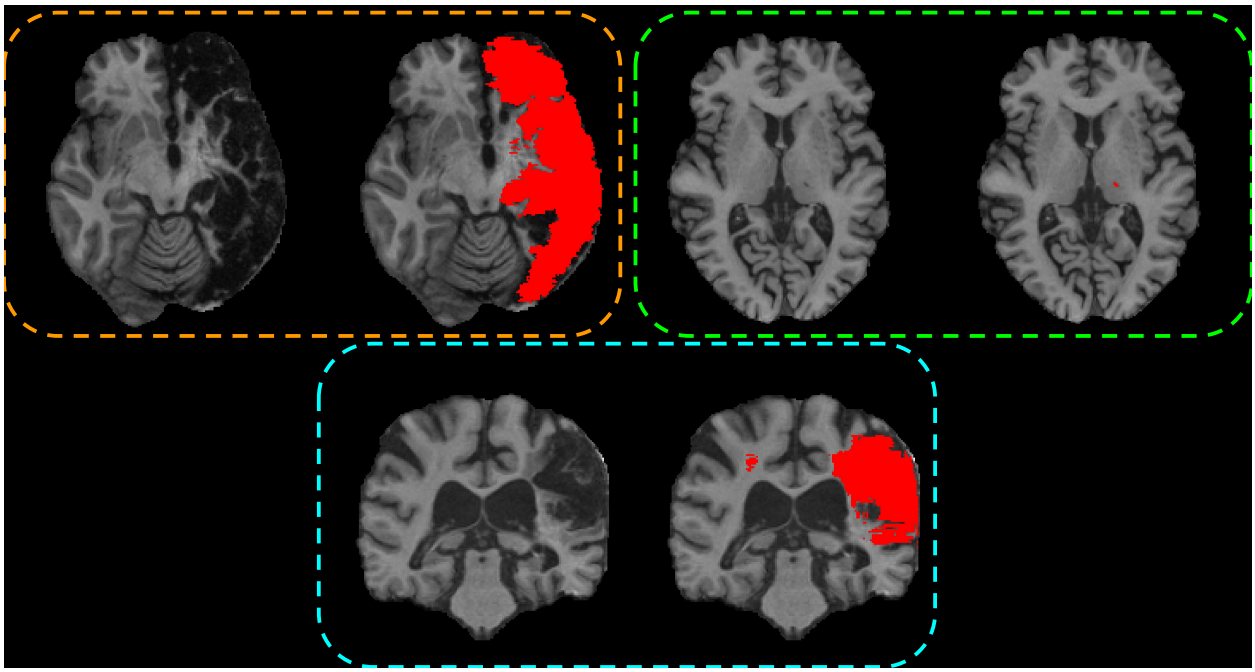


Figure II.6.1: Samples from the ATLAS-2 dataset. Orange box: a patient with a single large lesion. Green box: a patient with a single small lesion. Cyan box: a patient with two lesions. Label uncertainty can be visualized in the Orange and cyan examples.

II.6.1.3 Stroke lesion segmentation in T1w MRI

Pathology Description Stroke is a very common brain disorder, being the second-leading cause of death worldwide (6.55 million deaths in 2019) [239]. Most strokes (roughly 60%) are ischemic strokes, corresponding to the blocking of an artery by a clot. The principal treatment of ischemic strokes consists of the injection of a drug tissue plasminogen activator [240] that dissolves the plug. While the diagnosis of stroke mainly relies on CT scans for their wider availability [241], the follow-up of the patient requires accurate imaging of the stroke lesion to guide reeducation. This can be performed using T1w MRIs, as they offer accurate imaging of the damaged area [242]. Manual delineation of the damaged region is hand-consuming, thus automated solutions are desired.

Data Description The task consists of the automatic delineation of stroke lesions in T1w brain MRI. For this, the ATLAS 2 dataset is used (N=655), split into 475 for training, 30 for validation, and 155 for ID testing. This dataset is interesting as the lesions are heterogeneous concerning their size. Moreover, ground truth labels are noisy (see Figure II.6.1 for an illustration), as the exact contour of the lesion is often ambiguous. Thus this experiment is particularly interesting for uncertainty evaluation. All images are provided pre-processed, including resampling to $1\,\mathrm{mm}^3$ resolution and registration of the T1w sequences to the MNI template [243]. To match the pre-processing of the other tasks, the brain are further extracted using HD-Bet [232].

II.6.2 Benchmark implementation details

The benchmark uses the Dynamic U-Net implemented using the MONAI's library [244] as a base model, which has demonstrated high segmentation performance and robustness on various medical image segmentation tasks [17, 232, 245]. Its architecture is presented in detail in Figure II.6.2. The overall model contains 16.5 million of trainable parameters. The models are trained using the ADAM [37] optimizer with a learning rate of 2×10^{-4}, until the validation loss seizes to improve for a duration of 60 epochs. For WMH models, a patch training approach is employed, using a patch size of $128 \times 128 \times 128$ and a batch size of 4. This is because images are not registered to a common atlas, and thus exhibit heterogeneous dimensions, which make a patch approach more convenient. For tumor and stroke segmentation, a full 3D approach is used, with an image size of $208 \times 208 \times 144$ and $176 \times 176 \times 192$, respectively, and a batch size set to 1. Batch normalization [23] is used in WMH models; while instance norm [25] is used in 3D models that operate with a reduced batch size. Input images are normalized by applying a Zero Mean Unit Variance (ZMUV) normalization. A data augmentation scheme is implemented using TorchIO [246], using all augmentations provided in the library (random contrast, spatial, and artifact transformations).

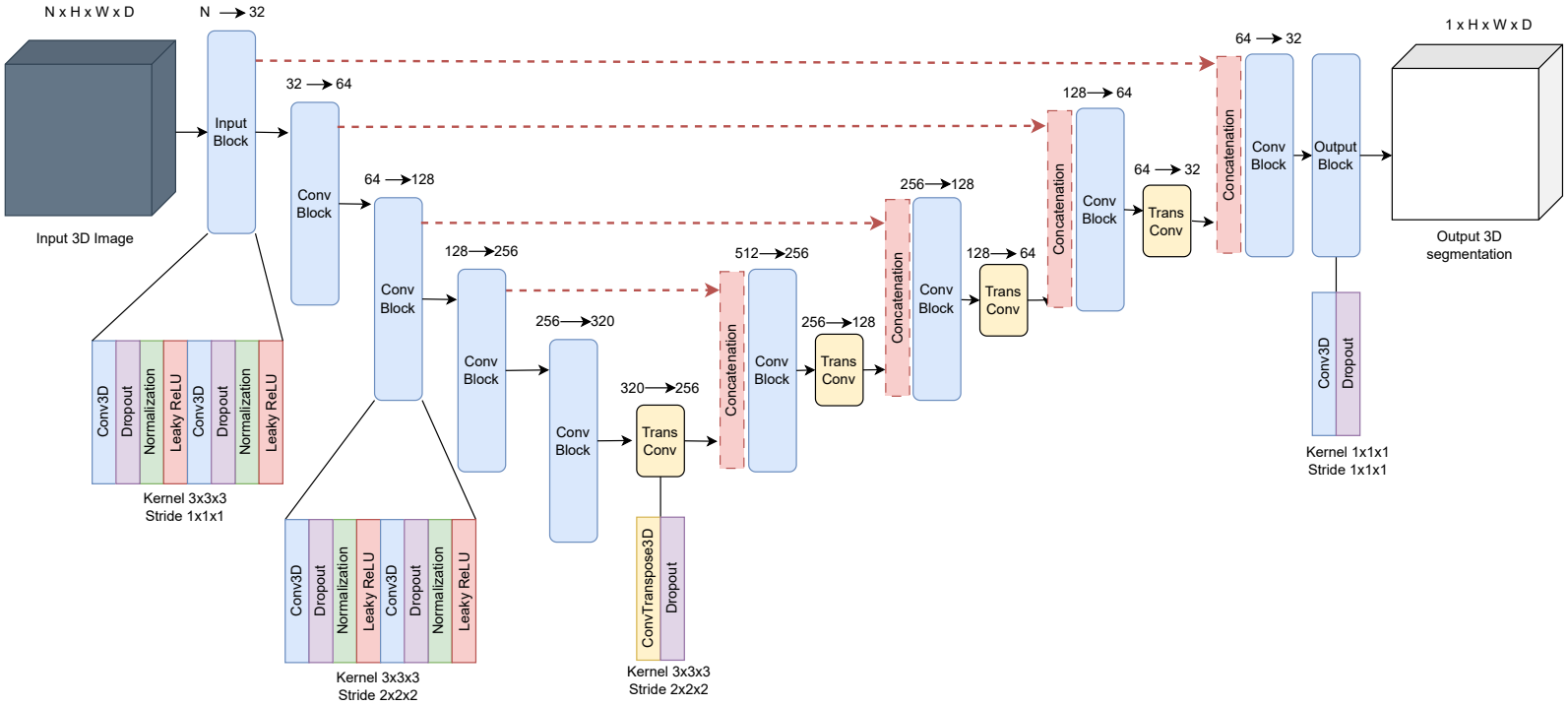


Figure II.6.2: Architecture of the Dynamic U-Net model used as segmentation backbone in the thesis experiments, representing 16.5 million trainable parameters.

II.6.3 Selection of a calibration-preserving segmentation objective

II.6.3.1 Considered Ad-hoc and Post-hoc calibration strategies

Throughout this thesis, multiple segmentation models will be trained. It has been discussed that the popular Soft Dice loss is known for heavily damaging the calibration of the NN [50]. Thus, the first important decision is the choice of the baseline calibration procedure that will be used throughout the thesis. More precisely, ad-hoc and post-hoc calibration techniques can be implemented. Ad-hoc calibration means that the training objective is modified so that the notion of calibration is enforced. Post-hoc calibration is a re-calibration strategy making use of a set-aside validation dataset to fix the calibration of the trained NN. Two ad-hoc techniques (Dice++ and Margin loss), as well as one post-hoc technique (Temperature Scaling), are compared to the baseline Soft Dice choice. They are presented below:

Soft Dice and Cross-Entropy This baseline loss objective corresponds to the sum of the Soft Dice loss (presented in Equation I.2) and the Cross-Entropy loss:

$$\mathcal{L}_1 = \mathcal{L}_{Dice} + \mathcal{L}_{CE} \tag{II.6.1}$$

It is expected that performing training with this loss yields to poorly calibrated NN, and it is used to get baseline calibration scores.

Margin Loss The margin loss [67] is a recent proposal that enforces calibration by adding constraints on the distance between logits l, corresponding to the pre-softmax NN outputs. More formally, the distances $\boldsymbol{d}$ between the winner class and the others in a K-classes segmentation problem is defined as $\boldsymbol{d}(l) = (\max_j(l_j) - l_k)$, with $k \in \{1, ..., K\}$. Then, the proposed loss is framed as:

$$\mathcal{L}_2 = \mathcal{L}_{Dice} + \mathcal{L}_{CE} + \lambda \sum_k \max(0, \max_j(l_j) - l_k - m) \tag{II.6.2}$$

wher m is a user-defined margin, set to the default value $m = 10$, and λ a weighting value set to the default value of $\lambda = 0.1$.

Dice++ and Cross-Entropy The Dice++ loss [50] is an amelioration of the Soft Dice loss that penalizes over-confident mistakes (FP and FN voxels). To achieve this, the Soft Dice objective is modified by adding a penalization factor γ:

$$\mathcal{L}_{Dice++} = \frac{\sum_{i=1}^{N} p_{1i} g_{1i}}{\sum_{i=1}^{N} p_{1i} g_{1i} + \beta \sum_{i=1}^{N} p_{0i} g_{1i}^{\gamma} + \alpha \sum_{i=1}^{N} p_{1i} g_{0i}^{\gamma}} = \frac{2TP}{2TP + \alpha FP^{\gamma} + \beta FN^{\gamma}} \tag{II.6.3}$$

$$\mathcal{L}_3 = \mathcal{L}_{Dice++} + \mathcal{L}_{CE} \tag{II.6.4}$$

The $\mathcal{L}_{Dice++}$ objective encourages the model to produce i) correct confident prediction and ii) unconfident incorrect predictions, via the over-penalization of confident mistakes controlled by γ. A value of $\gamma = 2$ is used, which was found as the optimal value to promote both segmentation quality and calibration [50].

Temperature Scaling Contrarily to the Margin and Dice++ losses, Temperature Scaling (TS) is a post-hoc calibration procedure [49]. The idea is to fit a scalar called the temperature T so that the NLL loss is minimized on a set of validation images. More formally, the temperature is used to adjust the entropy of the softmax output by scaling the logits z as follows:

$$p = \text{Softmax}(z/T) = \frac{e^{z/T}}{\sum_j e^{z_j/T}} \tag{II.6.5}$$

It is important to notice that all logits are scaled with the same scalar, including both correct and incorrect predictions, which can be seen as a limitation. However, TS has been shown to empirically reduce the calibration errors and it can be combined with any ad-hoc calibration objective to reach state-of-the-art calibration results [67]. A temperature superior to 1 indicates over-confidence, while a temperature inferior to 1 is indicated for under-confident network.

II.6.3.2 Corrected Calibration Metrics

To compare models trained with $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$, the ECE and Brier scores are computed. As mentioned in Section II.4.2, calibration metrics for medical images are biased because the vast majority of voxels are background voxels, correctly segmented with very high confidence, which greatly overestimate the calibration of the evaluated NN [67, 69]. In previous studies, authors proposed to exclude the image background of the metric computation by using a brain mask [69], or by skipping the background class in the metric computation [67]. However, these approaches have some limitations. Restricting computation to a foreground brain mask will still include a large quantity of confident non-foreground voxels if the foreground object is small (e.g. small lesion), thus will still overestimate calibration. Alternatively, skipping **all** background voxels signifies that False Positives are not included in the metric. To circumvent these limitations, a third option is here investigated, following which the calibration metrics are computed by excluding **confident correctly background voxels**, defined as voxels correctly assigned to the background class **and** with confidence above $p = 0.99$. This allows to efficiently exclude the large majority of confident background voxels while keeping uncertain background voxels.

Additionally to calibration metrics, we also provide the Dice scores and Surface Dice (SD) [247] scores to evaluate the quality of the segmentation.

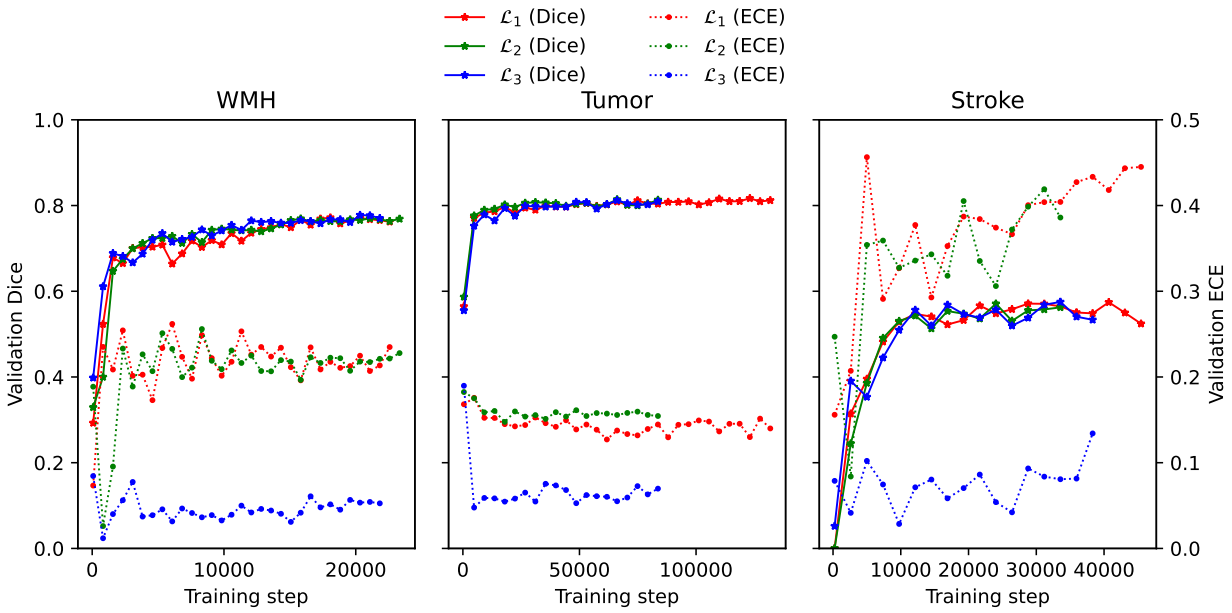


Figure II.6.3: Dice and Expected Calibration Error (ECE) at each training step for segmentation models trained with $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$. Dice scores are plotted with stars and continuous lines. ECE scores are plotted with dashed lines and dots.

II.6.3.3 Results

Tables II.2, II.3, and II.4 present the calibration and segmentation metrics for WMH, tumor and stroke lesion segmentation, respectively. The optimized temperatures are indicated in the tables. Figure II.6.3 displays the evolution of the Dice and ECE metrics during training for each tested loss and dataset.

First, it can be observed in Figure II.6.3 that for models trained using $\mathcal{L}_1$ and $\mathcal{L}_2$, the ECE increases during training as the Dice score increases. For models trained with $\mathcal{L}_3$, the ECE stagnates at a lower value. This is confirmed by the test calibration metrics *before* Temperature Scaling, which are minimized by models trained using $\mathcal{L}_3$, for each of the 3 segmentation tasks. Contrarily, poorer calibrations are achieved with models trained using $\mathcal{L}_1$ and $\mathcal{L}_2$, which exhibit similar miscalibration on the 3 segmentation tasks. No significative improvement can be observed when using the $\mathcal{L}_2$ loss as compared to the baseline $\mathcal{L}_1$. This may be due to the choice of the hyper-parameters in the Margin loss (margin m and weighting λ), which were kept at their default value for this benchmark. It can be concluded that $\mathcal{L}_3$ provides the best *intrinsic* calibration, before post-hoc calibration.

The post-hoc TS procedure yields similar temperatures across the datasets: a temperature around 2 for models trained with $\mathcal{L}_1$, around 1.5 for models trained with $\mathcal{L}_2$, and between 1.15 and 1.3 for models trained with $\mathcal{L}_3$. This may indicate that the miscalibration of the models is intrinsic to the neural network and loss choice, rather than linked to the dataset difficulty. Moreover, it appears that models trained with $\mathcal{L}_3$ are associated with temperatures close to 1, meaning that they are already well-calibrated, which confirms the trends observed

in Figure II.6.3. Overall, all models benefit from TS, with a significant reduction of the calibration error for each dataset and model. Interestingly, even if the model is initially poorly calibrated (e.g. models trained with $\mathcal{L}_1$), TS performs a very efficient post-hoc correction. For instance on WMH segmentation, $\mathcal{L}_1$ alone produces the worst calibration, but $\mathcal{L}_1 + TS$ achieves a very competitive one. For each of the three tasks, the best calibration is achieved after applying TS, and models trained with $\mathcal{L}_1$ are the ones that benefit the most from this post-hoc adjustment.

Regarding the DS settings (MSLUB and 1.5T for WMH segmentation, SSA for tumor segmentation), decreases in segmentation quality and calibration can be observed, as compared to the metrics obtained on ID data. This indicates that the trained models are not able to generalize very well to these datasets. Overall, the combination of $\mathcal{L}_3 + TS$ seems to provide the best and most robust results across the 3 segmentation tasks. It achieves the best segmentation performances on the 3 MS datasets, as well as on the SSA tumor dataset while having a very competitive calibration with and without post-hoc TS for each experiment. The fact that it provides by far the best intrinsic (prior TS) calibration is an interesting feature in situations where data is scarce and thus few images are allocated to validation. In these cases, training with $\mathcal{L}_3$ yields well-calibrated models without the need to perform TS, contrarily to the other tested losses. To conclude, in the following of this thesis, segmentation models will be trained using the $\mathcal{L}_3$ if not specified otherwise. Moreover, TS will systematically be performed using the validation dataset, after completion of the training.

II.6.4 Selection of a voxel uncertainty baseline estimator

Now that the segmentation objective was selected, several popular UQ frameworks are evaluated on the same 3 segmentation tasks. The retained UQ paradigms are Softmax uncertainty, MC dropout, Deep Ensemble, TTA, EDL and Learned uncertainty, representing all together more than 85% of the review UQ methods. Below, implementation details for each of the 6 techniques are presented.

Baseline Softmax uncertainty For this baseline approach, the Dynamic U-Net is trained using $\mathcal{L}_3$, followed by post-hoc TS calibration using the set of validation images. At test time, the uncertainty estimates is taken as the Maximum Softmax Probability (MSP) estimator [248], which allows to get one uncertainty score for each voxel from the predicted probability vectors. It is defined as:

$$\text{MSP} = 1 - \max_k(p_{i,k}) \tag{II.6.6}$$

where $p_{i,k}$ is the probability of class k at voxel i.

MC dropout To implement MC dropout, 3D dropout layers are inserted after each convolution in the Dynamic U-Net model, with a fixed dropout probability of $p = 0.10$. This dropout rate is selected in order to preserve the quality of the segmentation, as we observed that larger rates highly degraded the performance of the Dynamic U-Net. The MC model

		ECE ↓			Brier ↓			Dice (%) ↑			Surface Dice (%) ↑		
	Dataset	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI
$\mathcal{L}_1$	ID	0.19	0.01	$[-0.02, 0.02]$	0.22	0.01	$[-0.02, 0.02]$	77.6	1.4	$[-2.4, 2.3]$	94.2	0.8	$[-1.5, 1.3]$
	MSLUB	0.21	0.01	$[-0.01, 0.01]$	0.24	0.01	$[-0.01, 0.01]$	65.9	3.0	$[-5.0, 4.8]$	84.9	2.6	$[-4.5, 4.1]$
	1.5T	0.23	0.02	$[-0.03, 0.03]$	0.26	0.02	$[-0.03, 0.03]$	66.6	3.5	$[-6.0, 5.3]$	85.7	4.0	$[-7.4, 5.7]$
$\mathcal{L}_2$	ID	0.19	0.01	$[-0.02, 0.02]$	0.23	0.01	$[-0.02, 0.02]$	75.1	1.3	$[-2.2, 2.2]$	94.0	0.8	$[-1.3, 1.2]$
	MSLUB	0.22	0.02	$[-0.03, 0.03]$	0.25	0.02	$[-0.03, 0.03]$	63.0	3.8	$[-6.6, 6.0]$	80.7	3.8	$[-6.6, 5.8]$
	1.5T	0.23	0.02	$[-0.03, 0.03]$	0.26	0.02	$[-0.03, 0.03]$	65.2	3.4	$[-5.9, 5.1]$	86.8	4.1	$[-7.4, 5.7]$
$\mathcal{L}_3$	ID	0.03	0.01	$[-0.01, 0.01]$	0.09	0.01	$[-0.01, 0.01]$	78.3	1.3	$[-2.2, 2.2]$	94.7	0.8	$[-1.3, 1.2]$
	MSLUB	0.06	0.01	$[-0.01, 0.01]$	0.12	0.01	$[-0.01, 0.01]$	66.8	3.3	$[-5.6, 5.1]$	84.5	2.6	$[-4.4, 4.0]$
	1.5T	0.05	0.01	$[-0.01, 0.01]$	0.11	0.01	$[-0.01, 0.01]$	66.8	3.4	$[-5.9, 5.2]$	87.5	4.0	$[-7.3, 5.5]$
$\mathcal{L}_1$	ID	**0.02**	**0.00**	$[-0.00, 0.00]$	**0.04**	**0.00**	$[-0.01, 0.01]$	77.6	1.4	$[-2.4, 2.3]$	94.2	0.8	$[-1.5, 1.3]$
+ TS	MSLUB	**0.01**	**0.00**	$[-0.00, 0.00]$	**0.03**	**0.00**	$[-0.00, 0.00]$	65.9	3.0	$[-5.0, 4.8]$	**84.9**	**2.6**	$[-4.5, 4.1]$
($T = 2.21$)	1.5T	**0.01**	**0.00**	$[-0.00, 0.00]$	**0.03**	**0.00**	$[-0.01, 0.01]$	66.6	3.5	$[-6.0, 5.3]$	85.7	4.0	$[-7.4, 5.7]$
$\mathcal{L}_2$	ID	0.03	0.00	$[-0.01, 0.01]$	0.06	0.00	$[-0.01, 0.01]$	75.1	1.3	$[-2.2, 2.2]$	94.0	0.8	$[-1.3, 1.2]$
+ TS	MSLUB	0.02	0.00	$[-0.00, 0.00]$	0.05	0.00	$[-0.01, 0.01]$	63.0	3.8	$[-6.6, 6.0]$	80.7	3.8	$[-6.6, 5.8]$
($T = 1.61$)	1.5T	0.03	0.00	$[-0.01, 0.01]$	0.06	0.01	$[-0.01, 0.01]$	65.2	3.4	$[-5.9, 5.1]$	86.8	4.1	$[-7.4, 5.7]$
$\mathcal{L}_3$	ID	**0.02**	**0.00**	$[-0.00, 0.00]$	0.05	0.00	$[-0.01, 0.01]$	**78.3**	**1.3**	$[-2.2, 2.2]$	**94.7**	**0.8**	$[-1.3, 1.2]$
+ TS	MSLUB	**0.01**	**0.00**	$[-0.00, 0.00]$	0.06	0.00	$[-0.01, 0.01]$	**66.8**	**3.3**	$[-5.6, 5.1]$	84.5	2.6	$[-4.4, 4.0]$
($T = 1.30$)	1.5T	0.02	0.00	$[-0.00, 0.00]$	0.05	0.00	$[-0.01, 0.01]$	**66.8**	**3.4**	$[-5.9, 5.2]$	**87.5**	**4.0**	$[-7.3, 5.5]$

Table II.2: Results of the calibration benchmark for WMH segmentation. The mean (μ), the standard error of the mean (SEM) and the 90% confidence interval (CI) are computed using bootstrap (M=15000). ↓ indicates that we seek to minimize the metric, while ↑ indicates that it should be maximized. The best metrics for each dataset are indicated in **bold**. TS: Temperature Scaling. ECE: Expected Calibration Error.

		ECE ↓			Brier ↓			Dice (%) ↑			Surface Dice (%) ↑		
	Dataset	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI
$\mathcal{L}_1$	ID	0.11	0.01	$[-0.01, 0.01]$	0.06	0.00	$[-0.01, 0.01]$	84.6	0.9	$[-1.5, 1.4]$	91.1	0.9	$[-1.6, 1.5]$
	SSA	0.17	0.02	$[-0.03, 0.04]$	0.10	0.01	$[-0.02, 0.02]$	73.0	2.4	$[-4.1, 3.8]$	81.0	2.4	$[-4.1, 3.8]$
$\mathcal{L}_2$	ID	0.12	0.01	$[-0.01, 0.01]$	0.07	0.00	$[-0.01, 0.01]$	83.7	0.9	$[-1.5, 1.5]$	90.4	1.0	$[-1.7, 1.6]$
	SSA	0.19	0.02	$[-0.03, 0.03]$	0.11	0.01	$[-0.02, 0.02]$	71.3	2.6	$[-4.3, 4.1]$	78.8	2.7	$[-4.5, 4.3]$
$\mathcal{L}_3$	ID	0.04	0.00	$[-0.00, 0.00]$	0.04	0.00	$[-0.00, 0.00]$	83.6	0.9	$[-1.6, 1.5]$	90.3	1.0	$[-1.7, 1.6]$
	SSA	0.05	0.01	$[-0.01, 0.01]$	0.05	0.00	$[-0.01, 0.01]$	73.8	2.4	$[-4.1, 3.9]$	81.8	2.4	$[-4.1, 3.9]$
$\mathcal{L}_1$+TS	ID	**0.03**	**0.00**	$[-0.00, 0.00]$	**0.02**	**0.00**	$[-0.00, 0.00]$	**84.6**	**0.9**	$[-1.5, 1.4]$	**91.1**	**0.9**	$[-1.6, 1.5]$
$T = 2.06$	SSA	**0.04**	**0.00**	$[-0.01, 0.01]$	**0.04**	**0.00**	$[-0.00, 0.00]$	73.0	2.4	$[-4.1, 3.8]$	81.0	2.4	$[-4.1, 3.8]$
$\mathcal{L}_2$+TS	ID	0.07	0.00	$[-0.01, 0.01]$	0.05	0.00	$[-0.00, 0.00]$	83.7	0.9	$[-1.5, 1.5]$	90.4	1.0	$[-1.7, 1.6]$
$T = 1.45$	SSA	0.11	0.01	$[-0.02, 0.02]$	0.07	0.01	$[-0.01, 0.01]$	71.3	2.6	$[-4.3, 4.1]$	78.8	2.7	$[-4.5, 4.3]$
$\mathcal{L}_3$+TS	ID	**0.03**	**0.00**	$[-0.00, 0.00]$	0.03	0.00	$[-0.00, 0.00]$	83.6	0.9	$[-1.6, 1.5]$	90.3	1.0	$[-1.7, 1.6]$
$T = 1.15$	SSA	**0.04**	**0.00**	$[-0.01, 0.01]$	**0.04**	**0.00**	$[-0.01, 0.01]$	**73.8**	**2.4**	$[-4.1, 3.9]$	**81.8**	**2.4**	$[-4.1, 3.9]$

Table II.3: Results of the calibration benchmark for glioblastoma segmentation. The mean (μ), the standard error of the mean (SEM) and the 90% confidence interval (CI) are computed using bootstrap (M=15000). ↓ indicates that we seek to minimize the metric, while ↑ indicates that it should be maximized. The best metrics for each dataset are indicated in **bold**. TS: Temperature Scaling. ECE: Expected Calibration Error.

	ECE ↓			Brier ↓			Dice (%) ↑			Surface Dice (%) ↑		
	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI
$\mathcal{L}_2$	0.41	0.02	$[-0.04, 0.04]$	0.43	0.02	$[-0.04, 0.04]$	55.0	2.5	$[-4.1, 4.1]$	65.2	2.6	$[-4.4, 4.3]$
$\mathcal{L}_2$	0.34	0.02	$[-0.03, 0.03]$	0.36	0.02	$[-0.03, 0.03]$	59.0	2.4	$[-4.1, 3.9]$	68.8	2.5	$[-4.2, 4.1]$
$\mathcal{L}_3$	0.06	0.01	$[-0.01, 0.01]$	0.11	0.01	$[-0.01, 0.01]$	56.6	2.5	$[-4.1, 4.0]$	66.0	2.6	$[-4.3, 4.2]$
$\mathcal{L}_1$+TS ($T = 2.06$)	0.11	0.01	$[-0.02, 0.02]$	0.14	0.01	$[-0.02, 0.02]$	55.0	2.5	$[-4.1, 4.1]$	65.2	2.6	$[-4.4, 4.3]$
$\mathcal{L}_2$+TS ($T = 1.61$)	0.34	0.02	$[-0.03, 0.03]$	0.36	0.02	$[-0.03, 0.03]$	**59.0**	**2.4**	$[-4.0, 3.9]$	**68.8**	**2.5**	$[-4.1, 4.0]$
$\mathcal{L}_3$+TS ($T = 1.15$)	**0.04**	**0.01**	$[-0.01, 0.01]$	**0.08**	**0.01**	$[-0.01, 0.01]$	56.7	2.5	$[-4.1, 4.1]$	66.0	2.6	$[-4.4, 4.2]$

Table II.4: Results of the calibration benchmark for stroke lesion segmentation. The mean (μ), the standard error of the mean (SEM) and the 90% confidence interval (CI) are computed using bootstrap (M=15000). ↓ indicates that we seek to minimize the metric, while ↑ indicates that it should be maximized. The best metrics for each dataset are indicated in **bold**. TS: Temperature Scaling. ECE: Expected Calibration Error.

is trained using $\mathcal{L}_3$, followed by post-hoc TS calibration. At test time, $N = 20$ predictions are generated for each input image with dropout activated. Then, the entropy is used as an uncertainty marker, defined as:

$$H(p_i) = -\sum_{k=1}^{K} \frac{1}{T} \sum_{t=1}^{T} p_{i,k,t} \log(\frac{1}{T} \sum_{t=1}^{T} p_{i,k,t}) \tag{II.6.7}$$

$$\tag{II.6.8}$$

where $p_{i,k,t}$ is the probability that voxel i belongs to class k, for the t-th MC sample.

Deep Ensemble Deep Ensemble is implemented by training $N = 5$ identical Dynamic U-Nets, trained on the same subjects, with the $\mathcal{L}_3$ loss. After training, TS is performed for each member of the ensemble. Similarly to MC dropout, voxel-wise uncertainty is estimated using entropy.

Test Time Augmentation TTA is implemented on top of the standard Dynamic U-Net that has been trained to implement the Softmax uncertainty approach. At test time, the TorchIO library [246] is used to generate $N = 20$ alternative variants of the original image. The augmentation scheme comprises simple spatial transformations (rotations, flipping, affine, and elastic deformations) and contrast alterations (gamma, noise). From the pool of predictions, the entropy is computed.

Evidential Deep Learning The Dynamic U-Net used in this benchmark is transformed into an EDL network by replacing the output softmax with an exponential function, guaranteeing that the computed evidence are positive. Then, inspired by the TBraTS framework [157], the model is trained with the compound loss:

$$\mathcal{L}_{EDL} = \mathcal{L}_{Dice++} + \mathcal{L}_{ice} + \mathcal{L}_{KL} \tag{II.6.9}$$

where $\mathcal{L}_{ice}$ and $\mathcal{L}_{KL}$ are the losses introduced in Equations II.2.12 and II.2.12, respectively. The final uncertainty score for voxel i corresponds to $u_i = K/\sum_{k=1}^{K} \alpha_{i,k}$, where $\alpha_{i,k}$ are the estimated Dirichlet parameters for voxel i.

Learned uncertainty using the Labelflip loss The learned uncertainty framework is implemented using the Focal Labelflip loss that achieved state-of-the-art performance on the BraTS 2020 challenge [133]. The Dynamic U-Net is modified so that it has 2 outputs for each class k and voxel i: one for the segmentation probabilities $p_{i,k}$ and one for the learned uncertainty score, called the Labelflip probability $q_{i,k}$. Writing $y_{i,k}$ the one-hot label vector containing 1 for the correct class and 0 else, the Focal Labelflip loss is computed as:

$$w_{i,k} = q_{i,k}(1 - y_{i,k}) + y_{i,k}(1 - q_{i,k}) \tag{II.6.10}$$
$$\mathcal{L}_{\text{Labelflip}} = (p_{i,k} - w_{i,k})^2 w_{i,k}[\log(w_{i,k}) - \log(p_{i,k})] + \text{BCE}(q_{i,k}, z_i)$$

where z_i is the disagreement indicator for voxel i, computed by comparing the predicted segmentation and y_i, and BCE is the binary cross-entropy.

Following the original implementation, training is performed by combining the Focal loss [249] and the Focal Labelflip loss:

$$\mathcal{L}_{LU} = \text{Focal}(p, y) + \mathcal{L}_{\text{Labelflip}} \tag{II.6.11}$$

At test time, the modified Dynamic U-Net produces one uncertainty score $q_{i,k}$ for each voxel and class. To obtain a single uncertainty score $q_{total,i}$ per voxel, the class-wise estimates are summed at each voxel: $q_{total,i} = \sum_{k=1}^{K} q_{i,k}$.

II.6.4.1 Voxel-wise uncertainty metrics

For this benchmark, two popular voxel-wise uncertainty metrics are used to compare the approaches, namely the uncertainty-error overlap (UEO) [69] and Area under the error retention curves (R-AUC) [214].

Uncertainty-Error overlap To compute UEO, the voxel-wise uncertainty map is binarized using a threshold, producing a binary uncertainty map U highlighting unconfident voxels. Then, the segmentation error map E is computed by comparing the predicted segmentation and the ground truth. Finally, the UEO is taken as the Dice between the binarized uncertainty map and the error map:

$$\text{UEO} = \frac{2|U \cap E|}{|U| + |E|} \tag{II.6.12}$$

This metric requires the estimation of a voxel-wise uncertainty threshold to compute U. As in the original paper, this value is estimated for each UQ method by determining the optimal threshold maximizing the UEO on the validation dataset. While the UEO is easy to compute and interpretable, its downside is that it relies on the Dice and it is thus biased toward large volumes of errors. Thus, models that are under-performing with respect to the segmentation task can be attributed to large UEO values, as compared to models making very few mistakes.

Area under the Retention Curve The R-AUC metric is adapted from the Shift challenge on segmentation uncertainty quantification [214]. Formally, for a given test image, the voxels are ordered from the more uncertain to the most certain. Then, a fraction of the most uncertain voxels is removed and the performance (Dice) is estimated on the remaining ones.

This process is repeated for a set of thresholds, which produces a **fraction versus Dice** plot, from which the area under the curve is obtained. This metric will reward both accurate segmentation (high dice coefficient) and reliable uncertainty. As for the calibration metrics, the metric is enhanced by first removing the confident correct background voxels. This mitigates the important class imbalance between the background and foreground classes.

Inference Time Many voxel-wise uncertainty methods are based on sampling various predictions to compute uncertainty scores. However, this process is time-consuming. For real-world applications, the prediction should be provided as fast as possible and thus fast uncertainty estimates are preferred. To evaluate the efficiency of each technique, the average inference time per subject is provided. To carry the inferences, a NVIDIA RTX A5000 with a memory of 25 Gigabytes is used.

II.6.4.2 Results

Tables II.5, II.6, II.7 present the results of the voxel uncertainty benchmark on MS, tumor and stroke data, respectively. Additionally, Figure II.6.4 presents an example of voxel-wise uncertainty map obtained with the Deep Ensemble technique on a MS subject, with the associated UEO and R-AUC values. Similar examples for tumor and stroke lesion segmentation are provided in Appendix II.6.5 and II.6.6.

First, regarding **segmentation performance**, the Ensemble framework provides as expected the best quality predictions, being the top performer on WMH, tumor, and stroke segmentation. Interestingly, the gain in performance is clear for both ID and DS datasets, as compared to the single baseline Softmax model. Softmax and TTA achieve similar segmentation performance results, indicating that the augmentation strategy does not modify substantially the final segmentation, in good or bad. For EDL, results are similar to the ones achieved by the Softmax baseline for WMH and Tumor, but it performed better on stroke lesion segmentation, being close to the performance of the Deep Ensemble. Overall, the MC dropout models produced poorer segmentation results on each of the 3 tasks. This may be due to an under-fitting resulting from the 10% dropout applied after each convolution. Similarly, for the LU model, segmentation quality is degraded on WMH and tumor segmentation, as compared to the baseline Softmax model. For this framework, the modification of the learning objective may be responsible for the weaker results, as segmentation and uncertainty are learned jointly during optimization. Moreover, the Labelflip loss relies on the Focal loss instead of the Dice loss for stability, hence it does not directly optimize the segmentation metric used in this benchmark.

Second, regarding the **quality of voxel uncertainty estimates**, different tendencies are observed for the UEO and R-AUC metrics, respectively. For UEO, the LU framework achieves the best results on the 3 WMH datasets (ID, MSLUB, and 1.5T) and the ID and SSA tumor datasets. DE maximizes the UEO on stroke data. However, as presented earlier, UEO relies on the Dice score and it is thus biased in favor of large volume of errors. This may explain the good performance of Learned uncertainty with respect to the UEO, while it produces poorer segmentation results. Contrarily, R-AUC rewards both accurate segmentation and good uncertainty estimates, and it places DE as the best uncertainty estimator for the 3 MS

		UEO (%) ↑			R-AUC (%) ↑			Dice (%) ↑			Surface Dice (%) ↑			Time (s)
	Dataset	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ
Softmax	ID	42.0	0.5	$[-0.9, 0.8]$	93.7	0.6	$[-1.0, 0.9]$	78.3	1.3	$[-2.2, 2.1]$	94.7	0.8	$[-1.3, 1.2]$	**1.31**
	MSLUB	39.4	1.0	$[-1.7, 1.5]$	89.4	1.2	$[-2.0, 1.8]$	66.8	3.3	$[-5.5, 5.2]$	84.5	2.6	$[-4.4, 4.1]$	
	1.5T	40.9	1.8	$[-3.3, 2.3]$	87.0	3.7	$[-7.2, 4.6]$	66.8	3.5	$[-6.0, 5.2]$	87.5	3.9	$[-7.2, 5.4]$	
MC	ID	42.0	0.6	$[-1.0, 1.0]$	93.5	0.5	$[-0.9, 0.8]$	75.5	1.4	$[-2.3, 2.2]$	93.7	0.8	$[-1.4, 1.3]$	17.64
	MSLUB	38.7	1.1	$[-1.8, 1.7]$	89.9	1.0	$[-1.7, 1.6]$	63.9	3.2	$[-5.4, 5.2]$	82.0	2.5	$[-4.3, 3.9]$	
	1.5T	40.9	1.5	$[-2.7, 2.2]$	87.1	3.8	$[-7.2, 5.1]$	64.7	3.4	$[-6.0, 5.2]$	87.5	4.0	$[-7.4, 5.7]$	
DE	ID	42.1	0.5	$[-0.9, 0.8]$	94.2	0.5	$[-0.9, 0.8]$	**79.0**	**1.3**	$[-2.1, 2.0]$	**95.1**	**0.7**	$[-1.2, 1.1]$	6.82
	MSLUB	39.9	1.0	$[-1.6, 1.5]$	**90.2**	**1.1**	$[-1.9, 1.7]$	**68.0**	**3.2**	$[-5.5, 5.1]$	**86.2**	**2.4**	$[-4.2, 3.8]$	
	1.5T	42.5	0.9	$[-1.6, 1.4]$	87.6	3.7	$[-7.1, 4.6]$	**67.8**	**3.5**	$[-6.2, 5.3]$	**88.5**	**4.0**	$[-7.4, 5.5]$	
TTA	ID	41.9	0.5	$[-0.8, 0.8]$	94.4	0.5	$[-0.8, 0.8]$	78.0	1.3	$[-2.1, 2.0]$	94.7	0.7	$[-1.2, 1.1]$	38.89
	MSLUB	39.3	0.9	$[-1.6, 1.5]$	89.7	1.2	$[-2.0, 1.9]$	65.9	3.5	$[-5.9, 5.4]$	83.2	2.8	$[-4.8, 4.3]$	
	1.5T	41.7	1.8	$[-3.3, 2.5]$	**88.3**	**3.7**	$[-7.2, 4.5]$	67.0	3.5	$[-6.1, 5.2]$	88.5	3.9	$[-7.2, 5.4]$	
Learned	ID	**45.2**	**0.6**	$[-1.0, 0.9]$	92.6	0.6	$[-1.0, 0.9]$	74.0	1.5	$[-2.4, 2.4]$	93.9	0.8	$[-1.3, 1.3]$	1.61
	MSLUB	**43.1**	**0.6**	$[-1.0, 1.0]$	89.0	1.0	$[-1.7, 1.6]$	65.2	3.0	$[-5.1, 4.9]$	85.2	2.4	$[-4.1, 3.8]$	
	1.5T	**45.8**	**0.8**	$[-1.4, 1.3]$	85.7	3.7	$[-7.0, 4.9]$	63.4	3.5	$[-6.1, 5.3]$	85.0	4.1	$[-7.4, 5.9]$	
EDL	ID	41.9	0.5	$[-0.9, 0.9]$	**95.0**	**0.5**	$[-0.9, 0.8]$	78.8	1.4	$[-2.3, 2.2]$	94.7	0.8	$[-1.5, 1.3]$	1.65
	MSLUB	37.1	1.3	$[-2.3, 2.1]$	89.4	1.4	$[-2.5, 2.2]$	67.2	3.3	$[-5.6, 5.2]$	84.8	2.9	$[-5.1, 4.6]$	
	1.5T	42.8	1.0	$[-1.5, 1.7]$	**88.3**	**3.8**	$[-7.3, 4.6]$	67.4	3.4	$[-6.0, 5.1]$	87.4	3.9	$[-7.3, 5.4]$	

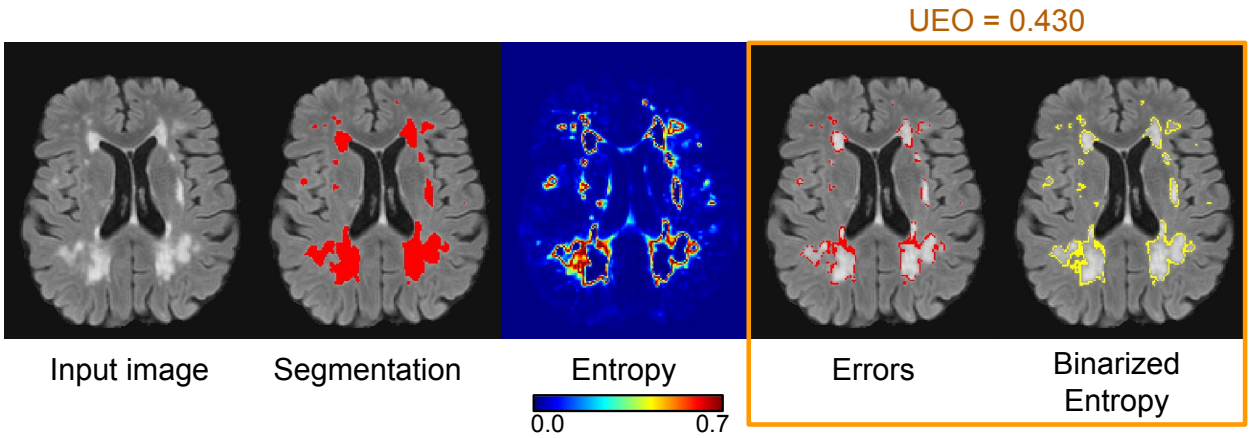
Table II.5: Results of the voxel-wise uncertainty benchmark for WMH segmentation. The mean (μ), the standard error of the mean (SEM) and the 90% confidence interval (CI) are computed using bootstrap (M=15000). ↑ indicates that the metrics should be maximized. The highest metrics for each dataset are indicated in **bold**. UEO: Uncertainty-Error Overlap.

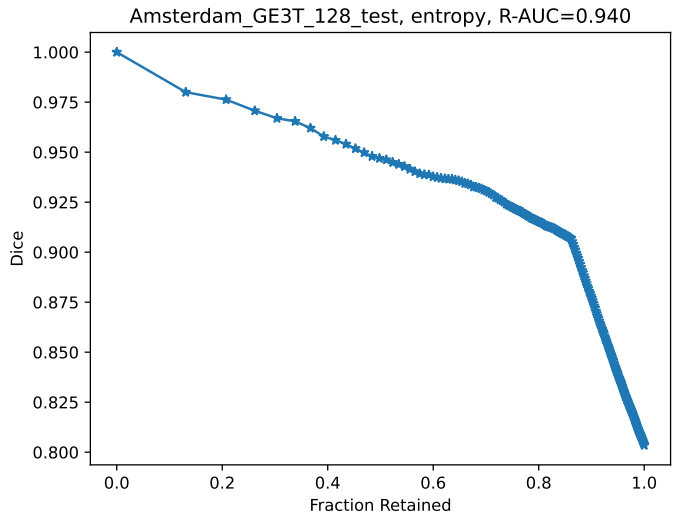
datasets, the SSA tumor dataset, and the stroke dataset. It is only overtaken by the MC dropout framework on the tumor ID dataset. Overall, disentangling uncertainty quality and segmentation performance during evaluation is complex. The UEO favors underperforming models, while contrarily the R-AUC favors high-performing models.

Regarding **robustness under domain-shift**, the R-AUC scores on DS data (MSLUB, 1.5T, SSA) are systematically lower than the scores obtained on the Test ID datasets. This is correlated with the observed decline in segmentation accuracy. As the volume of errors increases in DS settings, it could be expected that the UEO scores also increase in these settings. However, this is not always the case, as a reduction of the UEO scores can be observed on the MSLUB and the 1.5T datasets. This phenomenon can be due to the fact that the uncertainty threshold for the UEO is optimized on the set-aside validation images, which are a subset of the training images. Hence, this threshold may end up being sub-optimal in the presence of domain shifts.

Finally, regarding **inference speed**, the baseline Softmax approach is as expected the fastest. If time is crucial, the benchmark indicates that a calibrated segmentation model produces interesting uncertainty estimates (with respect to the UEO and R-AUC scores), while not being the top performer. The Learned Uncertainty and EDL frameworks are also very competitive, as they do not require sampling to compute the voxel uncertainty. However, we found no clear benefit of implementing the EDL or LU frameworks with respect to uncertainty quality, as compared to the baseline (calibrated) Softmax uncertainty paradigm. Sampling approaches, including MC, TTA, and DE, are less effective in terms of inference speed. TTA is the worst, as it implies generating augmented versions of 3D MRIs, a time-consuming process. This inference speed is of course dependent on the number of sampling steps, which were defined as 20 for MC and TTA, and 5 for DE. Interestingly, even if the number of sampled predictions is lower for DE, it still provides the best voxel-wise uncertainty quality. Moreover, the computational cost of DE is primarily at the training stage, which requires the repetition of the training multiple times. However, at test time it is the most efficient sampling approach, taking on average 1.77s on stroke data, 1.87s on multi-modal tumor data, and 6.82s for WMH data which requires patch sampling.

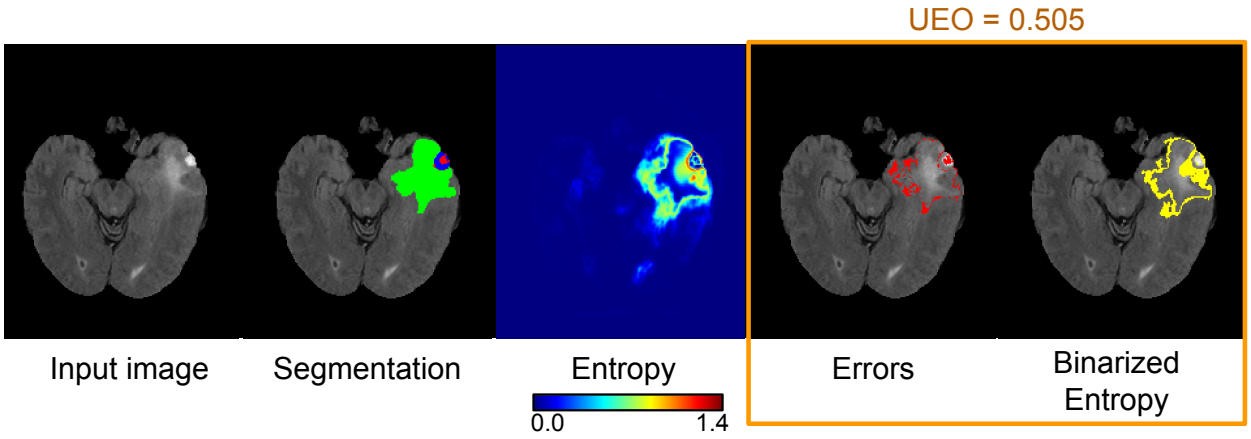
Overall, taking into account the segmentation performance, voxel-wise uncertainty quality, and inference speed, it appears that the Deep Ensemble framework is the most interesting one. First, it allows a gain in segmentation quality and robustness in DS settings. Second, the voxel-wise uncertainty estimates are indicative. Lastly, its computational overhead is meanly at the training stage, but once the pipeline is deployed, its inference time is not prohibitive. Thus, in the following of this thesis, the DE will be used as the baseline voxel-wise uncertainty method.

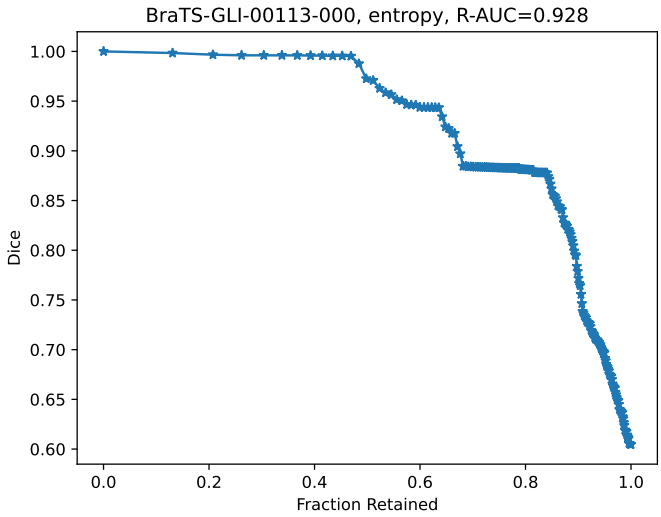




a

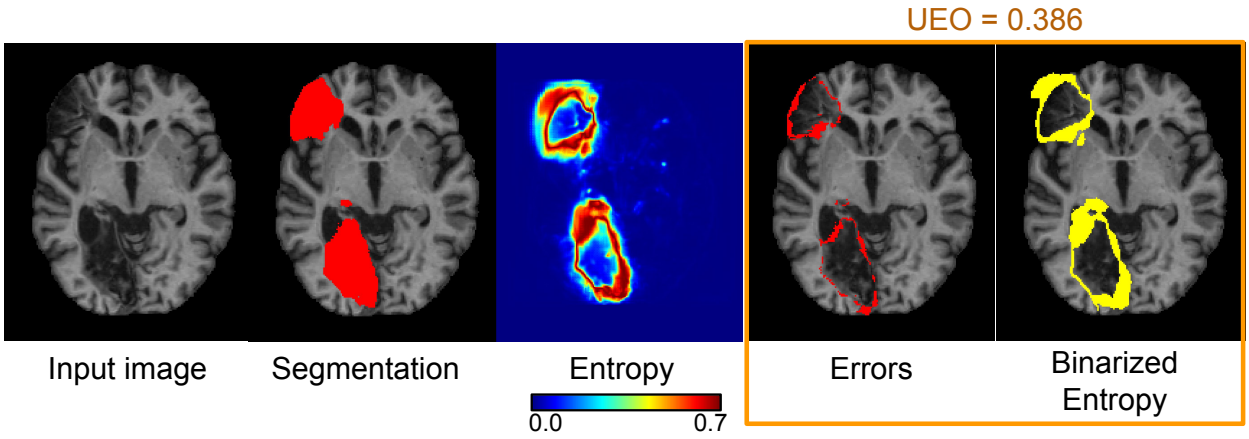
Figure II.6.4: Example of a voxel-wise entropy map generated using the Deep Ensemble technique on a MS patient. The segmentation achieves a Dice score of 0.803. Top: input image, prediction, entropy map, errors and binarized entropy. Bottom: associated R-AUC curve and score.

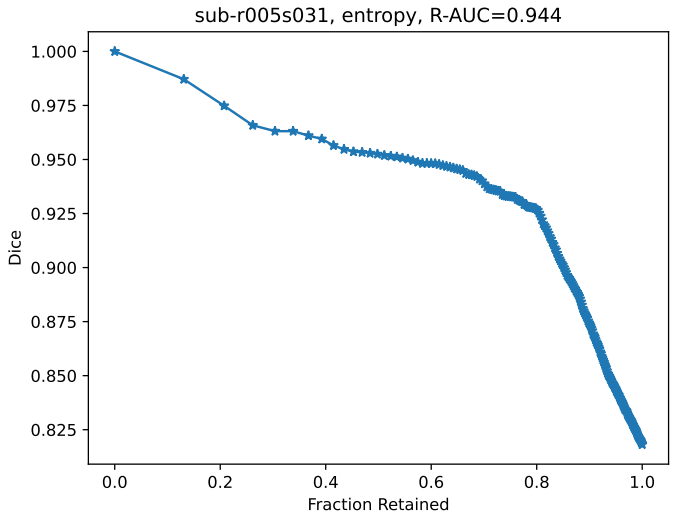




a

Figure II.6.5: Example of a voxel-wise entropy map generated using the Deep Ensemble technique a glioblastoma case. Top: input image, prediction, entropy map, errors, and binarized entropy. Bottom: associated R-AUC curve and score.





a

Figure II.6.6: Example of a voxel-wise entropy map generated using the Deep Ensemble technique on a stroke patient. Top: input image, prediction, entropy map, errors and binarized entropy. Bottom: associated R-AUC curve and score.

		UEO (%) ↑			R-AUC (%) ↑			Dice (%) ↑			Surface Dice (%) ↑			Time (s)
	Data	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ
Softmax	ID	43.1	0.3	$[-0.5, 0.5]$	92.7	0.8	$[-1.3, 1.2]$	83.7	0.9	$[-1.6, 1.5]$	90.8	0.9	$[-1.6, 1.5]$	**0.39**
	SSA	45.1	1.1	$[-1.8, 1.7]$	87.1	2.3	$[-3.9, 3.5]$	73.3	2.6	$[-4.4, 4.0]$	81.3	2.7	$[-4.6, 4.2]$	
MC	ID	39.3	0.5	$[-0.9, 0.9]$	**93.6**	**0.7**	$[-1.2, 1.2]$	83.3	1.0	$[-1.6, 1.5]$	90.2	1.0	$[-1.7, 1.6]$	4.63
	SSA	42.3	1.4	$[-2.4, 2.3]$	87.2	1.8	$[-3.0, 2.8]$	69.1	2.5	$[-4.3, 4.0]$	77.0	2.6	$[-4.4, 4.1]$	
DE	ID	41.5	0.4	$[-0.7, 0.7]$	93.5	0.7	$[-1.3, 1.2]$	**84.9**	**0.9**	$[-1.5, 1.5]$	**91.3**	**0.9**	$[-1.5, 1.5]$	1.87
	SSA	44.0	1.2	$[-2.0, 2.0]$	**88.8**	**2.2**	$[-3.9, 3.4]$	**74.4**	**2.5**	$[-4.3, 3.9]$	**81.9**	**2.6**	$[-4.4, 4.1]$	
TTA	ID	41.2	0.4	$[-0.7, 0.7]$	93.3	0.7	$[-1.3, 1.2]$	83.8	0.9	$[-1.5, 1.5]$	90.7	0.9	$[-1.6, 1.5]$	37.67
	SSA	43.2	1.2	$[-2.0, 1.9]$	87.9	2.2	$[-3.9, 3.4]$	73.4	2.6	$[-4.4, 4.1]$	81.1	2.7	$[-4.6, 4.2]$	
Learned	ID	**46.9**	**0.5**	$[-0.8, 0.8]$	92.7	0.7	$[-1.3, 1.2]$	81.6	1.0	$[-1.7, 1.6]$	89.1	1.0	$[-1.7, 1.6]$	**0.39**
	SSA	**50.2**	**1.5**	$[-2.6, 2.5]$	87.3	1.7	$[-2.8, 2.7]$	69.4	2.6	$[-4.3, 4.2]$	77.1	2.7	$[-4.6, 4.4]$	
EDL	ID	36.3	0.7	$[-1.1, 1.1]$	93.0	0.8	$[-1.3, 1.2]$	84.2	0.9	$[-1.4, 1.4]$	91.1	0.9	$[-1.5, 1.5]$	0.40
	SSA	40.8	1.3	$[-2.2, 2.0]$	88.5	1.9	$[-3.3, 3.0]$	73.9	2.5	$[-4.2, 3.8]$	81.7	2.6	$[-4.4, 4.1]$	

Table II.6: Results of the voxel-wise uncertainty benchmark for glioblastoma segmentation. The mean (μ), the standard error of the mean (SEM) and the 90% confidence interval (CI) are computed using bootstrap (M=15000). ↑ indicates that the metrics should be maximized. The highest metrics for each dataset are indicated in **bold**. UEO: Uncertainty-Error Overlap.

	UEO (%) ↑			R-AUC (%) ↑			Dice (%) ↑			Surface Dice (%) ↑			Time (s)
	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ
Softmax	36.0	1.1	$[-1.8, 1.8]$	85.7	1.2	$[-2.0, 1.9]$	56.6	2.5	$[-4.1, 4.0]$	65.9	2.6	$[-4.4, 4.2]$	0.27
MC	34.7	1.1	$[-1.9, 1.8]$	**87.6**	**1.2**	$[-2.0, 1.9]$	56.0	2.5	$[-4.2, 4.1]$	64.6	2.6	$[-4.4, 4.3]$	3.75
DE	**37.3**	**1.2**	$[-2.0, 1.9]$	**87.6**	**1.2**	$[-2.0, 1.9]$	**8.9**	**2.4**	$[-4.1, 3.9]$	**68.1**	**2.6**	$[-4.3, 4.3]$	1.77
TTA	36.3	1.2	$[-2.1, 2.0]$	86.8	1.2	$[-2.0, 1.9]$	57.4	2.5	$[-4.2, 4.0]$	66.8	2.6	$[-4.4, 4.2]$	14.39
Learned	37.2	1.2	$[-1.9, 1.9]$	85.7	1.3	$[-2.2, 2.1]$	57.6	2.5	$[-4.2, 4.2]$	67.1	2.7	$[-4.6, 4.4]$	**0.26**
EDL	35.8	1.1	$[-1.8, 1.7]$	84.1	1.2	$[-2.1, 1.9]$	58.3	2.5	$[-4.1, 4.0]$	68.0	2.5	$[-4.2, 4.0]$	0.32

Table II.7: Results of the voxel-wise uncertainty benchmark for stroke lesion segmentation. The mean (μ), the standard error of the mean (SEM) and the 90% confidence interval (CI) are computed using bootstrap (M=15000). ↑ indicates that the metrics should be maximized. The highest metrics for each dataset are indicated in **bold**. UEO: Uncertainty-Error Overlap.

II.7 Chapter conclusion

As presented in this chapter, the literature focusing on UQ in medical imaging applications is flourishing. For segmentation tasks, many different techniques can be applied to generate uncertainty maps, providing indicators of potentially incorrect voxels. This is because most popular approaches including MC dropout and Deep Ensemble were initially proposed for 2D classification tasks [107, 97]. Directly applying these methods without adaptation on 3D segmentation tasks thus produces one uncertainty estimate per voxel. In practice, voxel-level uncertainty is useful but it is not sufficient to fully measure the confidence of an automatic analysis. This leads to the development of emerging UQ methods operating at the instance level (e.g. lesion) and case level (OOD detection or segmentation quality assessment).

Many popular UQ frameworks (MC, DE, TTA or Softmax uncertainty) rely on the output probabilities of the model to compute an uncertainty score (e.g. MSP, variance, or entropy). The implicit pre-requisite is that the probabilities are indeed representative of the model's confidence, which, as shown in the calibration benchmark, is not usually the case. Training models with the usual Soft Dice loss indeed leads to poorly calibrated models. Simple modifications to the loss objective, as done with the Dice++ loss, can highly correct this pitfall. Then, a post-hoc scaling strategy such as Temperature Scaling can be adopted to obtain state-of-the-art calibrations.

Ultimately, popular voxel uncertainty estimators were compared, which can be divided into two groups: sampling-based approaches (MC, TTA, DE) and sampling-free approaches (Softmax, EDL, Learned uncertainty). DE appeared as the most interesting baseline as it allows the optimization of the segmentation accuracy while providing high-quality uncertainty estimates. It is linked to an overhead of computational during the model development stage, however it is efficient regarding inference speed. Overall, disentangling segmentation performance and uncertainty quality is particularly complicated, as evaluating voxel uncertainty generally involves detecting misclassified voxels.

In both the calibration and voxel uncertainty benchmarks, our experiments showed a drop in segmentation accuracy in domain-shift settings (MSLUB, 1.5 Tesla, and SSA datasets). More worryingly, the quality of uncertainty estimates also dropped, which indicates that the reliability of uncertainty estimates is reduced on domain-shift data.

As a conclusion to this chapter, we demonstrated how high-quality voxel uncertainty maps can be derived from properly calibrated models. However, inspecting the overall uncertainty map to identify error-prone areas is time-consuming and may not be aligned with the expectations of end-users. For applications involving the detection of lesions, the clinician's attention is rather situated at the lesion level. For a rapid validation of the automated prediction, a lesion uncertainty score would allow the clinician to directly review the ones flagged as uncertain, which may be false positive findings. However, translating voxel uncertainties to lesion uncertainty is not trivial, as the lesion uncertainty may not simply resume to the sum or mean of its voxel components. Thus, more complex aggregation techniques to quantify structural uncertainty are explored in detail in the following Chapter.

Chapter III

Quantifying lesion uncertainty using auxiliary classifiers

Contents

III.1 Problem definition

In the previous chapter, voxel-level uncertainty estimates have been presented, allowing the review of uncertain areas in the output segmentation that may be error-prone. However, these so-called uncertainty maps mainly highlight uncertainty at the border between classes. Moreover, the review of the 3D uncertainty map can be time-consuming and hard to interpret for non-experts. Finally, for pathologies such as Multiple Sclerosis, the attention of the clinician is at the lesion-level, rather than at the voxel level. To alleviate these limitations, lesion uncertainty scores are desired, which would allow the user of the software to directly review uncertain lesions in the prediction that may result in false positive findings. More formally, our objective is to obtain a module able to associate a single uncertainty score to each identified lesion instance in the segmentation.

In practice, the correctness of each lesion detection can be assessed using the ground truth segmentations, and three categories of lesions can be further defined:

- True Positive lesion (TP_{les}): the identified lesion has a non-null intersection with one or several ground truth lesions.
- False Positive lesion (FP_{les}): the identified lesion does not intersect any ground truth lesions.
- False Negative lesion (FN_{les}): the ground truth lesion does not intersect any predicted lesions.

In lesion-level UQ, uncertainty can only be quantified for lesions that have been detected, including TP_{les} and FP_{les} instances. FN_{les} are by definition not present in the output segmentation, and they thus fall outside the scope of the lesion-level uncertainty paradigm [168, 250]. Moreover, True Negative lesions (TN_{les}) are undefined. Lesion-level UQ also requires the development of specific metrics to evaluate the quality of uncertainty. Note that we use in this chapter the denominations FP_{les}, TP_{les}, TN_{les}, and FN_{les} to indicate the status of lesions to mark a distinction with FP, TP, FN, and TN voxels.

The experimental setting of this chapter, illustrated in III.1.1, is as follows. First, as the previous chapter demonstrated the relevancy of DE to obtain voxel-wise uncertainty estimates, a DE of 5 individually trained Dynamic U-Nets is composed, for each application. It is used to generate voxel-wise predictions (entropy and segmentation). Then, a connected component analysis (CCA) is carried out to identify each lesion in the raw segmentations, using a 26-connectivity — meaning that a lesion is defined by voxels that are interconnected by their faces, edges, or corners. Then, each lesion is processed by a Lesion Uncertainty Module to extract a structural uncertainty score. In this chapter, lesion-level uncertainty quantification is investigated through 3 lesion-oriented tasks: cross-sectional and longitudinal MS lesions segmentation in brain FLAIR MRI, which is the core of Pixyl expertise, as well as lung nodules detection in chest CT.

A direct and immediate approach to quantify lesion uncertainty would be to aggregate the voxel uncertainty scores (e.g. entropy) for each lesion, using a summary statistic such as the arithmetic mean. It supposes that each voxel equally contributes to the overall lesion score.

This baseline approach is called **Mean Entropy** in the following. In this chapter, we aim to determine if improvements over this baseline can be obtained by using more sophisticated models for lesion uncertainty. More precisely, the proposed paradigm is to use auxiliary classifiers to quantify lesion uncertainty. It operates as follows: a classifier is trained to distinguish between true and false positive lesions using the predictions on the training and validation images, using a standard binary classification setting. At test time, the probability $P(\text{FP}_{\text{les}})$ that the lesion is a false positive is used as the lesion uncertainty score. Interestingly, this uncertainty score is easily interpretable for clinicians, as compared to other metrics such as the average entropy. Implementing this classifier-based approach requires performing learning on lesion instances. This involves being able to build lesion representations that are suitable for training. Three different lesion representations are investigated in this thesis:

- The **Feature representation** (Section III.3) consists in extracting a set of meaningful features from the voxel maps (image, entropy, and segmentation) for each lesion. They are used to train a ML classifier (e.g. Random Forest, Logistic Regression, or Support Vector Machines) to predict the status of each lesion (TP_{les} or FP_{les}).
- The **Bounding box representation** (Section III.4) is obtained by extracting $32 \times 32 \times 32$ bounding boxes centered on each predicted lesion. Then, a classic CNN classifier is trained to predict $P(\text{FP}_{\text{les}})$ the probability that the lesion is a FP_{les}.
- The **Graph representation** (Section III.5) leverages a graph representation of the lesion, following which voxels are converted to nodes and node features are computed from the voxel-wise predictions. Then, a Graph Neural Network (GNN) is trained to predict $P(\text{FP}_{\text{les}})$.

Several research questions are open. First, is learning from lesion instances viable to quantify lesion uncertainty? Do the classifier approaches offer a better quality of lesion uncertainty than simpler aggregation methods such as the Mean Entropy? Which lesion representation (feature, bounding box, or graph) is the most appropriate to train the auxiliary classifiers? These different points will be investigated in this chapter.

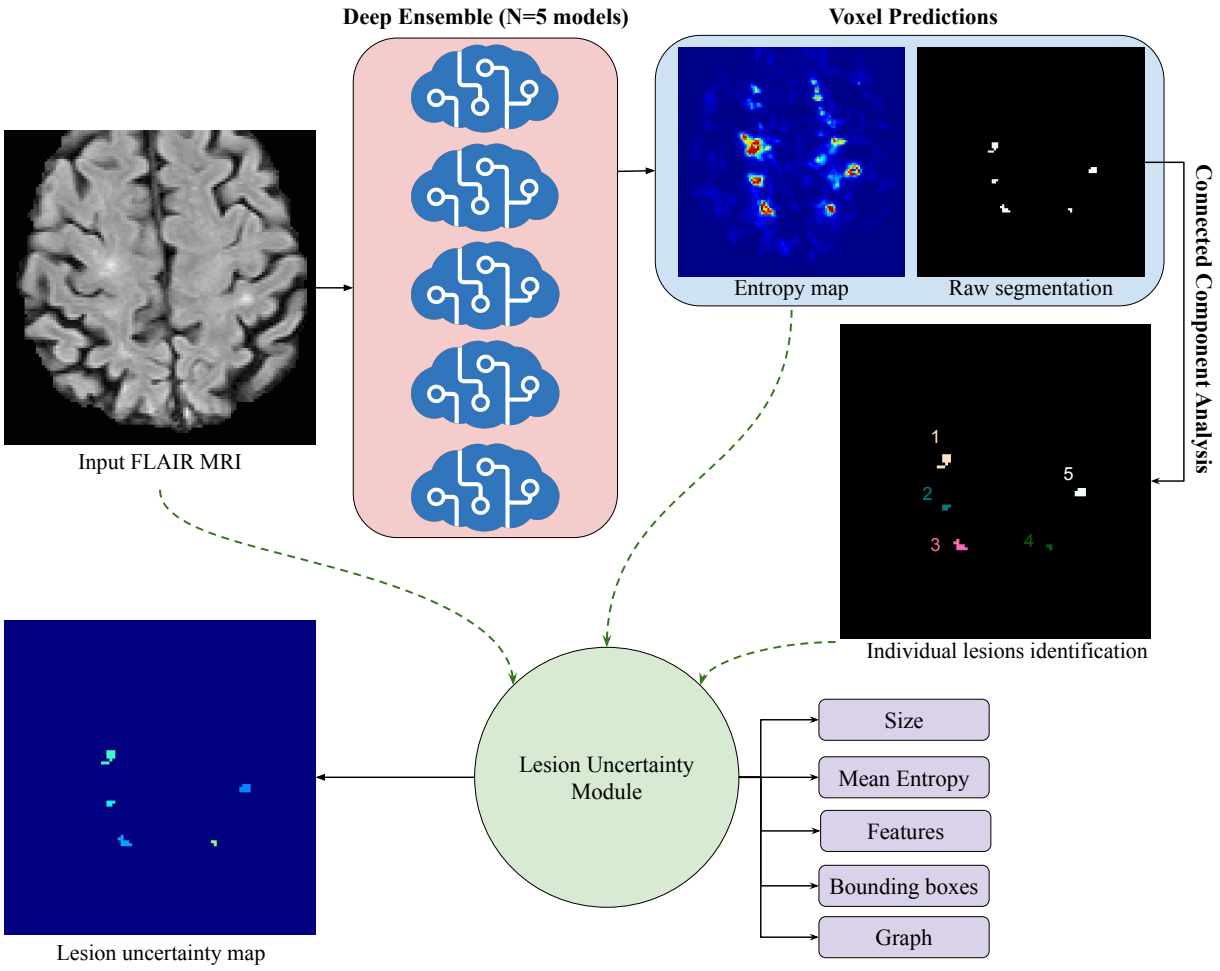


Figure III.1.1: Illustration of the inference process for lesion-level uncertainty experiments. Voxel predictions (entropy and segmentation) are first derived from a Deep Ensemble composed of 5 trained models. Then, Connected Component Analysis is performed to identify lesion instances in the raw segmentation. Each lesion is passed through a Lesion Uncertainty Module which computes lesion-level uncertainty scores.

III.2 Additional contributions to the paper "Beyond Voxel Prediction, identifying lesions you can trust"

This chapter is based on the work previously presented in the paper *Beyond Voxel Prediction, identifying lesions you can trust* [251]. However, several improvements and novelties are presented in this chapter. First, the paper used a MC dropout model as a voxel-wise prediction generator. Here, DE is used, as it is a stronger voxel uncertainty generator. Then, applications to lung nodules and longitudinal MS lesions detection are added. Moreover, robustness testing is added through the addition of the two domain-shift datasets (1.5 Tesla and MSLUB) for cross-sectional MS experiments. The feature-based model is improved using image radiomics in addition to uncertainty features. The bounding-box model is a novel addition. Moreover, the correlation between the computed and expert-derived uncertainty

scores (inter-rater variability and subtlety) is proposed for the lung nodules and longitudinal MS experiments.

III.3 A feature-based Machine Learning model for lesion-level uncertainty

The first presented application of lesion-wise uncertainty estimation is the detection of MS lesions in a cross-sectional setting. The datasets used in this part are identical to the ones described in the previous chapter (see Section II.6.1). In this setting, a single brain can present up to a hundred individual lesions. The different proposed methods and evaluation paradigms are introduced for this first application.

III.3.1 Training dataset generation

The **Feature, Bounding box, and Graph** approaches are all based on the concept of training an auxiliary classifier to predict the probability that the lesion is a FP_{les}, used as an uncertainty score. This requires the building of a labeled dataset comprising examples of TP_{les} and FP_{les} in sufficient amounts to allow for supervised training. In this thesis, the training strategy proposed by Bhat et al. [169, 170] is adopted. It consists in generating voxel-wise predictions on the training and validation images using the trained DE, which is now fixed. Each unique lesion is identified from this set of predictions using CCA, and a status (TP_{les} or FP_{les}) is attributed to each of them using the ground truth masks. To do so, a lesion pairing strategy is adopted and presented in further detail in Section III.6.1. For the MS cross-sectional experiment, this process generates a dataset composed of 8051 lesions for training (comprising 6854 TP_{les} and 1197 FP_{les}) and 1028 for validation (comprising 852 TP_{les} and 176 FP_{les}). In both the lesion-level training and validation datasets, the ratio of TP_{les} to FP_{les} is approximately $1:5$. Then, to train the feature-based ML model, features are extracted for each unique lesion, as presented in the next section.

III.3.2 Feature selection

One approach to fuse voxel-level information into lesion-level uncertainty is to use a feature extraction paradigm. For each identified lesion, a set of meaningful features is extracted and then used as inputs to train a ML model (e.g. Random Forest, SVM) to distinguish between TP_{les} and FP_{les}. The features should convey meaningful information for uncertainty quantification. In prior works, the feature vector included shape-based attributes, obtained from the binary lesion mask [169, 170]. This typically includes the size of the lesion and geometrical attributes such as flatness or sphericity. Alternatively, features can be collected from the voxel uncertainty maps, such as the average lesion uncertainty [252]. In addition to these 2 categories of features, one extra category is added here in the form of image intensity radiomics [171]. This builds on the intuition that for FP_{les} detection, important information can be obtained from the input image, such as the average lesion intensity.

In practice, we use PyRadiomics [171] to automatically compute a set of 107 features: 93 intensity features extracted from the input image, and 14 shape features extracted from the

binary lesion mask. We also add 3 features collected from the entropy map, namely the mean entropies of the overall lesion, the lesion contour, and the lesion interior. This process leads to a set of 110 features extracted from the voxel uncertainty maps, input image, and binary lesion mask. A summary of the extracted features is presented in Table III.1.

Input	Feature Type	N features
Image	First Order	18
	Gray Level Co-occurrence Matrix	24
	Gray Level Size Zone Matrix	16
	Gray Level Run Length Matrix	16
	Neighbouring Gray Tone Difference Matrix	5
	Gray Level Dependence Matrix	14
Lesion Mask	Size and Shape	14
Entropy Map	Mean overall, interior and boundary uncertainty	3
Total	-	110

Table III.1: Description of the 110 features used to train the ML lesion classifier. They are extracted from the input image, binary lesion mask, and entropy map, respectively.

III.3.3 Feature reduction

In Bhat et al. [169], authors demonstrated that radiomics extracted from voxel-wise uncertainty maps were heavily correlated in the context of liver lesion segmentation. They showed that the classifier can be simplified by resorting to a feature reduction step while maintaining the accuracy of the FP_{les} detection. Here, the same process is reproduced to reduce the number of features and monitor their correlation. The reduction algorithm works as follows:

- Features are collected on the training dataset samples, yielding to a feature vector $\boldsymbol{x} \in \mathcal{R}^m$, with m being the number of features.
- The correlation matrix $C \in \mathcal{R}^{m \times m}$ is constructed by computing Spearman's rank correlation coefficient between each pair of features. The matrix is transformed into a symmetric matrix by computing: $C' = 0.5 \times (C + C^T)$
- The correlation matrix is transformed into a distance matrix by taking: $D = 1 - |C'|$. As a result, a high correlation is linked to a low distance.
- Hierarchical clustering is performed from the distance matrix D. The result of this clustering can be visualized using a dendrogram.
- Clusters are formed by cutting the dendrogram at a given threshold, set to 1. Features below a cut form a single flat cluster. For each cluster, one single feature is kept by taking the one with the maximum mutual information with respect to the ground truth labels.

This process is here applied to the task of cross-sectional MS detection, which allows the reduction of the number of features from 110 to 7. The correlation matrix of the original set of 110 features is presented in Figure III.3.1. It can be noticed that many pairs of features exhibit strong positive (pale orange) or negative (pale blue) correlations. The 7 features

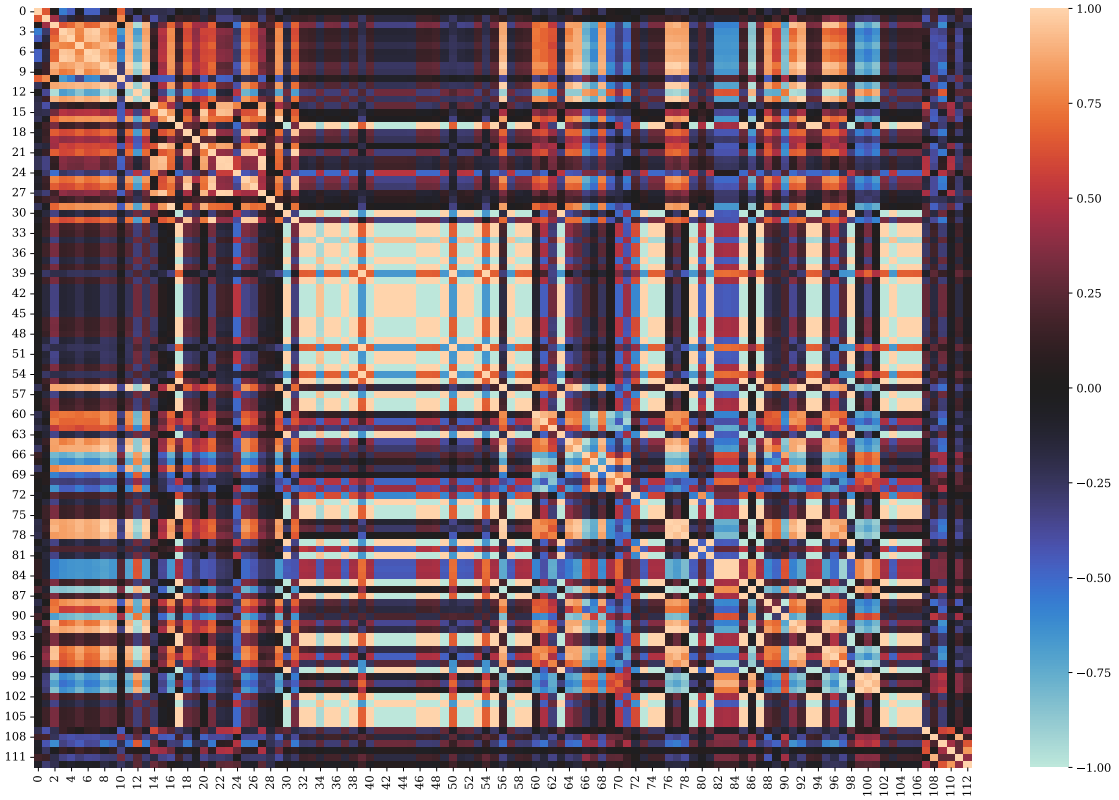


Figure III.3.1: Correlation matrix for the cross-sectional MS lesions classifier, on the 110 original features. The color map indicates Spearman's rank correlation coefficient, with pale blue and orange indicating high correlations between the features.

identified by the hierarchical clustering are Sphericity, Surface-Volume-Ratio, first-order Total Energy, first-order Uniformity, Mean Entropy, Interior Entropy, and Small Dependence Emphasis.

III.3.4 Machine Learning model development

The ML model is developed using the Sklearn framework [253]. Features are first normalized using a zero-mean unit-variance scaler. Then, a grid-search cross-validation paradigm is employed to identify the best hyper-parameters for the model, using 5 folds. Different models are tested: a Logistic Regression model, a Support-Vector Classifier, and a Random Forest model. Hyper-parameters tested during the cross-validation scheme are indicated in the Appendix A2. To identify a final ML classifier for the rest of the experiment, the balanced accuracy scores on the validation dataset are compared in Table III.2. The balanced accuracy score is a variant of the standard accuracy that takes into account data imbalance. It is suitable here as $\mathrm{FP}_{\mathrm{les}}$ are more rare than $\mathrm{TP}_{\mathrm{les}}$. It is defined as:

$$\mathrm{BAcc} = \frac{1}{2}\frac{TP}{TP+FN} + \frac{1}{2}\frac{TN}{TN+FP} \tag{III.3.1}$$

On the validation dataset, it appears that the Logistic Regression model with the full set of features (110) achieves the best balanced accuracy (0.78), while the Random Forest performed significantly worse. For each model, using the feature reduction strategy leads to a slight drop in the balanced accuracy, except for Random Forest. As a result, the Logistic Regression model with 110 features will be kept for the rest of the experiments.

Model	N features	BAcc
Logistic Regression	110	**0.78**
Random Forest	110	0.60
SVC	110	0.76
Logistic Regression	7	0.75
Random Forest	7	0.62
SVC	7	0.75

Table III.2: Classification performance on the MS validation dataset for the 3 tested ML classifiers, with and without feature reduction. BAcc: Balanced Accuracy score.

III.3.5 Feature contribution

Interestingly, the Logistic Regression model is interpretable as the importance of each feature is directly accessible. This provides some insights regarding the most meaningful features for lesion uncertainty quantification. In Appendix A.3.1, the weight of the top 10 most important features are provided. It appears that a mix of intensity, uncertainty, and shape features constitutes the top 10. The highest coefficient is attributed to the image radiomic **glrlm RunEntropy**, while the second is the **average interior entropy**. The third one is **Maximum2DDiameterRow**, a shape radiomics linked to the size of the lesion.

III.4 A bounding-box CNN for lesion uncertainty quantification

III.4.1 Concept

One downside of the feature-based approach previously introduced is that it requires a feature-engineering step to identify useful attributes to quantify lesion uncertainty. Moreover, the retained Logistic Regression model is a rather simple classifier, and it can be hypothesized that better FP_{les} detection performance could be attained by more complex classifiers. Another potential limit is the lack of global context to evaluate the lesion status. Indeed, the radiomics are extracted only within the lesion mask, thus excluding all the lesion surroundings that could be relevant to evaluate its certainty.

To alleviate these limits, a CNN approach is proposed. The framework is presented in Figure III.4.1. Cubic bounding boxes with a shape of $32 \times 32 \times 32$ are extracted from the voxel-wise volumes (input FLAIR, entropy map, and segmentation), centered on each lesion. Then, a CNN model is trained to predict the lesion status (TP_{les} or FP_{les}) using these 3 concatenated patches. This process i) alleviates feature-engineering and selection, as the CNN

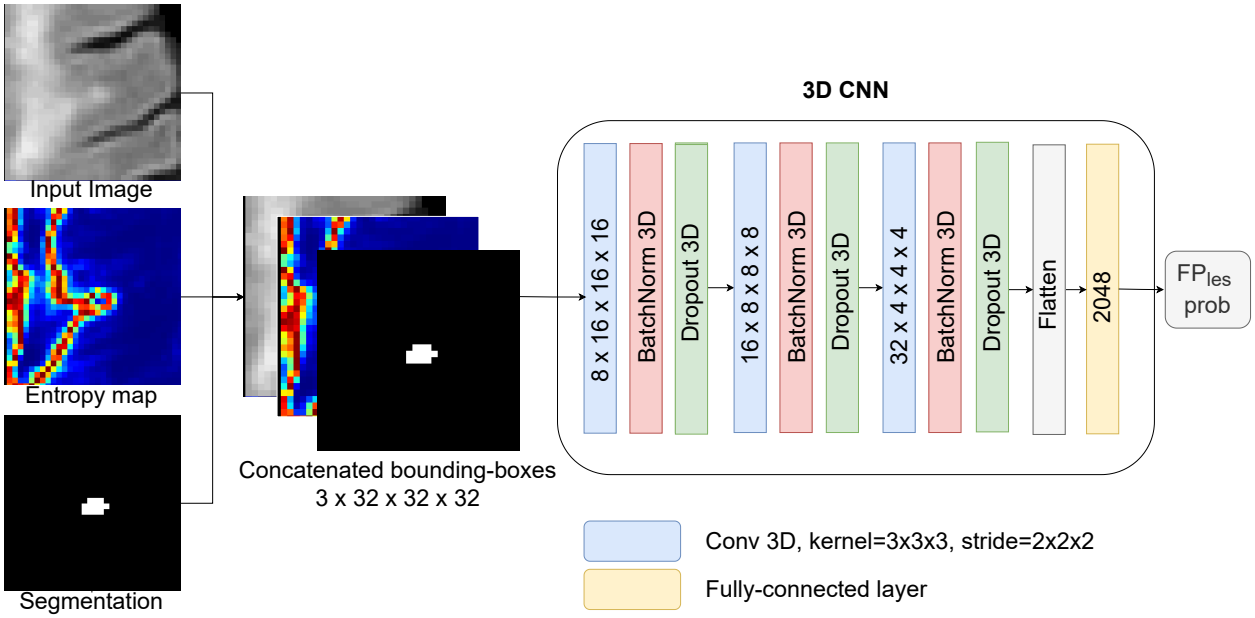


Figure III.4.1: Illustration of the bounding-box 3D CNN to quantify lesion uncertainty. The model receives as input bounding boxes centered on the lesion and predicts $\mathrm{FP}_{\mathrm{les}}$ the probability that the lesion is a false positive.

automatically learns meaningful parameters to perform its task, and ii) allows the inclusion of spatial context for each lesion, such as the proximity to other lesions or the location within the brain. The downside is the lack of interpretability, as it becomes cumbersome to identify which patterns in the bounding boxes contributed to the lesion classification. Another downside that can be anticipated is the need for larger lesion databases to allow for training, with a sufficient amount of samples from each class ($\mathrm{TP}_{\mathrm{les}}$ and $\mathrm{FP}_{\mathrm{les}}$) to allow for supervised learning.

III.4.2 CNN architecture

In terms of NN architecture, a simple configuration is used here (as presented in Figure III.4.1). Three convolutional blocks are stacked, each composed of i) a 3D convolution with an isotropic kernel size of $3 \times 3 \times 3$ and a stride of $2 \times 2 \times 2$ followed by ii) a 3D batch normalization layer and iii) a dropout layer. In each block, the bounding boxes are spatially downscaled by a factor of 2 in each direction, while the number of features gradually increases with 8, 16, and 32 features. Finally, the resulting feature representation is flattened to a 1D tensor of dimension 2048. A final FC layer transforms this latent vector into 2 output units, containing the probabilities that the input lesion is a $\mathrm{TP}_{\mathrm{les}}$ and a $\mathrm{FP}_{\mathrm{les}}$, respectively. In total, this lightweight 3D CNN contains 76600 parameters, which is rather low for a neural network classifier operating on 3D medical images. The choice of using a low-complexity CNN arises from the limited size of the medical-image databases used in the experiments.

III.4.3 Training setting

Training is performed in a standard supervised classification setting. The cross-entropy loss (Equation I.2.3) is used to train the 3D CNN to predict the status of the input lesion. However, as presented in Section III.3.1, TP_{les} samples are much more frequent than FP_{les} samples ($5:1$ ratio). To take into account this class imbalance during training, class weights are introduced in the loss formulation, corresponding to the inverse of the class frequency in the training dataset. This is akin to the **balanced weighting** implemented in Sklearn. More precisely, the weights w_c for each class are computed as follows:

$$
\begin{aligned}
w_c &= \frac{N}{C \times \sum_{n=1}^{N} \mathbf{1}\{\boldsymbol{y}^{(n)} = c\}} \\
w_1 &= \frac{8051}{2 \times 6854} = 0.587 \\
w_2 &= \frac{8051}{2 \times 1197} = 3.363
\end{aligned}
\tag{III.4.1}
$$

were $C = 2$ is the number of classes, N the number of train samples, w_1 and w_2 are the weights of the TP_{les} and FP_{les} classes, respectively. In practice the weight of the FP_{les} class is increased to take into account its low number of occurrences in the training set. Then, a weighted cross-entropy loss can be formulated:

$$
\mathcal{L}_{wce} = -\sum_{k=1}^{2} w_k \log \frac{\exp(x_k)}{\sum_{c=1}^{2} \exp(x_c)} \mathbf{1}\{\boldsymbol{y} = k\} \tag{III.4.2}
$$

where x_k are the scores predicted by the CNN. The ADAM optimizer [37] is used with a fixed learning rate of 2×10^{-4} and a batch size of 16 bounding boxes. Training is carried out until the validation balanced accuracy seizes to improve for 20 epochs. A dropout rate of 20% is employed in each convolutional block to reduce overfitting. A data augmentation strategy is adopted comprising flipping rotation, and gamma alterations. Note that intensity augmentations are only applied to the input FLAIR bounding box, as the entropy map should not be treated as a standard intensity map in the data augmentation pipeline.

Even though the model has a small number of parameters and that dropout is used in each block, overfitting is observed starting at the 50000-th training step, as illustrated in Figure III.4.2. However, the CNN achieves a satisfying validation balanced accuracy before the start of the overfitting.

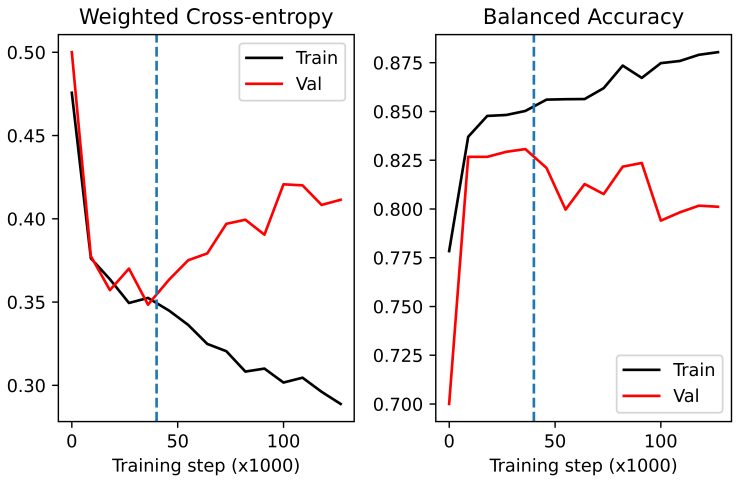


Figure III.4.2: Training and validation scores for the bounding box CNN trained on cross-sectional MS lesions. Left: weighted cross-entropy. Right: balanced accuracy. An overfit can be observed, starting around the 50K training step, indicated by the vertical dashed line.

III.5 A graph approach to lesion uncertainty quantification

III.5.1 Motivations

Finding an appropriate lesion representation to allow for the training of an auxiliary lesion classifier is complex, as lesions exhibit heterogeneous shapes, sizes, and appearances. Indeed, lesions can exhibit complicated, non-euclidean geometry as illustrated in Figure III.5.1. Here, Lewiners's marching cube algorithm [254] is employed to transform the 3D binary lesion masks into meshes, exhibiting the heterogeneous shapes they can display.

Because of this, the previously presented approaches may exhibit weak points. The feature-based model relies on features averaged over the lesion mask voxels (such as the average entropy, the average contour entropy, or the average lesion intensity...) which may discard subtle relationships between the lesion voxels. The bounding-box model relies on a fixed bounding box size of $32 \times 32 \times 32$, which although suitable for most lesions, may be suboptimal for very small or very large lesions. To circumvent this, a desirable representation should 1) preserve the entire available lesion voxel information and overall lesion structure and 2) be flexible enough to handle the non-euclidean geometry of the lesions. A third desirable property is a limited model complexity, due to the limited size of the medical-image datasets used in our experiments.

With these properties in mind, a natural way of handling the lesions is to use a recent DL framework suitable for non-euclidean data: Graph Neural Networks (GNN), which have increasingly gained interest for their performance and their ability to learn the structure of complex non-euclidean graph data. This way of modeling lesions allows to perform training

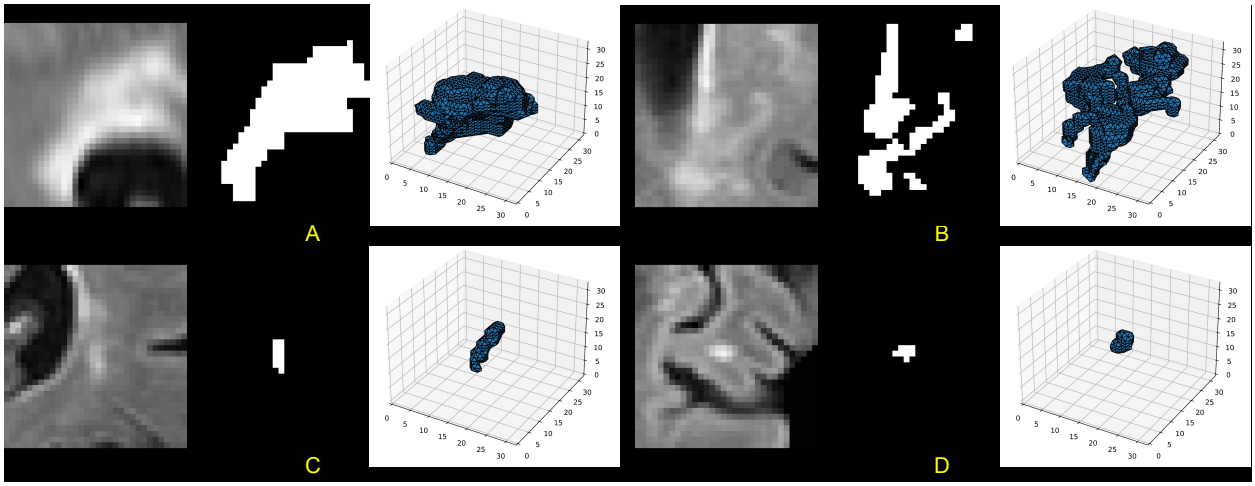


Figure III.5.1: Input image, predicted lesion masks and associated meshes obtained using the Marching's cube algorithm. It displays the high heterogeneity in the size and structure of the lesions.

directly on the lesion instances and offers a flexible framework for characterizing lesions through node features.

This graph-based approach is motivated by the recent success of GNNs in exploiting voxel uncertainties. More formally, Soberanis et al. [199] proposed to represent a medical image segmentation by an undirected graph and use the voxel uncertainty map derived by a MC dropout model to define node features. They then propose to classify each node (i.e. voxel) as correct, or incorrect, thus akin to a node classification task. The operative goal is to refine the segmentation by removing incorrect voxels. In this section, a similar idea is implemented, yet each unique lesion in the segmentation is represented by a distinct graph, and each graph is further classified as a TP_{les} or FP_{les} (thus the task is graph classification). To achieve this, lesions must first be represented by graphs, and then a GNN can be trained to classify them into TP_{les} and FP_{les}. These 2 steps are presented in the following.

III.5.2 Graph notations and Graph Neural Networks

III.5.2.1 Notations

In this section, the notation used to define graphs is introduced. Let $\mathcal{G}$ be an undirected graph defined by a set of nodes $\mathcal{V}$, connected by a set of edges $\mathcal{E}$. We write n the number of nodes in the graph, such as $|\mathcal{V}| = n$. The graph is defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \boldsymbol{A}\}$ where $\boldsymbol{A}$ is the adjacency matrix. If there exists a link between the nodes i and j, then matrix cell $\boldsymbol{A}(i, j) = w_{ij}$ where w_{ij} is the weight of the edge. For unweighted graphs, $w_{ij} = 1$. If there is no connection between nodes i and j, then w_{ij} is set to 0. The neighboring nodes of a node v are denoted by $\mathcal{N}(v)$. Finally, nodes are characterized by node attributes (or node features), defined by vectors $X_v \in \mathcal{R}^d$ for $v \in \mathcal{V}$, with d the number of attributes for each node.

III.5.2.2 Converting lesions to graphs

Representing lesions by graphs requires several design choices. The proposed approach is illustrated in Figure III.5.3. First, to incorporate voxels in the immediate vicinity of the lesion that contain meaningful information for uncertainty quantification, each lesion mask is dilated using a binary morphological operation. Then, each voxel of the dilated lesion mask is associated with a node in the graph. Thus, if the dilated lesion contains V voxels, its corresponding graph will include V nodes. As a result, each node in the graph is linked to a voxel of coordinates (x, y, z) in the medical image. Edges are added between each adjacent voxel, following a 26−connectivity. This indicates that a node will be connected to another node if the two associated voxels are touching via their faces, edges, or corners in the lesion mask. Undirected graphs are used here, meaning that edge weights are set to 1 when there is a link between nodes, and 0 else. Then, the last step is to define the set of node features used to train the GNN model. The following features for the node n_i associated to the voxel v_i positioned at (x, y, z) are used:

- The lesion intensity obtained at the (x, y, z) location in the input image.
- The voxel uncertainty obtained at the (x, y, z) location in the entropy map.
- The contour indicator, which is 1 if the voxel belongs to the contour of the lesion and 0 else.
- The degree of the node, corresponding to the number of edges linked to the node.
- The Euclidean distance to the contour, used to help the GNN locate the voxel in the overall lesion.
- The label at the (x, y, z) in the segmentation map, which is used to indicate if the voxel has been labeled as a lesion (label 1) or as background (label 0). Voxels labeled as 0 typically correspond to the voxels surrounding the lesion.

III.5.2.3 Using a Graph Isomorphism Network as auxiliary classifier

The task at hand here is graph classification, as the goal is to determine the label y of the lesion (TP_{les} or FP_{les}) based on its graph representation. To perform this task, a GNN can be employed to learn a representation of the entire graph h_G, such as $y = f(h_G)$. In the following, the operating principle of GNNs is presented.

A convenient way to introduce GNNs is to draw an analogy with the functioning of standard CNN models. Both principles are illustrated in Figure III.5.2. A CNN operates on images, which follow a grid-like, Euclidean data structure. In 2D, a pixel is associated with the coordinates (p_x, p_y) in the Euclidean space, where p_x represents the horizontal position and p_y is the vertical position. When applied to a pixel, a 2D convolutional filter has access to adjacent pixels to compute the filter response. For instance, a 3×3 convolutional kernel applied to a pixel C will have access to the 8 adjacent pixels (Figure III.5.2, left). In practice, the CNN can be seen as a special type of GNN where nodes correspond to the image pixels, and edges are added only between adjacent pixels in the image.

GNNs can be seen as a generalization of standard CNNs to handle any structure data, even if it cannot be represented in an Euclidean fashion. It is thus more versatile than the standard CNN. To learn the graph representation, GNNs operate using a neighbors aggregation scheme

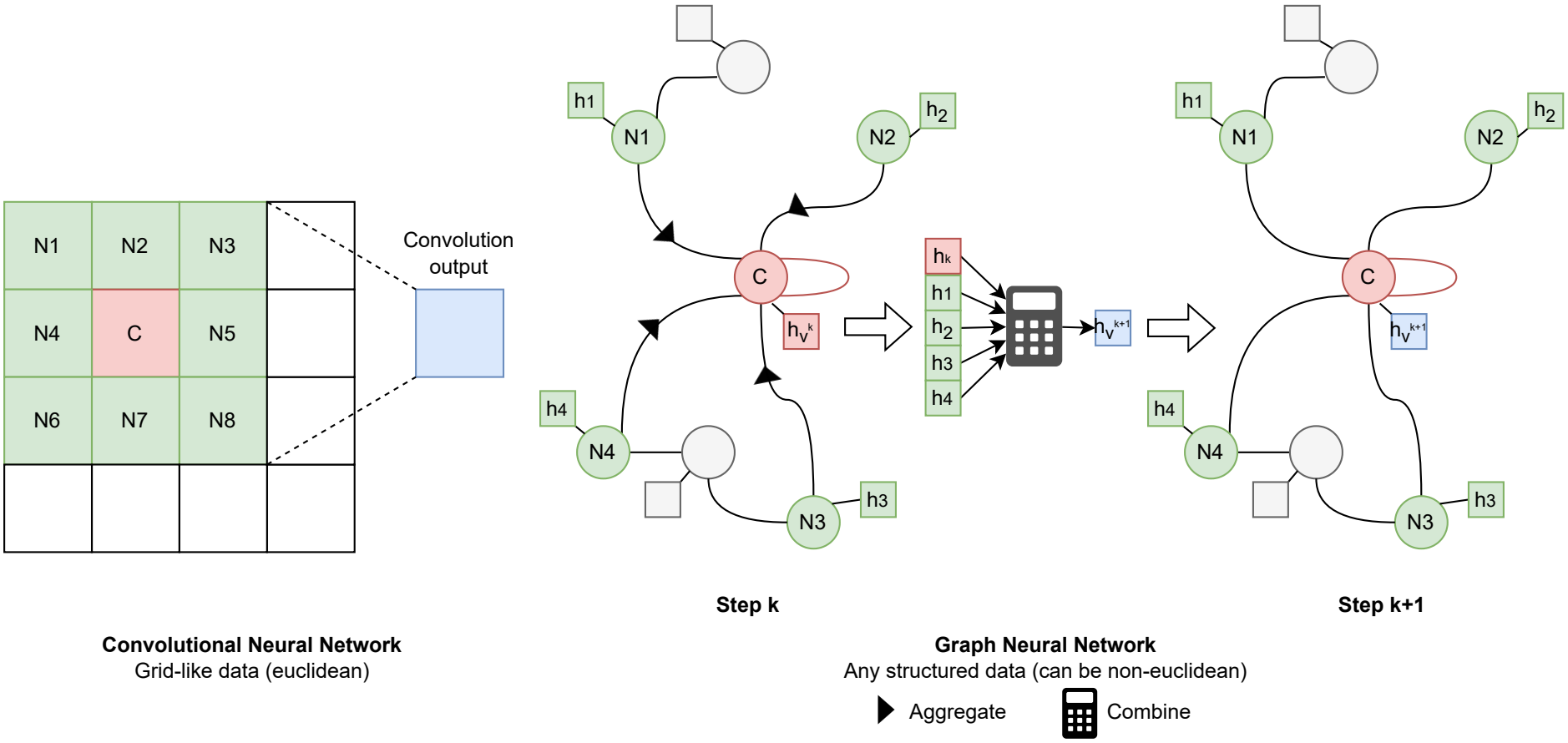


Figure III.5.2: Analogy between Convolutional and Graph Neural Networks. Left: a 3x3 convolution is applied at pixel C, belonging to a grid of 4x4 pixels. The convolution kernel output is obtained by aggregating information from neighboring pixels (N1-N8). Right: a graph composed of 5 nodes is represented (C, N1, N2, N3, N4). Each node is defined by a set of node features h. The update of the representation h_v^k of the central node C is further detailed. The first step is to aggregate the representations of the neighboring nodes. Then, a combine function is applied to compute h_v^{k+1}, the update node representation.

akin to the one performed in a CNN. More precisely, the node representation h_v is iteratively updated based on the representation of its neighbors (Figure III.5.2, right). This process is called message passing and is the core principle of GNN [255].

At the GNN input, the representation of the node is initialized to its node feature vector $h_v^0 = X_v$. Then, after k GNN layers, the node representation incorporates the information from the nodes k-steps away from it. More formally, the k-th layer of a GNN updates the representation h_v^k of the node v by performing the following operations [256, 257]:

$$a_v^k = \text{AGGREGATE}^k\{h_u^{k-1} : u \in \mathcal{N}(v)\} \tag{III.5.1}$$

$$h_v^k = \text{COMBINE}^k(h_v^{k-1}, a_v^k) \tag{III.5.2}$$

where $\mathcal{N}(v)$ is the set of nodes in the neighborhood of v. In practice, the AGGREGATE function allows the gathering of the representations of nodes in the vicinity of v to compute an aggregated feature vector a_v^k. Then, the COMBINE function updates the node representation at step k by relying on a_v^k and the preceding representation of the node (step $k-1$, h_v^{k-1}). Finally, at the last layer K of the GNN, a READOUT operation is necessary to aggregate the representations of each node and obtain a global graph representation allowing to perform classification:

$$h_G = \text{READOUT}(\{h_v^K | v \in G\}) \tag{III.5.3}$$

Several options can be found in the literature for the AGGREATE, COMBINE, and READOUT functions [256, 258]. Here, the Graph Isomorphism Network (GIN) paradigm is adopted, proposed by Xu et al. in 2018 for graph classification [256]. In GIN, the AGGREGATE and COMBINE functions are performed by MLP layers. To obtain the final graph representation (READOUT), the node representations at each different layer are first summed and then concatenated. This produces a graph representation h_G that is finally fed to a last FC layer that produces the class probabilities. The architecture of the GIN model used in the experiments is further detailed in Figure III.5.3.

$$h_v^k = \text{MLP}^k(h_v^{k-1} + \sum_{u \in \mathcal{N}_v} h_u^{k-1}) \tag{III.5.4}$$

$$h_g = \text{CONCAT}(\sum_{v=0}^{|\mathcal{V}|} h_v^0, ..., \sum_{v=0}^{|\mathcal{V}|} h_v^K) \tag{III.5.5}$$

III.5.3 Implementation details

The graph framework is implemented using the Deep Graph Library library [259], allowing the construction of graphs and the training of GNN models. The GIN model used here is

composed of 5 layers. Each layer is a succession of 2 FC layers interspersed with 1D batch normalization and ReLU activation (see Figure III.5.3). The hidden dimension is 32. To perform the graph readout, the input and hidden representations are concatenated, yielding a latent vector of dimensions 133. In total, the GIN model has 26700 parameters, which is low enough to authorize training to be performed on the CPU. To train the GIN model, the same training objective used for the bounding-box CNN is used, namely the weighted cross-entropy is used (Equation III.4.2).

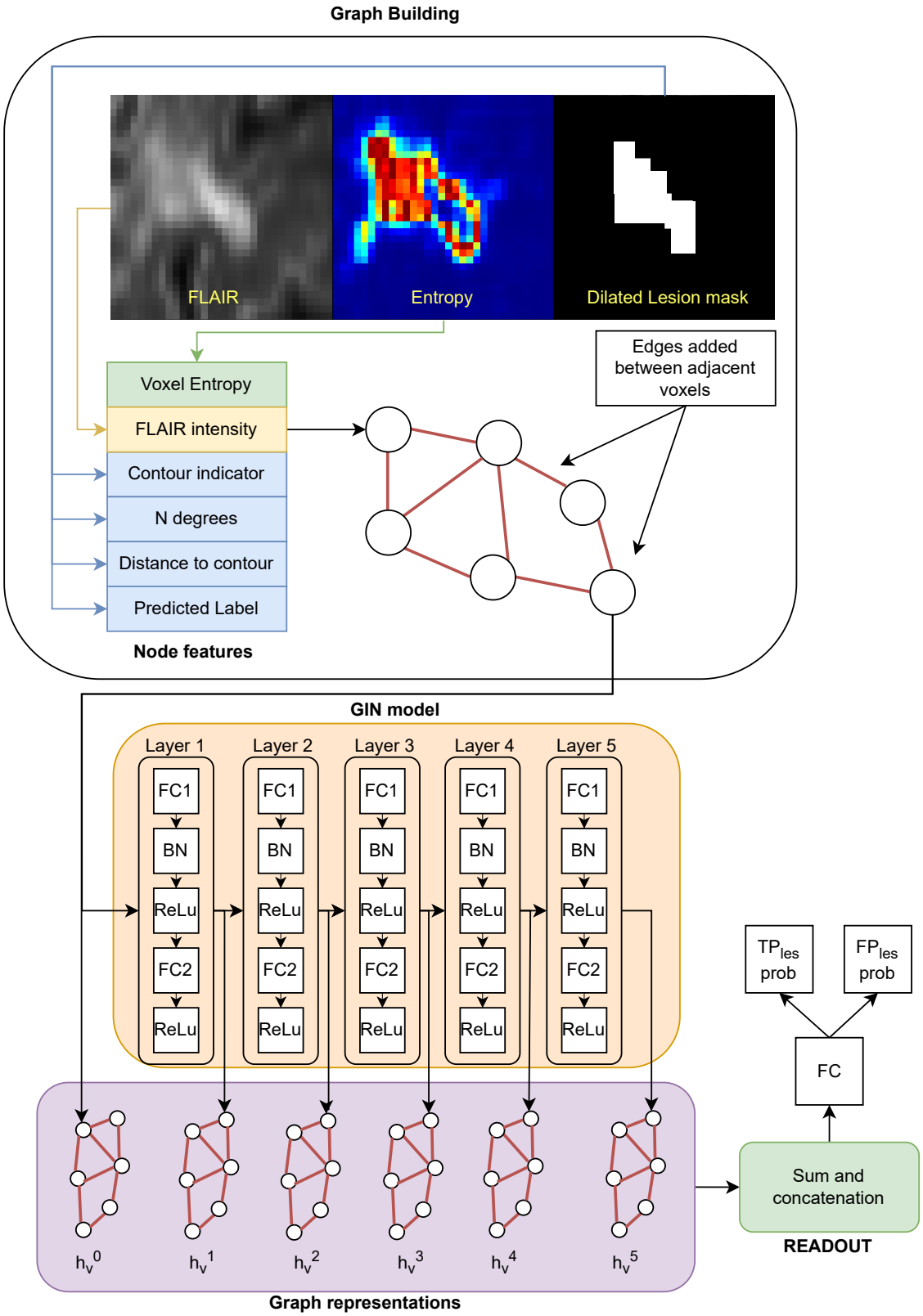


Figure III.5.3: Proposed pipeline for the graph-based approach, starting from the graph building using voxel-wise maps, followed by inference in the GIN model. FC: Fully-connected layer. BN: Batch Normalization.

III.6 Lesion-level metrics

III.6.1 Detection quality metrics

Voxel-level metrics are straightforward to compute as it simply involves comparing the predicted and ground truth value, for each voxel. At the lesion level, an extra step of lesion matching is required to identify TP_{les}, FN_{les}, and FP_{les}. This matching generally starts by performing a CCA of the predicted and ground truth masks. Then, from the two sets of individual lesions, the intersection between each possible pair is computed [129, 168, 170]. A predicted lesion is considered to be a TP_{les} if its overlap ratio with ground truth lesions reaches a sufficient amount. A loose definition is to set the threshold to any non-null intersection [170], but tighter definitions use a 25% [168] or 50% [129] minimum overlap ratio, estimated using the Dice or IoU scores. If the condition is not matched, then the lesion is flagged to be a FP_{les}. While these tighter definitions are motivated, there are some cases where they can mistakenly flag lesions as FP_{les}, although the detection is correct. This typically happens when there is a strong mismatch between the size of the predicted lesion and the reference lesion. In these cases, the Dice or IoU between the lesions will be low, leading to an estimation of a FP_{les}, even if the lesion is correctly detected. Examples of these edge cases are illustrated in Appendix A.1.1. Thus, to prevent these edge cases, the loose definition will be used in the experiments, and a lesion will be considered as TP_{les} if it has a non-null intersection with ground truth lesions.

When dealing with lesion segmentation, it is common that there is not a strict one-to-one relation between the predicted and ground truth lesions. Instead, many-to-one and one-to-many cases are frequent, especially for confluent Multiple Sclerosis lesions. In cases where *many* predicted lesions intersect *one* single ground truth lesion, the risk is to count each predicted lesion as distinct TP_{les}, thus inflating the count of true positives. To account for these situations, we opt for the graph-based lesion-matching algorithm proposed by Bhat et al. [169]. The concept is to build a graph connecting the predicted and reference lesions. Each lesion in the prediction and ground truth masks is associated with a node, and edges are added between lesions sharing a non-null intersection. This requires computing the intersection between each possible pair of predicted and reference lesions. Thus, the total number of comparisons is $n \times m$, with n and m the number of reference and predicted lesions, respectively. Figure III.6.1 presents two examples of many-to-one and one-to-many settings, with the corresponding lesion-matching graph. The IoU between each pair of overlapping lesions is presented on top of the edges. To count the number of TP_{les}, the algorithm counts the number of reference nodes with at least one edge, thus accounting for many-to-one cases. FP_{les} and FN_{les} then correspond to the count of predicted and reference solitary nodes, respectively.

From the counts of TP_{les}, FN_{les}, and FP_{les} extracted from the lesion matching graph, several lesion-wise detection metrics can be computed to estimate the lesion detection accuracy of the segmentation algorithm: the Lesion True Positive Rate (LTPR) and the Lesion False Discovery Rate (LFDR) [230]:

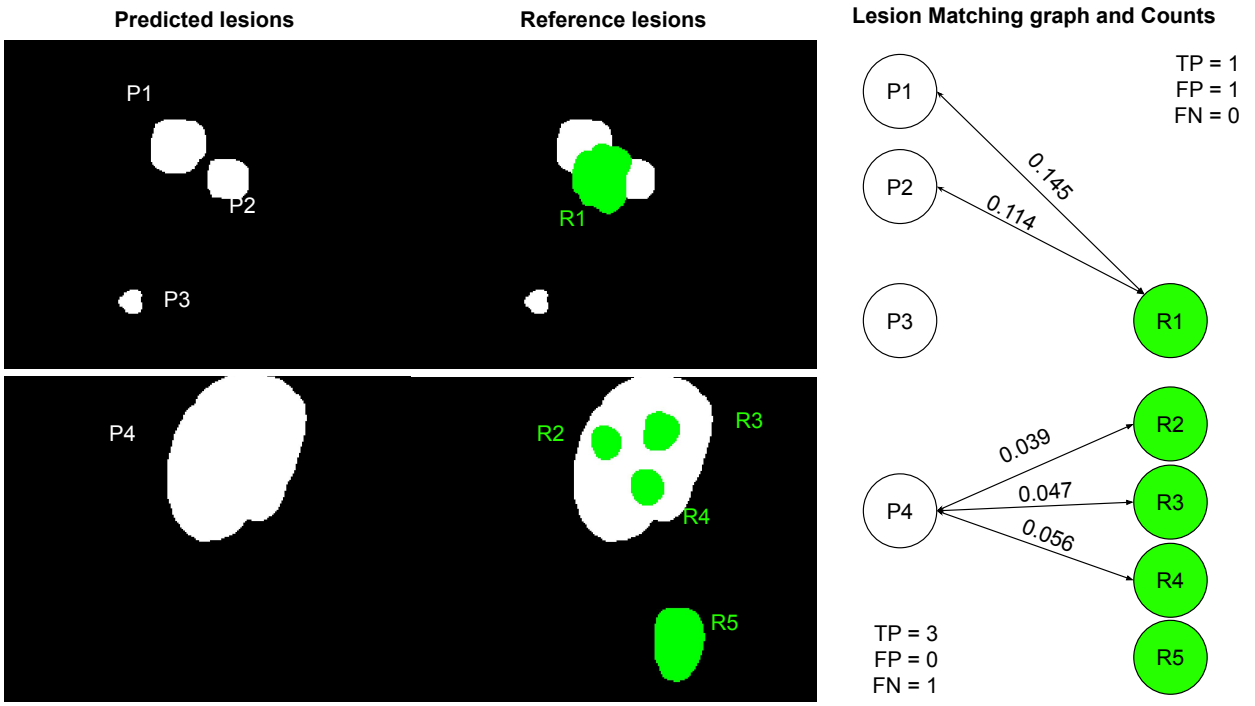


Figure III.6.1: Two examples of lesion matching. Top row: example of a many-to-one case plus one False Positive lesion. Bottom row: example of a one-to-many case plus one False Negative lesion. Predicted lesions are labeled as P_x and reference lesions as R_x. Edges are constructed between lesions that have a non-null intersection (estimated using the IoU score, indicated on top of the connecting edge).

$$\text{LTPR} = \frac{\text{TP}_{\text{les}}}{\text{TP}_{\text{les}} + \text{FN}_{\text{les}}} \tag{III.6.1}$$

$$\text{LFDR} = \frac{\text{FP}_{\text{les}}}{\text{TP}_{\text{les}} + \text{FP}_{\text{les}}} \tag{III.6.2}$$

III.6.2 Structural uncertainty quality metrics

The preferred approach to evaluate the quality of lesion-level uncertainty estimates is to measure the differentiability of TP_{les} and FP_{les} lesions based on their uncertainty scores [168, 250, 129]. Ideally, TP_{les} lesions should be associated with lower uncertainty than FP_{les} lesions, as illustrated in Figure III.6.2. Thus, by reviewing the most uncertain ones, the user of the software would be able to reject potential incorrect detections. In contrast, an uninformative lesion uncertainty score would not allow for this distinction. Nair et al. [129] and Molchaniva et al. [168, 250] assess this property through a lesion retention paradigm. Similar to the voxel-wise retention curves used previously, the idea is to rank the lesions from the most uncertain to the most certain. Then, each lesion in the scan is sequentially removed based on its uncertainty ranking, and the updated count of TP_{les} and FP_{les} is computed. Ideally, FP_{les} lesions should be associated with higher uncertainty than their TP_{les} counterparts, hence by

removing the most uncertain ones, the number of FP_{les} should decrease **quicker** than the number of TP_{les}. Lesion retention curves are obtained by plotting the number of TP_{les} versus the number of FP_{les} for each step of the stratification procedure. In this thesis, an evaluation paradigm based on the same criteria (FP_{les} lesions being associated with higher uncertainties than TP_{les} lesions) is proposed, based on standard classification metrics to ease clarity and interpretation. This evaluation strategy is presented in the following.

To evaluate the pertinence of the computed uncertainty score, a 2-class classification setting is adopted. FP_{les} are associated with the label 1, and TP_{les} with the label 0. At test time, each lesion in the test images is further associated with an uncertainty score, allowing to rank lesions from the most confident to the most uncertain. A binary decision (certain/uncertain) can be obtained by thresholding the uncertainty scores. By varying this threshold, standard classification metrics can be computed such as the Area under the Receiver Operating Characteristic curve (AUROC) or the Area Under the Precision-Recall curve (AUPR). For each threshold, the True Positive Rate (TPR) and False Positive Rate (FPR) are computed to draw the ROC curve:

$$\mathrm{TPR} = \mathrm{UTP}/(\mathrm{UTP} + \mathrm{UFN}) \tag{III.6.3}$$
$$\mathrm{FPR} = \mathrm{UFP}/(\mathrm{UFP} + \mathrm{UTN}) \tag{III.6.4}$$

whereas for the precision-recall curve, the FPR is replaced by the precision (Pre):

$$\mathrm{Pre} = \mathrm{UTP}/(\mathrm{UTP} + \mathrm{UFP}) \tag{III.6.5}$$

Here, the confusion matrix is obtained by matching the uncertainty status (certain/uncertain) and the correctness of the lesion detection (TP_{les}/FP_{les}):

- UTP: a lesion that is uncertain AND a FP_{les}.
- UFP: a lesion that is uncertain AND a TP_{les}.
- UTN: a lesion that is certain AND a TP_{les}.
- UFN: a lesion that is certain AND a FP_{les}.

Note that the precision-recall curve is more suitable for imbalanced settings, which is typically the case for cross-sectional MS lesions where FP_{les} are less common than TP_{les}. In this setting, the ROC curves tend to present an optimistic view of the performance, as it takes into account the UTN lesions (certain and true positive lesions) that are predominant in the cross-sectional MS experiment.

III.6.3 Results of the cross-sectional MS experiment

For the MS experiments, recall that three different test datasets are defined. Test ID contains images sharing the same distribution as the training images. MSLUB is a domain-shift dataset

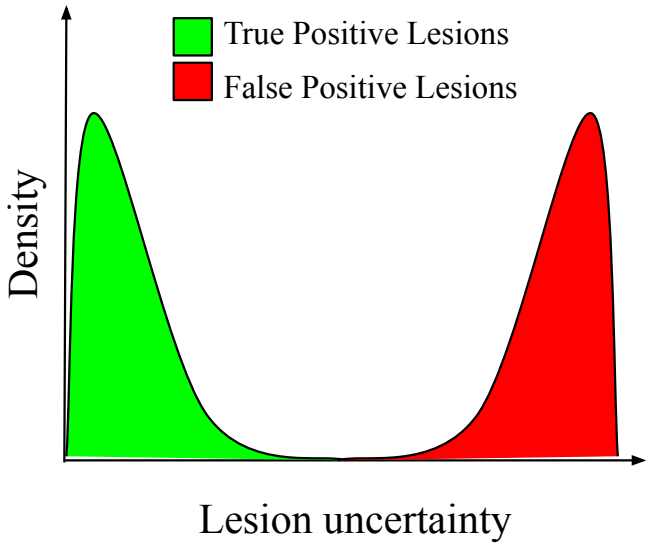


Figure III.6.2: Illustration of ideal lesion uncertainty quantification, where TP_{les} lesions are associated with lower uncertainties than FP_{les} lesions.

containing images acquired in a different imaging center, and with a different MRI device than the ones used in the training dataset. Finally, the 1.5 Tesla is a second domain-shift dataset with images acquired with a lower-quality MRI device. For each segmentation and uncertainty metric, bootstrapping [260] is performed to estimate the mean (μ), the standard error on the mean (SEM), and 90% confidence intervals (CI).

Table III.3 presents the quality of the segmentation provided by the Deep Ensemble (voxel-level and lesion-level metrics), as well as the counts of each type of lesions (TP_{les}, FP_{les}, and FN_{les}). Table III.4 and Figure III.6.3 present the quality of lesion-level uncertainty estimates (AUROC and AUPR) for each tested aggregation approach. Figure III.6.4 presents the distribution of lesion uncertainty scores for TP_{les} and FP_{les}, respectively, for each dataset and method. Finally, Figure III.6.5 provides illustrations of the computed lesion uncertainty scores.

Additionally to the **Mean Entropy** and classifier approaches (**Logistic**, **CNN** and **GNN**), the performance of a naive **Size** approach is presented for comparison. It is obtained by defining the lesion uncertainty U score as the inverse of its size S, expressed in the number of voxels, such that $U = 1/S$. The goal of this approach is to serve as a naive baseline, as small lesions are generally more likely to be FP_{les}. This simple technique does not consider any voxel-wise uncertainty score, hence it is expected to be outperformed by all other methods.

First, in terms of lesion-level segmentation metrics (Table III.3) on in-distribution data (Test ID), around 80% of the segmented lesions are TP_{les}, while the remaining 20% are FP_{les}. There are roughly the same amount of FN_{les} lesions as FP_{les} lesions. While the Dice scores are equivalent on the two shifted datasets (MSLUB and 1.5 Tesla), the lesion-level metrics tell a different story. On MSLUB, the DE produces very few FP_{les}, while the number of FN_{les} increases drastically (1413). On the opposite on the 1.5 Tesla dataset, the model has a low

number of FN_{les}, but an increased rate of FP_{les}. This highlights two different trends that are not visible when considering voxel-level metrics only.

Regarding lesion uncertainty quality (Table III.4), the **Size** approach provides as expected the worst lesion uncertainty scores on the Test ID and MSLUB datasets. This can be explained by the fact that in the context of MS lesions, small lesions can be segmented with high confidence. In contrast, the **Mean Entropy** score is a more robust baseline, achieving higher AUROC and AUPR scores in all settings, as compared to the **Size** scores. On the Test ID dataset, both the **Logistic** and **GNN** approaches offer a slight gain over the **Mean Entropy** baseline, on both metrics (AUROC and AUPR). This indicates that they offer a better distinction of TP_{les} and FP_{les} based on the predicted uncertainty scores, for test samples close to the training distribution. Additionally, the proposed **GNN** approach also outperforms **Mean Entropy** on the MSLUB dataset, on both metrics, showing that the GNN-based scores generalize well on this shifted dataset. On the 1.5 Tesla dataset, the GNN is also better on the AUROC metric, but not on the AUPR score. Overall, **CNN** offers weaker results than the other classification-based methods. It still outperforms the **Mean Entropy** baseline regarding AUPR scores on Test ID and MSLUB, but not on the AUROC scores. On the 1.5 Tesla dataset, it appears that the top AUPR scores are achieved by the **Size** and **Mean Entropy** scores, that the more sophisticated classifier-based approaches fail at outperforming. It may be due to the fact that for this dataset, most small lesions are actually FP_{les}, and thus the trivial **Size** score provides very competitive results. Finally, the density plots of lesion uncertainty scores (Figure III.6.4) provide interesting results. For TP_{les} lesions (blue histogram), the auxiliary classifiers (**Logistic**, **CNN**, and **GNN**) attribute uncertainty scores that are close to 0. In a clinical setting, lesions attributed with FP_{les} probabilities close to 0 could thus be overlooked as they predominantly correspond to TP_{les} lesions, allowing the clinician to focus on the lesions with high FP_{les} probabilities. This type of prioritization is less obvious for **Mean Entropy**, as the scores are much more uniformly distributed for TP_{les} samples.

	Voxel metrics			Lesion metrics						Lesion Counts		
	Dice (%)			LTPR (%)			LFDR (%)			TP_{les}	FP_{les}	FN_{les}
Dataset	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	Σ	Σ	Σ
Test ID	79.0	1.3	$[-2.1, 2.1]$	78.0	1.4	$[-2.3, 2.3]$	20.9	1.6	$[-2.6, 2.8]$	2196	528	563
MSLUB	68.0	3.2	$[-5.5, 5.1]$	54.2	2.4	$[-3.9, 3.9]$	10.9	1.9	$[-3.0, 3.3]$	1640	121	1413
1.5 Tesla	67.8	3.5	$[-6.1, 5.2]$	90.3	3.8	$[-7.3, 5.0]$	33.4	3.7	$[-5.7, 6.5]$	1016	382	71

Table III.3: Voxel-level and lesion-level performance for cross-sectional MS lesions segmentation. LTPR: Lesion True Positive Rate. LFDR: Lesion False Discovery Rate. SEM: Standard Error on the Mean. CI: 90% confidence intervals.

Method	**Dataset**	**AUROC** (%)			**AUPR** (%)		
		μ	SEM	CI	μ	SEM	CI
Size	Test ID	79.9	1.0	$[-1.7, 1.7]$	46.9	2.1	$[-3.6, 3.5]$
	MSLUB	78.0	2.1	$[-3.5, 3.4]$	20.2	2.6	$[-4.1, 4.3]$
	1.5 Tesla	86.5	1.1	$[-1.8, 1.7]$	69.7	2.6	$[-4.3, 4.1]$
Entropy	Test ID	85.7	0.8	$[-1.3, 1.3]$	52.3	2.3	$[-3.7, 3.7]$
	MSLUB	88.0	1.2	$[-2.1, 1.9]$	29.0	3.2	$[-5.1, 5.5]$
	1.5 Tesla	86.0	1.0	$[-1.7, 1.7]$	**69.8**	**2.5**	$[-4.2, 4.0]$
Logistic	Test ID	**86.8**	**0.8**	$[-1.3, 1.2]$	56.7	2.3	$[-3.9, 3.8]$
	MSLUB	87.8	1.2	$[-2.1, 2.0]$	33.8	4.1	$[-6.6, 6.9]$
	1.5 Tesla	86.6	1.0	$[-1.6, 1.6]$	66.1	2.8	$[-4.5, 4.5]$
CNN	Test ID	85.1	0.8	$[-1.3, 1.3]$	54.4	2.3	$[-3.9, 3.8]$
	MSLUB	86.1	1.4	$[-2.4, 2.3]$	**36.3**	**4.3**	$[-7.2, 7.2]$
	1.5 Tesla	**87.6**	**1.0**	$[-1.6, 1.5]$	66.7	2.7	$[-4.5, 4.4]$
GNN	ID	86.5	0.8	$[-1.3, 1.3]$	**57.2**	**2.3**	$[-3.9, 3.8]$
	MSLUB	**88.8**	**1.2**	$[-2.1, 2.0]$	35.7	4.2	$[-6.8, 7.1]$
	1.5 Tesla	87.2	1.0	$[-1.6, 1.5]$	66.8	2.8	$[-4.5, 4.5]$

Table III.4: Quality of lesion-wise uncertainty estimates for cross-sectional MS lesions. Top performing approaches are highlighted in **bold**, for each dataset. SEM: Standard Error on the Mean. CI: 90% confidence intervals.

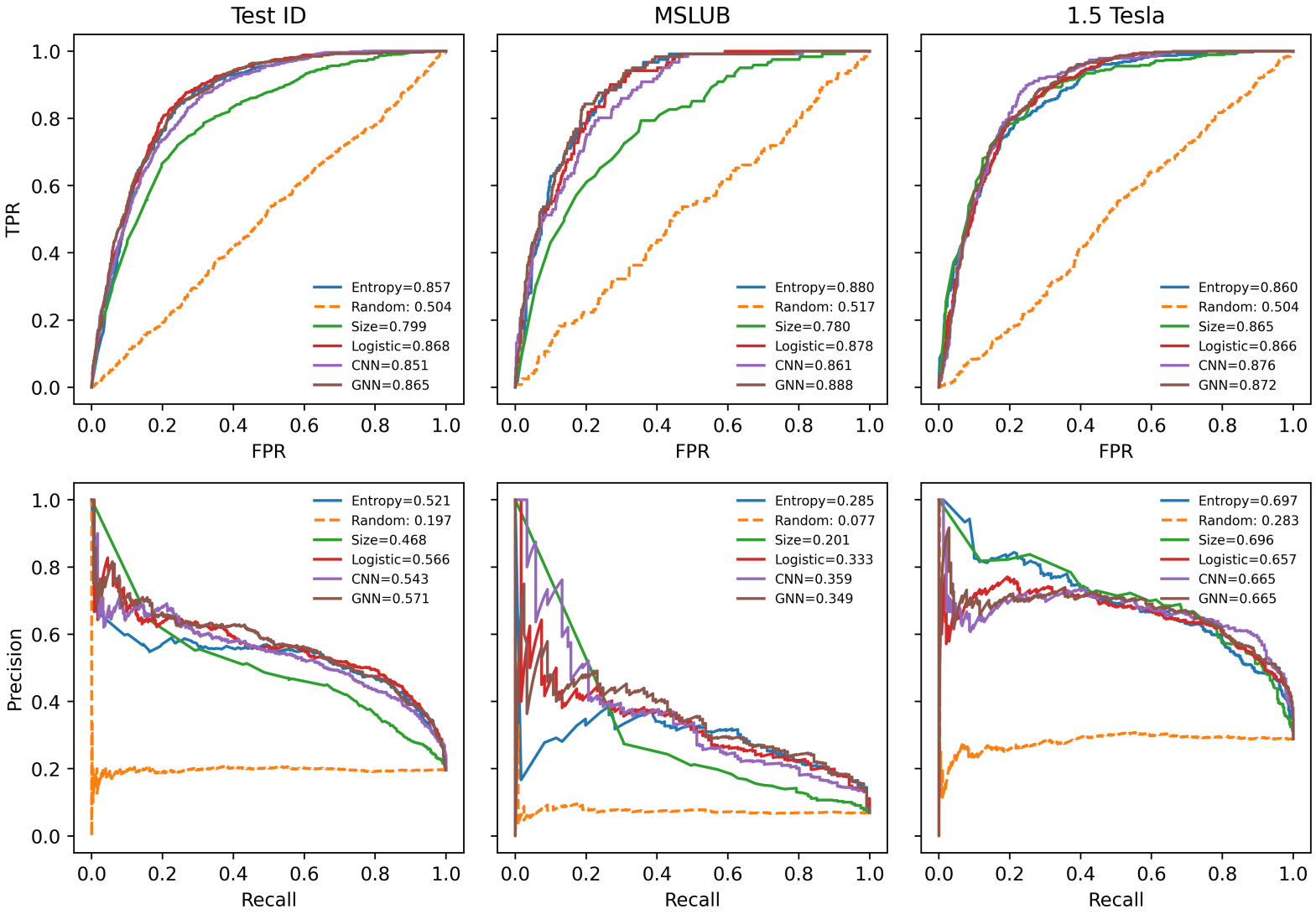


Figure III.6.3: Receiver operating characteristic (top row) and precision-recall curves (bottom row) for lesion uncertainty estimates on cross-sectional MS lesions. TPR: True Positive Rate. FPR: False Positive Rate.

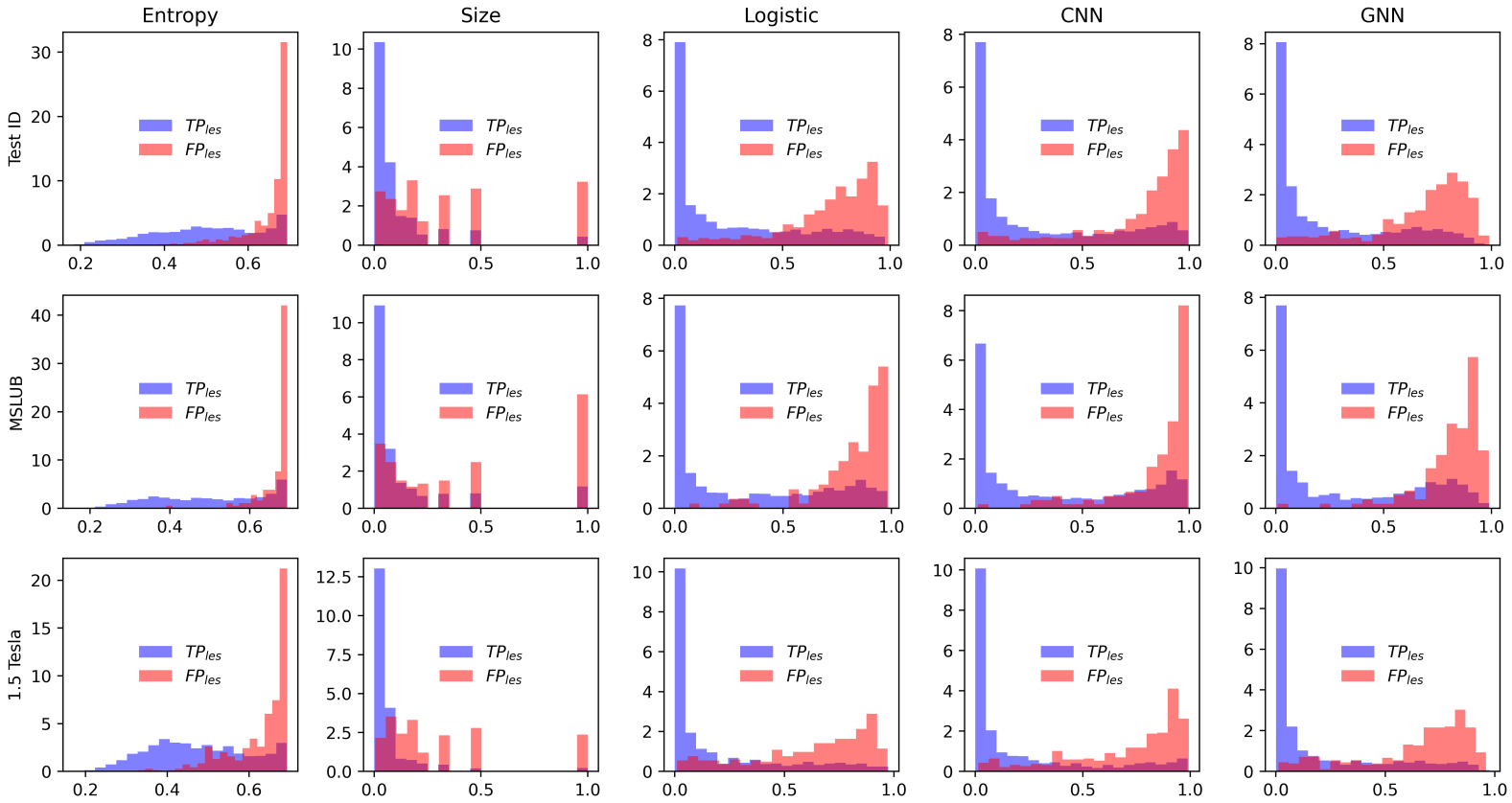


Figure III.6.4: Densities of lesion uncertainty scores for each dataset and approach on cross-sectional MS lesions. Blue indicates $\mathrm{TP}_{\mathrm{les}}$ samples, while red indicates $\mathrm{FP}_{\mathrm{les}}$ samples. An ideal uncertainty quantification module should enable a clear separation of the two classes.

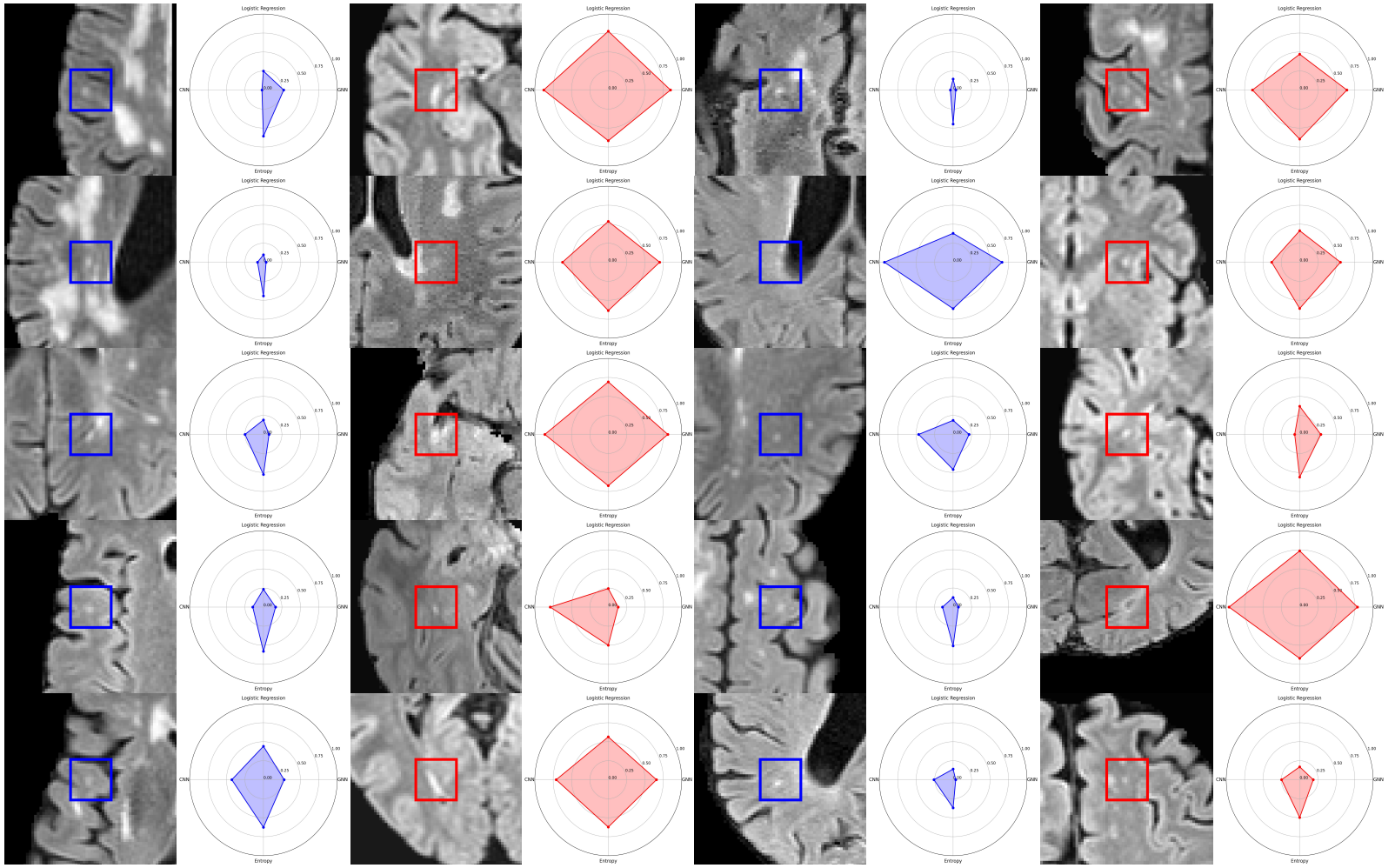


Figure III.6.5: Examples of lesion uncertainty quantification for the different tested approaches, for cross-sectional MS lesions detection. For each case, a red overlay indicates that the lesion is a FP_{les}, while blue indicates TP_{les}. Next to each image, a spider chart indicates the uncertainty scores estimated by each method.

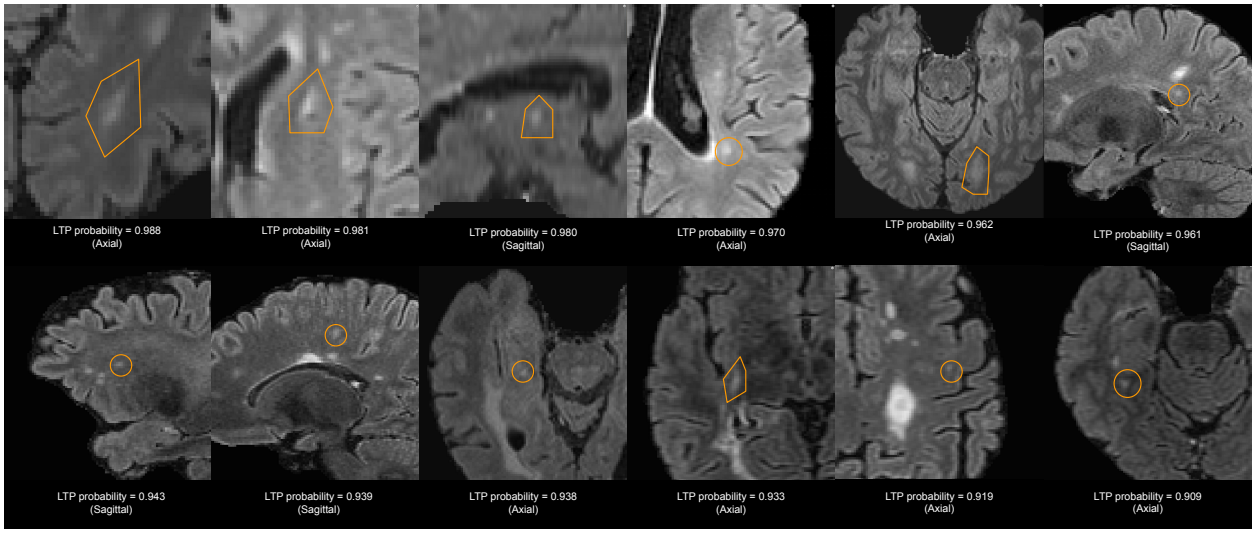


Figure III.6.6: Examples of lesions considered confident by the GNN model while being labeled as false positives. In each case, the FLAIR image presents a hyperintensity signal that may indicate the presence of a lesion and if so, incomplete ground truths.

III.6.4 Identification of annotation mistakes using lesion uncertainty scores

One limit of the explored classifier-based structural uncertainty quantification is that it is sensible to the accuracy of the ground truths. Indeed, if lesions are not seen by the annotator, it will result in a lesion labeled as a FP_{les}, although the model may be correct. This is particularly likely to happen for small lesions, that may be overlooked by the rater. As a result, the training of the auxiliary classifier becomes more unstable due to label noise. To highlight this phenomenon of lesions missed by the raters, the GNN model is used. More precisely, we propose to examine samples that are labeled as a FP_{les}, but that are associated with a low FP_{les} probability. In other words, we are interested here in lesions that the auxiliary classifier perceives as being likely a TP_{les}, while the actual status is FP_{les}. In Figure III.6.6, we present the 12 FP_{les} lesions in the Test ID dataset with the highest TP_{les} probability. It appears that in all cases, hyperintensity signals are visible in the FLAIR image, which may indicate that the ground truth is incomplete. By analyzing the errors of the auxiliary classifier on the test dataset, it could thus be possible to catch annotation oversights and proceed to the needed corrections.

III.7 Application to lung nodules segmentation in chest CT

The second investigated application of lesion uncertainty is the segmentation of lung nodules in chest CT. This experiment allows us to evaluate the protocol in a different imaging modality (CT instead of MRI) and with a different distribution of lesions.

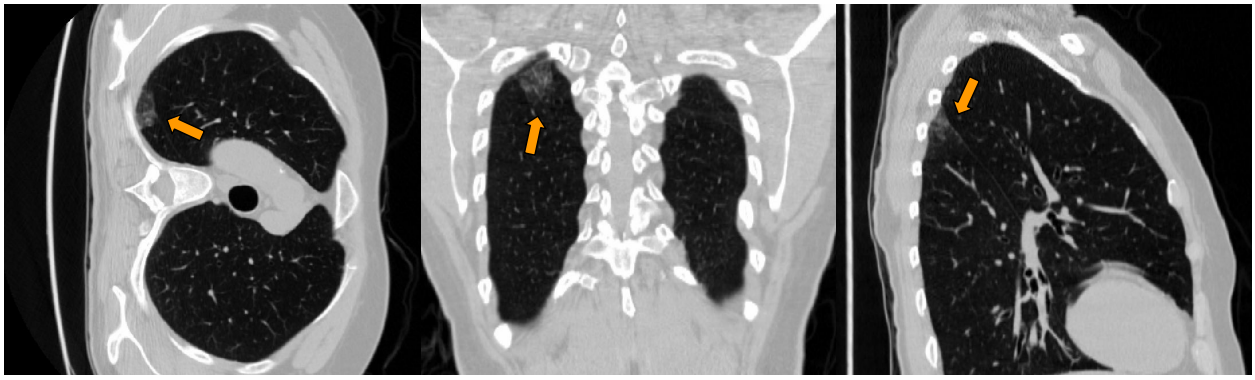


Figure III.7.1: Example of a CT scan of a patient presenting a lung nodule, indicated by the orange arrows.

III.7.1 Pathology description

Lung cancer is the first cause of cancer-related deaths in the world, responsible for 27% of cancer deaths in the USA, in 2014 [261]. The diagnosis of lung cancer is supported by imaging of the lungs, with volumetric chest CT being preferred to chest X-ray due to the superior spatial resolution [262]. Imaging helps identify suspicious findings, which are denoted as lung nodules if the shape is inferior to 3 cm in diameter, and lung masses else [263]. An example of a CT scan of a patient presenting a lung nodule is presented in Figure III.7.1.

Visually, lung nodules appear as round opacity appearing in chest CT scans. They are common in the adult population: approximately 30% of scans present such findings [264]. While the majority (95%) of these nodules are benign [264], detecting them remains crucial for early cancer detection. The malignancy of the identified nodules is evaluated in light of the patient's status: history of tobacco smoking, age, or past lung cancer. Yet, this detection is particularly difficult due to the small size of the nodules. A study carried out in the USA indeed concluded that 35% of lung cancers were associated with nodules with a diameter smaller to 10 mm [265]. A chest CT typically contains millions of voxels, meaning that nodules may occupy only an infinitesimal part of the imaged volume [266]. Thus, automated detection algorithms could greatly assist the clinician examine the scan. However, computer-aided nodules detection tools are known to produce numerous false positives [172, 267], hence a proper structural uncertainty quantification is desired to help the clinician review the segmentation.

III.7.2 Materials and preprocessing

This experiment relies on the LIDC-IDRI dataset [268], a large-scale dataset focusing on lung nodules in chest CT. It consists of a collection of 1018 helical thoracic CT scans of 1010 patients, acquired in seven imaging centers. Images were labeled through a two-phase annotation process performed by four experts, yielding voxel-level annotations of lung nodules larger than 3 mm. This dataset is particularly interesting as nodule-level scores were provided by the experts, including a subtlety score. This subtlety score corresponds to the nodule

detection difficulty perceived by the rater. The score is an integer in the range 1 to 5, where 1 indicates **Extremely Subtle**, and 5 indicates **Obvious**. Moreover, the inter-rater variability (IRV) can be computed for each nodule, corresponding to the number of experts (between 1 and 4) that annotated the finding as a nodule. Both these nodule-level scores (subtlety and inter-rater variability) can be interpreted as a form of structure-level uncertainty ground truth for the lesion uncertainty evaluation. In this experiment, a single CT scan per patient is used, reducing the number of available images from 1018 to 1010. The dataset is randomly split into a training part (710 images), a validation part (50 images), and a testing part (250 images).

A simple preprocessing strategy is adopted here, similar to Yu et al. [269]. First, the image intensity (expressed in Houndfield units for CT scans) is clipped in the range $[-1000, 400]$. Second, chest CTs are resampled to a voxel spacing of $3\,\text{mm} \times 1.5\,\text{mm} \times 1.5\,\text{mm}$ (in the axial, coronal and sagittal planes, respectively).

III.7.3 Experimental protocol

The experimental protocol used for lung nodule detection is strictly identical to the one employed for the cross-sectional MS experiments. A DE composed of 5 identical Dynamic U-Nets is used to generate predictions (entropy and segmentation) for each image in the training, validation, and test datasets. For training and validation nodules, the three proposed structural representations (features, bounding box, and graph) are extracted, leading to a nodule-level training dataset of 3058 instances (1467 TP_{les} and 1591 FP_{les}) and a nodule-level validation dataset of 192 (78 TP_{les} and 114 FP_{les}). It can be noted that the proportion of FP_{les} is different from the one obtained for cross-sectional MS lesions. Here, the number of FP_{les} is slightly superior to the number of TP_{les}, in contrast to the previous experiment.

From these nodule datasets, the three different auxiliary classifiers are trained (the Logistic Regression model, the CNN model, and the GNN). No modifications are made regarding their implementation, hyper-parameters, and training strategy. Additionally to AUROC and AUPR scores, the Spearman's correlation between the predicted uncertainty scores and the nodule uncertainty ground truths (subtlety and IRV) are also provided. These correlations can however only be computed for TP_{les} lesions. Indeed, FP_{les} lesions do not have corresponding subtlety and IRV scores, and FN_{les} lesions do not have associated predicted lesion uncertainty scores.

Appendix A.3.2 presents the top 10 features ranked by their order of importance for the Logistic Regression model trained to detect FP_{les} nodules. As for the cross-sectional MS experiment, the top 10 features include a mix of intensity features (4 features), shape features (5 features), and an uncertainty feature. Interestingly, 4 features are common to the Logistic Regression model trained on cross-sectional MS lesions: the entropy of the lesion interior, Maximum2DDiameterSlice, Maximum2DDiameterColumn, and Maximum3DDiameter. This mainly indicates that small lesions with high average interior entropy are likely to be FP_{les}, for both pathologies.

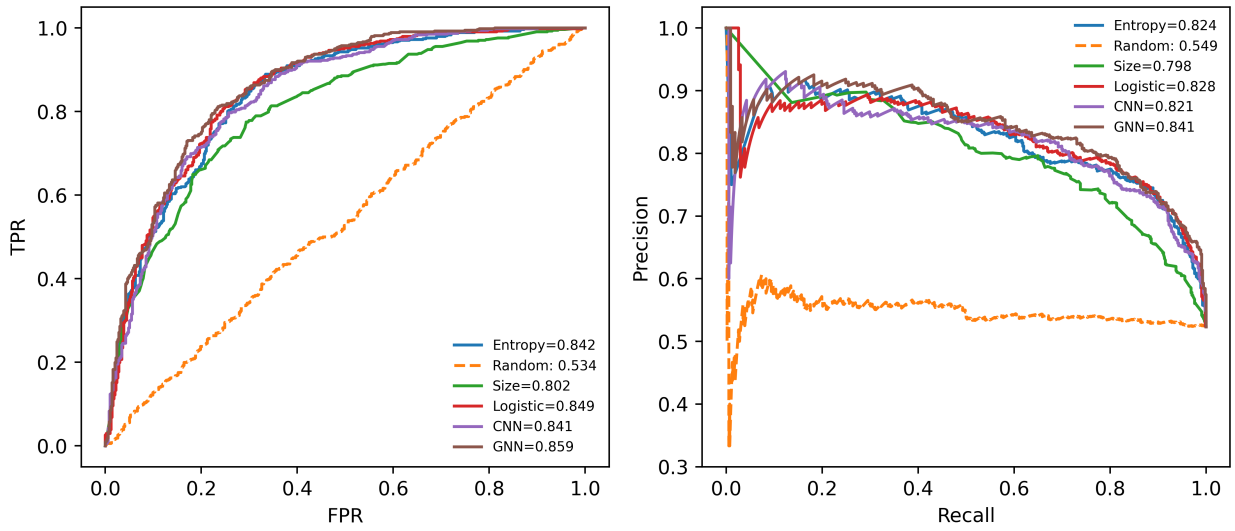


Figure III.7.2: Receiver operating characteristic (left) and precision-recall curves (right) for lung nodules uncertainty estimates. TPR: True Positive Rate. FPR: False Positive Rate.

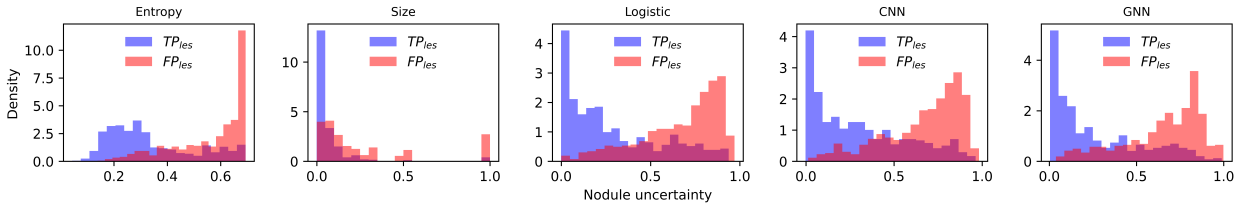


Figure III.7.3: Densities of lung nodule uncertainty scores for each approach. Blue indicates TP_{les} while red indicates FP_{les}. The best separation of the two classes is provided by the GNN approach.

III.7.4 Results of the lung nodule experiment

Table III.5 presents the segmentation quality metrics of the DE trained to detect lung nodules. Table III.6 and Figure III.7.2 present the nodule-level uncertainty quality for each tested approach. Figure III.7.4 presents the correlation between the computed structural uncertainty scores and the ground truth uncertainty scores (IRV and subtlety). Figure III.7.5 presents examples of nodule-level uncertainty scores computed with each approach. Finally, Figure III.7.3 displays the densities of nodule-level uncertainty scores for each approach.

First, the DE produces numerous FP_{les} on the test dataset, following the same trend observed on the training and validation datasets, with slightly more FP_{les} than TP_{les}. However, the number of FN_{les} is low, with only 121 FN_{les} for the 250 test subjects (roughly 0.5 FN_{les} per scan). The Dice is rather low which can be explained by the small size of nodules and the sensibility of the metric to the size of the objects [31]. Regarding the quality of nodule-level uncertainty estimates, the Size approach remains the weakest estimator, as expected. The

	Voxel metrics			Lesion metrics						Lesion Counts		
	Dice (%)			LTPR (%)			LFDR (%)			TP_{les}	FP_{les}	FN_{les}
Dataset	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	$\sum$	$\sum$	$\sum$
Test ID	47.0	1.9	$[-3.1, 3.1]$	70.3	2.5	$[-4.1, 4.0]$	50.3	2.2	$[-3.6, 3.6]$	497	542	121

Table III.5: Voxel-level and lesion-level performance for lung nodules segmentation in CT scans. LTPR: Lesion True Positive Rate. LFDR: Lesion False Discovery Rate. TP_{les}: True Positives lesions. FP_{les}: False Positive lesions. FN_{les}: False Negative lesion. The mean (μ), the Standard Error on the Mean (SEM), and 90% confidence intervals (CI) are estimated using bootstrap.

Method	**AUROC** (%)			**AUPR** (%)			**Correlation Subtlety** (↓)			**Correlation IRV** (↓)		
	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI
Size	80.2	1.4	$[-2.3, 2.2]$	79.8	1.9	$[-3.1, 3.0]$	-0.46	0.04	$[-0.06, 0.07]$	-0.56	0.03	$[-0.05, 0.06]$
Entropy	84.2	1.2	$[-2.1, 2.0]$	82.5	1.9	$[-3.1, 3.0]$	-0.39	0.04	$[-0.06, 0.06]$	-0.51	0.03	$[-0.06, 0.06]$
Logistic	84.9	1.2	$[-2.0, 2.0]$	82.9	1.9	$[-3.1, 3.0]$	-0.49	0.04	$[-0.06, 0.06]$	-0.59	0.03	$[-0.05, 0.05]$
CNN	84.1	1.2	$[-2.0, 2.0]$	82.2	1.9	$[-3.2, 3.1]$	**-0.51**	**0.04**	$[-0.06, 0.06]$	-0.58	0.03	$[-0.05, 0.05]$
GNN	**85.9**	**1.2**	$[-1.9, 1.9]$	**84.1**	**1.8**	$[-3.0, 2.9]$	-0.46	0.04	$[-0.06, 0.06]$	**-0.61**	**0.03**	$[-0.05, 0.05]$

Table III.6: Quality of lesion-wise uncertainty estimates for lung nodules. Top performing approaches are highlighted in **bold**, for each dataset. The mean (μ), the Standard Error on the Mean (SEM), and 90% confidence intervals (CI) are estimated using bootstrap.

Mean Entropy provides more informative nodule-level uncertainty scores. Following the same trend as the cross-sectional MS experiment, it is yet outperformed by both the **Logistic** and **GNN** models, on both AUROC and AUPR. Overall, the best quality of uncertainty is obtained by the **GNN**, with the top AUROC and AUPR metrics. As in the cross-sectional MS experiment, the CNN is the weakest classifier-based approach, with a performance below **Mean Entropy**.

The high performance of the **GNN** can be explained by the cleaner ground truths available in this experiment. Indeed, in the LIDC-IDRI dataset, the annotation procedure contains 2 steps. In the first one, the raters annotate the scans independently, while in the second step, they have access to their colleague's annotations to correct their decision if needed. As a result, overlooked nodules can be caught. Then, the training of the auxiliary classifiers is more efficient, as label noise is reduced. Another possible reason is that the ratio between $\mathrm{TP}_{\mathrm{les}}$ and $\mathrm{FP}_{\mathrm{les}}$ is much more balanced in the nodules experiments, contrary to the cross-sectional MS experiment.

The correlation study (Figure III.7.4), in which the link between predicted and ground truth uncertainty scores is studied, shows interesting findings. It appears that the average entropy is weakly correlated with the expert scores, being even outperformed by the Size method. All auxiliary classifiers (**Logistic**, **CNN**, **GNN**) demonstrate superior Spearman's correlation scores, indicating that they more accurately match with the human notion of uncertainty. Concerning the IRV, the best correlation is achieved by the GNN model (SP= -0.611), while for the Subtlety scores the CNN offers the best correlation (SP=-508).

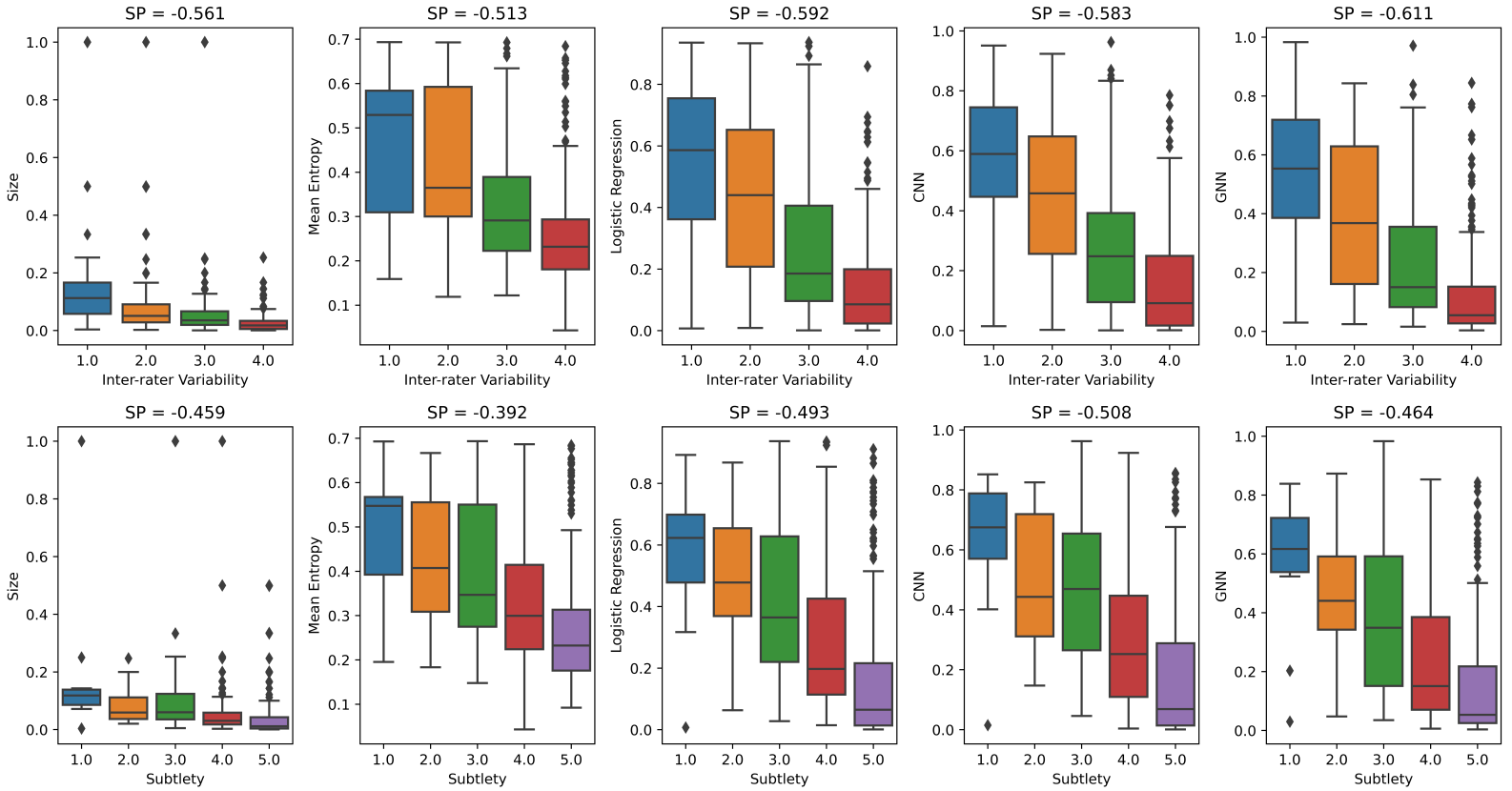


Figure III.7.4: Relationship between predicted nodule and ground truth nodule uncertainty scores (subtlety and Inter-rater Variability) on the 250 test subjects. SP = Spearman's correlation.

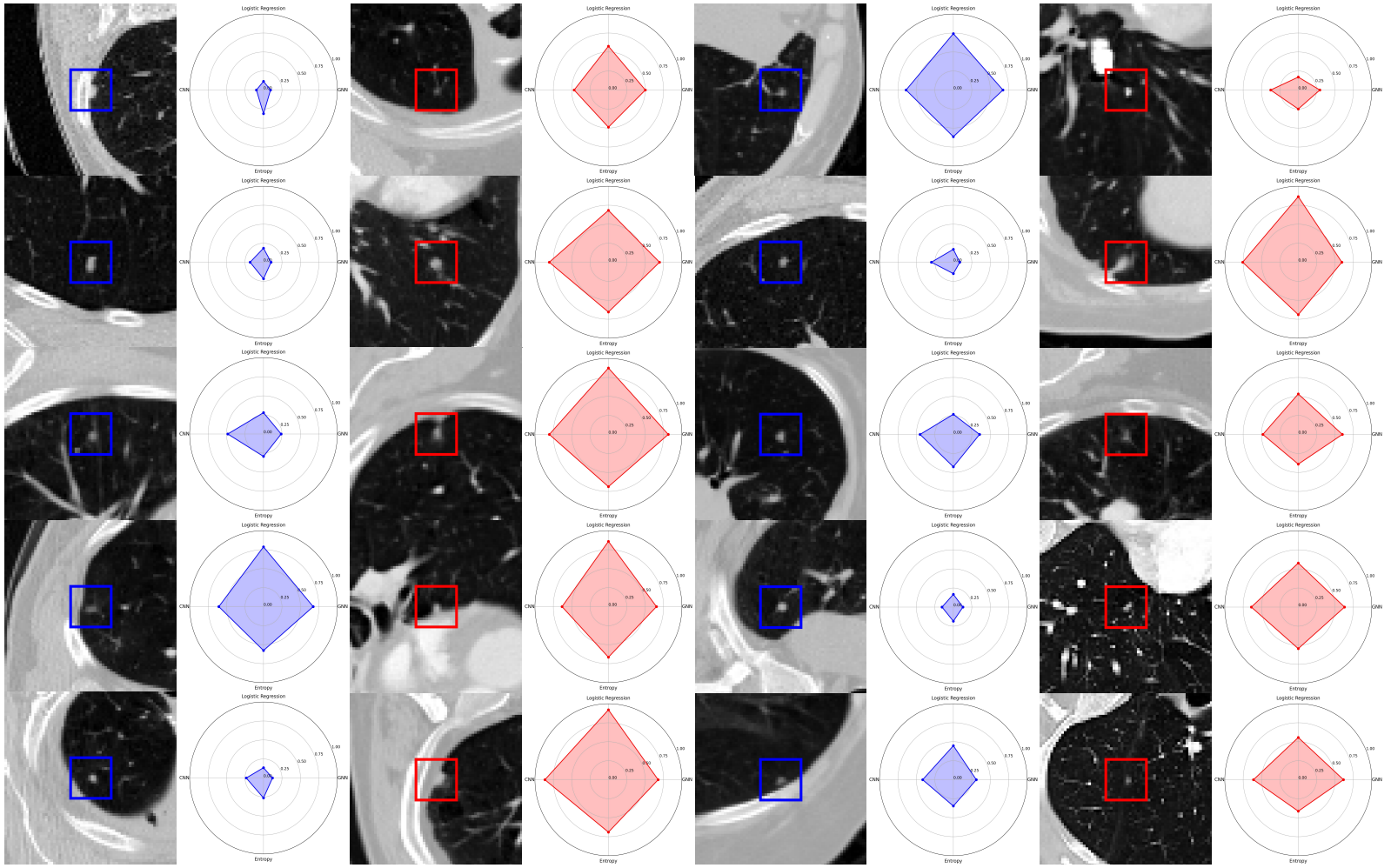


Figure III.7.5: Examples of lesion uncertainty quantification for the different tested approaches, for lung nodules detection. For each case, a red overlay indicates that the lesion is a FP_{les}, while blue indicates TP_{les}. Next to each image, a spider chart indicates the uncertainty scores estimated by each method.

III.8 Application to longitudinal Multiple Sclerosis lesions segmentation in brain MRI

Analyzing MR images of MS patients in a cross-sectional setting is interesting as it allows to quantify the extent of the disease (number of lesions and total lesion load). However, to monitor the progression of the disease, longitudinal evaluation has to be carried out. In this setting, the goal is to detect evolving lesions between two consecutive time points, usually separated over several months or years. More particularly for MS, **new** lesions are a crucial biomarker. Indeed, the absence or appearance of new lesions allows clinicians to determine the efficiency of the anti-inflammatory drug, and modify the treatment if necessary [63].

These new lesions are usually very small, subtle, and rare. To detect them automatically, two approaches can be adopted. The first approach consists of the separate analysis of each time point independently, using a cross-sectional segmentation model such as the ones used in the previous section [270]. Then, by comparing the two segmentations, changes can be detected. However, this approach may lack accuracy. For instance, a stable lesion can be considered as new if it is a False Negative in the first visit, and a True Positive in the second visit. Moreover, this approach requires the precise pairing of individual lesions between each visit, which is not trivial. This motivated the development of direct longitudinal models, that take as inputs both co-registered visits, and directly provide a delineation of the new lesions. In the MSSEG-2 challenge [63] focusing on new MS lesions segmentation, this approach was the most popular one, being implemented by 21 out of 28 participating teams.

Developing longitudinal new MS lesions segmentation models is not straightforward because of the scarcity of open-source longitudinal data. Three datasets can be found in the literature. The more recent dataset is the MSSEG-2 dataset [63] that comprises 100 patients (40 for training, 60 for testing) with two imaging visits. The ground truth annotations correspond to the delineation of the new lesional voxels. However, 40% of the patients are stable and thus the corresponding ground truth masks are empty. Second, the ISBI 2015 dataset comprises multiple visits (4 or 5) of 5 different patients. However, ground truth masks do not correspond directly to the new lesions, but instead contain the total WMH load for each visit (thus including a majority of stable lesions). The MSLUB-LONG dataset [230] is a longitudinal dataset comprising 2 visits of 20 MS patients. However, the ground truths correspond to the WMH changes between the 2 visits, comprising new but also shrinking, growing, and disappearing lesions. Due to this inhomogeneity in annotation policies, using these datasets altogether is impractical. To solve this problem, a data synthesis approach can be adopted to train longitudinal MS lesions segmentation networks. This strategy is presented in Section III.8.1.

Once a model has been developed using real and synthetic datasets, the evaluation is carried out using the testing split of MSSEG-2. Interestingly for this dataset, the ground truth annotations of 4 experts are available. This enables the computation of the inter-rater variability as a form of ground truth lesion uncertainty, as done in the lung nodules experiments.

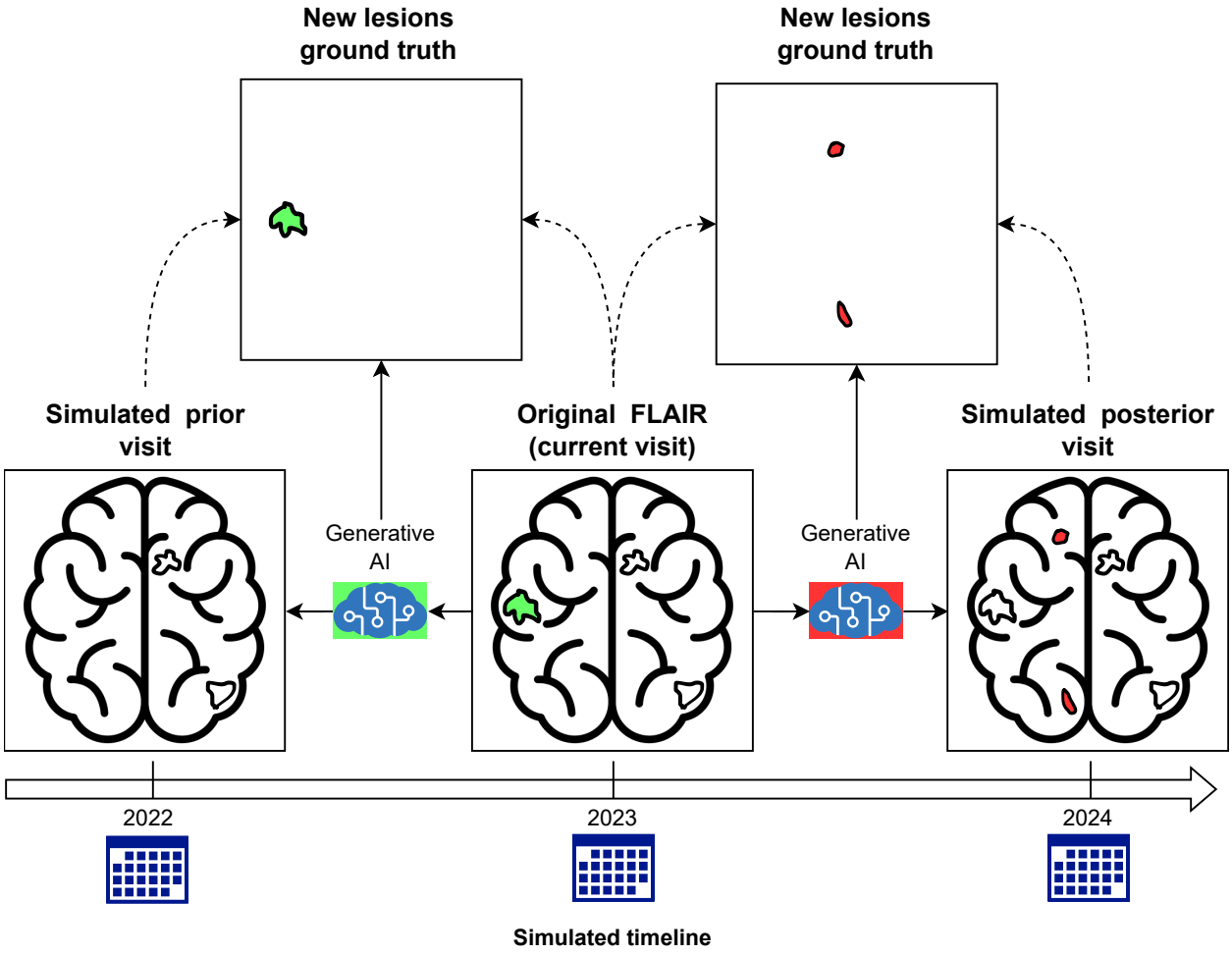


Figure III.8.1: Principle of the synthetic creation of longitudinal cases from a single MRI visit. By erasing one or several lesions (green), a prior visit can be simulated. Alternatively, by injecting new lesions (red), a posterior visit is simulated. In both cases, a perfect ground truth mask can be obtained to eventually train a segmentation model.

III.8.1 Longitudinal cases synthesis using a Generative Adversarial Network

In contrast to longitudinal annotated data, annotated cross-sectional MS data is more abundant. A possible way to circumvent the lack of annotated longitudinal MS cases is to use cross-sectional MS data to create synthetic longitudinal examples. This concept is illustrated in Figure III.8.1. More formally, from a single cross-sectional case labeled as visit T, a prior visit $T-1$ can be created by *erasing* lesions in the original scan, or alternatively a posterior visit $T+1$ can be obtained by *adding* lesions. This concept was initially presented in Manjon et al. [271] and further developed in Kamraoui et al. [272], where authors develop two different inpainting autoencoders to perform the erasing and addition operations. This technique is particularly interesting as multiple longitudinal cases can be obtained from a single cross-sectional image, by varying the number and the location of the added/removed

lesion. Moreover, the resulting ground truth is perfect, because the synthetic image matches perfectly the inpainting mask used as input, thus alleviating the problem of noisy, partially incorrect ground truths. Adding lesions in a cross-sectional image is more cumbersome than erasing lesions, as it requires 1) creating an artificial lesion mask to be added or using one from another scan, and 2) injecting it in a realistic area. In Kamraoui et al., the authors used a probabilistic atlas to identify realistic regions. However, this approach implies registering images to a common atlas, which requires an extra preprocessing step. For simplicity, this work thus focuses on the erasure of lesions, that can be easily implemented without the need for atlases.

A lesion erasure model operates as follows. During training, healthy tissue voxels are randomly inpainted with Gaussian noise. The model receives as input the noisy image as well as the inpainting mask S, corresponding to the binary mask of altered voxels. The task of the autoencoder is to recover the intensity of the inpainted healthy tissue, thus behaving as a denoising autoencoder. During inference, the process is different as this time, white matter hyperintensities are filled with Gaussian noise. By carrying inference with the denoising autoencoder trained to convert noise into healthy white matter, the lesion is removed. The model is only allowed to modify the image inside the inpainting mask: $\text{Image}_{out} = \text{Model}_{out} \times S + \text{Image}_{in} \times (1 - S)$.

Intuitively, the realism of the generated longitudinal cases seems crucial to guarantee that the model benefits from these synthetic data during training. If the inpainted area is unrealistic and presents artifacts, then the model could simply learn this pattern during training instead of focusing on meaningful differences between the two visits. By reproducing the autoencoder proposed in [272], it appeared that the inpainted area lacked details and the erasing was visually obvious (see Figure III.8.3). It could be hypothesized that a longitudinal model trained on this synthetic data could minimize the error function by learning to recognize this blurred area in place of details more relevant to the detection of new lesions. Thus, an enhancement of the synthesis method is proposed here as a contribution, by incorporating an adversarial training strategy, illustrated in Figure III.8.2 and presented in the following.

III.8.2 Adversarial training with voxel-level counterfactual scores

To improve the quality of the generated samples, an adversarial training paradigm based on an auxiliary discriminator model is proposed. The task of the discriminator during training is to distinguish between real and inpainted areas, produced by the erasure model (here, the generator). Usually, a discriminator is a classifier that outputs an image-level realism score [273]. Alternatively, a patch discriminator [274] can be used, that predicts a realism score for each sub-patch of the input image. However, in this setting, both options are not satisfying. Indeed, most voxels of inpainted patches are real voxels, as the erasure model can only modify the input image inside the inpainting mask, which only occupies a limited portion of the image. Thus, in inpainted patches, most voxels are actually real, unaltered voxels. Thus, training the discriminator to predict an image-wise score may result in unstable learning. To circumvent this, an encoder-decoder discriminator is proposed here, which produces a voxel-level counterfactual score. Ideally, the discriminator should predict high counterfactual scores in the inpainted region of predicted patches and low counterfactual scores elsewhere.

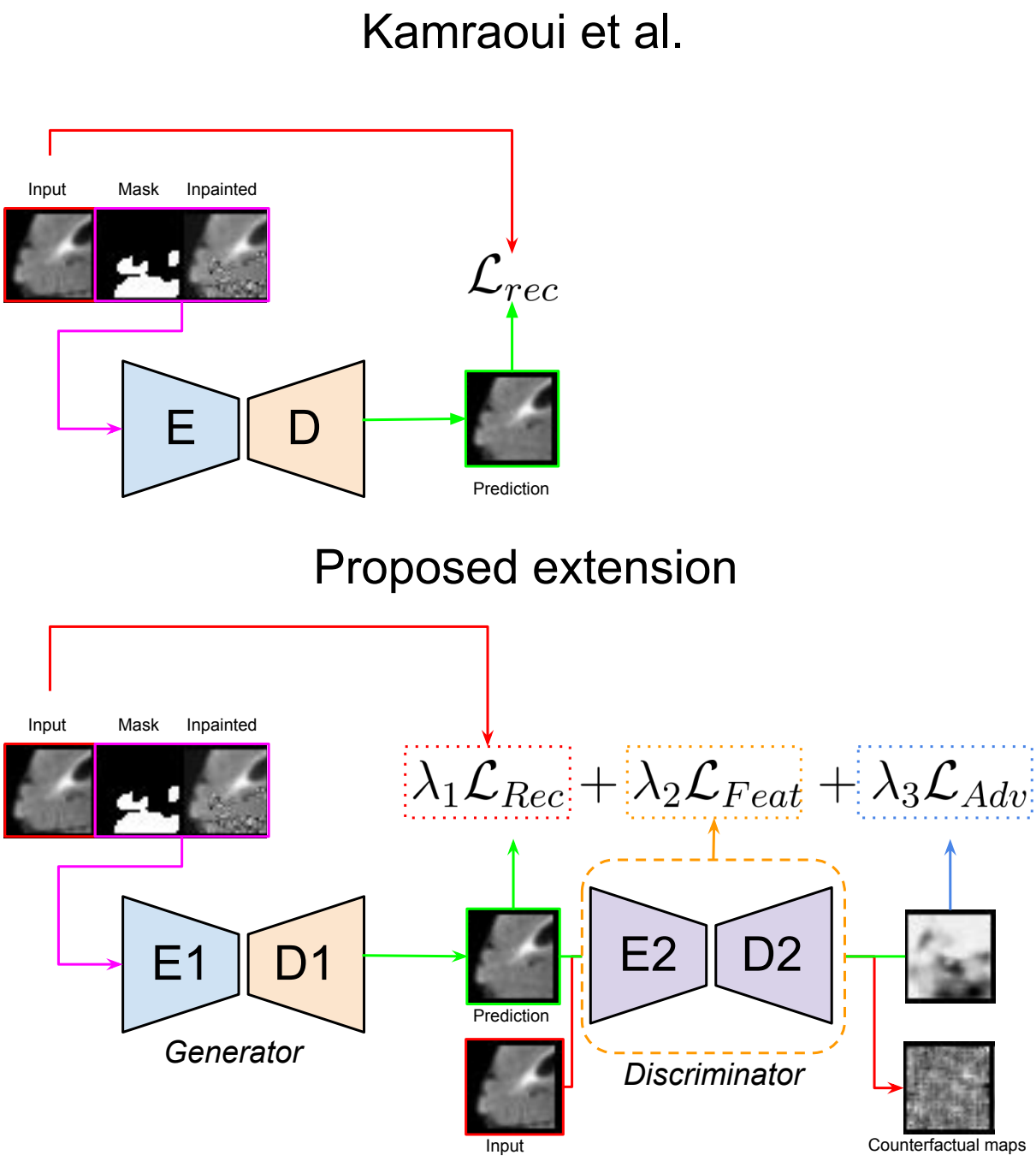


Figure III.8.2: Proposed inpainting framework using an adversarial approach. The Generator (E1-D1) receives a patch inpainted with noise along with the binary inpainting mask. It produces a reconstruction of the original patch. This prediction is presented to the Discriminator model (E2-D2) which predicts a realism score for each voxel in the patch, with a score of 1 for realistic voxels and 0 for unrealistic voxels.

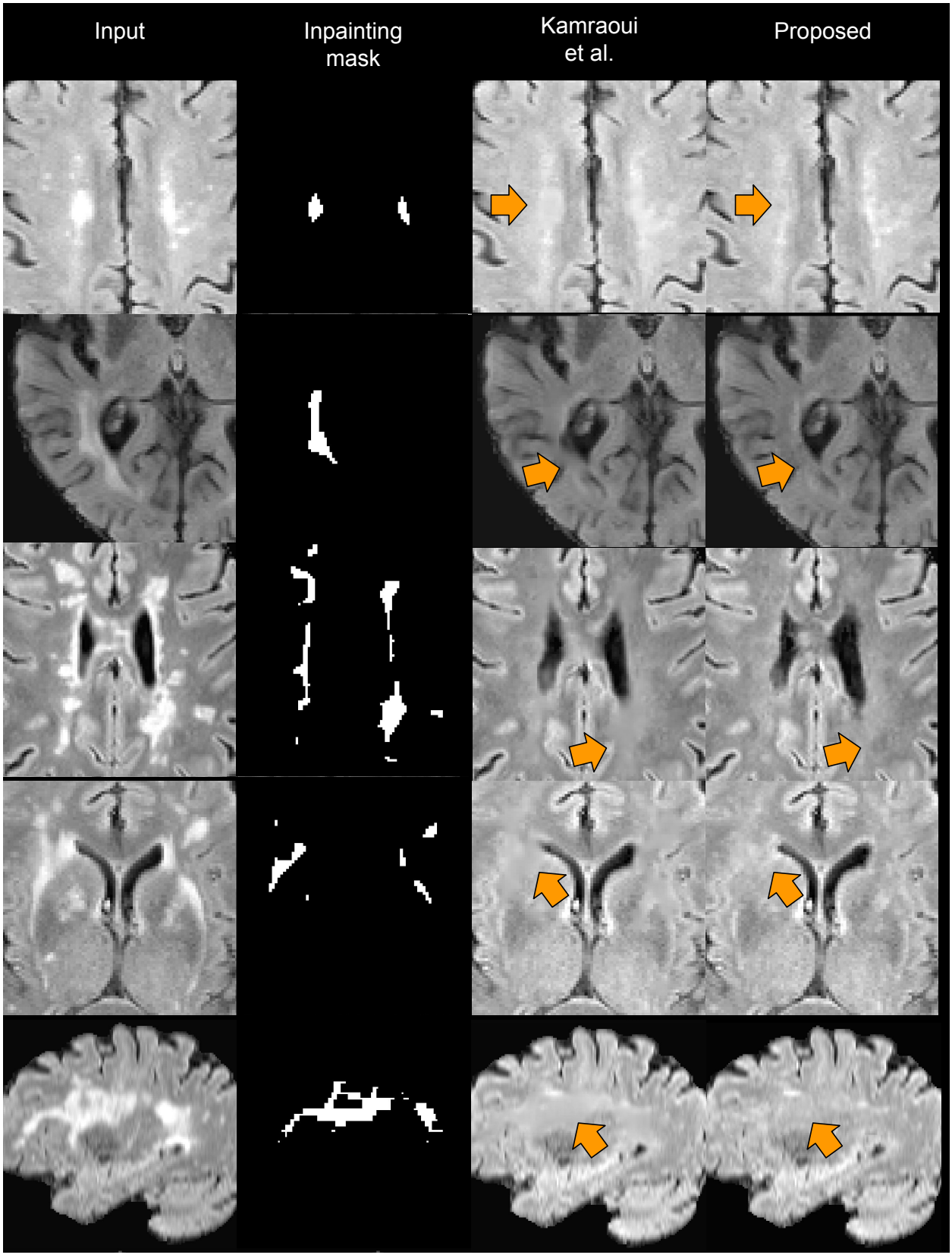


Figure III.8.3: Comparison of lesion erasing approaches. The top 4 rows are axial views, while the last one is a sagittal view. Orange arrows indicate blur artifacts in the baseline model, that are eliminated with the proposed one.

More formally, given a 3D input image $X \in \mathcal{R}^{H \times W \times D}$, the voxel-level discriminator produces a realism map of dimension $\hat{R} \in \mathcal{R}^{H \times W \times D}$ where a value close to 1 indicates a voxel judged as **real**, whereas a value close to 0 indicates a voxel judged as **fake** (meaning generated by the erasure model).

Adding a discriminator to guide training requires a modification of the training objective. In its original formulation, the lesion inpainting model is trained solely using a reconstruction loss corresponding to the Mean Squared Error (MSE) between the original input and the output of the denoising model [271]. However, using this loss for image prediction is known to be suboptimal in terms of realism, as the MSE is not directly linked to the perception of image quality of human raters [275]. Here, a mode sophisticated reconstruction loss is adopted, corresponding to the sum of the MSE, the Mean Average Error (MAE), and the Structural Similarity Image Index (SSIM) loss [276]:

$$\mathcal{L}_{MAE}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} |X_i - Y_i|$$
$$\mathcal{L}_{MSE}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} |X_i - Yi|^2$$
$$\mathcal{L}_{SSIM}(X, Y) = \frac{[2\mu_X \mu_Y + (k_1 L)^2][2\sigma_{XY} + (k_2 L)^2]}{[\mu_X^2 + \mu_Y^2 + (k_1 L)^2][\sigma_X^2 + \sigma_Y^2 + (k_2 L)^2]}$$

where μ_X (respectively μ_Y) is the voxel mean of X (respectively Y), σ_X (respectively σ_Y) is the variance of X (respectively Y), σ_{XY} is the covariance of X and Y, L is the dynamic range of the voxel values, and k_1 and k_2 are constants set to 0.01 and 0.03 respectively. For the SSIM loss, the score is computed using a moving window approach, using a window size of $(5, 5, 5)$. The final reconstruction loss $\mathcal{L}_{rec}$ is the sum of these three terms.

The voxel-level discriminator is trained using a MSE loss. This is thus akin to the Least Squares GAN proposed in Mao et al. [277], aiming at alleviating the instability of GAN training. Here, we set the ground truth realism map R as the opposite of the inpainting mask S: $R = 1 - S$. In other words, the goal of the discriminator is to detect the parts of the input patch that have been altered by the generator. Writing D the proposed voxel-level discriminator, Y the original input image and X the output of the generator, the discriminator loss is expressed as:

$$\mathcal{L}_{D,real} = \text{MSE}(D(Y), J)$$
$$\mathcal{L}_{D,fake} = \text{MSE}(D(X), R)$$
$$\mathcal{L}_{D,total} = \mathcal{L}_{D,real} + \mathcal{L}_{D,fake}$$

where J is a matrix filled with ones, as for real input images, all voxels are real. The discriminator predictions are also used to guide the generator G training, using the following adversarial loss:

$$\mathcal{L}_{Adv} = \text{MSE}(D(X), J)$$

For the generator, the task is to fool the discriminator into predicting the inpainted voxels as real (target equals to one). Thus, for this loss, the target is the J matrix. To further improve the realism of the inpainted zones, a feature matching loss is employed, which has been introduced in the Pix2pix HD model [278]. The idea is that the discriminator model should produce similar features for unaltered images and realistic generated images. This loss thus forces the generator to produce images that appear natural at multiple scales, as the features are collected from different layers of the discriminator. More specifically, the latent representations produced by the original input image $Z_{real,k}$ and the generated on $Z_{fake,k}$ at the k-th layer of the discriminator are gathered. Then, the discrepancy between both representations is estimated using a MAE term:

$$\mathcal{L}_{Feat} = \frac{1}{k} \sum_{k} \text{MAE}(Z_{real,k}, Z_{fake,k}) \tag{III.8.1}$$

For the proposed voxel-level discriminator, the features are collected in each layer of the encoder and decoder. The total generator loss is finally expressed as:

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{Rec} + \lambda_2 \mathcal{L}_{Feat} + \lambda_3 \mathcal{L}_{Adv} \tag{III.8.2}$$

where $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_3 = 10$. The resulting training curves are provided in Figure III.8.4.

III.8.3 Implementation details

The framework is implemented using MONAI's generative AI library [279]. A Residual U-Net is employed, enhanced with SPADE normalization layers [280] to conserve the semantic information (here, the inpainting binary mask) throughout the network. The discriminator follows the same residual architecture, except for the SPADE normalization layers that are replaced with standard batch normalization. Training is performed using a patch-based setting with a patch size of $32 \times 32 \times 32$ and a batch size of 16. This reduced patch size has two motivations. First, most MS lesions fit in this reduced bounding box as new lesions are usually small. Second, this helps reduce the imbalance between real voxels (indeed most of the image) and inpainted voxels. This has a beneficial impact on training stability.[37]

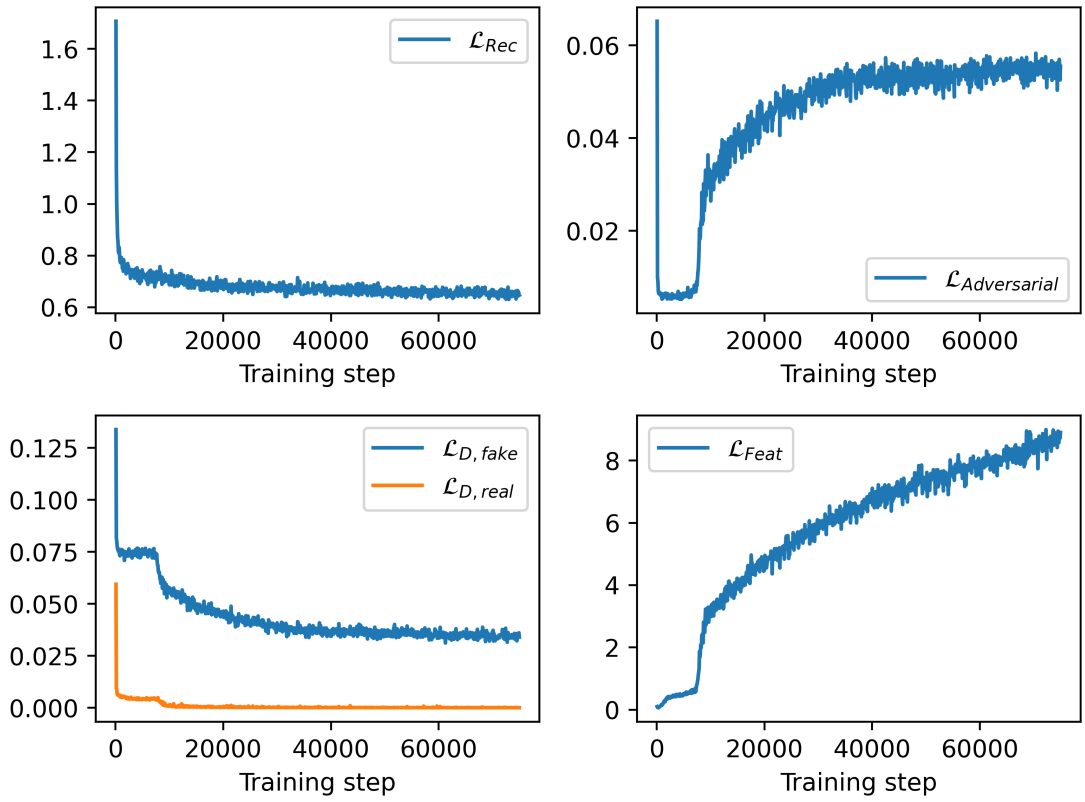


Figure III.8.4: Training losses for the proposed inpainting Generative Adversarial Network. It can be noticed that in the early steps of training, the adversarial loss ($\mathcal{L}_{Adversarial}$) has a value close to 0, indicating that at the beginning of training, the generator successfully fools the discriminator. However, as training progresses, the discriminator becomes more performant in detecting altered voxels, and as a result the adversarial loss augments.

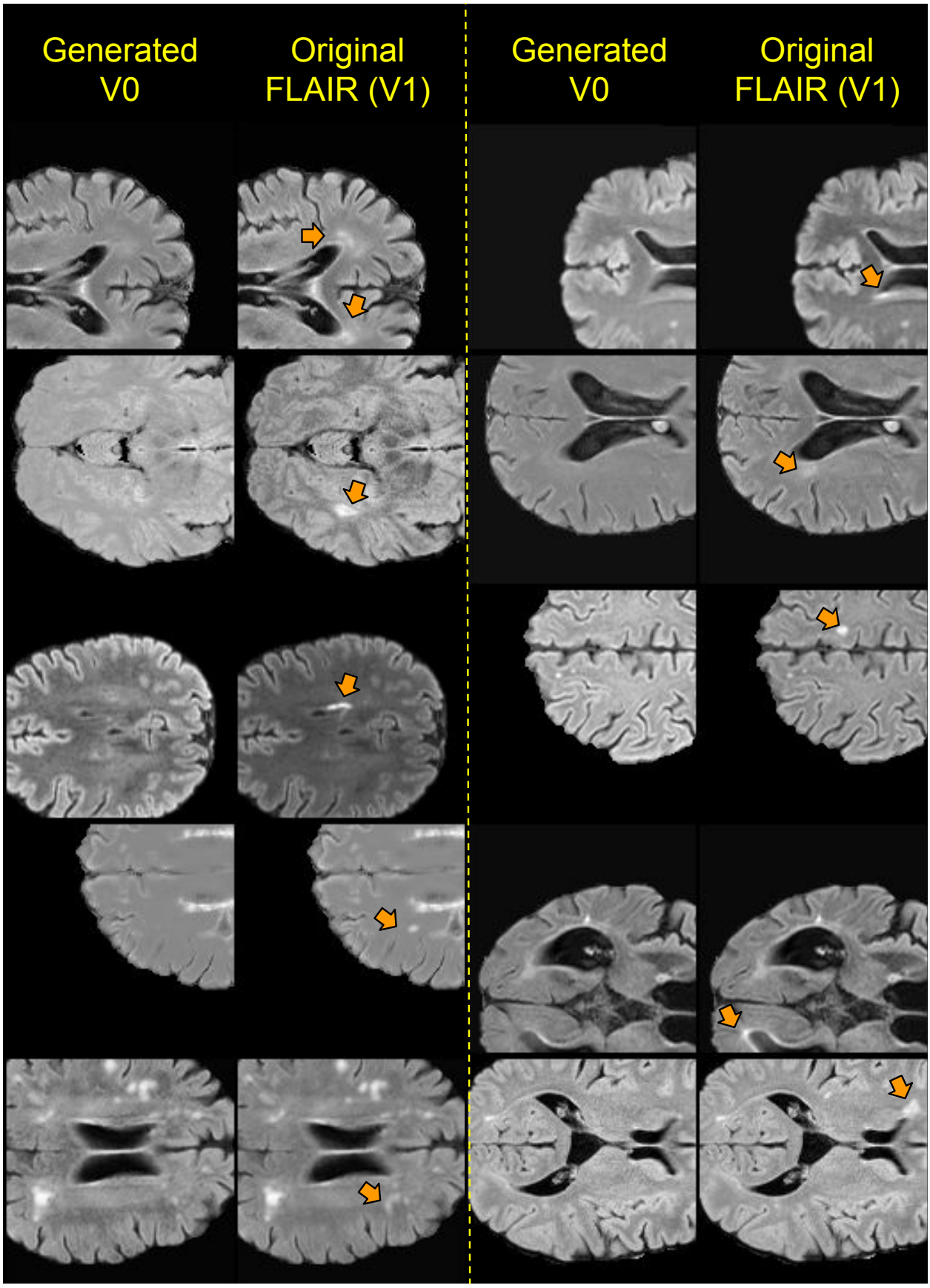


Figure III.8.5: Examples of generated longitudinal cases obtained with the proposed approach. The orange arrows indicate the lesion that is removed. V0: prior generated visit. V1: original current visit.

are used for the generator and discriminator, respectively, with a learning rate of 2×10^{-4}. Training is performed until the loss seizes to improve on the validation set for 50 epochs.

III.8.4 Generation parameters

In this section, the design choices for the synthesis of longitudinal cases from cross-sectional MS data are presented. First, all available cross-sectional images (N=274) are collected. For each cross-sectional image[1], 20% of the lesions are randomly inpainted, focusing on lesions in the range [10 mm^3, 500 mm^3] as new lesions are usually small. The original cross-sectional FLAIR will correspond to the second time point, while the output of the inpainting model will correspond to the first one. To mimic the natural variability in image contrast between the two visits, random gamma alterations are further applied to the generated first visit. This process generates a total of 274 synthetic longitudinal cases. Several examples of synthetic longitudinal cases are presented in Figure III.8.5 for our GAN approach.

Two versions of the synthetic dataset are generated, one using the baseline AE model (Kamraoui et al. [272]), and one with the proposed GAN extension. By fixing the random seed used to generate the scans, two strictly equivalent datasets are obtained, in terms of new lesion masks and gamma alterations. Thus, if noticeable differences are observed between the two synthetic datasets, it will be possible to conclude that it is because of the inpainting model only. The two datasets are referred to as **SynAE** and **SynGAN** in the following.

The longitudinal ensembles are trained using the same hyperparameters as the ones previously used in this thesis: 5 Dynamic U-Nets are trained with a combination of the cross-entropy and Dice++ losses (Equation II.6.3), followed by post-hoc Temperature Scaling on the validation dataset. A patch training is adopted with a patch size of 128^3 and a batch size of 6. The learning rate is set to 2×10^{-4}.

III.8.5 Performance of the longitudinal MS lesions segmentation

Table III.7 presents the segmentation performance of the different longitudinal Deep Ensembles trained with and without synthetic data. More specifically, several settings have been tested: training using only MSSEG-2, training using only synthetic data (**SynAE** or **SynGAN**), and pretraining on synthetic data followed by finetuning on the training split of MSSEG-2. Figure III.8.6 displays the training curves for models trained with each strategy.

Interestingly, it can be observed that the models trained using only synthetic data achieve an interesting performance on the real test data. More particularly, the models have a high detection rate (high number of TP_{les} and low number of FN_{les}). However, they also make a lot of FP_{les} detection, especially for the models trained only using the SynGAN dataset. This can be due to the fact that the pairs of visits used in the synthetic datasets only differ on new

[1]The reader may find it surprising that all available cross-sectional images are used to generate the synthetic longitudinal dataset, as a proportion of this cross-sectional dataset has been used to train the inpainting models. However, the inpainting tasks are totally disjoint during training and testing. During training, inpainting areas correspond to healthy white matter, whereas during inference the inpainting areas are white matter hyperintensities. Thus, there is no risk of overfitting.

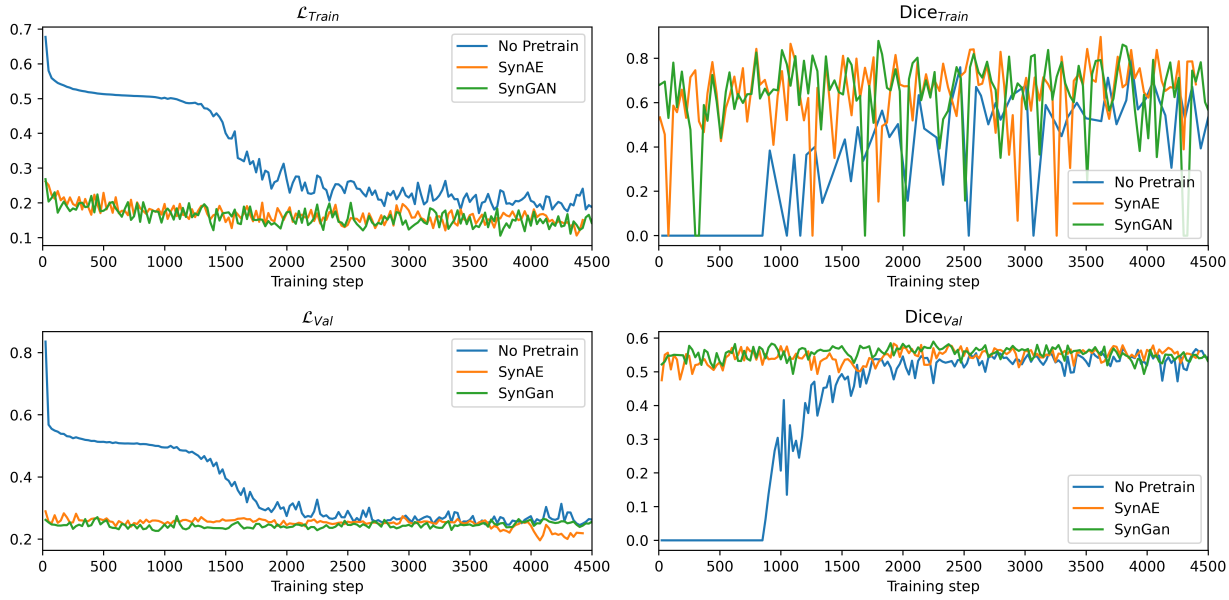


Figure III.8.6: Loss functions and Dice scores monitored during training and validation for the longitudinal models trained without pretaining (**No Pretrain**), with pretraining on **SynAE**, and with pretraining on **SynGAN**. It can be seen that models trained with synthetic data converge faster than models trained without synthetic data.

MS lesions voxels, as the inpainting models can only modify the image within the erasure mask. This can make the models trained on synthetic data too sensitive when confronted with real longitudinal cases, where intensity changes can be observed in both visits, unrelated to the presence of a new lesion. Thus, these models are of little practical use, but it is interesting to note that models trained without real longitudinal data manage to generalize on real data.

Second, it can be observed in the training curves (Figure III.8.6) that models that are not pretrained on synthetic data take about 1000 training steps to reach a validation Dice superior to 0, while models pretrained on synthetic data achieve a high validation Dice right from the start of the finetuning. Moreover, both training and validation losses reach a lower value, as compared to the baseline model without pretraining. Then, on the testing dataset, the models that were pretrained on synthetic data achieved a higher Dice than the baseline model (see Table III.7). The increase is statistically significant when comparing the baseline model and the model pretrained on Syn-GAN (p-value=0.002, Wilcoxon test). Lesion-wise, a gain with respect to the LTPR is also observed, paired with a decrease in the subject-level LFDR. At the dataset level, the models pretrained on synthetic data can detect about 15 additional TP_{les}, and miss 15 lesions less (reduction of the count of FN_{les}). When comparing the DE pretrained on **SynAE** and **SynGAN**, a slight advantage is observed for **SynGAN** in terms of Dice and LTPR. However, a higher volume of FP voxels for lesion-free subjects is observed for the **SynGAN**, as compared to **SynAE**. Overall, the improvement of the realism of the synthetic data achieved by the proposed adversarial model (**SynGAN**) does not lead to a clear boost in terms of new MS lesions segmentation performance, as compared to the

baseline approach (**SynAE**). This may indicate that although the synthetic areas produced by the baseline inpainting AE model are blurry, it still helps the model learn meaningful features that are useful for real longitudinal data processing, such as the difference in contrast between the two visits.

As the MSSEG-2 challenge results are provided[2], these results can be compared to competition participants. The higher Dice obtained by the participants was 0.507, achieved by the MedICL team, while the second best Dice score was 0.500, achieved by team LaBRI. Regarding the average FP volume on lesion-free subjects, the best result (i.e. lower volume) was achieved by team LYLE with a volume of 0.470 mm^3, and the second best was 0.498 mm^3, achieved by Neuropoly-2 team. To conclude, the proposed strategy making use of synthetic data allows to slightly outperform the challenge participants concerning the Dice, yet exhibiting an increase for the FP detections on lesion-free subjects.

III.8.6 Quality of lesion-level uncertainty for new MS lesions

Now that an effective longitudinal segmentation model has been developed with the help of synthetic data, the lesion uncertainty quantification pipeline can be implemented. However, one important challenge arises concerning the training of the auxiliary classifiers, which require a sufficient amount of TP_{les} and FP_{les} in the training dataset to allow for training. The training split of MSSEG-2 only includes 33 subjects, and inference with the trained DE model yields to a total of 103 TP_{les} and only 28 FP_{les} instances for the training of the auxiliary classifiers. This seems inappropriate for learning-based approaches, more particularly the CNN and GNN approaches which contain more trainable parameters than the Logistic Regression model. As an attempt to circumvent this limitation, Data Augmentation is used to generate 10 variants of each of the 33 training images, leading to an extended dataset of 330 images. The same augmentation strategy used during training of the segmentation models is employed, comprising contrast, spatial, and artifact augmentations. This allows the construction of a lesion training dataset containing 983 TP_{les} and 317 FP_{les}.

Another particularity of this longitudinal experiment is that here two MRI sequences are used as inputs to the segmentation model. Thus, for the Logistic Regression model, intensity-based radiomics are extracted from both MRIs, for each lesion. This increases the number of features from 110 to 203. Similarly, the CNN model now receives a total of 4 bounding boxes for each lesion (one for each MRI visit, one for the entropy map, and one for the lesion mask) and the GNN model now receives an extra intensity node feature. Besides this change, all hyper-parameters are kept identical for the auxiliary classifiers. Appendix A.3.3 presents the top 10 features of the Logistic Regression model trained on the new MS lesions dataset. It presents a mix of intensity features from both MRI visits, as well as shape and uncertainty features (average interior entropy).

The qualities of lesion-wise uncertainty estimates for each proposed technique (Size, Mean Entropy, Logistic Regression, CNN, and GNN) are presented in Table III.8. The correlation of lesion uncertainty with respect to inter-rater variability is presented in Figure III.8.7. Finally, Figure III.8.8 reports the densities of lesion uncertainty scores for each method. It

[2] https://files.inria.fr/empenn/msseg-2/Challenge_Day_MSSEG2_Results_2021.pdf

	Voxel metrics			Lesion metrics						Lesion Counts		
	Dice (%) ↑			LTPR (%)			LFDR (%)			TP_{les}	FP_{les}	FN_{les}
Strategy	μ	SEM	CI	μ	SEM	CI	μ	SEM	CI	$\sum$	$\sum$	$\sum$
MSSEG-2 Only	46.8	5.4	$[-9.0, 8.8]$	33.7	5.5	$[-8.8, 9.2]$	72.2	4.8	$[-8.1, 7.8]$	168	**106**	65
SynAE Only	42.8	4.7	$[-7.9, 7.6]$	**39.6**	**5.9**	$[-9.7, 9.8]$	82.7	3.0	$[-5.1, 4.9]$	197	687	36
SynGAN Only	33.0	4.4	$[-7.3, 7.3]$	44.6	6.2	$[-10.2, 10.3]$	93.8	1.4	$[-2.3, 2.2]$	**215**	2771	**18**
Pretain on Syn-AE Finetune on MSSEG-2	49.5	5.4	$[-9.2, 8.6]$	36.1	5.8	$[-9.2, 9.7]$	**70.5**	**5.0**	$[-8.3, 8.0]$	185	121	48
Pretain on Syn-GAN Finetune on MSSEG-2	**50.9**	**5.4**	$[-9.0, 8.7]$	37.2	5.7	$[-9.4, 9.3]$	70.7	4.9	$[-8.1, 7.8]$	186	120	47

Table III.7: Performance of the Deep Ensembles trained with and without synthetic longitudinal data for new Multiple Sclerosis lesions segmentation. LTPR: Lesion True Positive Rate. LFDR: Lesion False Discovery Rate. TP_{les}: True Positives lesions. FP_{les}: False Positive lesions. FN_{les}: False Negative lesion. The mean (μ), the Standard Error on the Mean (SEM), and 90% confidence intervals (CI) are estimated using bootstrap.

Method	**AUROC** (%)			**AUPR** (%)			**Correlation IRV**		
	μ	SE	CI	μ	SE	CI	μ	SE	CI
Size	70.5	3.0	$[-5.0, 4.9]$	61.7	4.6	$[-7.7, 7.5]$	-0.42	0.06	$[-0.10, 0.11]$
Entropy	**77.9**	**2.6**	$[-4.4, 4.2]$	**66.4**	**4.7**	$[-8.0, 7.6]$	**-0.51**	**0.06**	$[-0.09, 0.10]$
Logistic	74.1	2.9	$[-4.9, 4.7]$	61.8	4.8	$[-8.0, 7.8]$	-0.43	0.06	$[-0.10, 0.11]$
CNN	69.0	3.0	$[-5.1, 4.9]$	56.2	4.6	$[-7.7, 7.7]$	-0.44	0.06	$[-0.10, 0.10]$
GNN	73.4	2.8	$[-4.8, 4.6]$	64.8	4.3	$[-7.4, 6.9]$	-0.48	0.06	$[-0.10, 0.10]$

Table III.8: Quality of lesion-level uncertainty estimates for new Multiple Sclerosis lesions detection. The top-performing scores are highlighted in bold for each metric. IRV: Inter-rater variability. The mean (μ), the Standard Error on the Mean (SEM), and 90% confidence intervals (CI) are estimated using bootstrap.

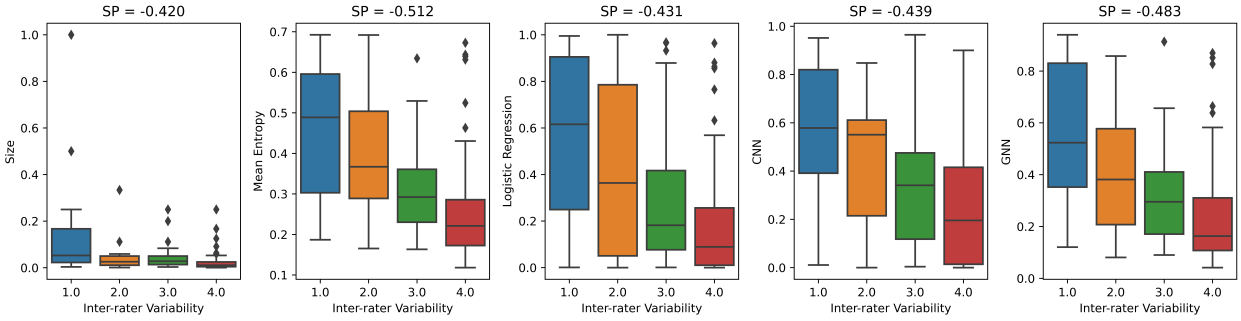


Figure III.8.7: Correlation of lesion uncertainty with respect to inter-rater variability for new MS lesions segmentation on the 60 test subjects. SP = Spearman's Correlation.

appears that the scores derived from auxiliary classifiers fail to outperform the Mean Entropy baseline which ranks first for each metric (AUROC, AUPR, correlation with IRV). This lack of performance can be explained by the reduced FP_{les} examples, which Data Augmentation alone fails to fully alleviate. This highlights one major limitation of the investigated classifier approaches, which is the need for a sufficient amount of TP_{les} and FP_{les} examples to allow for proper training. Regarding the correlation with the lesion-level inter-rater variability (Figure III.8.7), a moderate Spearman's correlation is observed for each of the compared estimators, with an advantage for the Mean Entropy, achieving a correlation of -0.512. The second best in terms of correlation is the GNN model.

As a side note, we can remark that the ground truths annotations of new MS lesions are noisy, similar to what was observed in the cross-sectional MS datasets (Section III.6.4). This is visible in Figure III.9.1. For lesions labeled as FP_{les} (red overlays), we can often see a hypersignal in the second image that was absent from the first image. This could thus indicate incomplete ground truths, although validation by a radiologist is necessary for these suspected cases.

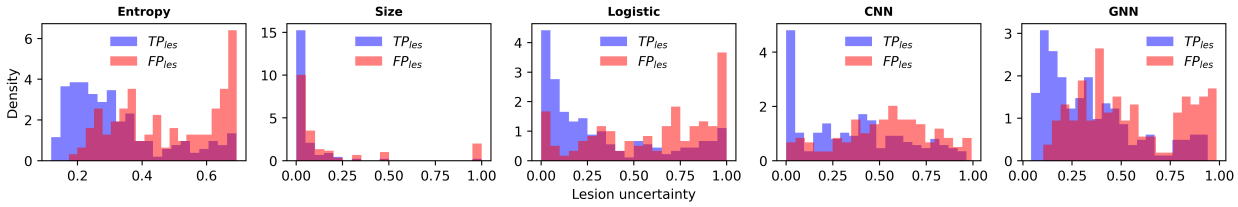


Figure III.8.8: Densities of lesion uncertainty scores for each approach on new MS lesions segmentation. Blue indicates the density of $\mathrm{TP}_{\mathrm{les}}$ lesions, while red stands for $\mathrm{FP}_{\mathrm{les}}$ lesions. The classifier approaches (Logistic, CNN, and GNN) fail at accurately distinguishing $\mathrm{TP}_{\mathrm{les}}$ from $\mathrm{FP}_{\mathrm{les}}$, which manifest by overlapping densities.

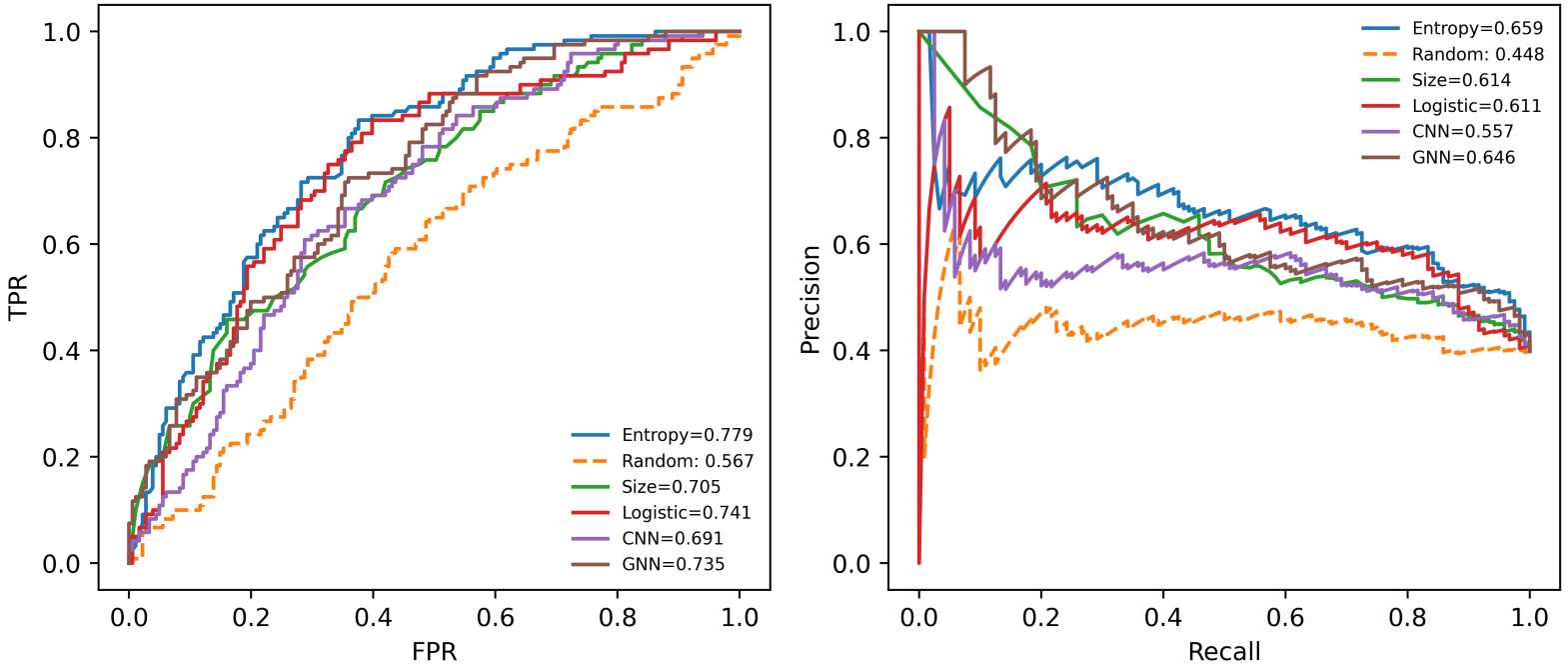


Figure III.8.9: Receiver operating characteristic (left) and precision-recall curves (right) for lesion uncertainty estimates on new MS lesions segmentation. In this experiment, the classifier-based approaches (Logistic, CNN, GNN) fail at outperforming the baseline entropy method.

III.9 Chapter conclusion

In this chapter, the problem of quantifying uncertainty at the lesion level was investigated, through three different lesion-oriented applications: cross-sectional and longitudinal MS lesions detection in FLAIR MRI, and lung nodules detection in CT. The goal was to determine if sophisticated lesion models based on features, bounding boxes, or graphs, can offer a better quantification of lesion uncertainty, as compared to a simple aggregation of the voxel uncertainties (Mean Entropy). To this aim, the ability to distinguish TP_{les} and FP_{les} instances based on their uncertainty score is monitored.

The results collected from the experiments show that the **GNN** approach is an efficient framework to quantify lesion uncertainty, with successful results on the cross-sectional MS and lung nodules experiments. It is particularly powerful when lesion instances are abundant and the annotation of lesions is consistent, which is the case on the LIDC-IDRI lung nodules dataset. In this setting, there is a performance gain as compared to the **Mean Entropy**. This translates into a gain in terms of AUROC and AUPR, as well as an increased correlation with human-level uncertainties (IRV and subtlety). Modeling lesion instances by graph is very convenient, as it can handle smoothly the complex geometry of these objects. This gain was also observed in the cross-sectional MS experiment, although it does not translate in all domain-shift settings (1.5 Tesla dataset). In contrast, the **Logistic** model performed well on the MS cross-sectional experiment but had weaker results on lung nodules. Finally, the **CNN** model provides overall weaker results in terms of classification-based lesion uncertainty quantification. This may indicate that using a standard DL classification approach is not suitable for this task, due to the limited size of the training data which leads to overfitting.

Performing training directly on lesion instances, as done for the **Logistic**, **CNN**, and **GNN** approaches, seems like an intuitive way to quantify uncertainty at the instance level. However, it is based on the assumption that a training dataset of lesion instances can be created with enough examples of TP_{les} and FP_{les}. When it is not the case, as in the longitudinal MS experiment, the performance of the auxiliary classifiers degrades and they no longer present a gain in performance compared to the baseline **Mean Entropy**. Moreover, these methods seem to be sensitive to label noise, which can occur when a TP_{les} is labeled as FP_{les} because of errors in the ground truth segmentations, which perturbates their learning. In these situations, the **Mean Entropy** that does not require any auxiliary training, offers a very competitive baseline. It is also paired with low complexity, as the **Mean Entropy** baseline only requires a CCA followed by lesion-wise averaging, a very efficient procedure.

Yet, one advantage of the developed classifier-based approaches is that they offer an interpretable uncertainty score, which is the probability that the lesion is a false detection. For clinicians, this may be easier to grasp than averaged entropy scores. Finally, the evaluation procedure (AUROC and AUPR scores) focused on the ranking of the uncertainty scores, which may not fully grasp the usefulness of the lesion-level uncertainty scores in clinical routine. An open lead would be to monitor the performance of the AI-clinician pair with and without these lesion-level uncertainty scores to determine their real clinical benefit.

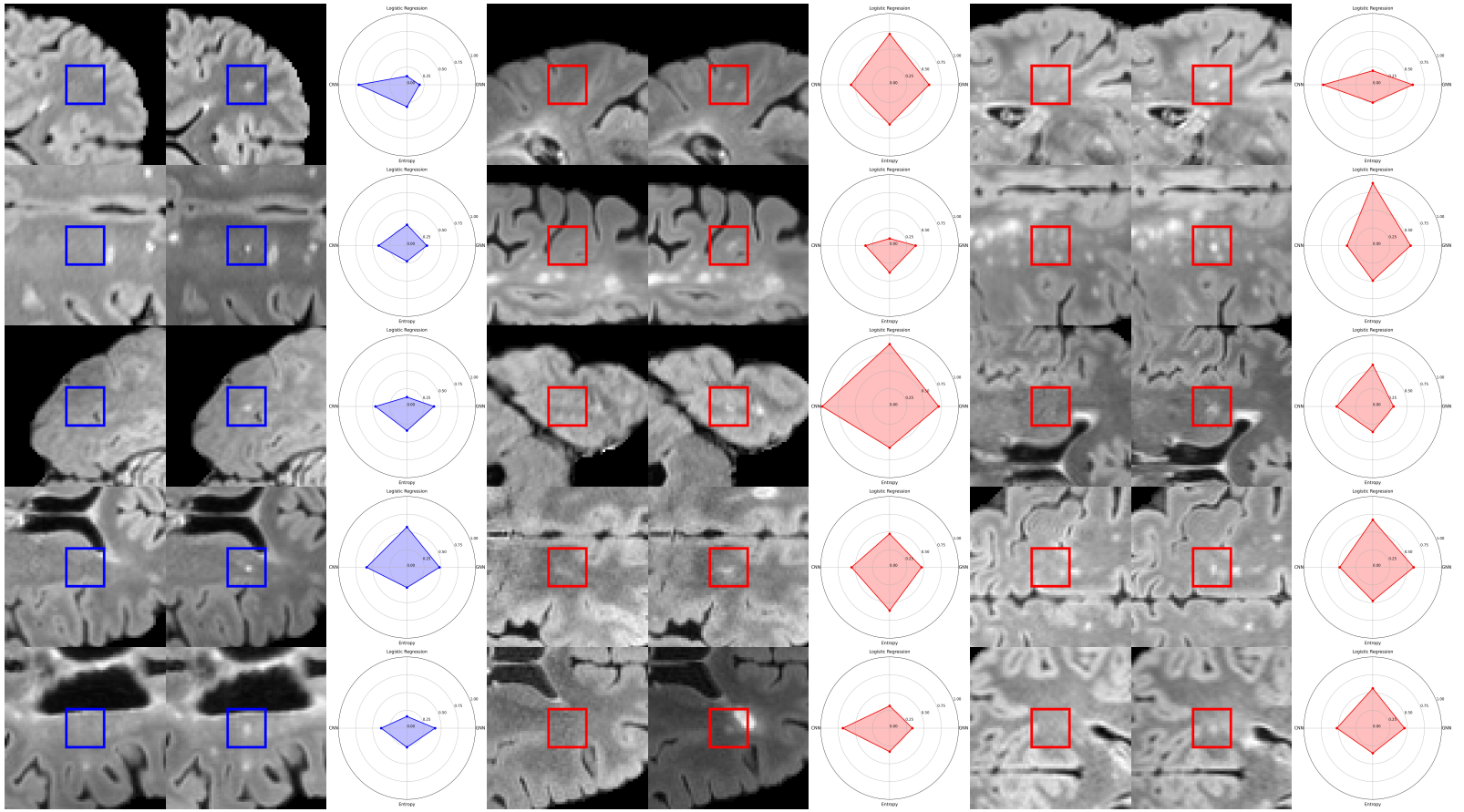


Figure III.9.1: Examples of lesion uncertainty quantification for the different tested approaches, for new MS lesions detection. For each case, from left to right, we present the first visit, the second visit, and a spider chart indicating the uncertainty scores estimated by each method. Red overlays indicates that the lesion is a $\mathbf{FP}_{\text{les}}$, while blue overlays indicates TP_{les}. Note that in many cases, lesions labeled as $\mathbf{FP}_{\text{les}}$ present a clear hyperintensity signal in the second visit that was not present in the first one, which may indicate errors in the ground truth segmentations.

Chapter IV

Out-of-distribution detection and quality control for medical image segmentation

Contents

IV.1 Motivations

The previous voxel and lesion-level experiments highlighted one major weakness of DL models, which is the lack of robustness in the presence of domain shifts. More formally, this occurs when test samples differ significantly from the training samples, in which case the segmentation quality generally drops, leading to misleading predictions. By default, DL segmentation models are not equipped with an abstention mechanism, meaning that when confronted with an image far from their training distribution they will still produce a segmentation, sometimes with high confidence. An illustration of this phenomenon is provided in Figure IV.1.1, in which a DL ensemble trained to segment glioblastoma in brain T1w MRI is applied to a lumbar T1w MRI. The ensemble classified an important part of the image as brain tumor, even though there is no clear pattern explaining the error in the input image. More worryingly, the entropy map highlights an area of low voxel uncertainty (blue areas) within the predicted foreground. This area is highlighted by the yellow arrow overlayed on the entropy map.

At Pixyl, the daily number of analyses is quickly increasing, with a number approaching 200 analysis per day at the moment of writing of this thesis (see Figure IV.1.2). This means that a visual inspection of each input image is not humanly feasible before sending the result of the analysis to the clients. Moreover, due to human mistakes or misunderstanding of the software specifications, it is frequent that a non-conform image is sent to be analyzed, such as a FLAIR MRI for a model expecting a T1w MRI. One promise of UQ is to be able to automatically detect these pathological cases using the model uncertainty. Ideally, image-level uncertainty scores should be higher for out-of-distribution (OOD) images than for in-distribution (ID) images, allowing for their detection. This would allow automatically alerting the user that the sent image does not conform to what the model was trained for, and warn that results may be suboptimal. In this chapter, we aim to investigate how reliable uncertainty estimates are for this task of OOD detection. To do so, a complete benchmark comprising various scenarios of data shifts is proposed, allowing benchmarking of standard UQ techniques as well as more recent OOD detectors, including reconstructed-based and latent-based approaches. For the latter, we propose an investigation on the sensitivity of the approach with respect to the choice of the layer used for feature extraction, and the choice of the segmentation architecture.

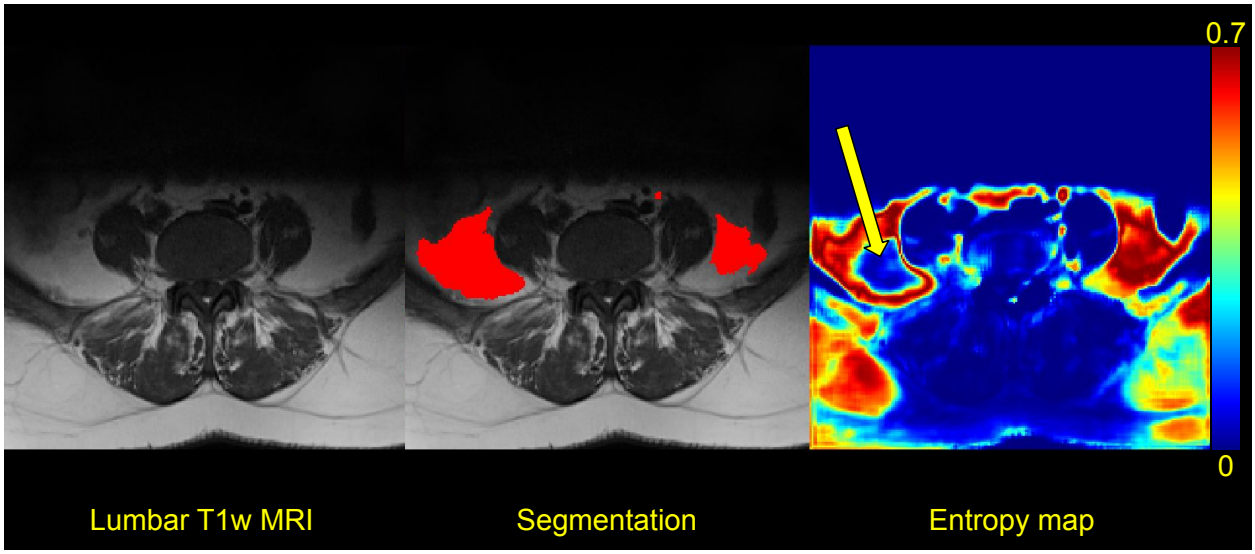


Figure IV.1.1: Illustration of an extreme OOD case. A T1w MRI brain tumor segmentation ensemble is applied to an image far from its training distribution (tumor-free lumbar MRI). Without an abstention mechanism, it detects a tumor volume of 136mL. The yellow arrow on the entropy map indicates a region of overconfident error.

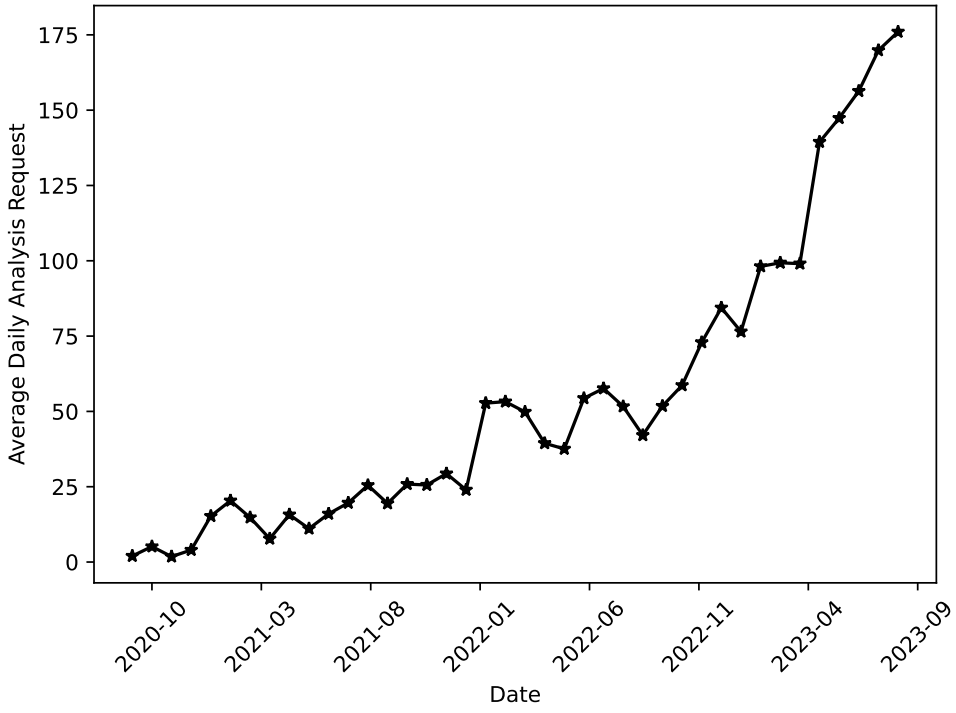


Figure IV.1.2: Evolution of the average daily number of analysis requests at Pixyl, over the period 2020 to autumn 2023. At the time of writing this thesis, approximately 200 automatic analyses are performed daily.

IV.2 Additional contributions to the paper "Multi-layer Aggregation as a Key to Out-of-distribution Detection"

This chapter is based on the work previously presented in the paper *Multi-layer Aggregation as a Key to Out-of-distribution Detection* [155]. Several additions are presented here. First, a new category of OOD data is included, namely **Population shifts**, corresponding to shifts in the imaged population or disease subtype. The uncertainty baseline, which was the MC dropout model in the original paper, is replaced by the DE approach for a stronger uncertainty quality. The most recent version of the BraTS dataset is used (BraTS 2023) instead of BraTS 2021. A reconstruction-based baseline is added as an additional contribution, namely the Memory-guided Normality for Anomaly Detection (MNAD) model, for which a 3D implementation is proposed. For latent-space OOD detection, the discussion focuses on the use of the Mahalanobis Distance due to its high performance on several medical image applications. A deeper dive into the dependence of the layer selection, multi-layer aggregation function, and incidence of the neural network architecture is presented. Finally, the second part of the chapter focusing on the link between OOD detection and segmentation quality assessment is a novel addition.

IV.3 Out-of-distribution detection for medical-image segmentation

IV.3.1 In and out-of-distribution datasets

In this chapter, we aim to explore in detail various domain-shift scenarios, including variations in the imaged population, disease, modality, and noise level. To allow for many OOD settings, the choice is made to focus on T1w MRI as the expected modality, which is widely available in open-source repositories, as compared to T2-weighted FLAIR MRI for example. In terms of predictive task, the brain tumor segmentation exercise is selected, as it allows the usage of a large open-source, multi-center imaging dataset (BraTS 2023 [226, 281]), allowing robust conclusions to be drawn. Thus, **in-distribution** data will correspond to brain T1w MRI of adult patients with glioblastoma. In our experimental setting, the tumor delineation has to be carried from a single T1w MRI sequence. Thus, the task is made easier by focusing on the binary segmentation of the whole tumor, corresponding to the concatenation of the 3 tumor tissue classes (necrosis, edematous, and enhancing tumor). Additionally, control samples are added in the experimental protocol, corresponding to a cohort that shares the same properties as the training samples (same modality, organ, and pathology), but that were acquired in a different imaging center. An effective model should be able to generalize to these images and thus, the OOD detection module should identify them as ID samples to prevent false alarms. In this direction, we propose to use the LUMIERE glioblastoma dataset [282] as a Control dataset, from which we select 74 pre-operative T1w brain MRIs. The images were acquired at the University Hospital of Bern, Switzerland.

To develop the segmentation models, we use the same train/validation/test stratification of

the BraTS 2023 dataset as the one used in the voxel-level experiments (Section II.6.1): 876 subjects for training, 30 for validation, and 227 for testing. The testing split is called Test ID in the following. Then, 24 OOD datasets, categorized into 5 types of shifts, are investigated. Each OOD dataset is presented in the following.

IV.3.1.1 Transformation shifts

A very frequent type of *shifted* data in medical image analysis corresponds to noisy images, presenting artifacts. However, gathering a sufficient amount of these noisy data points for the OOD experiments is cumbersome. Fortunately, recent advances in data augmentation allow the generation of artifact images from clean images with high realism [72]. Interestingly, this allows total control over the strength of the injected artifacts. To generate Transformation shifts datasets, the 227 Test ID images are corrupted with artifacts. Several augmentations are investigated:

- **Motion artefact.** Acquiring a 3D MRI is a time-consuming process, making it sensitive to the subject's motion during the acquisition. Motion artifacts can then appear, manifesting primarily as the blurring of sharp edges in the image. To simulate motion artifacts, the k-space of the image can be altered [283]. Specifically, 2 head movements are emulated, as if the head had moved with a rotation comprised in the range $[-10, 10]$ degrees and with a translation comprised in the range $[-10, 10]$ mm, in a random direction.
- **Ghost artifacts** correspond to a replication of the imaged region along one or several axes of the image. They mainly originate from periodic motion during the MRI acquisition, including cardiac or respiratory movements.
- **Bias artifact** are very frequent MRI artifacts that cause nonuniform illumination within the acquired image. As a result, some parts of the image can appear darker (respectively brighter) than the rest of the image. This is due to the inhomogeneity of the MRI magnetic field that yields variations in low frequencies across the volume. This can be emulated using a linear combination of polynomial basis functions [284].
- **Spike artifacts** (also called Herringbone artifact) corresponds to periodic stripes appearing in the image, due to aberrant points in the k-space, that can be caused by, among other things, Radio-Frequency pulse abnormalities.
- **Gaussian Noise** perturbation is an easy and popular approach to degrading the quality of an image. Here, images are first normalized so that their mean intensity is 0, and the standard deviation equals to 1. Then, Gaussian noise with a mean of 0 and a standard deviation of 0.5 is added to the image.
- **Downsampling.** Due to time constraints, MRIs can be acquired at low resolutions or with anisotropic voxel sizes. For instance, MRIs are often acquired with a slice thickness superior to the axial resolution (e.g. a voxel resolution of $1 \times 1 \times 3$, meaning that the slice thickness is 3 times superior to the in-plane resolution). This anisotropy can be simulated by downsampling the image along one or several axes, then interpolating the image back to its original resolution [285], effectively decreasing the image resolution in the concerned directions, making the analysis more ambiguous.
- **Scaling Perturbation** leverages scaling modification of the image to augment or shrink the appearance of the imaged region. In 50% of cases, the brain doubles in size,

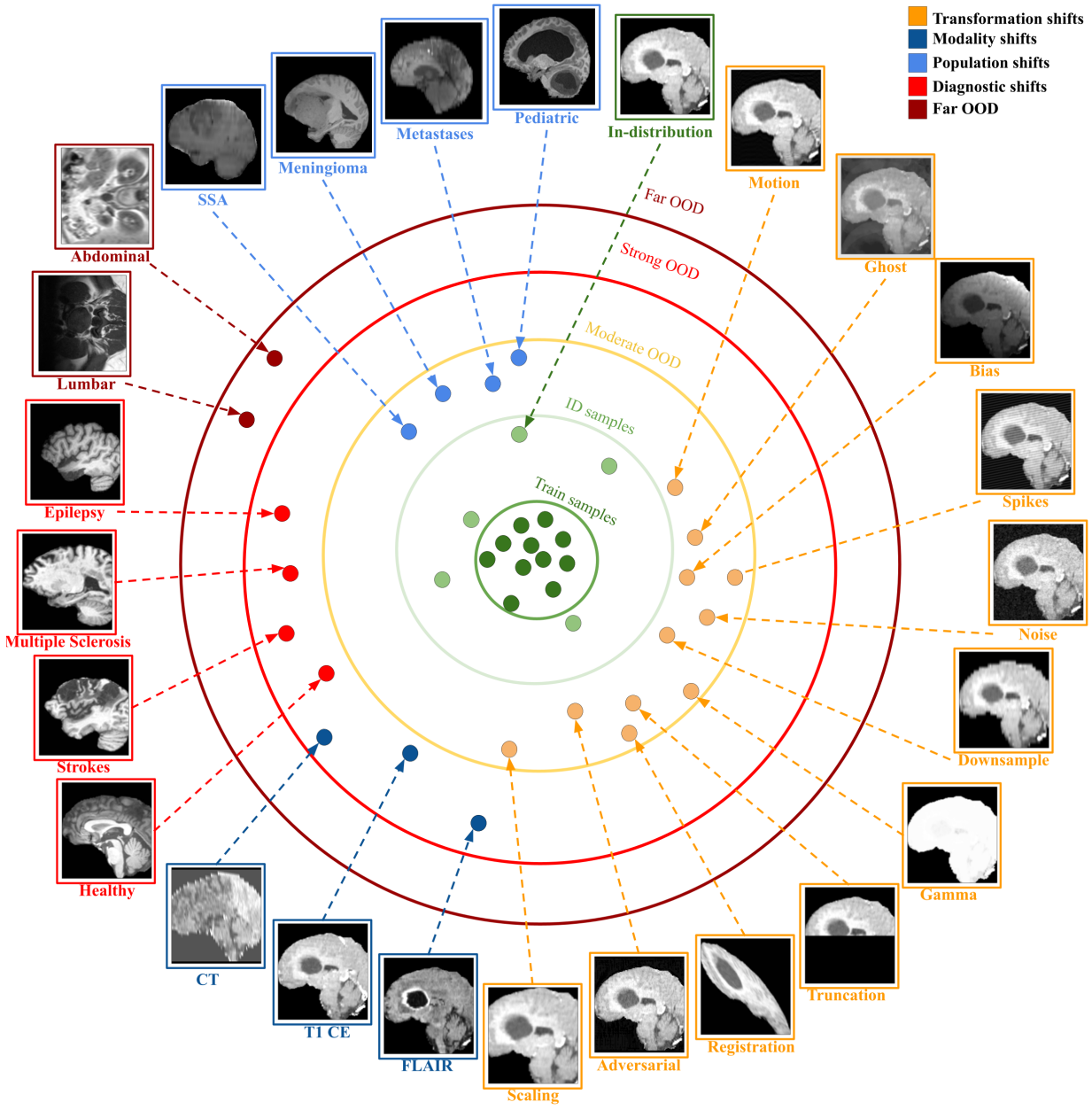


Figure IV.3.1: Illustration of the different out-of-distribution datasets used in the experiments. Five categories are explored: Transformation, Modality, Population, Diagnostic, and Far OOD.

whereas in the rest of cases, the brain shrinks by a factor of half.

Additionally to these standard MRI perturbations that are implemented in the TorchIO library [246], several additional types of perturbations are implemented for the OOD experiments:

- **Gamma alterations** correspond to extreme contrast alteration of the image, obtained by raising the intensity values to the power γ. More specifically, two gamma values are employed, respectively $\gamma_1 = 4.5$ and $\gamma_2 = -4.5$.
- **Truncation** corresponds to a random cropping of half of the image, in a random direction. This emulates errors in the file transfer or download, which can thus contain missing slices.
- **Erroneous Registration.** Image registration is a widespread preprocessing step of medical images, aiming at aligning images to a common reference atlas. For instance, the BraTS images are registered to the SRI24 atlas of healthy adult brains [286]. However, this registration can be erroneous, yielding to deformed brains. To mimic this process, the registration matrix is corrupted by adding noise to it. More precisely, the T1w on the Test ID are registered on the SRI24 to obtain a 3D affine registration matrix A, which corresponds to a 4×4 matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

  where blue indices quantify the rotation, shearing, and scaling applied to the original image, whereas the red indices stipulate the amount of translation. From this matrix, an erroneous registration matrix is obtained by applying noise to the matrix elements. For rotation, shearing and scaling indices, the perturbation n_{ij} follows a Gaussian distribution $\mathcal{N}(0, 0.1)$. For translation, the perturbation u_{ij} is drawn from an uniform distribution $\mathcal{U}(-5, 5)$. The final perturbed registration matrix is obtained via:

$$A_{noise} = \begin{bmatrix} a_{11} + n_{11} & a_{12} + n_{12} & a_{13} + n_{13} & a_{14} + u_{14} \\ a_{21} + n_{21} & a_{22} + n_{22} & a_{23} + n_{23} & a_{24} + u_{24} \\ a_{31} + n_{31} & a_{32} + n_{32} & a_{33} + n_{33} & a_{34} + u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

  Finally, registration is recomputed using A_{noise} instead of A on both the T1w and ground truth segmentation to generate the OOD dataset.
- **Adversarial Attacks** corresponds here to perturbations added to the input image to hack the functioning of the DL model and yield overconfident mistakes. More particularly, the Fast Gradient Sign Method (FGSM) [287] is employed, following which the adversarial noise η is proportional to the gradient of the image x with respect to the loss used to train the network $\mathcal{L}(\boldsymbol{y}, \boldsymbol{x}, \theta)$:

$$\eta = \epsilon \text{sign}(\Delta_x \mathcal{L}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \tag{IV.3.1}$$
$$\tilde{x} = x + \eta \tag{IV.3.2}$$

where y is the ground truth, θ the parameters of the trained network, and $\tilde{x}$ the altered image. Note that the original FGSM implementation, proposed for 2D image classifiers, expects that the ground truth y is available at the moment of the attack, akin to a *white-box* attack. However, for a more realistic setting, it is considered here that the ground truth is not accessible at the moment of inference. Thus, the prediction of the model $\hat{y}$ is used as pseudo ground truth. As a result, the FGSM will apply a destructive perturbation pushing the prediction far from the one obtained on the unaltered image.

IV.3.1.2 Population shifts

DL models are usually trained with a dataset representative of a single population, for example, adult patients with glioblastoma in the context of the BraTS Adult Glioblastoma dataset. This population represents the model's optimal operating domain, and deviation from this target population may result in suboptimal results. However, when the model is deployed, all test samples will likely not exactly match the training population, especially in industrial software. Thus, several population shift settings are included in the OOD experiments, relying on the BraTS 2023 auxiliary datasets:

- **Pediatric subjects** [288] corresponds to MR images of pediatric subjects diagnosed with glioblastoma. The dataset includes 99 cases of infants older than one month of age. Although these two distributions (adult population on one side, pediatric population on the other) appear to be very different, the images in the pediatric database are pre-processed in the same way as the images in the adult database, including the registration to the SRI25 healthy adult template [286]. As a result, variations in brain size are partly eliminated.
- **Metastase** is a type of brain tumor that originates from cancer cells migrating from their original organ to the brain [289]. The BraTS 2023 Metastases dataset comprises 238 cases. Brain metastases are generally smaller than glioblastoma, hence their segmentation represents an important challenge for the segmentation models.
- **Meningioma** corresponds to brain tumors originating from meninges in the brain, thus distinguishing themselves from glioblastoma from their location in the brain. For this setting, a subset of 250 cases from the BraTS 2023 Meningioma dataset [290] is employed.
- The **Sub-Saharan Africa** BraTS 2023 dataset [238] dataset comprises 60 subjects, which mainly differ from the BraTS Adult Glioblastoma distribution from the lower quality of MRI and the more advanced stage of the disease, due to late diagnosis.

Samples from each dataset are provided in Appendix A.3.1 to highlight the heterogeneity between the populations.

IV.3.1.3 Modality shifts

Medical images are generally saved in DICOM formats, whose meta-data (headers) may be incorrectly filled [291]. As a consequence, mismatches between the expected input modality (here, brain T1w) and the test image modality (e.g CT or T2w) may be undetected. To represent this scenario, 3 different Modality shift OOD datasets are proposed:

- The FLAIR and T1ce sequences corresponding to the 227 Test ID subjects are used to mimic errors in the input MRI sequence. While FLAIR sequences are visually very different from the expected T1w sequences, the injected T1ce should be more difficult to distinguish automatically from the standard T1w images.
- To illustrate a more drastic modality error, brain CT scans are employed in place of brain MRIs. To implement this, 250 subjects from CQ500 dataset [292] are employed, which contains brain CT scans of patients undergoing intracranial hemorrhage or cranial fractures.

IV.3.1.4 Diagnostic shifts

DL segmentation models are usually trained to handle a single pathology (e.g. brain tumor, MS, strokes). Yet, once the model is deployed in the real world, it can be confronted with images exhibiting unseen anomalies, which can lead to misleading predictions. To test OOD detection methods on this scenario, T1w brain MRIs that **do not** present a tumor are used:

- The **Multiple Sclerosis** dataset corresponds to the 170 T1w MRIs from the WHM 2017 dataset [293].
- The **Stroke** dataset corresponds to 250 T1w MRIs selected from the ATLAS-2 dataset [242].
- The **EPISURG** dataset [294] corresponds to the patients who underwent brain resection as a treatment for epilepsy. It corresponds to 162 brain T1w MRIs.
- The **Healthy** dataset corresponds to 250 T1w brain MRIs from young and healthy adults collected from the IXI dataset [295].

IV.3.1.5 Far Out-of-Distribution

Finally, OOD methods are evaluated in extreme cases that correspond to non-brain MR data. It should be imperatively detected as OOD data. Two settings are tested:

- The **Abdominal** dataset corresponds to 80 abdominal T1w MRI from the CHAOS dataset [296].
- The **Lumbar** dataset corresponds to 250 images extracted from the Lumbar Spine MRI dataset [297].

IV.3.2 Data preprocessing

All brain OOD data follow a preprocessing pipeline similar to the one applied to BraTS data. More specifically, all brain data are skull-stripped using the HD-Bet algorithm [232]. Then, images are registered to the SRI-24 atlas [286], resulting in an isotropic voxel resolution of $1\,\mathrm{mm}^3$ and an image size of $240 \times 240 \times 155$. For non-brain data (abdominal and lumbar MRI), a resampling to $1\,\mathrm{mm}^3$ is used followed by a center cropping to match the target volume size of $240 \times 240 \times 155$. This ensures that the images share the same spatial dimensions as the in-distribution data.

IV.3.3 OOD detection metric

OOD detection is usually cast as a binary classification problem, where ID samples are labeled as 0, and OOD samples as 1. The tested OOD detection methods provide a continuous non-conformity score, where higher values should correspond to higher degrees of non-conformity. To compute classification metrics, the scores on the OOD data are compared to the scores obtained on the Test ID dataset, allowing the computation of the areas under the ROC curve (AUROC) and PR curve (AUPR) for each of the OOD datasets. Additionally, Dice scores are reported for images where the manual delineation of the whole tumor is available (i.e. all Transformation and Population shifts, as well as the FLAIR and T1ce images in the Modality shifts). For the rest of the OOD datasets (Diagnostic shifts, Far OOD, the CT scan dataset), all segmented voxels will correspond to false positive voxels. Thus, the average false positive volume per subject is reported in place of the Dice.

IV.3.4 Pitfalls of classic UQ methods for OOD detection

As a first attempt at this benchmark, an OOD score derived from the standard UQ methodology is tested. The DE approach proved to be a powerful voxel and lesion-level uncertainty estimator. Naturally, an immediate idea for OOD detection is to derive the entropy maps computed by the DE to obtain an image-level OOD score. Following the experimental protocol used throughout this thesis, we train 5 Dynamic U-Nets using the cross-entropy and Dice++ losses (Equation II.6.3) to build a DE. The particularity is that data augmentation is reduced here to simple spatial (flip, mirroring) and gamma alterations. Artefact augmentations are discarded as they are used to simulate OOD data (e.g. Bias, Motion, Ghosting, Spikes).

A simple image-level score derived from the DE is adopted, which consists of the computation of the average entropy over the entire volume as an image-level OOD score. Another solution would be to use the average foreground entropy, however, this supposes that the DE will classify as foreground at least one voxel in the image, which may not be always the case. Moreover, focusing on the foreground voxels will discard all uncertain voxels that can be present in the background. Thus, the following image-level OOD score is used for the DE:

$$\text{OOD}_{DE} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{H}_n \tag{IV.3.3}$$

where $\mathcal{H}_n$ is the entropy value at the n-th voxel of the image.

Figure IV.3.4 presents the segmentation performance for all ID and OOD datasets for the Dynamic U-Net ensemble. Dice scores are reported for datasets where expert manual delineations of tumors are available. For the rest, the average FP volume per subject is reported. In Figures IV.3.2 and IV.3.3, the performance of this score concerning OOD detection on the different proposed scenarios is illustrated, along with the performance of a random classifier.

First, the AUROC and AUPR scores on **Control** samples are similar to the one of a random

classifier, which indicates that Control and Test ID samples are indistinguishable based on the uncertainty-based OOD score. This is an expected property of the OOD detector, as **Control** samples should indeed be considered as ID samples. Then, it appears that heavy shifts are well detected, including **Adversarial**, **FLAIR**, **CT**, **Lumbar** and **Abdominal** datasets. Several OOD datasets are moderately well-detected (**Motion**, **Ghost**, and **Bias**). The remaining OOD samples are poorly detected, with performance being sometimes inferior to the one of a random classifier (**Diagnosis shifts** and most **Population** shifts). Averaging the OOD detection performance on the 24 OOD datasets yields an average AUROC of 0.66 and an average AUPR of 76, which is not satisfying enough for our OOD detection module.

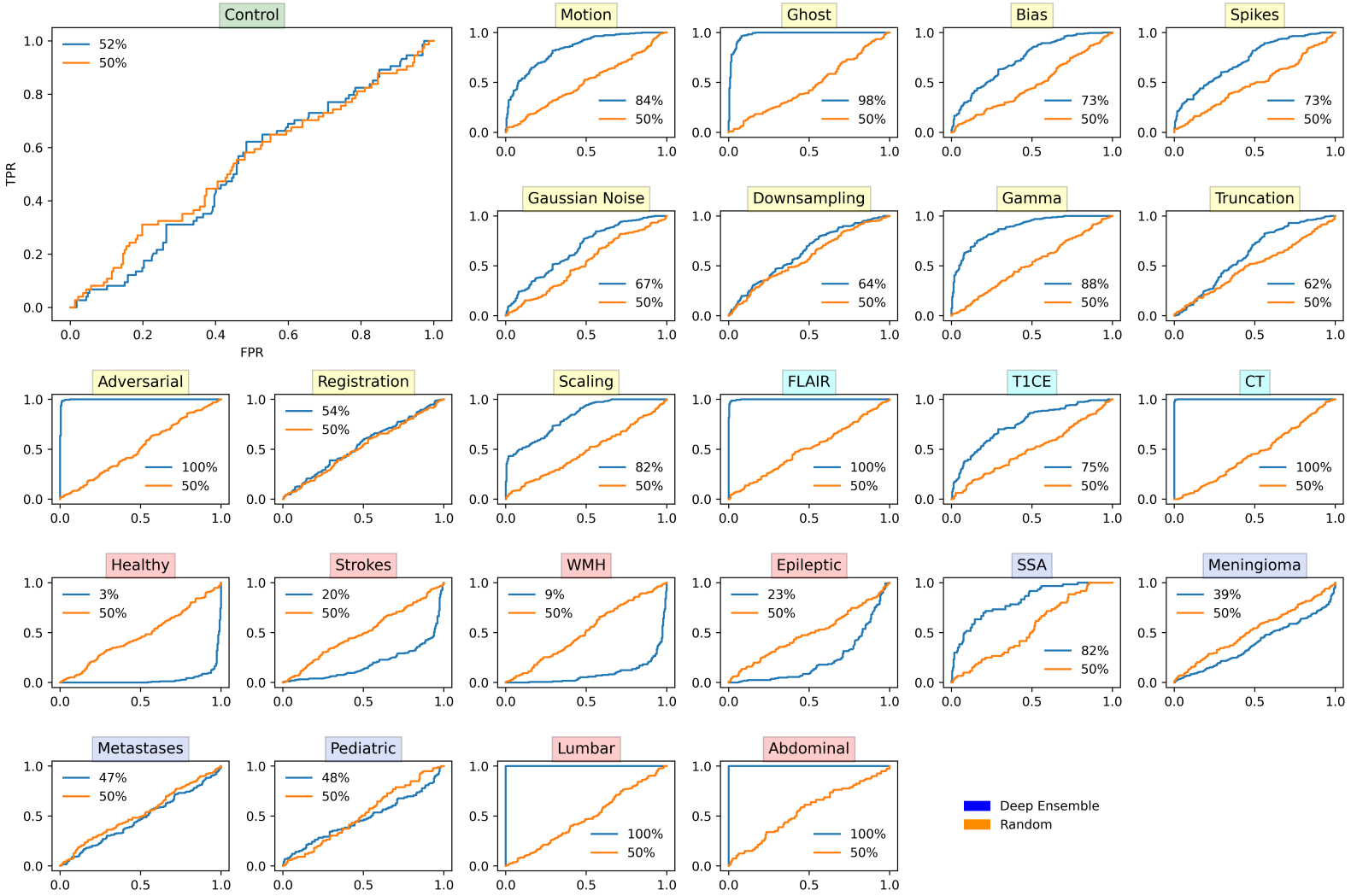


Figure IV.3.2: Receiver Operating Characteristic curves for the Deep Ensemble (blue) on the OOD benchmark, along with the performance of a random classifier (orange).

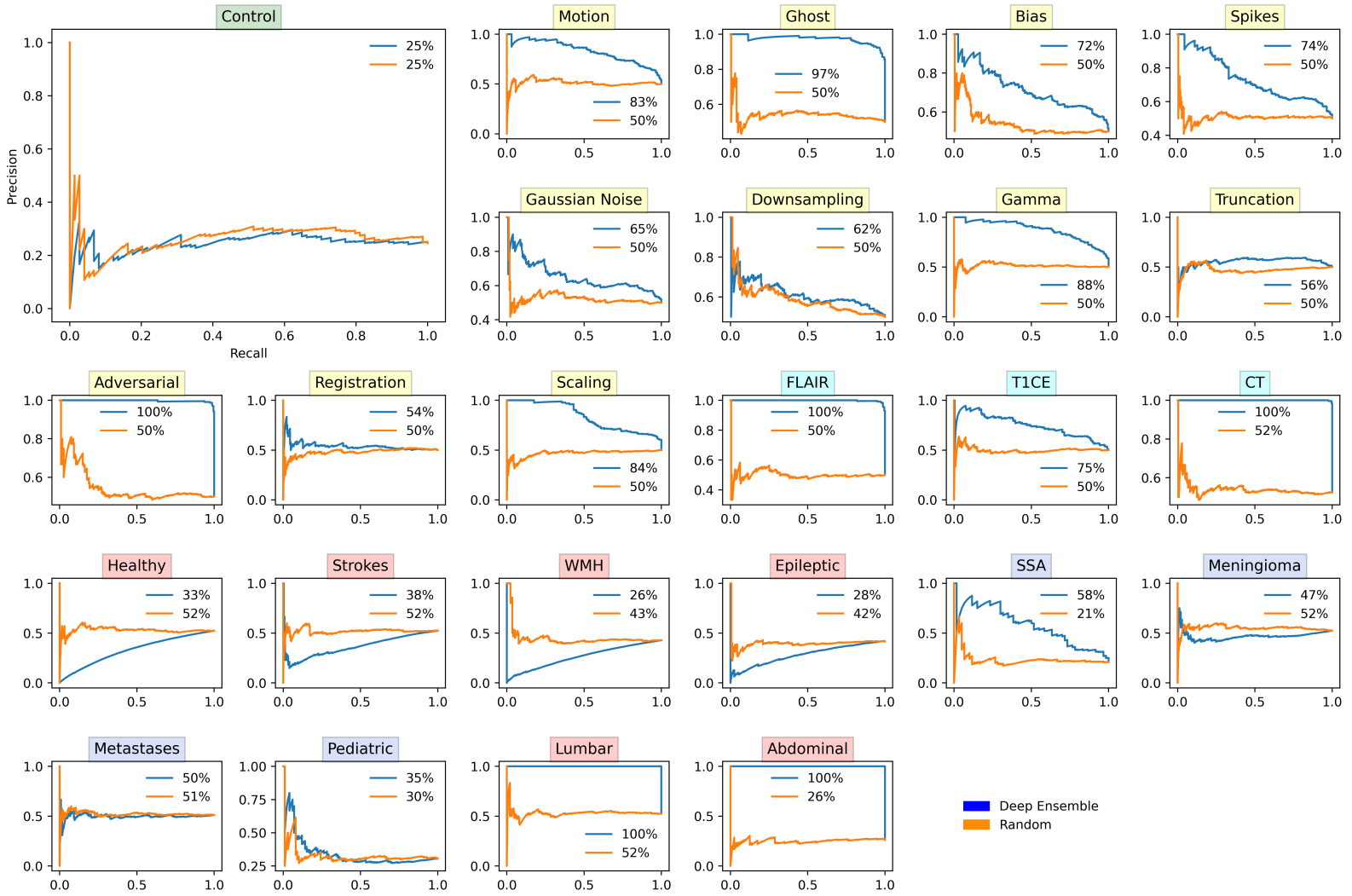


Figure IV.3.3: Precision-recall curves for the Deep Ensemble (blue) on the OOD benchmark, along with the performance of a random classifier (orange).

These observations are in line with previous work on OOD detection using uncertainty approaches. For instance, Ulmer et al. [298] showed that ReLU models fail at OOD detection as the confidence level of the NN extrapolates to out-of-distribution data, which hinders their usefulness for OOD detection tasks. For real-world use, the OOD detection algorithm should not only detect extreme OOD cases but also more subtle shifts (such as population and diagnosis shifts). This first observation thus motivates the evaluation of other OOD detection solutions.

IV.3.5 An unsupervised anomaly detection baseline for OOD detection

IV.3.5.1 Concept

A popular paradigm for anomaly detection is to use the reconstruction error of an autoencoder (AE) model, trained on ID data only, as an OOD score. The hypothesis is that the reconstruction error on OOD data will be higher than for ID data, as the model was not trained to reconstruct OOD data during training. This approach requires building a model dedicated to OOD detection, contrary to the uncertainty paradigm that uses the outputs of the segmentation models. Interestingly, it is a fully unsupervised approach as the AE only needs ID images to be trained, alleviating the need for manual annotations, and does not require access to OOD data during training.

However, one limitation of this approach is that DL AEs demonstrate a powerful generalization capability, making them able to reconstruct OOD at test time with high fidelity, thus violating the basic assumption of reconstruction-based approaches. As an attempt to alleviate this weakness, several improvements have been proposed to hinder the generalization of AE, including memory-augmented AEs [299, 300]. In *Learning Memory-guided Normality for Anomaly Detection* (MNAD for short), Park et al. [300] propose to build an AE model that explicitly learns prototypes of the ID images during training. At test time, the model reconstructs the input using the learned prototypes, thus lessening the generalization capacity of the DL model. These patterns are stored using a dedicated module, called memory module.

The MNAD architecture has been proposed for video images, thus making use of 2D convolutions. Here, a 3D adaptation is proposed for 3D image processing, illustrated in Figure IV.3.5. As in standard AE models, the MNAD model is composed of an encoder and a decoder. Additionally, a memory module is added to the bottleneck of the model. The memory receives as input the feature maps computed by the encoding part. This corresponds to a 4D array of shape $64 \times H \times W \times D$ where $H \times W \times D$ is the size of the 3D image. This array contains the so-called queries of shape $64 \times 1 \times 1 \times 1$, each being associated with a voxel in the feature map. The memory modules then operate two distinct operations, namely reading and updating, using the M memory items $\boldsymbol{p}_m \in \mathcal{R}^{64}$ containing the ID data prototypes learned during training. Reading consists of the computation of matching probabilities $w_{m,k}$ for each memory item and each query, where high weights indicate that the currently observed query is similar to the memory item. This is followed by the computation of updated features $r_k \in \mathcal{R}^{64}$ using a weighted average (one r_k per voxel in the feature map output by the encoder):

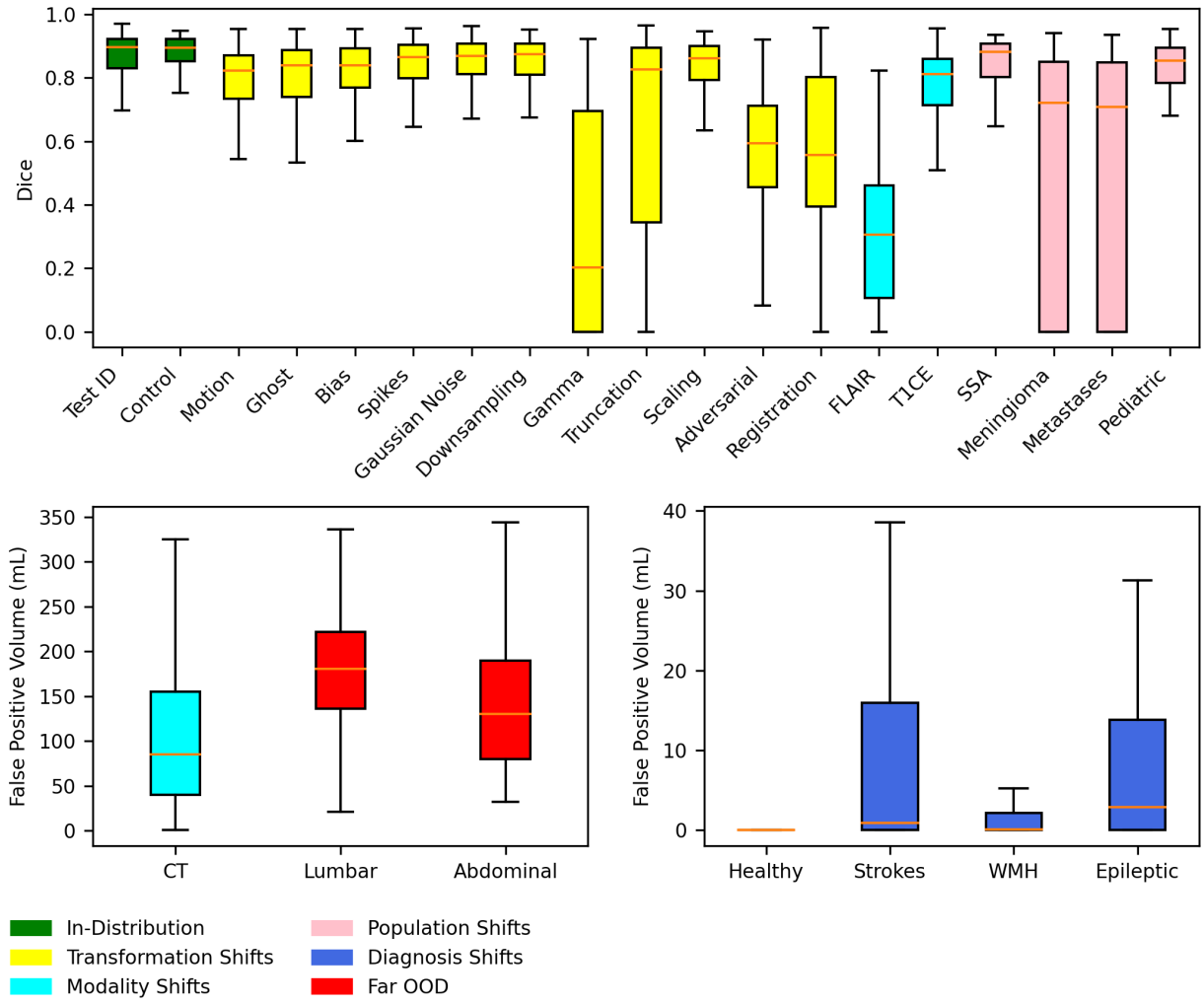


Figure IV.3.4: Segmentation performance of the Dynamic U-Net Ensemble on the different datasets used in the OOD experiments. The average Dice is presented for datasets where the ground truth delineation of the whole tumor is available (top row). For the rest of the datasets, we present the average False Positive volume per subject, in milliliter.

$$w_{m,k} = \frac{\exp((\boldsymbol{p}_m)^T \boldsymbol{q}_k)}{\sum_{m'=1}^{M} \exp((\boldsymbol{p}'_m)^T \boldsymbol{q}_k)} \tag{IV.3.4}$$
$$\text{and } r_k = \sum_{m'=1}^{M} w_{k,m'} \boldsymbol{p}_{m'}$$

The queries and updated features are concatenated and serve as inputs to the decoder. Finally, the memory items are updated using queries as follows:

$$\boldsymbol{p}_m \leftarrow \|\boldsymbol{p}_m + \sum_{k=1}^{K} v'_{k,m} \boldsymbol{q}_k\|_2 \tag{IV.3.5}$$
$$\text{with } v'_{k,m} = \frac{v_{k,m}}{\max_k v'_{k,m}}$$
$$\text{and } v_{k,m} = \frac{\exp((\boldsymbol{p}_m)^T \boldsymbol{q}_k)}{\sum_{k'=1}^{K} \exp((\boldsymbol{p}'_m)^T \boldsymbol{q}_k)}$$

In practice, the memory items are updated by using a weighted sum of the queries, emphasizing the queries that are near to the memory item. Training is performed using a 3-terms loss objective, composed of a reconstruction term (L2 distance) as in standard AE models, as well as two auxiliary losses, namely the feature compactness and separateness losses. The compactness loss aims at ensuring that the queries will be close to the nearest memory item. It is defined as:

$$\mathcal{L}_{compactness} = \sum_{k=1}^{K} \|\boldsymbol{q}_k - \boldsymbol{p}_i\|_2 \tag{IV.3.6}$$
$$\text{with } i = \arg\max_{m \in M} w_{k,m}$$

where i is the index of the nearest memory item for query $\boldsymbol{q}_k$. However, this compactness term can push queries and memory items to be alike, as the former is used to update the latter. To prevent this, a separateness term is computed between the queries and $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, its closest and second closest memory items:

$$\mathcal{L}_{separateness} = \sum_{k=1}^{K} \max(\|\boldsymbol{q}_k - \boldsymbol{p}_i\|_2 - \|\boldsymbol{q}_k - \boldsymbol{p}_j\|_2 + \alpha, 0) \tag{IV.3.7}$$
$$\text{with } j = \arg\max_{m \in M, m \neq i} w_{k,m}$$

where α corresponds to a *margin* set to 0.1. The overall loss is finally obtained by computing the weighted sum of the three terms:

$$\mathcal{L}_{MNAD} = \mathcal{L}_{Rec} + \lambda_1 \mathcal{L}_{compactness} + \lambda_1 \mathcal{L}_{separateness} \tag{IV.3.8}$$

where the weighting factors λ_1 and λ_2 are set to 0.01.

IV.3.5.2 Implementation details and training parameters

The proposed MNAD 3D builds on the original 2D implementation[1], and is obtained by replacing 2D convolutions per their 3D counterparts. The number of memory items M is kept to its default value of 10. Instance normalization is used in each block. Training is carried out in a full 3D manner, using a batch size of 1 and the ADAM optimizer [37] with a learning rate of 2×10^{-4}. The training dataset corresponds to the T1w of the BraTS 2023 training split. As for the DE, data augmentation is kept to a minimal setting comprising spatial (rotation and mirroring) as well as gamma alterations. Figure IV.3.6 presents the loss functions monitored during training for the MNAD model.

IV.3.5.3 Performance of the MNAD model in the OOD benchmark

Figure IV.3.7 displays two examples of reconstruction predicted by the MNAD model, for an ID and an OOD image. First, the reconstruction error on ID data is low, and the reconstruction is accurate (although a little bit blurry, a known limitation of AE models). For the OOD data that exhibit a spike artifact, the MNAD model did not reconstruct the artifact, resulting in a high reconstruction error in the background of the image. Overall, the reconstruction error of the OOD sample is much higher than the one obtained on the ID image (see Figure IV.3.10), allowing for its detection.

The performance of the MNAD model on the OOD benchmark is presented in Figures IV.3.8 and IV.3.9, along with the performance of the DE. In terms of OOD detection (AUROC and AUPR scores, Figure IV.3.8 and IV.3.9), the MNAD model outperforms the DE on 8 out 10 **Transformation Shifts**, 2 out of 3 **Modality Shifts**, 3 out of 4 **Diagnosis Shifts**. Detection of **Far OOD** is perfect, as for the DE. However the performance on **Population Shifts** is still disappointing, with a performance close to the one of a Random Classifier. Finally, it can be noticed a performance above chance is obtained on the Control dataset, which is an undesired property of the OOD score. It indicates that Control samples are associated with higher reconstruction errors than Test ID samples, although they present similar properties (T1w brain MRI of adults with glioblastoma). This may indicate that the reconstruction-based OOD score is too sensitive, which would result in FP detection at test time. Averaging the performances on the 24 test datasets yields to an average AUROC of 0.78 and an average AUPR of 0.76, which although being a net increase compared to Deep Ensemble, could still be improved. Figure IV.3.10 presents a polar visualization of reconstruction errors for ID and OOD data points, on each tested setting. The distance of the point to the center (0, 0) indicates the reconstruction error. For extreme OOD settings (e.g. Lumbar, Abdominal), the reconstruction error allows to perfectly separate ID (blue)

[1] https://github.com/cvlab-yonsei/MNAD

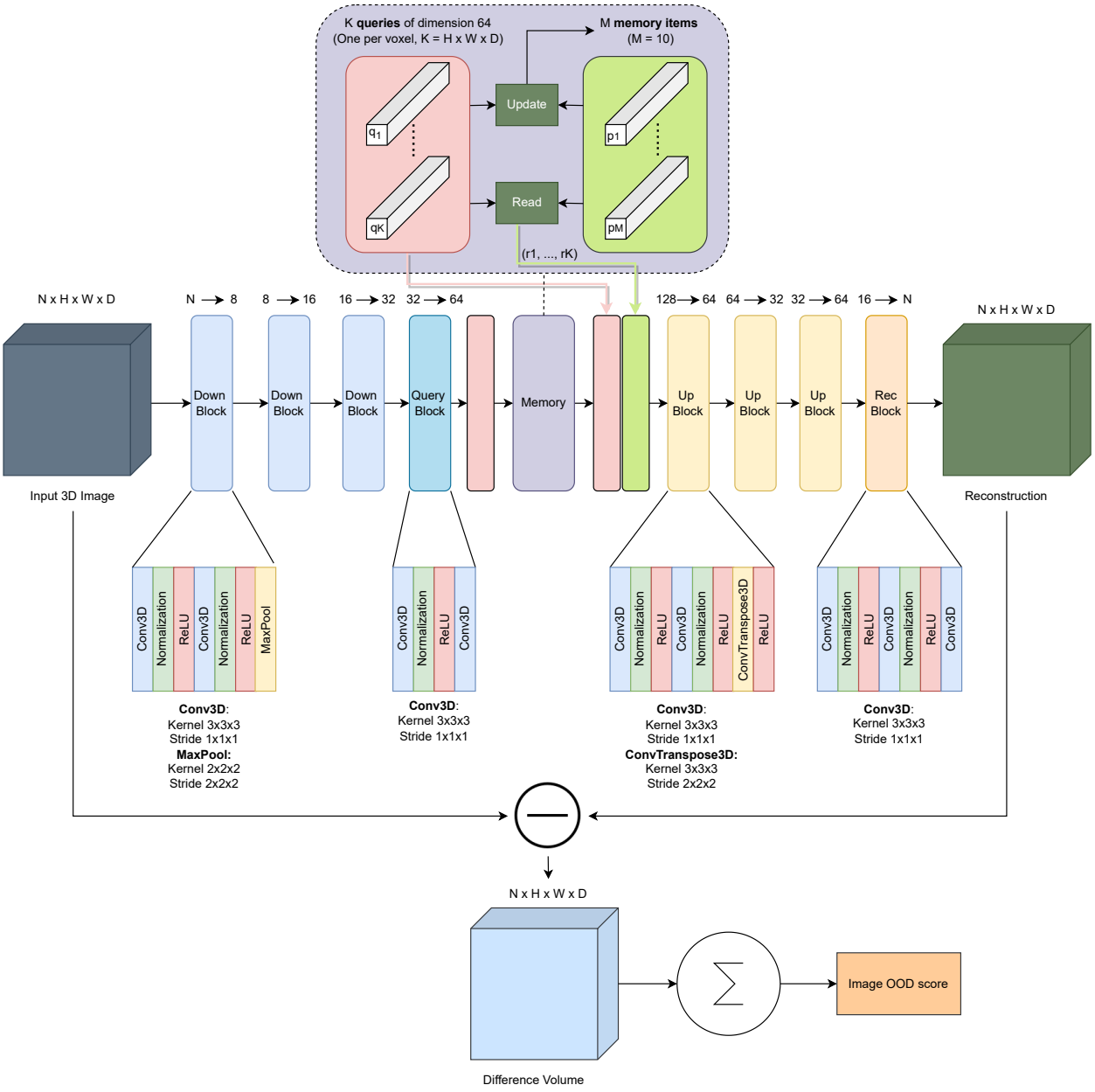


Figure IV.3.5: Architecture of the 3D MNAD model for Unsupervised OOD Detection. An encoder part (blue) maps the input volume into a set of queries, combined with memory items to produce the input to the decoder. These memory items correspond to learned prototypes. The decoder (yellow blocks) then produces a reconstruction of the input.

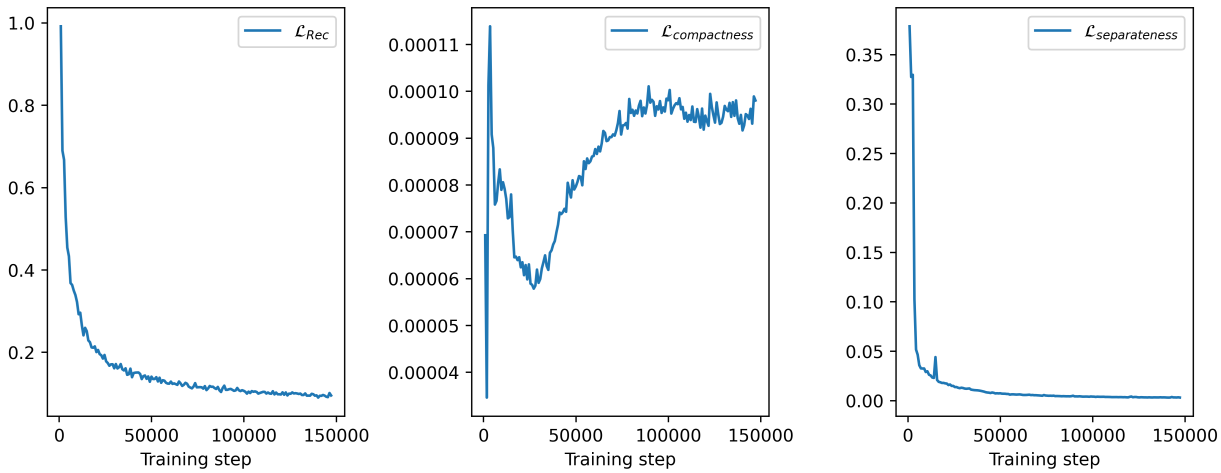


Figure IV.3.6: Training loss functions for the MNAD model trained on T1w data of glioblastoma subjects.

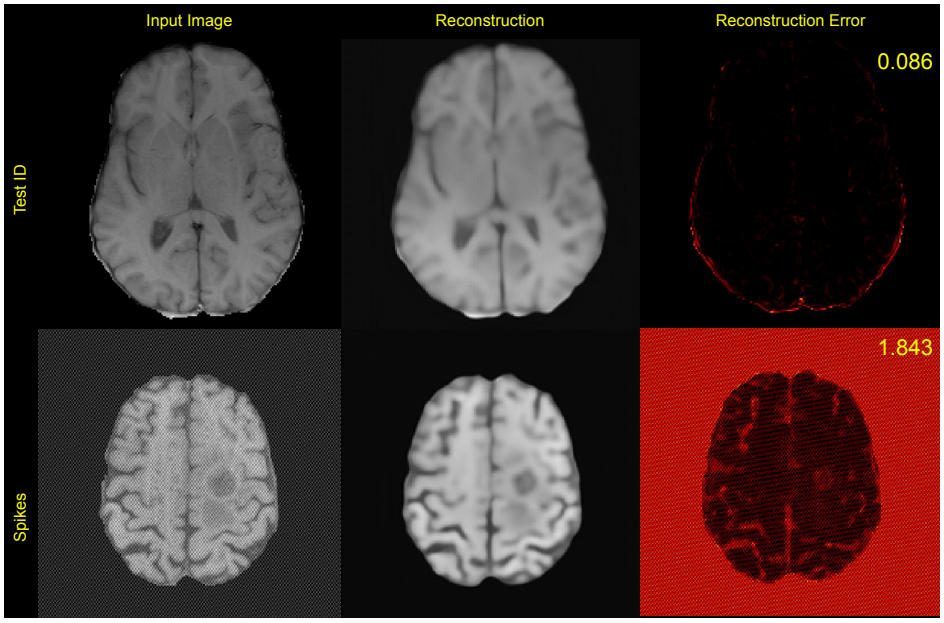


Figure IV.3.7: Reconstruction examples of the 3D MNAD model. The top row displays an ID test image, for which the reconstruction is accurate. As a result, the reconstruction error is low. In contrast, the OOD sample (bottom row) is poorly reconstructed as the spike artifact is not present in the training dataset. As a result, the reconstruction error for this image is high.

and OOD (red) data points. However, this is not the case for more subtle anomalies (e.g. WMH, Epileptic, Meningioma).

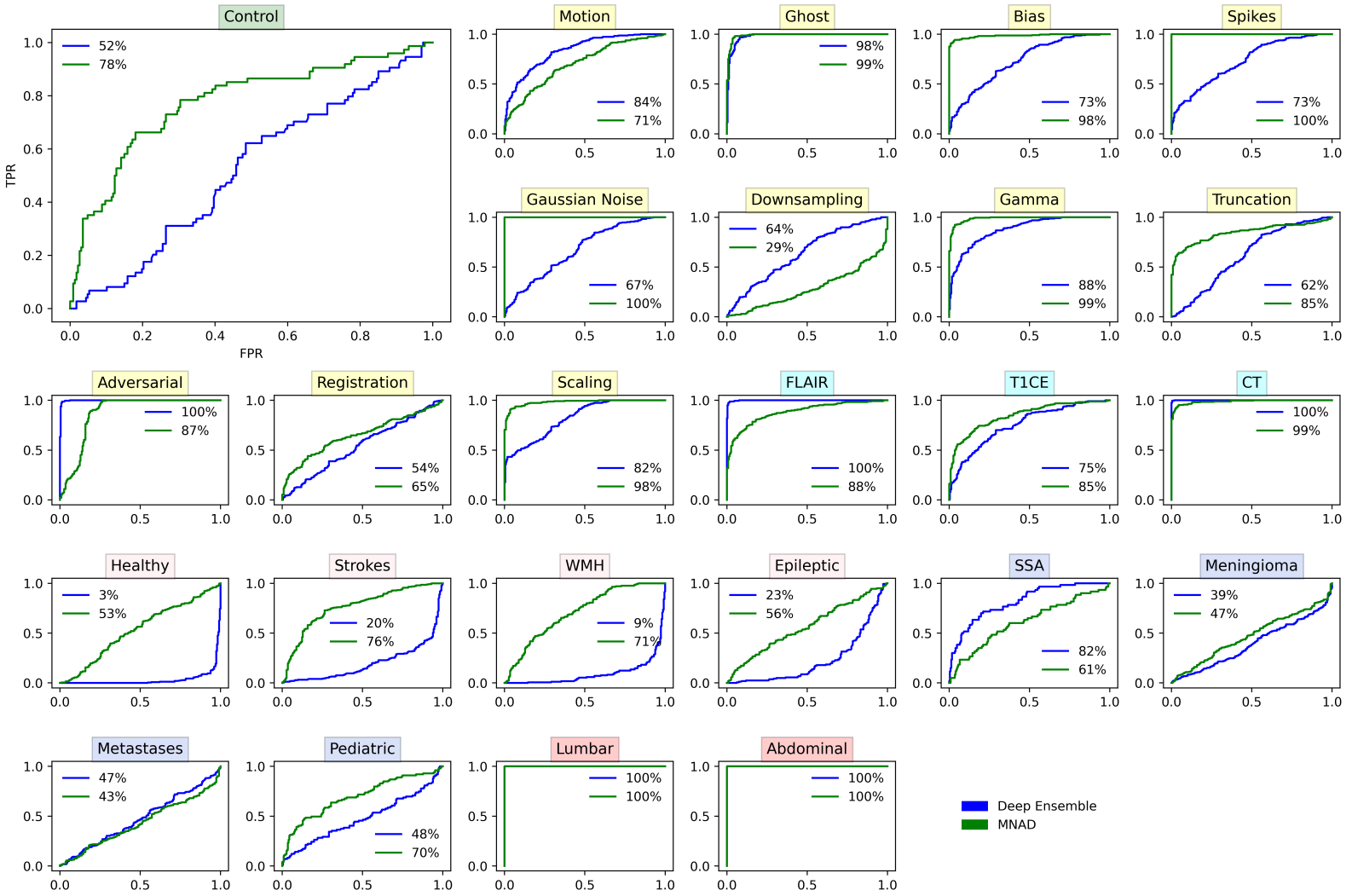


Figure IV.3.8: Receiver Operating Characteristic curves for the MNAD model (green) on the OOD benchmark, along with the performance of the Deep Ensemble (blue).

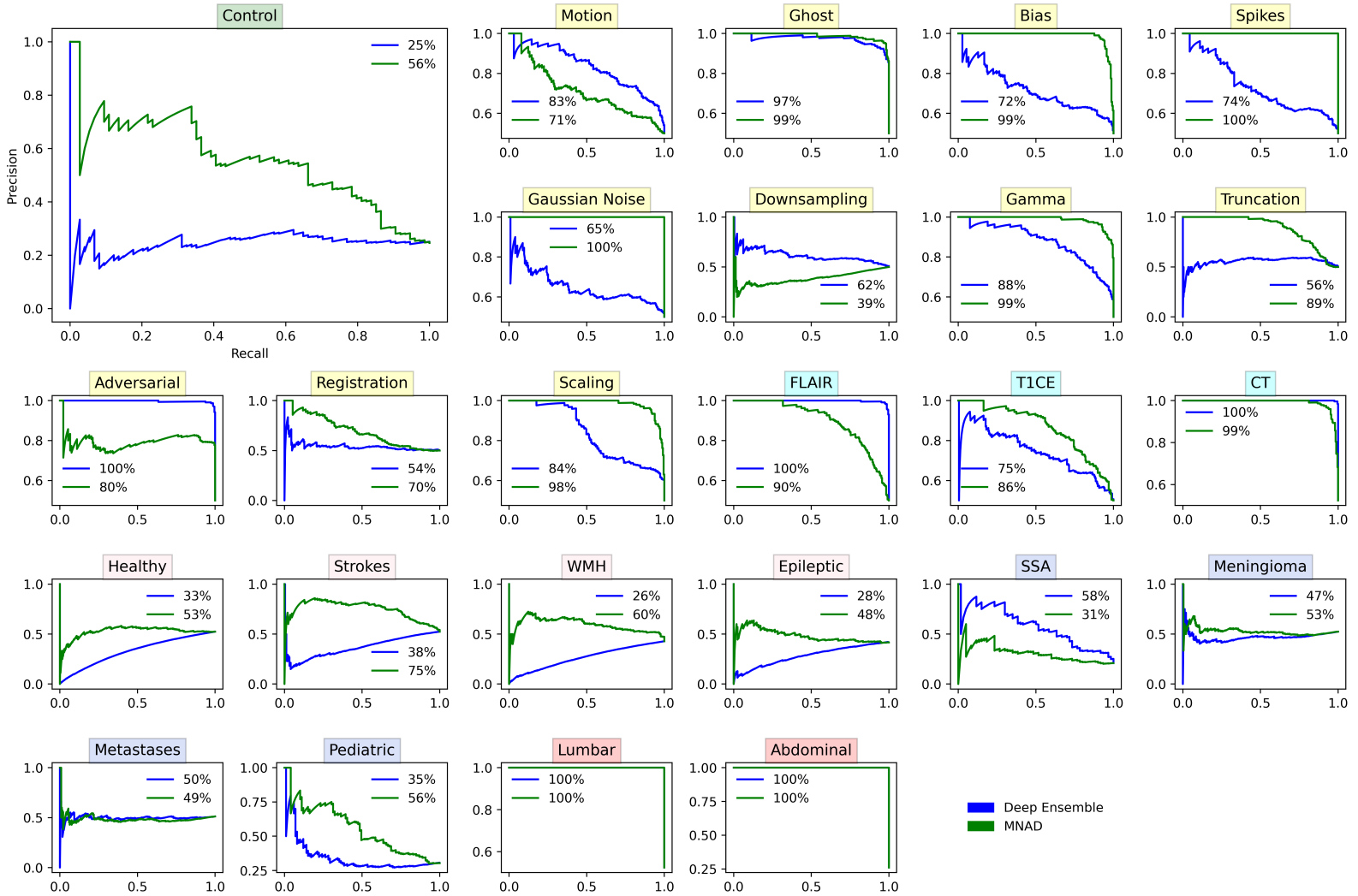


Figure IV.3.9: Precision-recall curves for the Deep Ensemble (green) on the OOD benchmark, along with the performance of the Deep Ensemble (blue).

Optimal OOD detection performance could not be achieved with either uncertainty-based or reconstruction-based approaches. This motivates the exploration of a novel efficient OOD approach that has extensively gained attention lately, in and outside the domain of medical image processing: latent-space OOD detection. This framework is introduced in the next section.

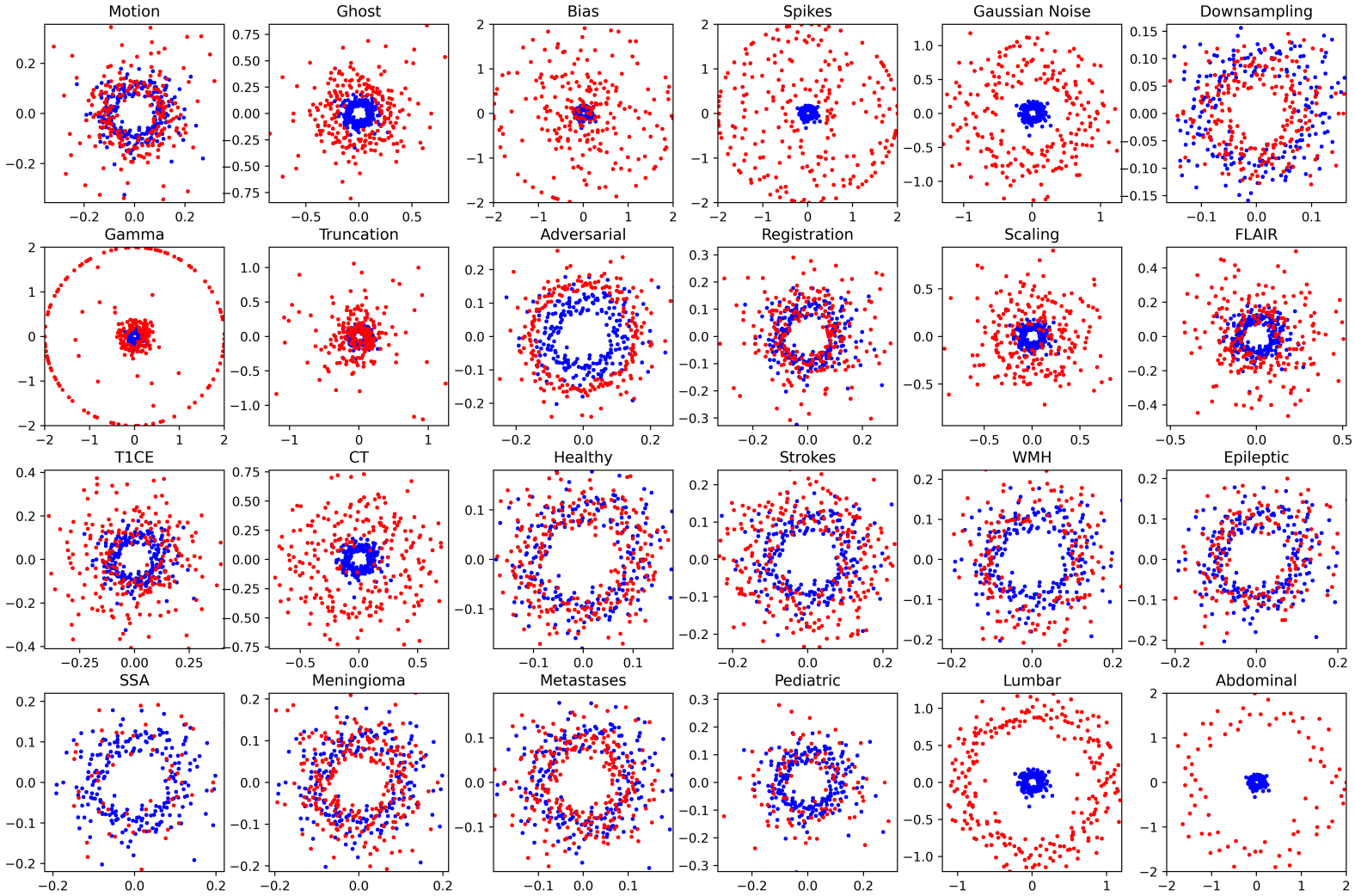


Figure IV.3.10: Reconstruction errors for in and out-of-distribution samples, for each tested setting. Blue points represent the Test ID samples, while red indicates the OOD images. The distance to the center (0, 0) indicates the reconstruction error. For ease of visualization, the reconstruction errors are clipped to a range of $[-2, 2]$.

IV.3.6 Latent-space OOD detection

When a neural network processes an image, it generates lower-dimensional representations of the input due to the consecutive downsampling operations applied in the convolutional layers. In principle, these representations should capture the essential features of the input data. Two samples sharing similar latent representations should thus be equivalent in the original image space. Latent-space OOD detection builds on this principle. The main hypothesis of latent-space OOD detectors is that ID and OOD samples are more easily distinguishable from the latent-space of a trained model, rather than in the input image domain (reconstruction-based model) or output model's uncertainty (uncertainty-based score).

Latent-space OOD detection is appealing on paper, as it allows the detection of non-conform inputs using the trained model only. In contrast to reconstruction-based approaches, no additional model dedicated to OOD detection is needed. Moreover, it can be implemented on a single model, alleviating the need to train several models as for the Deep Ensemble approach.

All latent-based OOD detectors rely on the same principle. First, they are plugged on top of a trained DL model, here the segmentation model that performs the delineation of the whole tumor from brain T1w MRI. To operate, the latent-based detectors require access to a dataset of ID samples, which is generally taken as the training datapoints [151, 153, 154]. By making inferences on these ID data points with trained models, a set of feature maps $F_i \in \mathbb{R}^{N_i \times H_i \times W_i \times D_i}$ is collected for one specific convolution layer i (single-layer methods) or all convolution layers (multi-layer methods). Here, N_i corresponds to the number of convolutional filters in the i-th layer, and $H_i \times W_i \times D_i$ to the spatial dimensions of the feature map. These features can be seen as embeddings of the training images in the latent space of the trained model. Second, at inference time, a metric is computed to estimate the distance between the test features and the train features to detect OOD samples. Recently, this approach to OOD detection has gained a lot of interest, and various ways of computing this feature-based distance have been proposed. Section II.2.11 presented an overview of these methods. In this chapter, the focus is on the Mahalanobis Distance which is becoming the preferred latent-space framework for OOD detection in medical image processing [72, 301, 151, 153, 150].

IV.3.7 The Mahalanobis distance

IV.3.7.1 Mathematical definition

The Mahalanobis Distance (MD) is a popular distance metric in ML allowing to compute the distance between a distribution and a test point. MD essentially estimates the distance of the test datapoint x_{test} to the center of a training distribution given its mean $\mu \in \mathbb{R}^M$ and covariance matrix $\Sigma \in \mathbb{R}^{M \times M}$, estimated using D training data points $(x_1, ..., x_D)$:

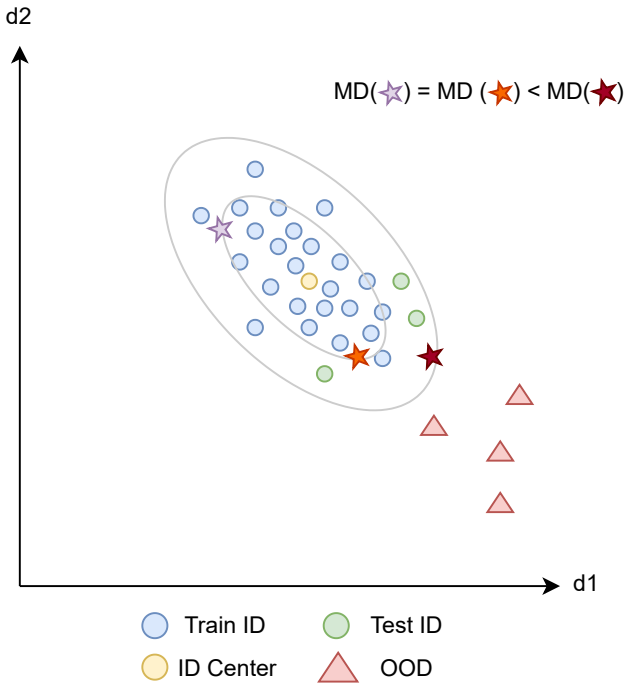


Figure IV.3.11: Illustration of the Mahalanobis distance (MD) in a two-dimensional setting. Points on the same ellipse share identical MDs.

$$\mu = \frac{1}{D}\sum_{i=1}^{D} x_i \tag{IV.3.9}$$

$$\Sigma = \frac{1}{D}\sum_{i=1}^{D}(x_i - \mu)(x_i - \mu)^T \tag{IV.3.10}$$

$$\mathrm{MD}(x_{test}; \mu, \Sigma) = (x_{test} - \mu)^T \Sigma^{-1}(x_{test} - \mu) \tag{IV.3.11}$$

In a simple one-dimensional setting, the MD simplifies to $MD = \frac{x_{test}-\mu}{\sigma}$, which is equivalent to the number of standard deviations σ the test sample x_{test} is away from the mean μ. With two dimensions, MD can be visualized in a plot (Figure IV.3.11) as the distance of the test point to the center of the training dataset. Points can be placed onto an ellipse whose main directions are determined by the train samples. Points on the same ellipse share the same Mahalanobis distances [302].

IV.3.7.2 Mahalanobis distance on latent representations

In our setting case, we seek to compute the MD on latent representations generated by the segmentation model. Thus, the training distribution corresponds to the distribution of the latent representations of training images, and the test point is the latent representation of the test image. For 3D CNNs, the feature maps are 4D tensors (3 spatial dimensions plus a

dimension equal to the number of convolution filters in the layer). Computing the MD from these high-dimensional matrixes is cumbersome, as the estimation of the inverse covariance matrix becomes intractable. Thus, a dimensionality reduction step is generally carried out. For instance, Gonzalez et al. [72] used consecutive average pooling to reduce the size of the feature map, until the number of elements falls below a defined threshold, set to 10^4 elements. More drastic reductions can be applied. For instance, Woodland et al. [153] used principal component analysis to further reduce the size of the feature maps. Finally, Calli et al. [150] and Anthony et al. [151] propose to perform a spatial averaging of the feature map. Given a 4D feature map of shape $F_i \in \mathbb{R}^{N_i \times H_i \times W_i \times D_i}$, the compressed latent representation $z_i \in \mathcal{R}^{N_i}$ is obtained as:

$$z_i = \frac{1}{H_i}\frac{1}{W_i}\frac{1}{D_i}\sum_{h=1}^{H_i}\sum_{w=1}^{S_i}\sum_{d=1}^{D_i} F_i(h,w,d) \tag{IV.3.12}$$

IV.3.7.3 Layer selection

One crucial challenge when computing the MD is the choice of the convolution layer to gather the latent representations. Generally, a single layer is selected to perform the OOD detection, selected for its sensitivity to non-conform inputs. Gonzalez et al. [301, 72] and Woodland et al. [153] used the feature map from the bottleneck layer of the U-Net, which is generally the layer with the higher number of convolutional filters in encoder-decoder architectures. Other studies focusing on OOD detection in the latent space rather used the penultimate convolution layer [33, 68]. There is no clear consensus, and several studies have thus focused on the impact of layer selection for OOD detection in 2D image processing. For natural 2D image classification, Wang et al. [303] argue that the first layers of a DL architecture essentially focus on low-level features such as texture and shape, while the ultimate layers extract more complex features. Consequently, they show that latent-based OOD detectors plugged into early layers in the net are efficient at detecting color or texture-based OOD samples, while more complex OOD samples are best detected for the last hidden layers. Similarly, Anthony et al. [151] showed that there is not one single best layer to perform latent-based OOD detection, but rather the optimal layer depends on the type of abnormality. Their study focuses on 2D medical image classification. Drawing similar conclusions, Calli et al. [150] showed that fitting one detector by layer and computing the final OOD score as the average of the individual layer's scores was more robust. The authors observed that the MD scales with the number of features in the layer. Thus, to avoid deep layers dominating the averaged score, the layer scores are first divided by the number of features in the layer before being averaged.:

$$\text{OOD}_{\text{multi}} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_l}\text{MD}_l \tag{IV.3.13}$$

where N_l and MD_l are the number of features in the layer L and the corresponding MD, respectively. However, these 3 studies are limited to 2D image classification, for which the

neural architectures are sensitively different from the ones used for 3D segmentation. It is thus unclear how that translates to our 3D medical image segmentation setting. Moreover, previous studies on latent-space OOD detection usually use a single neural network in the experiment [151, 153, 154]. However, it seems that latent-space OOD detection performance may be dependent on architectural choices within the model.

Several questions thus remain open. First, can latent-space OOD detection based on the MD outperform the uncertainty and reconstruction baselines? If so, what is the optimal layer to perform the detection? Can any gain be obtained by aggregating the scores of multiple layers, as compared to the single-layer baseline? Is the choice of the segmentation model architecture determining OOD detection performance?

IV.3.8 Multi-layer aggregation of Mahalanobis distances

Intuitively, it seems that each convolution layer extracts different sets of features, and thus their relevance regarding OOD detection can be heterogeneous. In the Dynamic U-Net used throughout this thesis, there are 9 blocks: one input block, followed by 3 downsampling blocks, a bottleneck block, and 4 upsampling blocks (see Figure II.6.2 for the overall architecture). Two convolution layers are used in each block, resulting in a total of 18 convolutions, excluding the final one responsible for producing the probabilities. There are thus two ways of computing the MD. The first approach would consist of the selection of one of these 18 convolutions to perform the OOD detection. The second approach would be to fit the parameters required for MD computation (mean and covariance) for each layer separately. Then, at test time, one MD is computed by layer and a final aggregation step is responsible for computing the overall image conformity score. Two aggregation techniques have been investigated in the literature, namely the mean or the max. The mean approach is investigated by Calli et al. [150], which computes one MD score per layer before computing the average. It essentially supposes that each layer contributes equally to the overall image score. However, some layers may be not provide interesting information for OOD detection, for each type of OOD setting.

Alternatively, the max aggregation was investigated by Wang et al. [303] for 2D natural image classification. Their work is not based on the MD, but rather they fit a one-class SVM (OCSVM) model for each layer. Briefly, OCSVMs are outlier detection models that can be fit using only samples from the normal class, here the features of the ID points. At test time, they thus obtain one abnormality score per layer in their classification network. Finally, the overall score is taken as the max of the layer's scores. It is based on the intuition that there is **one optimal** layer for each OOD setting. It keeps the highest non-conformity score among all the layers and drops out the contribution of the others.

To evaluate how these different techniques perform on our proposed 3D MRI segmentation benchmark, the MD is computed in different ways. The first one is the standard single-layer setting, where the MD is computed for each of the convolution layers, independently. It will allow to determine if there exists an optimal layer to perform OOD detection. Then, as illustrated in Figure IV.3.12, the **Mean** and **Max** of the layer's scores are also computed. To do so, the layer scores are first scaled by the number of features in the layer before computing the mean or the max. This will allow us to determine if multi-layer aggregation can improve

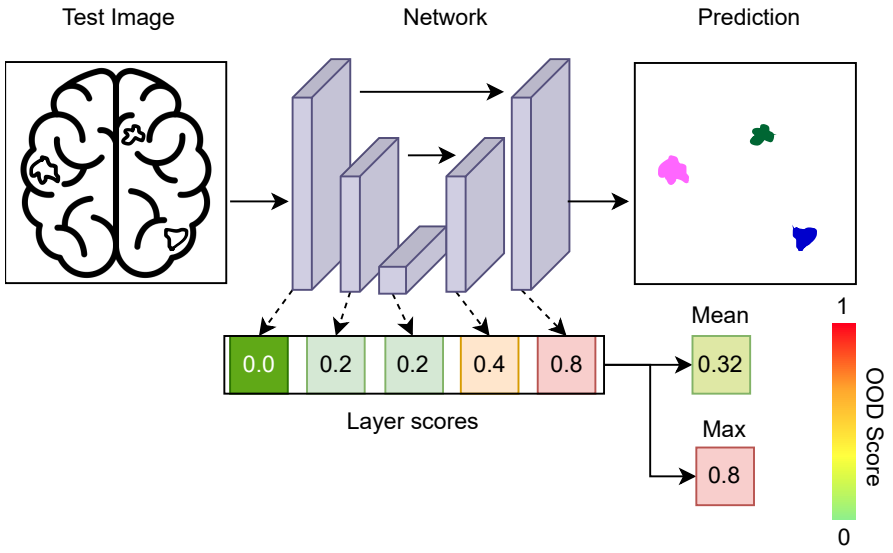


Figure IV.3.12: Illustration of mean and max aggregation for multi-layer OOD detectors, yielding to different grading of the test case.

OOD detection.

Finally, one last important design choice, generally underlooked in latent-space OOD detection studies, is the choice of the neural network architecture. To determine the incidence of this choice, the experiment is replicated for 4 popular medical image segmentation backbones: the Dynamic U-Net, the Attention U-Net [16], the V-Net [15] and the Residual U-Net. These architectures notably differ by the number of convolution layers (18 for Dynamic U-Net, 14 for Attention U-Net, 19 for V-Net, 23 for Residual UNet) as well as the number of parameters (16.5 million for Dynamic U-Net, 16.7 million for Attention U-Net, 45.6 million for V-Net and 18 million for Residual U-Net). For each architecture, 5 models are individually trained, with the same hyper-parameters, to compose the ensemble. To fit the MD detectors, inferences are carried out on the training images once training ends, to gather the latent representations of ID data points.

IV.3.9 Aggregated Mahalanobis distances for Deep Ensembles

The experimental setting used throughout this thesis relies on the use of Deep Ensemble for both its positive impact on predictive performance, as well as the state-of-the-art voxel uncertainties it can provide. However, all the presented latent-based detectors have been initially proposed for single models. A straightforward approach is used here, following which an OOD score is computed for each model in the ensemble. Then, the final ensemble's score is taken as the average of the model's scores.

IV.3.10 Results

Tables IV.1 and IV.2 present the performance (AUROC and AUPR) of each MD detector, on each OOD setting. The same tables for the Attention U-Net ensemble, the V-Net ensemble and the Residual U-Net are provided in Appendix A4. Figure IV.3.13 shows the OOD detection metrics (AUROC and AUPR) averaged over the 24 test datasets depending on the convolution layer selected for the MD computation, for each ensemble (Dynamic U-Net, Attention U-Net, V-Net, and Residual U-Net).

First, the latent-space OOD detection based on the MD offers excellent overall OOD detection performance. The highest quality is obtained for the MD computed at the 18th convolution of the Dynamic U-Net model, with an average AUROC of 0.91 and an average AUPR of 86. It represents an impressive gain over the uncertainty and reconstruction baselines. Overall, MD detectors provide excellent detection capabilities on **Transformation shifts**, **Modality**, and **Far** OODs. For these settings, the OOD is global, meaning that the entire image is OOD. It can thus be expected that latent representations of these OOD samples are significantly different from the ones of ID samples. The performance on **Population** and **Diagnosis** shifts is largely improved as compared to the Deep Ensemble and MNAD, although not reaching the same level of performance as the other types of OOD. For **Diagnosis** and **Population**, the OOD is actually more subtle and does not cover the entire scan. More precisely, the OOD area is restricted to a limited region of the input MRI (generally, the lesion area). The rest of the brain conforms to what has been observed during training. Thus, it can be expected that the latent representations of these images are closer to the ones of ID images, making them less easily detectable. This can also be a side-effect of the dimensionality reduction that we operate on the feature maps before calculating the MD, which may discard subtle anomalies in intermediate activations.

Second, Figure IV.3.13 highlights the importance of the layer choice on the robustness of the OOD detection. Indeed, the performance is heterogeneous for each backbone, based on the convolutional layer used to compute the MD. For each backbone, the MD computed at the first layer provides poor results, which can be explained by the fact that the first convolution extracts very generic features that may not be useful for outlier detection. Bottleneck layers (in yellow) do not exhibit the highest OOD detection quality, although they are popular choices for latent representation extraction [301, 153]. The highest performance is obtained at the penultimate convolutional layer in the decoder (18-th layer) for the Dynamic U-Net. This last layer is particularly efficient in detecting Population and Diagnosis shifts. However, for the other architectures, the optimal detection is not reached at the penultimate layer. For Attention U-Net, the top performance is reached by the fifth convolution of the encoder, while for V-Net it's the 16th convolution, located in the decoder. This highlights the fact that layer selection is crucial to the final OOD detection quality, and the optimal layer is not consistent across segmentation backbones.

To prevent having to determine the optimal layer for each model, the multi-layer aggregations (Mean and Max) are promising. Indeed, it can be noticed that these approaches obtain high AUROC and AUPR scores for each backbone, although they are outperformed by some single-layer scores. There is no clear advantage for one or the other, as the Max

aggregation outperforms the Mean for the Dynamic U-Net, Residual U-Net, and Attention U-Net ensembles, while the opposite is observed for the V-Net ensemble. To conclude, multi-layer aggregation can be used to alleviate the cumbersome layer selection problem as it performs consistently well and outperforms most of the single-layer scores.

Regarding the backbones, our experiments indicate that architecture choices have an impact on latent-based OOD detection performance. More precisely, the highest OOD detection performance is obtained by the Dynamic U-Net, while lower AUROC and AUPR values are obtained for the V-Net, Attention U-Net, and Residual U-Net. Moreover, the Dynamic U-Net exhibits a slightly superior segmentation performance as compared to the other backbones (Tables A.4.1, A.4.2, and A.4.3). This may indicate a link between the performance on the downstream task (here, brain tumor segmentation) and the performance of outlier detection.

Lastly, we analyze in more detail the OOD scores computed by the MD Max aggregation approach for the Dynamic U-Net ensemble. It achieves top-quality performance on the OOD benchmark, with an AUROC above 90 for 14 out of the 24 OOD datasets. Figure IV.3.16 presents a visual representation of the OOD scores computed with this approach. In the figure, the distance to the center (0, 0) is representative of the Mahalanobis Distance, with higher values indicating higher degrees of non-conformity. In most settings, the OOD samples (red) are clearly separated from the ID samples (blue) based on their MD scores, which is particularly visible for the **Spikes**, **Bias**, or **Gaussian Noise** datasets. Finally, we also present the ROC and PR curves associated with this approach in Figures IV.3.14 and IV.3.15, along with the performance of the Deep Ensemble baseline, which exhibits the drastic improvement achieved with the proposed latent-space detector.

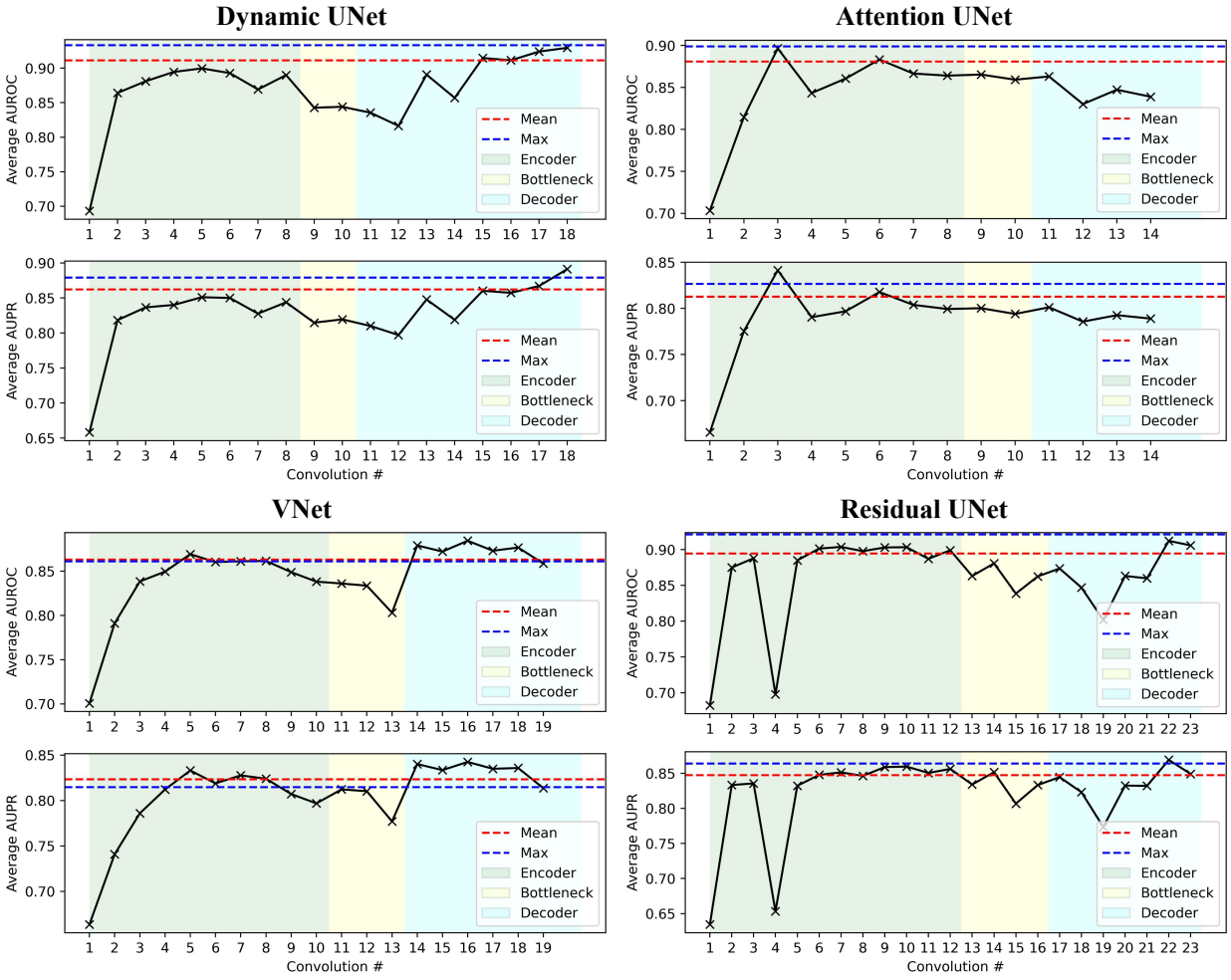


Figure IV.3.13: Average OOD detection performance for Mahalanobis Distance detectors depending on the selected convolutional layer, for each architecture. The performance of the multi-layer aggregation (Mean and Max) is indicated with dashed lines.

Type	ID	Transformation											Modality			Diagnosis				Population				Far		Avg
Dataset	Control	Motion	Ghost	Bias	Spikes	Noise	Downsample	Gamma	Truncation	FGSM	Registration	Scaling	FLAIR	T1CE	CT	Healthy	Strokes	WMH	Epilepsy	SSA	Meningioma	Metastases	Pediatric	Abdominal	Lumbar	Average
Deep Ensemble	52	84	98	73	73	67	64	88	62	**100**	54	82	**100**	75	**100**	3	20	9	23	82	39	47	48	**100**	**100**	66
MNAD	78	71	99	98	**100**	**100**	29	99	85	87	65	98	88	85	99	53	76	71	56	61	47	43	70	**100**	**100**	78
MD Conv 1	32	52	71	98	88	76	44	97	99	54	66	**100**	69	64	60	33	53	76	45	58	60	40	60	**100**	**100**	68
MD Conv 2	49	94	96	**100**	**100**	**100**	**94**	97	99	99	76	99	94	93	**100**	45	58	85	59	85	63	74	63	**100**	**100**	85
MD Conv 3	49	93	99	**100**	**100**	**100**	91	99	**100**	96	75	**100**	93	95	99	55	71	94	58	88	64	77	68	**100**	**100**	87
MD Conv 4	54	**96**	**100**	**100**	**100**	**100**	91	**100**	**100**	98	**83**	**100**	95	96	99	50	64	86	80	91	68	78	73	**100**	**100**	88
MD Conv 5	60	**96**	**100**	**100**	**100**	**100**	93	**100**	**100**	99	80	**100**	97	**98**	98	48	65	85	84	**93**	67	78	79	**100**	**100**	89
MD Conv 6	56	**96**	**100**	**100**	**100**	**100**	91	**100**	**100**	99	80	**100**	99	**98**	99	40	65	86	76	**93**	66	76	82	**100**	**100**	88
MD Conv 7	62	**96**	**100**	**100**	**100**	**100**	86	100	**100**	98	74	**100**	99	96	99	37	54	68	61	88	66	79	86	**100**	**100**	86
MD Conv 8	59	95	**100**	**100**	**100**	**100**	86	100	**100**	98	81	**100**	99	96	**100**	52	68	77	67	90	67	78	85	**100**	**100**	88
MD Conv 9	42	92	**100**	**100**	99	98	70	**100**	**100**	88	**83**	**100**	**100**	93	**100**	30	55	67	52	82	64	64	86	**100**	**100**	83
MD Conv 10	49	91	**100**	**100**	**100**	99	68	**100**	**100**	87	82	**100**	**100**	92	**100**	33	47	60	54	91	68	71	84	**100**	**100**	83
MD Conv 11	46	92	**100**	**100**	98	97	67	**100**	**100**	91	81	**100**	**100**	92	**100**	28	46	56	58	85	65	65	86	**100**	**100**	82
MD Conv 12	44	85	**100**	98	97	92	59	**100**	**100**	84	80	**100**	**100**	91	**100**	32	41	55	49	83	64	63	85	**100**	**100**	80
MD Conv 13	57	**96**	**100**	**100**	**100**	99	84	**100**	**100**	99	76	**100**	**100**	96	99	50	60	73	76	91	70	78	**92**	**100**	**100**	88
MD Conv 14	57	91	**100**	**100**	98	98	71	**100**	**100**	95	72	**100**	**100**	97	**100**	42	57	71	53	87	70	70	87	**100**	**100**	85
MD Conv 15	54	95	**100**	**100**	**100**	**100**	88	**100**	99	98	80	**100**	99	97	99	67	74	88	84	92	73	78	84	**100**	**100**	90
MD Conv 16	48	94	**100**	**100**	99	**100**	87	**100**	99	97	70	**100**	99	97	99	75	77	89	83	90	74	76	83	**100**	**100**	89
MD Conv 17	49	91	99	**100**	**100**	**100**	90	99	**100**	92	68	**100**	96	96	99	92	87	96	86	90	77	81	79	**100**	**100**	91
MD Conv 18	46	91	**100**	**100**	**100**	99	81	99	99	95	65	100	99	94	99	**96**	**91**	**97**	**93**	86	**79**	80	85	**100**	**100**	91
MD Mean	52	95	**100**	**100**	**100**	**100**	90	**100**	**100**	98	79	**100**	99	97	99	62	71	90	78	92	74	81	84	**100**	**100**	90
MD Max	51	94	**100**	**100**	**100**	**100**	87	**100**	**100**	99	77	**100**	98	97	99	88	82	94	85	91	**79**	**83**	87	**100**	**100**	**92**

Table IV.1: OOD detection performance (AUROC, expressed in percentage) for each OOD detector and dataset, for the Dynamic U-Net ensemble. The highest score for each dataset is indicated in **bold**. MD: Mahalanobis Distance.

Type	ID	Transformation											Modality			Diagnosis				Population				Far		Avg
Dataset	Control	Motion	Ghost	Bias	Spikes	Noise	Downsample	Gamma	Truncation	FGSM	Registration	Scaling	FLAIR	T1CE	CT	Healthy	Strokes	WMH	Epilepsy	SSA	Meningioma	Metastases	Pediatric	Abdominal	Lumbar	Average
Deep Ensemble	25	83	97	72	74	65	62	88	56	**100**	54	84	**100**	75	**100**	33	38	26	28	58	47	50	35	**100**	**100**	66
MNAD	56	71	99	99	**100**	**100**	39	99	89	80	70	98	90	86	99	53	75	60	48	31	53	49	56	**100**	**100**	76
MD Conv 1	18	50	65	93	89	72	45	97	98	53	66	99	64	59	61	41	52	58	38	27	59	46	47	**100**	**100**	64
MD Conv 2	24	88	91	**100**	**100**	**100**	86	97	98	98	72	98	92	93	99	48	59	76	51	53	59	71	38	**100**	**100**	80
MD Conv 3	23	87	98	99	**100**	**100**	84	98	**100**	94	72	**100**	89	93	97	51	71	89	51	57	62	71	44	**100**	**100**	81
MD Conv 4	26	92	98	99	**100**	**100**	81	99	**100**	96	78	**100**	93	95	98	49	61	71	62	61	63	71	50	**100**	**100**	82
MD Conv 5	30	94	98	99	**100**	99	85	**100**	99	98	77	**100**	94	96	97	46	63	70	66	69	65	71	57	**100**	**100**	83
MD Conv 6	29	93	98	99	**100**	99	81	99	99	98	77	**100**	97	96	98	44	64	73	60	69	64	69	62	**100**	**100**	83
MD Conv 7	33	93	98	99	**100**	98	76	**100**	99	97	72	**100**	97	95	98	43	55	54	45	62	64	76	67	**100**	**100**	81
MD Conv 8	30	92	98	99	**100**	99	76	**100**	**100**	96	79	**100**	99	94	98	50	62	59	49	69	65	73	69	**100**	**100**	82
MD Conv 9	21	91	**100**	**100**	99	98	64	**100**	**100**	83	82	**100**	**100**	92	**100**	40	55	53	41	58	64	62	74	**100**	**100**	79
MD Conv 10	25	92	**100**	**100**	**100**	99	62	100	**100**	81	84	**100**	**100**	92	**100**	41	49	46	41	74	68	69	72	**100**	**100**	80
MD Conv 11	23	92	**100**	**100**	99	96	61	**100**	**100**	87	81	**100**	**100**	91	**100**	39	50	46	43	63	64	60	75	**100**	**100**	79
MD Conv 12	21	87	**100**	98	98	91	57	**100**	**100**	82	81	**100**	**100**	91	**100**	40	46	44	38	63	64	59	74	**100**	**100**	77
MD Conv 13	29	94	98	99	99	98	74	**100**	99	97	74	**100**	99	95	98	49	59	58	59	66	68	73	79	**100**	**100**	83
MD Conv 14	27	89	98	**100**	98	96	63	99	**100**	92	70	**100**	99	96	98	45	57	56	43	62	66	65	73	**100**	**100**	80
MD Conv 15	27	92	98	99	**100**	99	79	**100**	99	95	76	**100**	97	96	98	59	68	74	67	69	67	70	64	**100**	**100**	84
MD Conv 16	24	92	98	99	99	98	79	99	98	95	68	99	98	96	98	65	73	76	67	63	69	69	63	**100**	**100**	83
MD Conv 17	24	84	96	99	**100**	**100**	81	98	**100**	86	66	**100**	94	94	98	81	81	88	72	60	73	74	55	**100**	**100**	84
MD Conv 18	22	90	99	**100**	**100**	99	73	98	99	92	66	**100**	98	92	98	89	88	92	85	56	78	77	71	**100**	**100**	86
MD Mean	25	92	98	99	**100**	**100**	81	99	**100**	97	76	**100**	97	96	98	55	68	79	62	67	69	73	65	**100**	**100**	84
MD Max	25	90	98	99	**100**	**100**	77	99	**100**	97	74	**100**	97	95	98	76	76	84	70	64	73	74	68	**100**	**100**	85

Table IV.2: OOD detection performance (AUPR, expressed in percentage) for each OOD detector and dataset, for the Dynamic U-Net ensemble. The highest score for each dataset is indicated in **bold**. MD: Mahalanobis Distance.

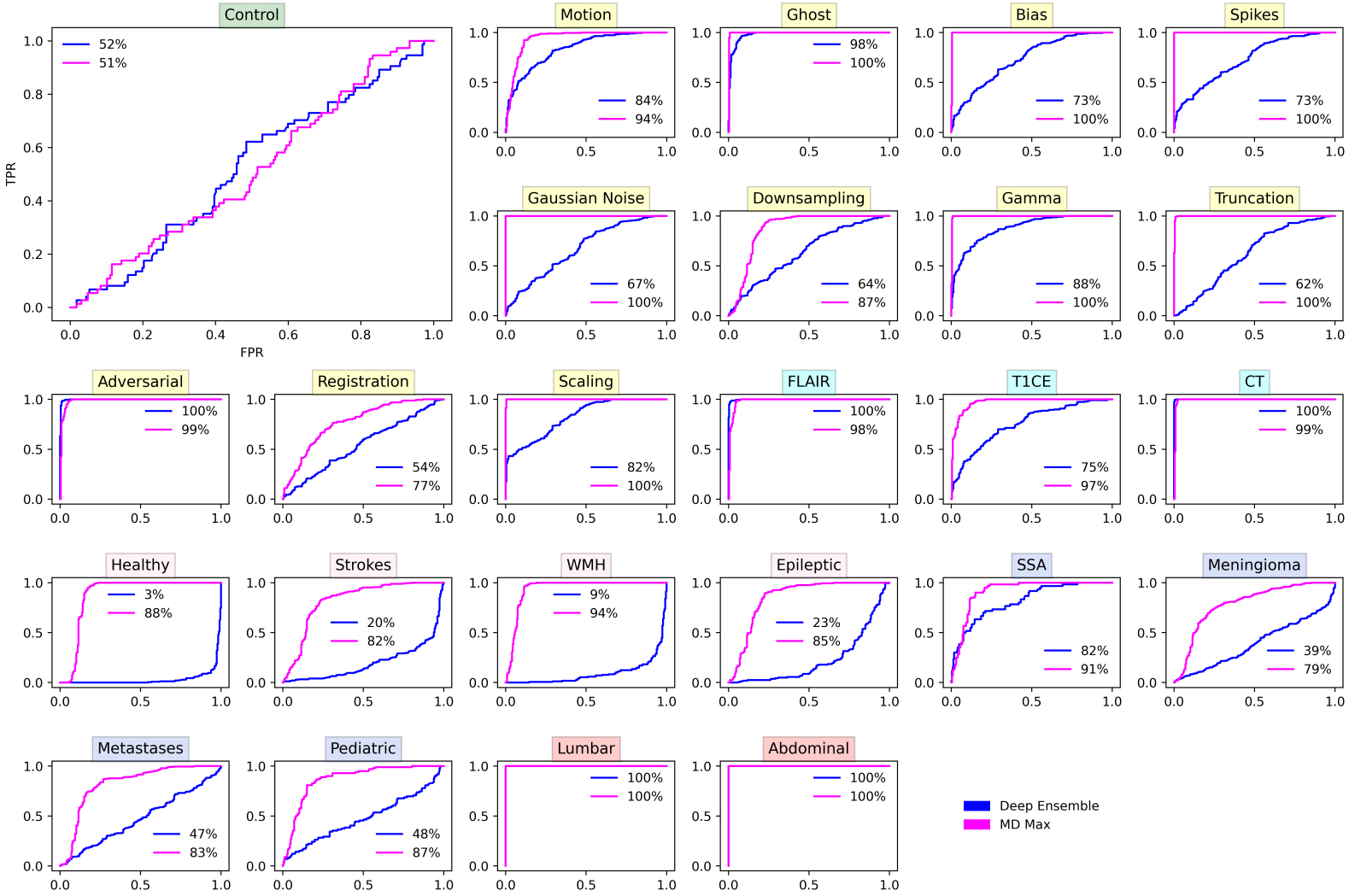


Figure IV.3.14: Receiver Operating Characteristic curves for the Mahalanobis Distance detector with Max aggregation (magenta) on the OOD benchmark, along with the performance of the Deep Ensemble (blue).

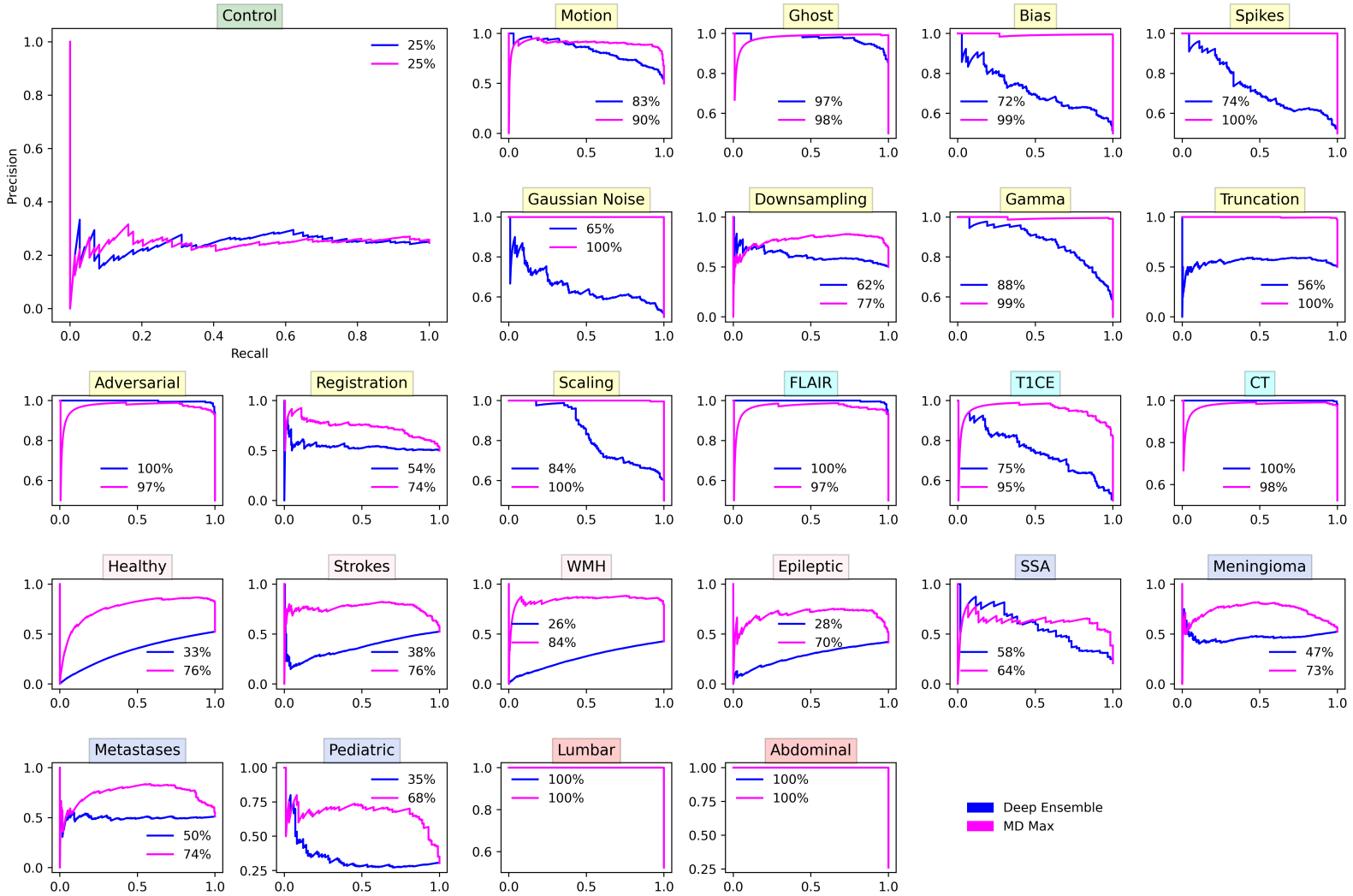


Figure IV.3.15: Precision-recall curves for the Mahalanobis Distance detector with Max aggregation (magenta) on the OOD benchmark, along with the performance of the Deep Ensemble (blue).

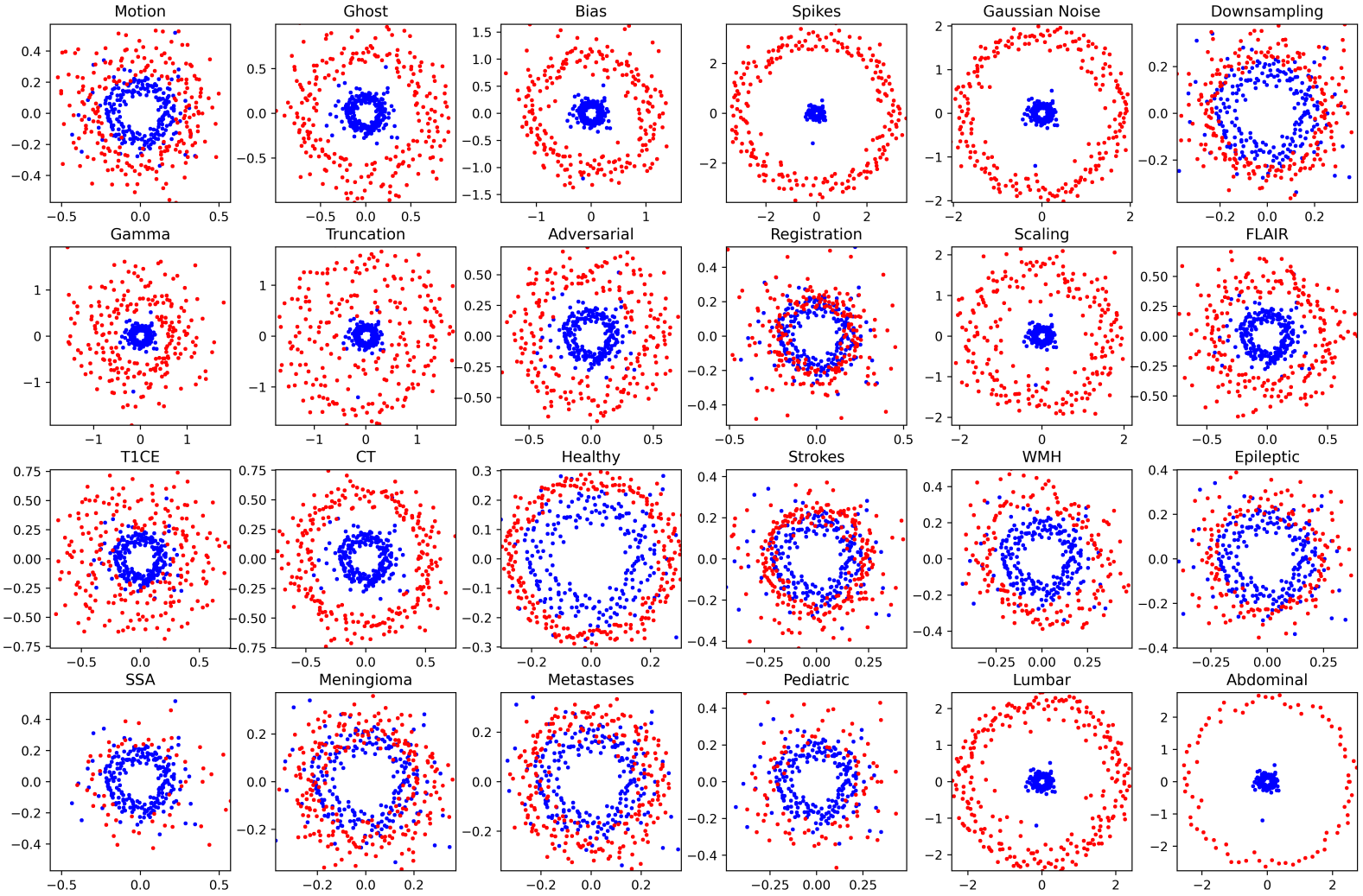


Figure IV.3.16: Mahalanobis distances for in and out-of-distribution samples, for each tested setting. Blue points represent the Test ID samples, while red indicates the OOD images. The distance to the center (0, 0) indicates the Mahalanobis distance.

IV.4 From out-of-distribution detection to quality control

In the previous section, OOD detection has been defined as the task of detecting input samples that are far from the training distribution, for which the functioning of the model is expected to be degraded. In this setting, an image is flagged as OOD if it presents a characteristic not represented in the training dataset (e.g. unseen pathology or MRI artifact). However, this definition does not take into account the performance of the model on the OOD data points. For example, in the previous section, the Deep Ensemble is able to provide satisfying segmentations on several OOD datasets, including **Downsample**, **Motion**, **SSA**, and even **T1ce** (see Figure IV.3.4). Yet these samples are classified as OOD with high accuracy. This is because the latent space methods do not explicitly consider the output of the model, but rather the representation of the input in the latent space. As an effect, an image can be flagged as OOD because of an unusual pattern, despite being segmented with high accuracy. To get a better view of this phenomenon, we investigate in Figure IV.4.1 the relationship between MD scores and segmentation adequacy (Dice scores) for each test image where the ground truth of brain tumors is available, representing 3825 MRI volumes. The scatter plot highlights a lack of clear correlation between both quantities.

It could be argued that if the input image includes an artifact that does not prevent the proper functioning of the DL model, it **should not** be labeled as an OOD sample to prevent false alarms. The concept of defining OOD images with respect to the performance of the downstream task (here, segmentation) was first proposed in Shaw et al. [43], where authors estimate the conformity of an MRI image with respect to the model's ability to provide the correct output. The same redefinition is explored in three recent studies focusing on OOD detection in medical image segmentation [304, 153, 305]. Instead of defining OOD inputs as images presenting artifacts or missing attributes, they cast OOD images as cases for which the associated segmentation is poor, thus allowing them to take into account the generalization power of the network. On the other hand, it will also be considered as OOD an ID image poorly segmented by the model. Moreover, this definition is task and model-dependent. For two models A and B, an image can be OOD for model A but not for model B, based on the chosen performance threshold.

Thus, it appears that there are two possible definitions of OOD samples, one considering the conformity of the input of the model using user-defined rules (e.g. image resolution, presence of artifacts), and one considering the associated output prediction provided by the model. So far, the presented methods have been introduced for input QC, and thus are not optimal for assessing the quality of the segmentation. More generally, QC methods proposed in the literature either focused on the conformity of the input image, or alternatively on the quality of the output prediction, and the connection between these two levels of QC is overlooked. In the second part of this chapter, we will investigate how QC decision-making can be enriched by considering simultaneously both input and output-level quality.

IV.4.1 Unified input-output QC for medical image segmentation

In this section, we aim to investigate the relationship between input and output QC in the context of medical image segmentation. The input score will be the multi-layer MD with

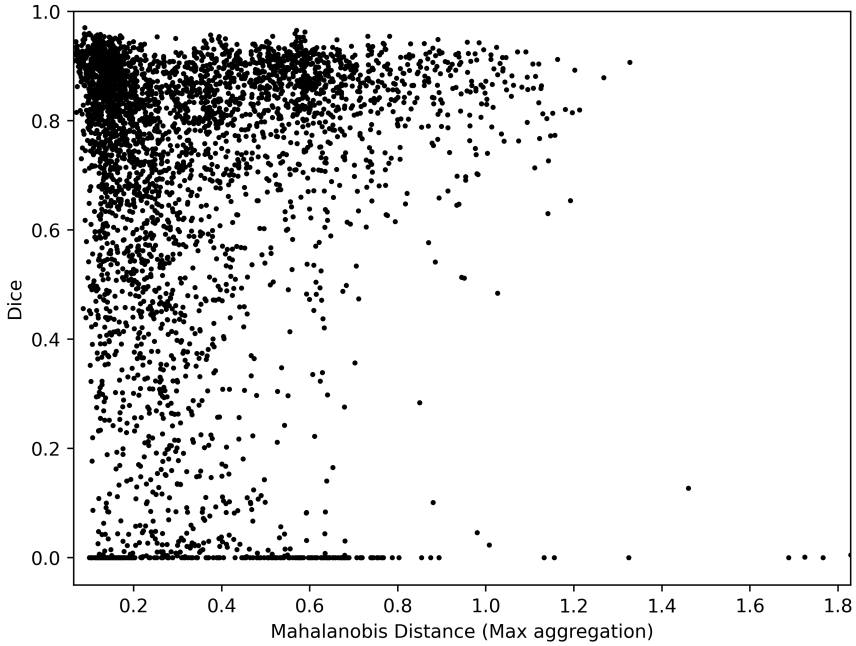


Figure IV.4.1: Scatter plot of Mahalanobis distances (with max multi-layer aggregation) with respect to Dice scores, for the Dynamic U-Net Ensemble. It exhibits a lack of correlation between the two metrics, with a fraction of images being well segmented (high Dice scores) while being attributed with high Mahalanobis distances.

max aggregation, which previously proved to be a performant input-level OOD detector. It remains to decide on an estimator to perform output QC. As presented in the literature review, the most preferred approach is to consider the variability among a set of plausible masks, for each input sample. These samples are typically generated using standard UQ methodologies, such as MC dropout, TTA, or DE. The selected pipeline is illustrated in Figure IV.4.2. First, the majority vote MV is obtained from the 5 segmentation masks generated by the Ensemble. Then, the Dice scores between each individual segmentation mask and the majority vote are computed. The final output QC score, called Ensemble Prediction Agreement (EPA) [306] corresponds to:

$$\text{EPA} = \frac{1}{K}\sum_{i=1}^{K} \text{Dice}(S_k, MV) \tag{IV.4.1}$$

The intuition is that a high-quality segmentation should be associated with a high level of agreement between the individual ensemble members, leading to a high EPA. In contrast, if the segmentation deviates significantly from one model to the other, the prediction is likely uncertain, and its overall quality should be rather low. In practice, to have a score that grows as the non-conformity increases, we use $1 - EPA$ as the output QC score.

Now that the DE is equipped with an input-level QC score (MD with max aggregation)

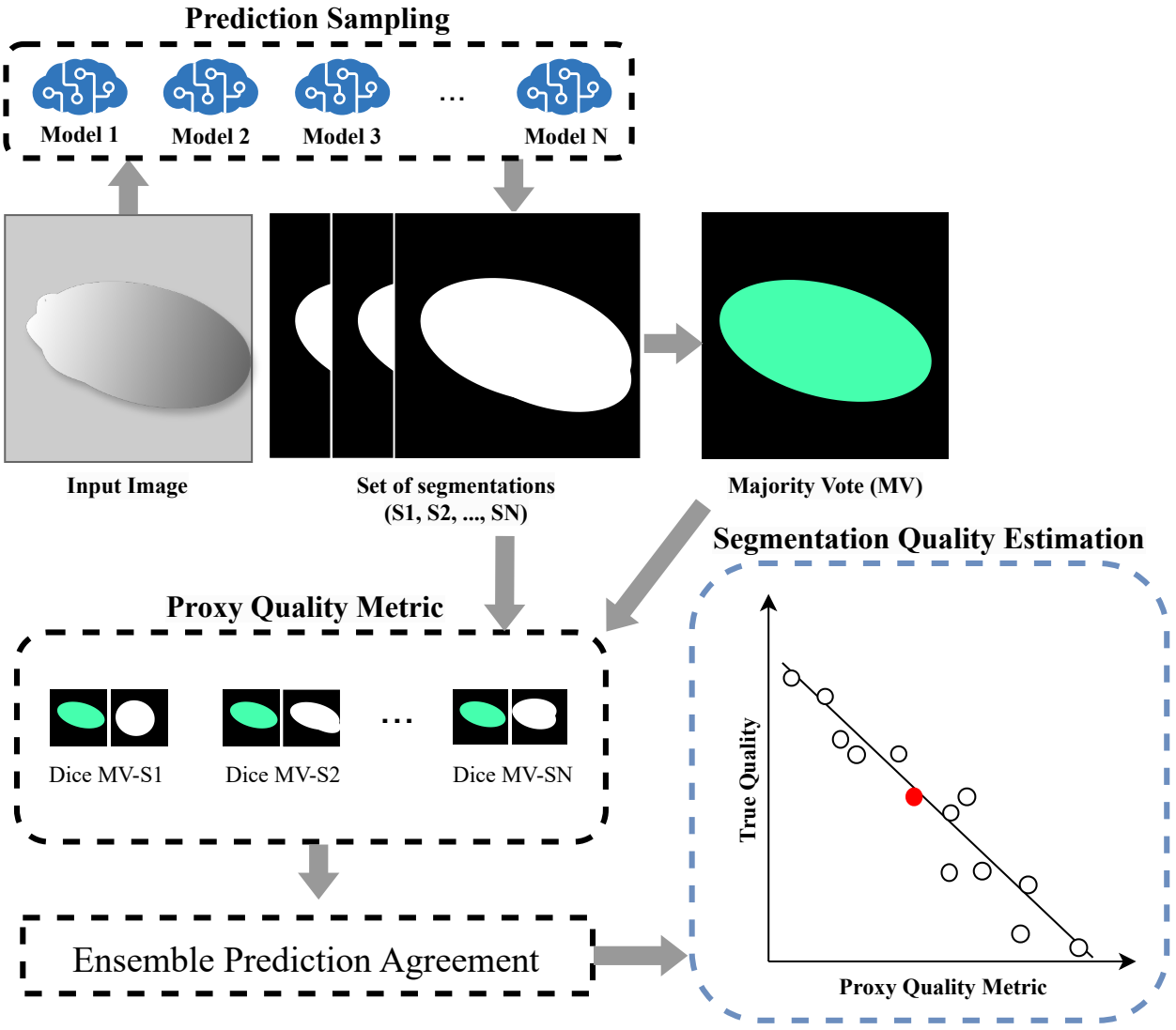


Figure IV.4.2: Illustration of the proxy output QC score derived from the Deep Ensemble.

and an output-level QC score (1 - EPA), each pair of test image and segmentation can be positioned in the QC prediction space. More specifically, 4 cases are possible using the proposed protocol (see Figure IV.4.3), listed below in increasing priority:

- **Region A - Input QC ✓and Output QC ✓:** optimum operating regime; corresponding to the ideal setting where the image is conform and the output prediction is estimated as performing. It is expected that this subgroup will contain the top-quality predictions of the model.
- **Region B - Input QC ✗ and Output QC ✓:** Robust operating regime; corresponding to images that may contain an anomaly (artifact), but for which the output QC is successful. This could represent images that the model is able to process even though their quality is not perfect.
- **Region C - Input QC ✓and Output QC ✗:** Dysfunctional regime, corresponding to images that have passed the input QC, but for which the disagreement in the ensemble is high (low EPA). This could represent images that are conform in terms of quality but are still poorly segmented.
- **Region D - Input QC ✗ and Output QC ✗:** Divergent regime, corresponding to the worst-case scenario where both input and output QC failed. This could represent out-of-distribution images for which the prediction is highly sub-optimal. This subgroup should be reviewed with top priority.

Building this confusion matrix requires setting two thresholds on the input and output quality scores, respectively. To determine them automatically, the validation dataset split can be used. More specifically, the scores (MD and 1-EPA) are computed for each validation image. The thresholds are then taken as the 95-th percentiles on the validation images. It signifies that if the MD is superior to the 95-th percentile of validation MDs, the input QC will be considered as failed. Similarly, if the test image has a (1-EPA) score above the 95-th percentile, the output QC score is considered as failed. Selecting the 95-th percentile relies on the underlying assumption that abnormal occurrences (poor quality image or poor quality segmentation) are rare, and thus most of the data points should pass both input and output QCs [304].

IV.4.2 Prediction space stratification for cross-sectional MS lesions segmentation

To test the relevance of the proposed prediction space stratification, we employ the ensemble of Dynamic U-Nets trained to perform MS lesions segmentation. This ensemble has been previously employed in the voxel-level and lesion-level experiments (Chapter II and III). In this setting, 3 test datasets of MS subjects are available (Test ID, MSLUB, and 1.5 Tesla datasets, introduced in Table II.1), and previous experiments showed that the segmentation performance decreased on the MSLUB and 1.5 Tesl datasets due to the generalization gap (Tables II.5, III.3). The protocol is as follows: input and output-level QC scores (MD and 1-EPA, respectively) are computed for each of the 103 test images with non-empty ground truth segmentations, and predictions are assigned to one of the 4 regions using the thresholds optimized on validation images. Then, the segmentation performance in each region is assessed using 3 segmentation metrics: the Dice score, the Surface Dice, and the 95%

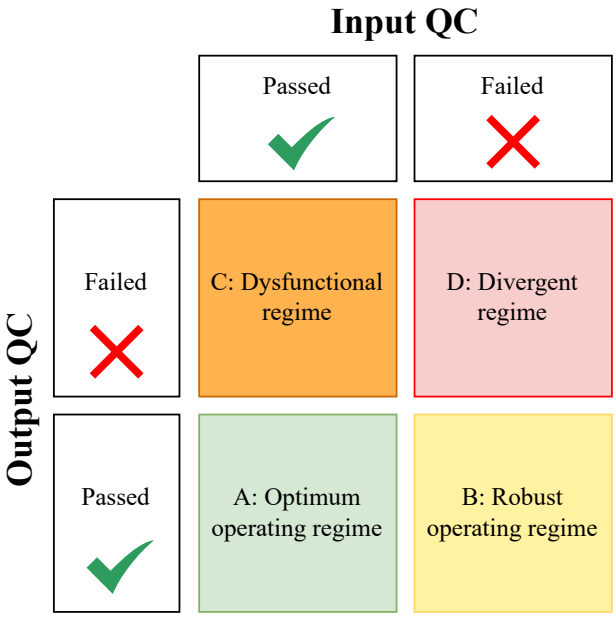


Figure IV.4.3: Proposed stratification of the prediction space using input and output QC estimates.

Haussdorf distance. For this application, a slight modification is needed for the computation of the MD score (input-level QC). Indeed, the MS models are patch-based, meaning that they process the input 3D MRI by first dividing the volume into patches of a fixed size of $128 \times 128 \times 128$. As a result, for a single input MRI, several latent representations are extracted (one per patch). As in Gonzalez et al. [72], the final volume score is obtained by averaging the Mahalanobis distances of the patches.

The resulting stratification of the prediction space is presented in Figure IV.4.4. Four colormaps are proposed, 3 indicating the value of the segmentation metric (Dice, Surface Dice, and Haussdorf) as well as a colormap indicating the dataset source of the test point (Test ID, MSLUB or 1.5 Tesla datasets). The input and output QC thresholds fit on the validation dataset are represented by black dashed lines. Figure IV.4.5 presents boxplots of segmentation metrics in each region (A: optimum, B: robust, C: dysfunctional, D: divergent), and corresponding averaged metrics are presented in Table IV.3. First, the MDs are moderated for each test image, with a max value around 0.30. For comparison, the **Far OOD** images in the OOD benchmark received MD of around 2 (see Figure IV.3.16). Then, it appears that region A, corresponding to the expected optimum operating regime, includes most data points (roughly 50% of test samples). This region contains the top-quality predictions, for each metric. It appears in Figure IV.4.4 that most Test ID data points in practice fall within this region. Then region B (robust operating regime) contains approximately 25% of the test samples, with a slight decrease in segmentation performance. Region C (dysfunctional regime) is the less populated bin with about 5% of test points. Here, the performance is inferior to the one obtained in regions A and B. Finally, region D, corresponding to the divergent regime, regroups the remaining data points. Performance in this bin is weak (average Dice around 0.50). It can be seen in Figure IV.4.4 that this region mostly regroups images from

Region	Proportion	Dice ↑	Surface Dice ↑	95% Haussdorf ↓
A	53/103 (51.46%)	0.791 ± 0.086	0.960 ± 0.039	4.718 ± 4.763
B	27/103 (26.21%)	0.754 ± 0.090	0.930 ± 0.069	6.722 ± 5.710
C	6/103 (5.83%)	0.732 ± 0.105	0.902 ± 0.070	10.073 ± 5.091
D	17/103 (16.50%)	0.541 ± 0.163	0.766 ± 0.149	21.545 ± 12.552

Table IV.3: Average segmentation performance in each region of the prediction space for the cross-sectional MS lesions Deep Ensemble. A: optimum operating regime. B: robust operating regime. C: Dysfunctional regime. D: Divergent regime.

the MSLUB and 1.5 test data points. These images are both distant from the training distribution (high MDs) and the intra-ensemble variability is high (low EPA).

Thus, it appears that empirically the proposed stratification of the prediction space respects the expected behavior, with top-quality predictions in region A (optimum), and progressively decreasing quality of segmentation in regions B (robust), C (dysfunctional), and D (divergent). Figure IV.4.6 presents 3 examples of test data points in regimes A, B, and D, respectively. They correspond to the points P1, P2, and P3 indicated in the lower right plot in Figure IV.4.4. P1 (optimum operating regime) corresponds to a data point with low input and output non-conformity scores, and it is indeed associated with a high-quality prediction (high Dice, Surface Dice, and low 95% HD). P2 (robust operating regime) is the datapoint with the higher MD over the test samples (MD = 0.28). Inspection of the FLAIR indeed reveals the presence of an important motion artifact, particularly visible in the ventricles. However, it does not prevent the correct functioning of the model (acceptable segmentation performance). Finally, P3 (divergent regime) is the point with the worst output QC score ($1 - \text{EPA} = 0.34$). The associated segmentation is poor, with a Dice score of only 0.23. The FLAIR is also noisy, explaining the high estimated MD.

IV.4.3 Prediction space stratification for glioblastoma segmentation

The second experiment relies on the ensemble of Dynamic U-Nets used for glioblastoma segmentation in the previous voxel-level experiments (Section II.6). As a reminder, this ensemble takes as input 4 brain MRI sequences (FLAIR, T2, T1, T1 with contrast agent) and outputs 3 tumor classes: necrosis, edematous, and GD-enhancing tumor. A total of 5 test datasets are used here: the in-distribution test split (Test ID, N=227), and 4 domain-shift datasets originating from auxiliary BraTS 2023 dataset: Sub-Saharan Africa [238](N=60), Meningioma [290](a subset of N=250 cases is used), Metastases [289] (N=238), and Pediatric [288] (N=99). All these datasets are challenging to segment because of shifts in the appearance of the tumor and/or the tumor size and location.

The protocol is similar: input and output QC scores are computed on the validation images to determine decision thresholds for the input and output QC scores, set to be the 95-th percentiles. It allows to associate each test image with a regime (A, B, C, or D). The resulting stratification of the prediction space is presented in Figure IV.4.7. In some cases, the Haussdorf Distance was undefined because of an empty prediction mask. To allow the

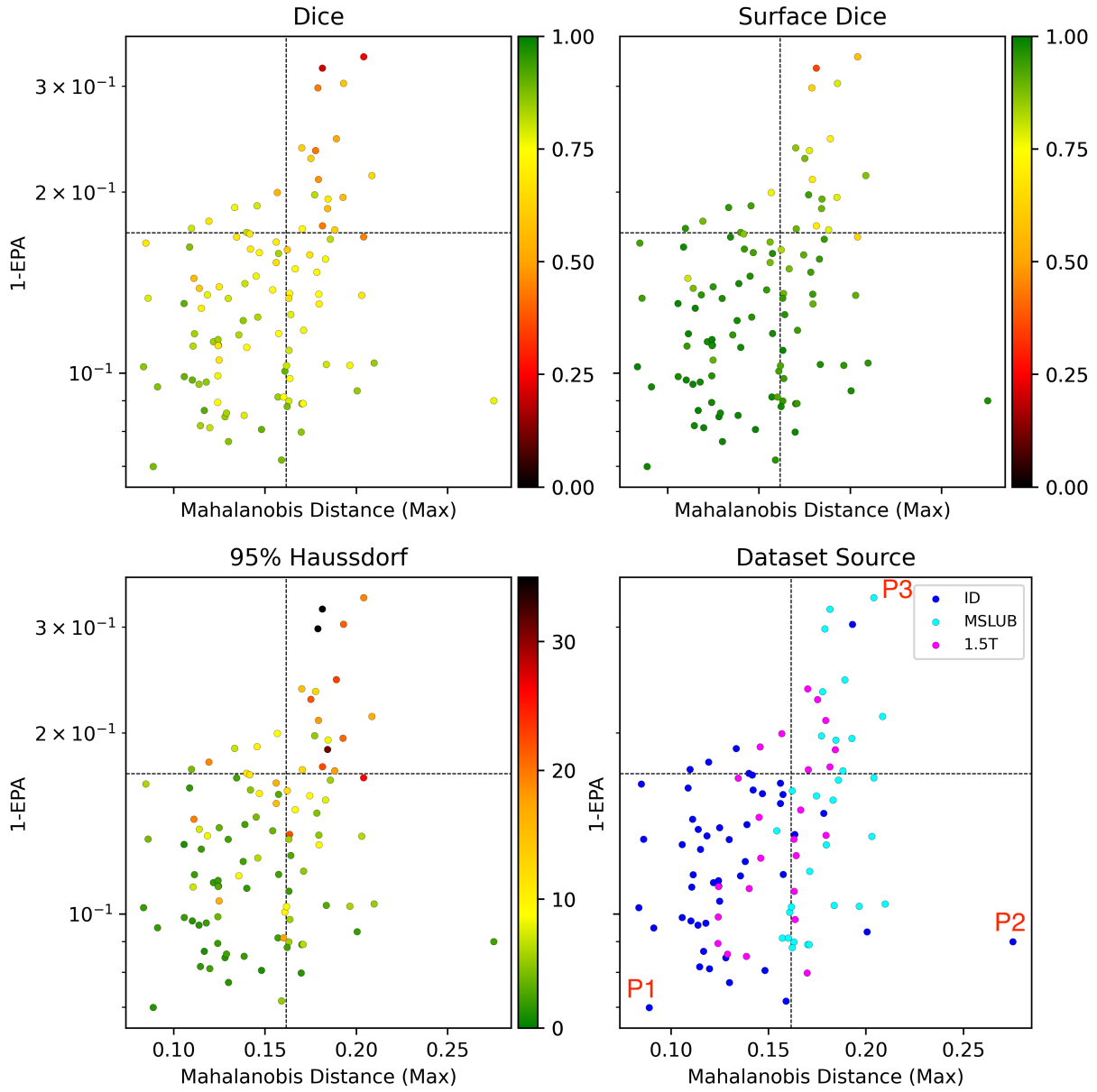


Figure IV.4.4: Prediction space stratification for the cross-sectional MS ensemble. The y-axis is plotted on a logarithmic scale. P1, P2, and P3 indicate cases that are further detailed in Figure IV.4.6.

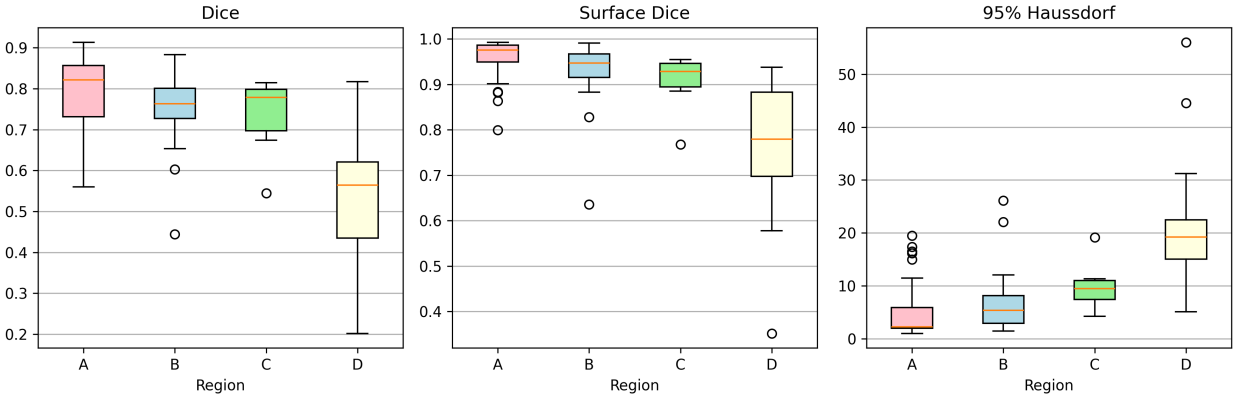


Figure IV.4.5: Box-plots of segmentation metrics (Dice, Surface Dice, 95% Haussdorf Distance) for each region of the prediction space, for cross-sectional MS lesions segmentation.

Regime	**Proportion**	**Dice** ↑	**Surface Dice** ↑	95% **Haussdorf** ↓
A	264/874 (30.21%)	0.828 ± 0.141	0.886 ± 0.152	8.410 ± 14.49
B	400/874 (45.77%)	0.707 ± 0.206	0.732 ± 0.226	12.851 ± 15.544
C	20/874 (2.29%)	0.678 ± 0.196	0.575 ± 0.151	16.536 ± 22.189
D	190/874 (21.74%)	0.334 ± 0.355	0.259 ± 0.264	24.552 ± 28.279

Table IV.4: Average segmentation performance in each region of the prediction space for the glioblastoma Deep Ensemble. A: optimum operating regime. B: robust operating regime. C: Dysfunctional regime. D: Divergent regime.

visualization of these cases, the distances were mapped to the maximum Haussdorf distance obtained on the test images (HD = 50). Figure IV.4.8 presents boxplots of segmentation metrics in each region (A, B, C, D), and corresponding averaged metrics are presented in Table IV.4. Figure IV.4.9 presents 2 examples of test data points in regime B and D, respectively.

As for the cross-sectional MS experiment, the 4-regimes stratification presents a gradually decreasing level of quality. Regime A (optimum regime, 30.21% of samples) mostly contains images from the in-distribution test split, with top-quality segmentations. Regime D (divergent regime, 21.74%) interestingly regroups the majority of extremely poor predictions (Dice and Surface Dice of 0, represented as black dots). Regimes B (robust regime) and C (dysfunctional) present intermediate performance levels. Note that in this experiment, only 227 images out of 874 are in-distribution samples, which explains why regime A only contains roughly 30% of the data points. Regarding the two examples provided in Figure IV.4.9, sample P1 is representative of the robust functioning regime. A massive artifact may be responsible for the high associated MD, however the artifact does not intersect the tumor. Thus, the ensemble provides a high-quality segmentation. Example P2 is a sample from the SSA dataset with a low image quality. The tumor is small and located in the infratentorial region of the brain. The ensemble predicts an empty tumor mask for this subject, thus associated with a null

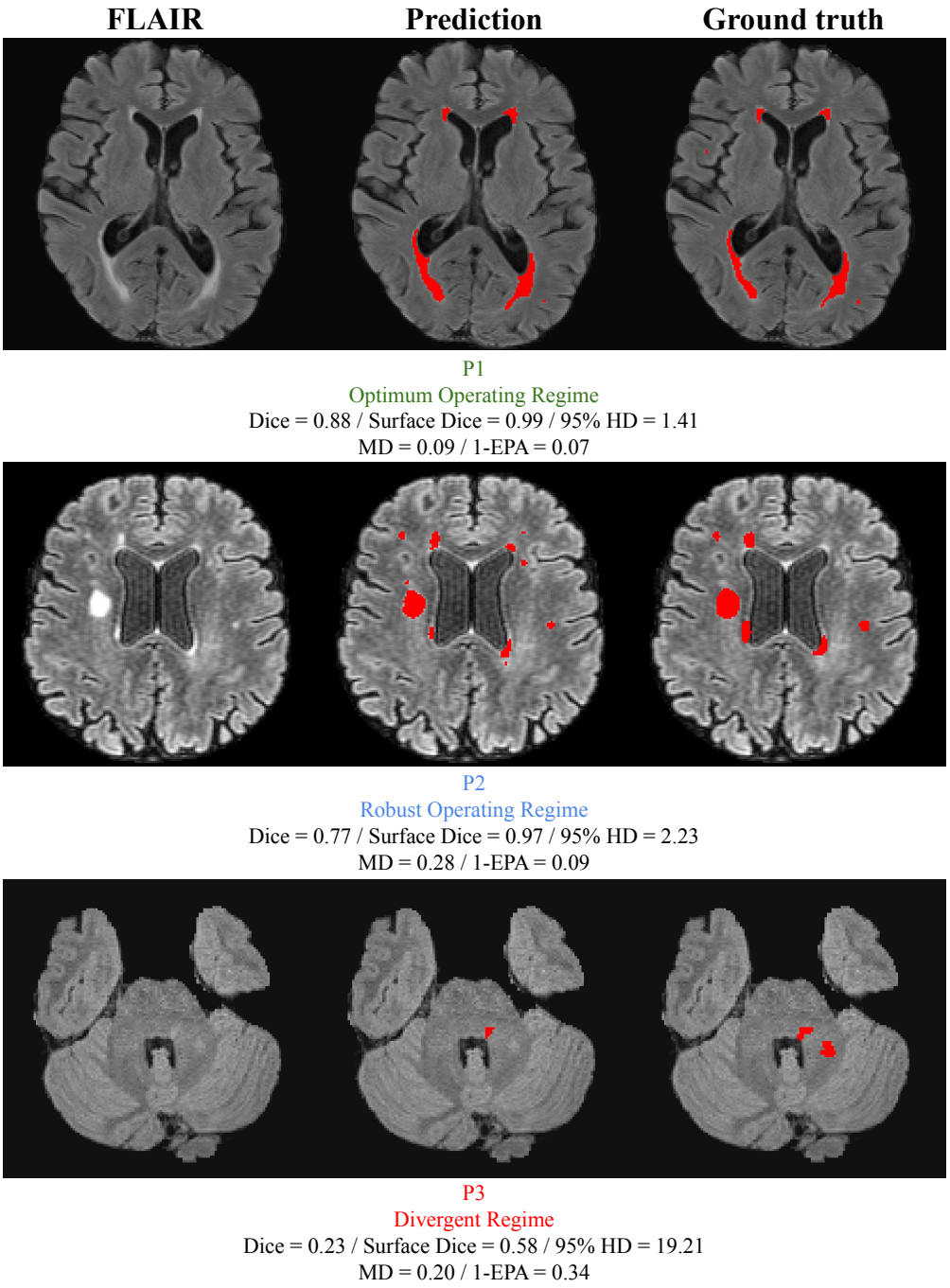


Figure IV.4.6: Examples of data points in Regime A, B, and D for the cross-sectional MS ensemble. P1 is a high-quality FLAIR MRI, associated with low input and output QC scores. It is also nearly perfectly segmented. On the contrary, P2 has a lower quality, particularly visible in the ventricles. Yet, the associated segmentation is valid which explains its location in the robust operating regime. Finally, P3 is a sample from the MSLUB dataset, with a lower quality and a poor prediction.

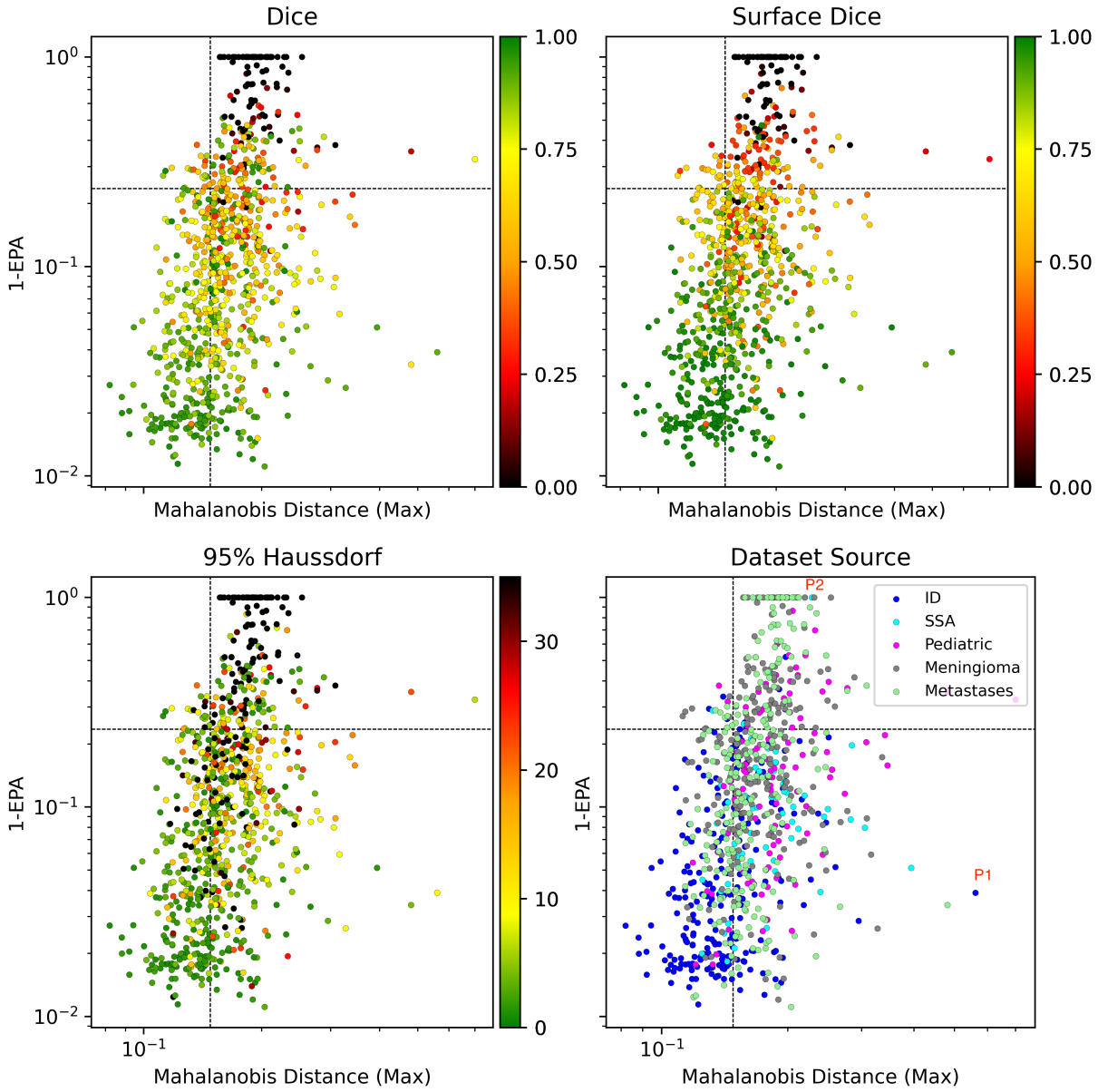


Figure IV.4.7: Prediction space stratification for the glioblastoma ensemble. Both the x-axis and y-axis are represented on logarithmic scales. P1 and P2 correspond to cases that are further detailed in Figure IV.4.9.

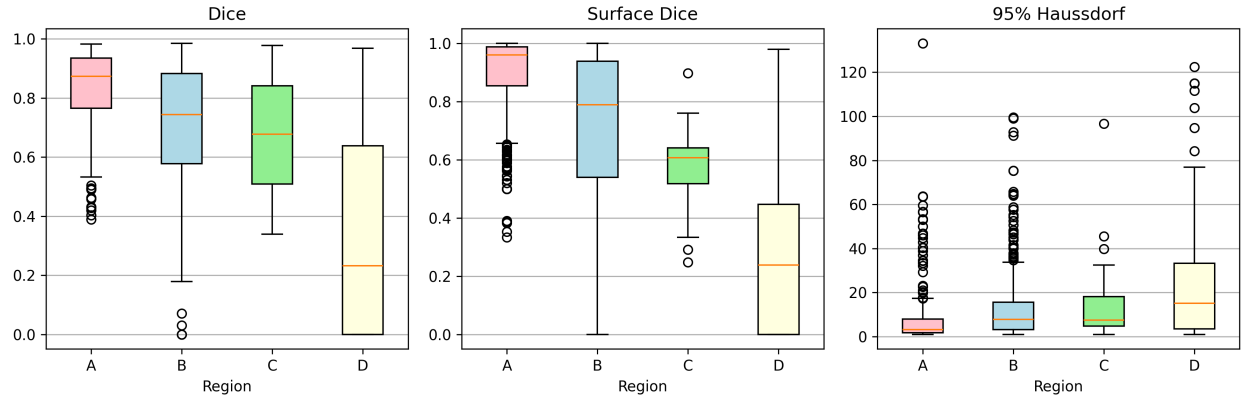
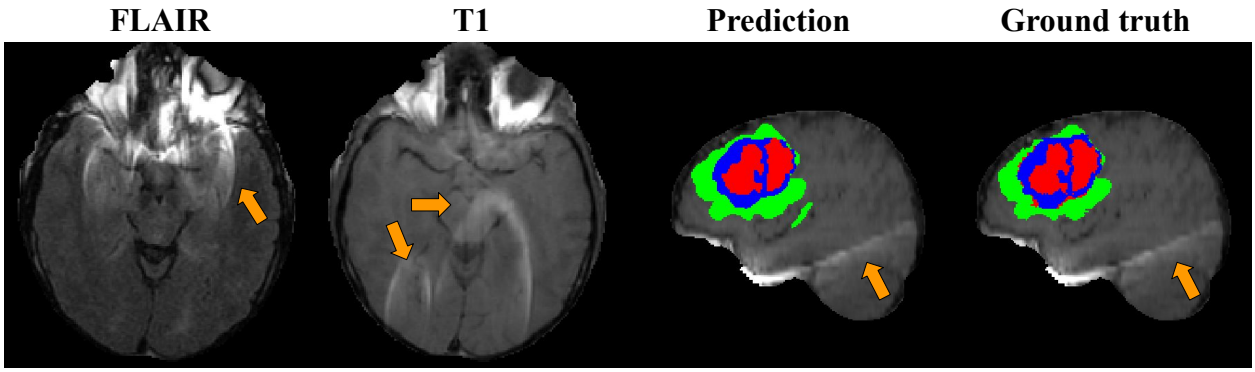
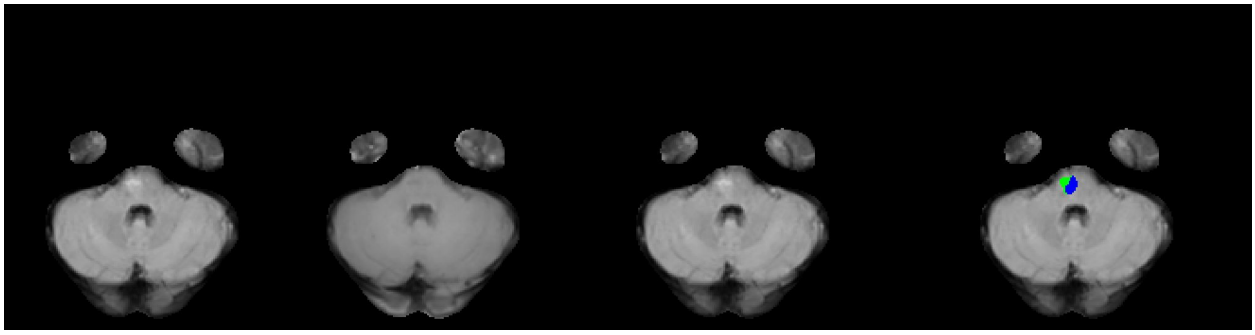


Figure IV.4.8: Box-plots of segmentation metrics (Dice, Surface Dice, 95% Haussdorf Distance) for each region of the prediction space, for glioblastoma segmentation.



P1
Robust Operating Regime
Dice = 0.91 / Surface Dice = 0.93 / 95% HD = 9.25
MD = 0.56 / 1-EPA = 0.04



P2
Divergent Regime
Dice = 0.00 / Surface Dice = 0.00 / 95% HD is undefined
MD = 0.23 / 1-EPA = 1.00

Figure IV.4.9: Examples of data points in Regime B and D for the glioblastoma ensemble. The first subject (P1) presents a heavy artifact visible in both FLAIR and T1w sequences, indicated by orange arrows. However, it does not prevent the proper functioning of the image, being associated with high metrics. The second example (P2) is a subject from the SSA dataset, with a missed tumor lesion. It is associated with a high output QC score.

Dice and Surface Dice, as well as an undefined Haussdorf distance. It is also linked with the highest possible output QC score ($1 - \text{EPA} = 0$).

IV.4.4 Prediction space stratification for polyp segmentation in 2D colonoscopy

Until now, all experiments have focused on 3D medical image processing. To verify that the results hold for 2D applications, a novel task is introduced here: polyp segmentation in 2D colonoscopy images. This also allows us to evaluate the approach on large-scale datasets, as 2D images are more widely available than their 3D counterparts.

IV.4.4.1 Pathology description and datasets

Polyps correspond to abnormal growths of a mucous membrane and are most often located in the colon and rectum. The majority of polyps are benign, but they can change over time and become cancerous, leading to colorectal cancer. Thus, the early detection and staging of polyps is a key to early cancer treatment. Polyp detection is generally carried out using colonoscopy, a technique following which a flexible tube called a colonoscope is inserted into the rectum and advanced through the entire colon, providing 2D images of the large intestine. Automated tools are needed to assist clinicians in detecting polyps during colonoscopy, as they may overlook polyps due to fatigue or lack of experience [307].

To explore this task, a training dataset is created using data collected from different open data hubs: Kvasir [308] (1000 images), ETIS-LaribPolyp [309] (196 images), CVC-ColonDB [310] (380 images) and CVC-ClinicDB [311] (612 images). This results in a set of 2188 endoscopic images with associated binary polyp mask, from which a random split is made: 60% for training (1312 images), 20% for validation (438 images) and 20% (438 images) for in-distribution test (Test ID). All images are resized to a shape of 768×512. To simulate domain-shift scenarios, the PolypGen dataset [312] is employed. This dataset comprises endoscopy images from 6 different centers, exhibiting a heterogeneous population and acquired with different endoscopic systems. 251 samples are used from Center 1, 270 for Center 2, 456 for Center 3, 146 for Center 4, 206 for Center 5, and 83 for Center 6. With the Test ID split, this represents a test set of 1849 samples to evaluate the QC strategy.

IV.4.4.2 2D polyp segmentation ensemble

To process the 2D colonoscopy images, a dedicated 2D segmentation architecture is used. The backbone is in all points equivalent to the 3D Dynamic U-Net used so far, except 3D convolutions are replaced with their 2D counterparts. The model contains an input block followed by 4 downsampling blocks and 4 upsampling blocks. Two convolutions are used in each block, for a total of 18 convolutions. A last convolution is in charge of producing the class probabilities. The model takes as input the 3-channel 2D images (for red, green, and blue channels). Batch normalization is used in each block. In total, the 2D UNet contains 2 million trainable parameters. An ensemble of polyp segmentation models is formed by training 5 individual U-Nets, with the $\mathcal{L}_3$ loss (Chapter II, Equation II.6.3). The rest of the pipeline is strictly identical: input and output level scores are computed on the validation

Regime	Proportion	Dice ↑	Surface Dice ↑	95% Haussdorf ↓
A	748/1849 (40.45%)	0.897 ± 0.104	0.748 ± 0.210	52.68 ± 73.68
B	508/1849 (27.47%)	0.850 ± 0.169	0.663 ± 0.257	76.890 ± 94.558
C	99/1849 (5.35%)	0.656 ± 0.211	0.388 ± 0.190	162.959 ± 102.400
D	494/1849 (26.72%)	0.470 ± 0.286	0.278 ± 0.221	217.790 ± 122.244

Table IV.5: Average segmentation performance in each region of the prediction space for the polyp Deep Ensemble. A: optimum operating regime. B: robust operating regime. C: Dysfunctional regime. D: Divergent regime.

split, allowing to define thresholds as the 95-th percentiles. Then, each test image is associated with a regime (A, B, C, or D).

IV.4.4.3 Results

The stratification of the prediction space for polyp segmentation is presented in Figure IV.4.10. Figure IV.4.11 presents boxplots of segmentation metrics in each region (A, B, C, D), and corresponding averaged metrics are presented in Table IV.5. Figure IV.4.12 presents 3 examples of test data points in regimes A, B and D, respectively. Similar to the previous experiments, the optimum operating regime (regime A) in practice regroups the top-quality predictions, with an average Dice score of around 0.90. It can be observed in Figure IV.4.10 that most Test ID samples fall within this regime. For domain-shift data points (PolypGen), the segmentation quality is heterogeneous depending on the imaging center. For instance, images from Center 3 (C3) are generally associated with high-quality segmentation metrics, while images from Center C4 are for the majority in the divergent regime, associated with low-quality segmentations. Three examples of pairs of images and predictions are presented in Figure IV.4.12. The sample P1 is a data point originating from the Test ID, close to the training distribution. It is associated with a high-quality segmentation, representative of the expected functioning of the model. Sample P2 is a data point from Center 2 of the PolypGen dataset, attributed to the robust functioning regime (high MD but high EPA). Its color is more unusual, with a blue tone. This probably perturbates the segmentation models, as the associated segmentation is poor. Finally, sample P3 originates from Test ID but is associated with the divergent regime (high MD and low EPA). Indeed, the image is extremely unusual. The resulting segmentation is poor, as can be expected for this extremely particular image.

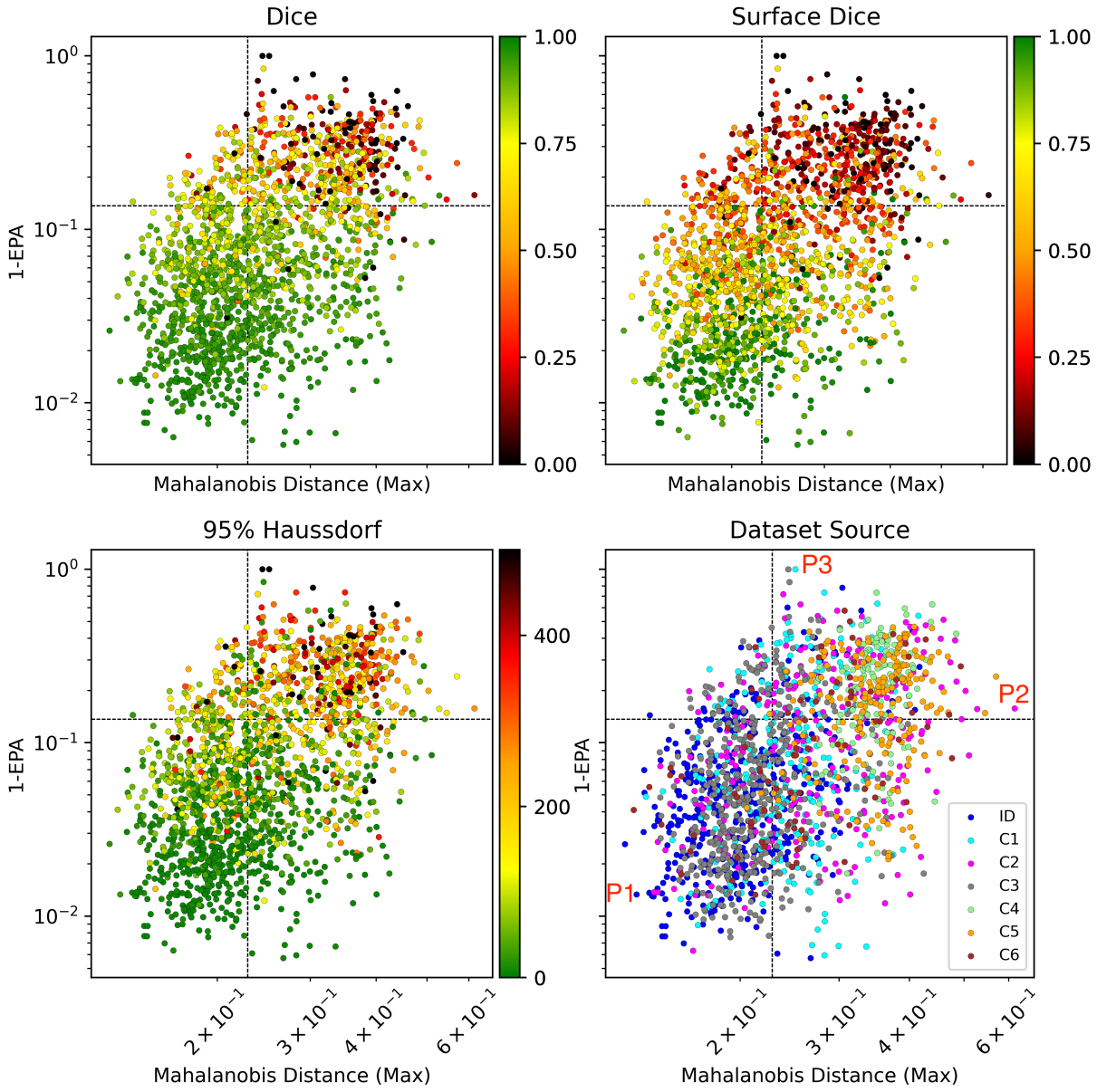


Figure IV.4.10: Prediction space stratification for the polyp ensemble. Both the x-axis and y-axis are represented in logarithmic scales. P1, P2, and P3 correspond to cases that are further detailed in Figure IV.4.12.

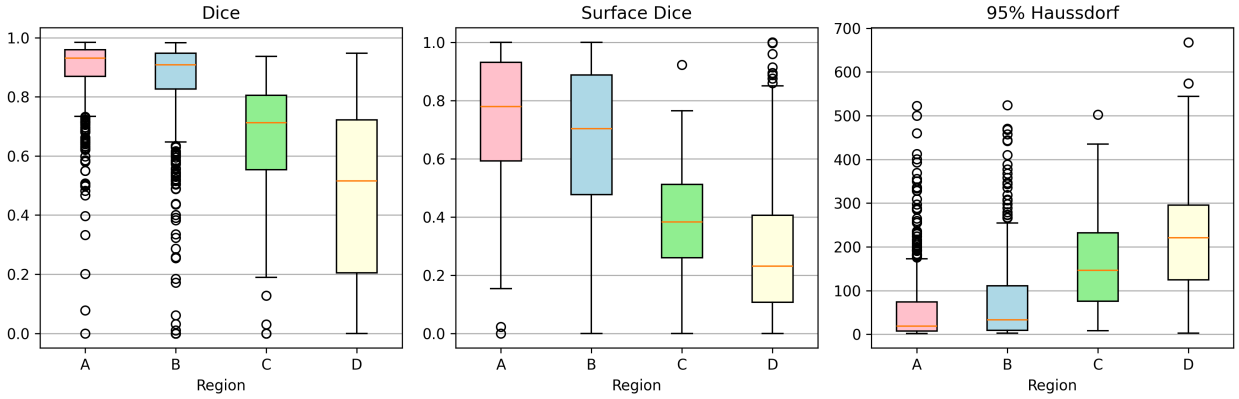
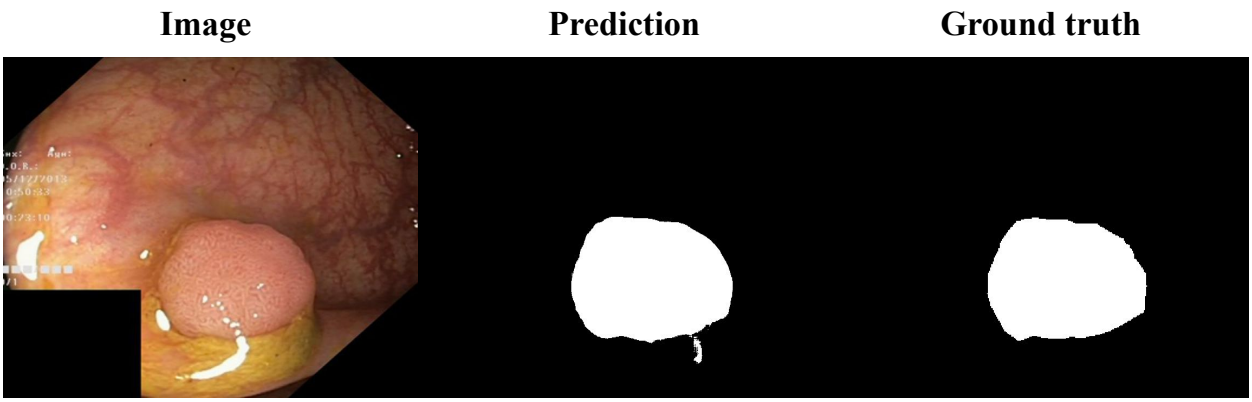
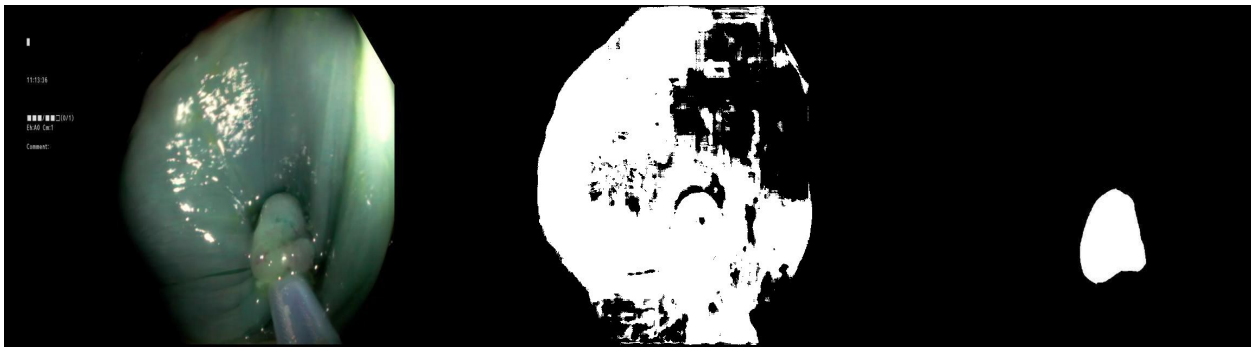


Figure IV.4.11: Box-plots of segmentation metrics (Dice, Surface Dice, 95% Haussdorf Distance) for each region of the prediction space, for polyp segmentation.



P1
Optimum Operating Regime
Dice = 0.98 / Surface Dice = 0.92 / 95% HD = 20.94
MD = 0.12 / 1-EPA = 0.03



P2
Robust Operating Regime
Dice = 0.15 / Surface Dice = 0.04 / 95% HD = 264.89
MD = 0.61 / 1-EPA = 0.16

P3
Divergent Regime
Dice = 0.0 / Surface Dice = 0.0 / 95% HD = 341.48
MD = 0.30 / 1-EPA = 0.74

Figure IV.4.12: Examples of data points in Regime A, B and D for the polyp ensemble. The first subject (P1) presents a clear delineation of the polyp, allowing for a top-quality segmentation. In contrast, the second example (P2) presents an unusual illumination with a blue-ish tone. As a result, the segmentation is erroneous, with a large quantity of false positive voxels. The associated MD is extremely high, however, the output QC score (1-EPA) is not excessive, which explains the attribution to regime B. The last example (P3) is a sample with a high output QC score. The image is artifactual and blurry, and the segmentation doesn't intersect the ground truth.

IV.5 Chapter conclusion

In this chapter, the crucial problem of detecting non-conform inputs has been investigated. In commercial solutions such as Pixyl.Neuro, the volume of analysis is too large to perform a visual validation of the input image before sending the automated analysis to the radiologist. Moreover, non-conform inputs are common due to human error or misunderstanding of the software specifications. In such cases, it is essential to inform the radiologist analyzing automatic results that the inputs sent were non-conform. To tackle this challenge, image-level conformity scores were evaluated, comprising uncertainty-based, reconstruction-based, and latent-based approaches. On a wide benchmark comprising 24 OOD scenarios built around the task of whole tumor segmentation in brain T1w, the superiority of latent-space OOD detection has been demonstrated. This approach, built around the Mahalanobis Distance, achieves top OOD detection performance on various types of shifts, such as artifacts, modality, or pathology shifts. It is also particularly interesting as it can be implemented in any trained model, requiring only access to the intermediate feature maps generated by the model. One pitfall of the MD is its sensitivity to the layer selection. More particularly, the optimal layer is dependent on the neural network architecture, which requires a layer selection procedure using OOD test data. However, experiments showed that it can be alleviated by using a multi-layer strategy, consisting of the combination of the individual layer scores using the max or mean operation. These aggregation strategies achieve high and consistent performances for each segmentation backbone.

Yet, it has been shown that OOD detection alone is not sufficient. Indeed, OOD detection focuses on the conformity of the input image, not on the quality of the output segmentation. This has two weaknesses: conform images that are poorly segmented are not flagged to the user, and OOD images correctly segmented can lead to false alarms. Additionally, we showed that the MDs are poorly correlated with the actual quality of the output (assessed using Dice scores, see Figure **??**). It highlights an important gap in the UQ literature, that is the distinction between input-level and output-level QC. Both goals are generally pursued separately in the literature. We argue that benefit can be gained by talking both QC simultaneously. To implement this, the MD is complemented with an output-level score, namely the level of agreement between the members of the deep ensemble (EPA). By combining these two scores, we propose a stratification of the prediction space in 4 areas: optimum operating regime, robust operating regime, dysfunctional regime, and lastly divergent regime. The proposed stratification has been challenged on 3 tasks: cross-sectional MS lesions segmentation in FLAIR brain MRI, glioblastoma segmentation in multi-modal brain MRI, and polyp segmentation in colonoscopy images. For each experiment, the optimum operating regime contains the top-quality predictions of the model, generally associated with in-distribution images. The divergent regime, in contrast, contains the poorest predictions. They are associated with images distant from the training dataset. Thus, this dual-level QC strategy has the potential to automatically assess the quality of a prediction, while providing knowledge about the distance of the test point to the training distribution. This could bring additional information to the user, by alerting them about pitfalls of the algorithm, in which case the reviewing of the case is imperative. Interestingly, both metrics (MD and EPA) are efficient to compute using a Deep Ensemble, allowing a seamless integration in AI-based software.

Chapter V

Conformal Prediction for predictive intervals on lesion volumes

Contents

V.1 Motivations

Segmentation tasks are often an initial step to a more in-depth analysis. More particularly, segmentation masks can be used to derive several high-level metrics meaningful for the clinician, such as volumetry measures. In the context of Multiple Sclerosis, segmentation is used to assess the total lesion volume (called lesion load), and the count of the lesions. These metrics are useful to determine the extent and progression of the disease, and can even be used to predict the disability of the patient [313]. The lesion volume is also an important imaging biomarker to predict the patient's neurological outcome after a stroke [314] or to assess the grade of a glioblastoma [315]. In the case of neurodegenerative diseases such as Alzheimer's disease, brain atrophy is quantified by estimating the volume of different anatomical regions (e.g. hippocampus or amygdala) compared to normative values [316]. Apart from neurological applications, volumetry can also be applied to organs to detect abnormal growth or in aging studies [317].

In the context of Pixyl software, each automated analysis yields an automatic report in which these high-level metrics (volumes and/or lesion counts) are reported. This is the first information presented to the user, who can then check the automatic segmentation to confirm its suitability. However, to this date, these metrics are reported without predictive intervals (PIs), which affects the trustworthiness, reliability, and accountability of the automatic results. For quality insurance, equipping the reported high-level metrics with proper PIs appears as a required property to improve the usefulness of automatic reports.

This form of uncertainty quantification, focusing on high-level metrics, has gained little attention in the medical image UQ literature. Indeed, PI construction for ML models has been mainly studied in the context of 1D regression tasks [318, 319, 320] and applications in the context of medical image processing are very scarce. Reference work by Eaton et al. [81] proposes either a sampling approach or a regression model to compute PIs for lesion counting in 2D medical images. In the former, several plausible and diverse segmentation masks are generated for the same input image, forming a distribution over the quantity of interest (e.g. lesion volume or number), from which the mean and the standard deviation can be extracted to define a PI. For this, standard UQ methods such as Test Time Augmentation (introduced in Section II.2.10) or MC dropout (introduced in Section II.2.6) can be employed. With the regression approach, a network is trained to directly predict the PI's components: the mean value as well as the lower and upper bounds from the data themselves, using a dedicated loss function, the Quantile loss (also called Pinball loss) [321].

Recently, the Conformal Prediction framework [75, 76] has been gaining attention, as a distribution-free, model-agnostic uncertainty quantification tool offering statistical guarantees in finite samples. CP has been essentially applied in classification and regression problems, where predictive sets are constructed around the notion of *coverage*. Coverage refers to the probability that a given interval contains the true value of the quantity being predicted. In other words, it measures how well the interval captures the uncertainty associated with the prediction. In practice, CP can be implemented as a post-processing step of the predictive intervals to make sure that they will encompass the desired fraction of the true test scores, with popular choices being 90%, 95%, or 99% PIs. Due to its promising results outside

the medical image field, this chapter proposes to explore the usage of CP to equip volume estimation with predictive intervals.

V.1.1 Additional contributions to the paper "TriadNet: Sampling-free predictive intervals for lesional volume in 3D brain MR images"

The work presented in this chapter is based on the paper *TriadNet: Sampling-free predictive intervals for lesional volume in 3D brain MR images* [322], which has been presented at the UNSURE workshop, a satellite event of the MICCAI 2023 conference. Several additions are presented here. Firstly, the framework is tested on cross-sectional MS lesions segmentation in addition to the brain tumor application. Secondly, a deeper discussion on the impact of the size of the calibration dataset is proposed, and robustness testing is investigated by using two domain-shift datasets for the MS experiment, and one domain-shift dataset for brain tumors. Finally, weighted conformal prediction to tackle domain shifts is investigated in the last section.

V.2 Conformal prediction for lesion volumes

V.2.1 Problem formulation

In this section, CP is explored to obtain PIs associated with volumes in medical images. We are addressing a 3D segmentation problem involving $N-1$ foreground classes, excluding the background class. Our objective is to estimate the true volumes, denoted as $Y \in \mathbb{R}^{N-1}$, for each foreground class based on the predicted segmentation. In this scenario, considering an estimation $X \in \mathbb{R}^{N-1}$ of the volumes as a random variable, we define a predictive interval, denoted as $\Gamma_\alpha(X)$, as a range of values intended to encompass Y, the actual volumes, with a specified degree of confidence, typically denoted as 1-α (e.g., 90% or 95%). In essence, given a set of estimated volumes $X_1 \dots X_n$ and their corresponding ground truth volumes $Y_1 \dots Y_n$, $\Gamma_\alpha(\cdot)$, the predictive interval $\Gamma_\alpha(\cdot)$ should be modeled to satisfy the following condition:

$$1-\alpha \leq P(Y_{\text{test}} \in \Gamma_\alpha(X_{\text{test}})) \leq 1-\alpha+\frac{1}{n+1} \tag{V.2.1}$$

for any $(Y_{\text{test}}, X_{\text{test}})$ following the same distribution as the (Y_i, X_i)'s. This property is called the *marginal coverage*, as the probability takes into account the randomness in the calibration and test dataset. That is to say, by randomly sampling multiple calibration and test datasets, the coverage is guaranteed to be at least $1-\alpha$ in average [76]. In this section, 90% confidence intervals are used, corresponding to an error rate $\alpha = 0.1$. In practice, a PI is composed of 3 components: the estimated volume X_i, the lower bound l_i, and the upper bound u_i. There are two main methodologies to estimate the bounds, namely sampling-based or direct estimation, presented in the following.

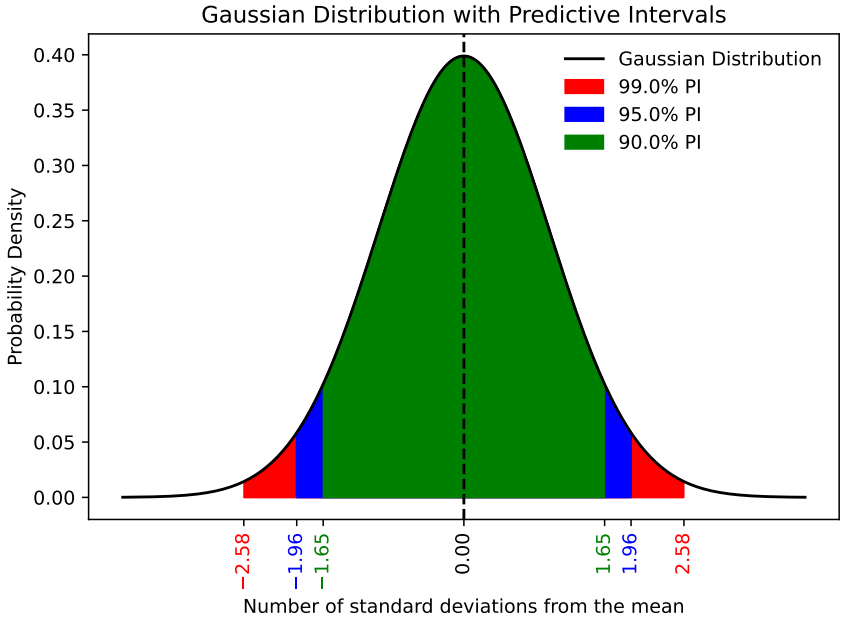


Figure V.2.1: A Gaussian distribution with a mean of 0 (represented by the black dashed line) and a variance of one, with different confidence intervals represented: 90% (green), 95% (blue), and 99% (red). The corresponding numbers of standard deviations are indicated on the x-axis.

V.2.1.1 PI estimation via sampling

Standard UQ methodologies (e.g. TTA, MC dropout) allow to sample a set of plausible segmentation masks for each input image. A natural idea to compute a predictive interval is to exploit this set of predictions. In practice, the mean and the standard deviation of the target volume can be derived using the set of segmentations. This allows to obtain predictive intervals in the form:

$$\Gamma_\alpha(X) = [\mu(X) - z\sigma(X), \mu(X) + z\sigma(X)] \tag{V.2.2}$$

where z stipulates the degree of confidence of the interval. For instance, for a 90% confidence interval, z corresponds to 1.65 (see Figure V.2.1 in green). In practice, this approach supposes that $Y_{test}|X_{test} = X$ follows a Gaussian Distribution $\mathcal{N}(\mu(X), \sigma(X))$, which may be a simplifying assumption. However, it allows the computation of PIs using any sampling approach (MC dropout, DE, TTA) which is thus convenient.

V.2.1.2 Direct PI estimation

In contrast to sampling-based PI estimation, direct PI estimators are models that directly predict the mean value X as well as the $\hat{t}_{\alpha/2}$ and $\hat{t}_{1-\alpha/2}$ quantiles, allowing the computation of $(1-\alpha\%)$ PIs:

$$\Gamma_\alpha(X) = [\hat{t}_{\alpha/2}(X), \hat{t}_{1-\alpha/2}(X)] \tag{V.2.3}$$

For example, with 90% PIs, the ground truth volume Y is supposed to land below $\hat{t}_{0.05}(X)$ with 5% probability and above $\hat{t}_{0.95}(X)$ with 5% probability.

It is interesting to note that with direct PI estimation, PIs are not necessarily symmetrical with respect to the mean, and are thus in principle more flexible than sampling-based PI estimators.

V.2.2 Conformal calibration of predictive intervals

At this stage, there is no statistical guarantee that the computed PIs will achieve the user-defined level of coverage, *i.e.* 90% or 95%, on the test dataset. Indeed, To ensure this, Conformal Prediction (CP) can be used. It operates by first defining a score function $s(X, Y)$ [76], to estimate the degree of conformity of the estimate X with respect to the true quantity Y, with larger scores indicating larger deviations. The score function takes the following form for sampling-based PI:

$$s(X, Y) = \frac{|Y - \mu(X)|}{\sigma(X)} \tag{V.2.4}$$

and the following one for direct PI estimators:

$$s(X, Y) = \max\{\hat{t}_{\alpha/2}(X) - Y, Y - \hat{t}_{1-\alpha/2}(X)\} \tag{V.2.5}$$

These scores reflect the ability of the interval to capture the ground truth quantity. In the CP framework, they are used to estimate a corrective value $\hat{q}$ to be applied on the PIs to match the target coverage level on a calibration dataset $(X_i, Y_i)_{i=1,...,n}$ comprising images and associated ground truth volumes. In practice, the corrective value $\hat{q}$ is computed as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$-th quantile of the empirical scores $\hat{q} = \text{Quantile}(s_1, s_2, ..., s_n; \frac{\lceil (n+1)(1-\alpha) \rceil}{n})$. At test time, the calibrated PI is computed as follows for sampling-based PI:

$$\Gamma_\alpha(X) = [\mu(X) - \hat{q}\sigma(X), \mu(X) + \hat{q}\sigma(X)] \tag{V.2.6}$$

and as follows for direct PI estimates:

$$\Gamma_\alpha(X) = [\hat{t}_{\alpha/2}(X) - \hat{q}, \hat{t}_{1-\alpha/2}(X) + \hat{q}] \tag{V.2.7}$$

As $\hat{q}$ increases, the intervals expand. Supposing the test samples are exchangeable with the calibration samples, the marginal coverage property is guaranteed. To recap, CP can

be essentially seen as a post-hoc calibration step applied to the PIs to achieve the desired level of coverage. A larger level of coverage will yield to larger interval widths. Additional mathematical details about the link between the score functions and the target coverage are provided in Appendix A5.

V.2.3 TriadNet: sampling-free predictive intervals

Sampling-based PI estimation is straightforward when a standard UQ methodology such as TTA or MC dropout can be employed. However, they require sampling multiple predictions to allow the estimation of the mean and standard deviation, which is not in line with real-world applications where inference time is crucial. Moreover, it is based on the underlying hypothesis that the sampling distribution of the volume follows a normal distribution, which may not in practice be systematically valid. This motivates the development of a direct PI approach specially tailored for medical image processing.

The starting point is to notice that the PI can be constructed using 3 different delineations of the same lesion, one permissive (higher volume), one restrictive (lower volume), and a balanced one (mean volume). One can think of using an ensemble of 3 models to generate these 3 masks: the first one with a high precision and a low recall to generate the lower volume estimation; the second with a low precision and a high recall to generate the upper volume estimation, and a last one with balanced precision and recall for the mean volume estimation. However, this implies reproducing the training three times, increasing the computational cost. Here, we argue that the same result can be obtained with a single multi-head architecture, trained to output the three different masks simultaneously. This architecture, which we name TriadNet, is represented in Figure V.2.2.

Essentially, TriadNet follows a classic encoder-decoder architecture, except it possesses 3 different output heads: one for each element of the PI. The lower-bound volume is obtained by computing the sum of the voxels segmented as lesions in the restrictive mask. Similarly, the upper-bound volume is obtained by summing lesion voxels in the permissive mask. Finally, the same process is applied to the balanced mask for the average volume estimation. Interestingly, any segmentation backbone (e.g. V-Net, Attention U-Net) can be turned into a triad-like version by simply duplicating the output convolution 3 times. In the following experiment, the Dynamic U-Net backbone is adopted to be consistent with the rest of the thesis. Interestingly, this has a very low impact on the complexity of the segmentation model, adding just a few thousand learnable parameters. Thus, the training and inference time is not increased. Then, a special loss function is defined to train TriadNet to output the three distinct masks. The key idea is to adjust the penalties applied to False Positive (FP) and False Negative (FN) voxels to obtain restrictive or permissive masks. To achieve this, we propose to employ the Tversky loss [29], as it provides a direct control on the trade-off between recall and precision.

The Tversky loss $T_{\alpha,\beta}$ is a variant of the standard soft Dice loss [15], with two additional hyperparameters α and β which respectively control the weighting of FP and FN predictions. With $\alpha = \beta = 0.5$, the Tversky loss is strictly equivalent to the standard Dice loss. Moreover, the Tversky loss can be implemented with a third hyper-parameter γ to penalize over-confident errors and thus favor calibration, as done in the Dice++ loss (see Equation II.6.3). This

yields to the following implementation of a Tversky++ loss ($\mathcal{T}^{+}$):

$$\mathcal{T}^{+}_{\alpha,\beta,\gamma} = \frac{\sum_{i=1}^{N} p_{1i} g_{1i}}{\sum_{i=1}^{N} p_{1i} g_{1i} + \beta \sum_{i=1}^{N} p_{0i} g_{1i}^{\gamma} + \alpha \sum_{i=1}^{N} p_{1i} g_{0i}^{\gamma}} = \frac{2TP}{2TP + \alpha FP^{\gamma} + \beta FN^{\gamma}} \tag{V.2.8}$$

The γ is set to 2 as done in the original Dice++ implementation [50]. In the following, the γ is omitted to simplify the notations. Writing p_{lower}, p_{mean} and p_{upper} the outputs of each head and y the ground truth segmentation, we defined the Triad loss as:

$$\text{TriadLoss} = \mathcal{T}^{+}_{1-\epsilon,\epsilon}(p_{lower,y}) + \mathcal{T}^{+}_{0.5,0.5}(p_{mean,y}) + \mathcal{T}^{+}_{\epsilon,1-\epsilon}(p_{upper,y}) \tag{V.2.9}$$

with ϵ a hyper-parameter in the range $[0, 0.5]$ controlling the penalties applied to FP and FN during the training of the lower and upper bound heads. With ϵ values close to 0, the lower and upper-bound masks will diverge significantly from the mean mask. In contrast, a value of ϵ close to 0.5 will increase the similarity between the three masks. In practice, we found that a value of $\epsilon = 0.2$ worked well in practice and provided sufficient variability to efficiently approximate the bounds of the intervals.

In summary, the mean decoder is trained using a conventional Dice Loss. To generate more restrictive masks (resulting in smaller volumes), the lower bound decoder is trained to minimize FP at the expense of a higher FN rate. Conversely, aiming for more permissive masks (resulting in larger volumes), the upper bound decoder is trained to minimize FN at the expense of a higher number of FP voxels. To visualize the differences between the three delineations, the three delineations can be superposed, as shown in Figure V.2.3. It provides interpretability concerning the uncertainty about the exact delineation of the target object (here, MS lesions).

V.2.4 Comparison with known approaches

The proposed TriadNet framework is compared with 3 sampling-based approaches: Confidence Thresholding (CT), Monte Carlo dropout (MC), and Test Time Augmentation (TTA). For sampling-based approaches, the same budget of $T = 50$ samples is allocated, which allows robust estimation of the mean and standard deviation.

Confidence Thresholding [81] is a straightforward approach to obtain PI's from the output probability estimates produced by a trained segmentation model. For each class, the probability map is binarized with progressively increasing thresholds. As the threshold increases, fewer voxels are segmented, thus the volume decreases. More specifically, T different thresholds uniformly distributed in the range $[0.05, 0.95]$ are used to binarize the probability maps, for each class. This method only works if the model is properly calibrated. Otherwise, if the produced probabilities are binary (extreme over-confidence), the sampled volumes will be homogenous and will not allow for a proper estimation of the standard deviation. Thus,

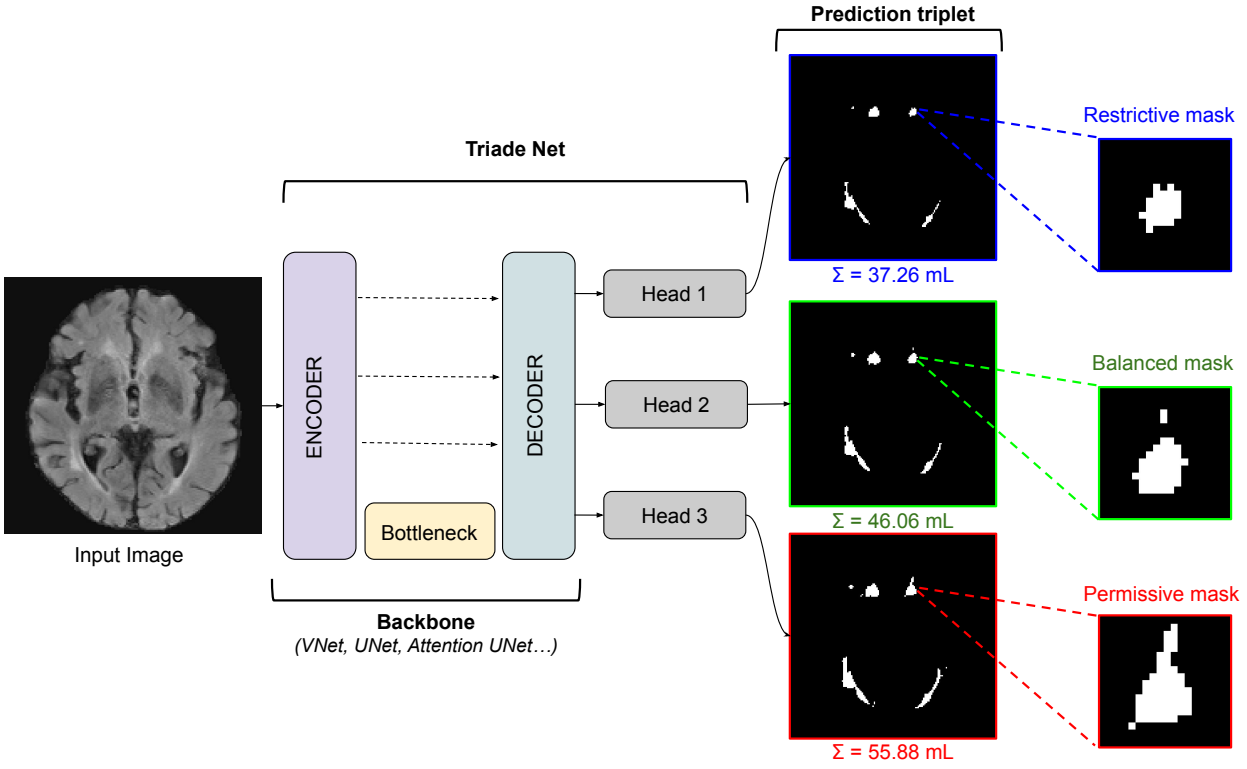


Figure V.2.2: Illustration of the TriadNet model. A segmentation backbone is enhanced by adding 3 output heads at the end of the decoder. Each one is responsible for the computation of an element of the predictive interval: the lower bound (blue), the upper bound (red), and the average volume (green). The lower-bound mask is more restrictive than the balanced mask, which is itself more restrictive than the upper-bound mask.

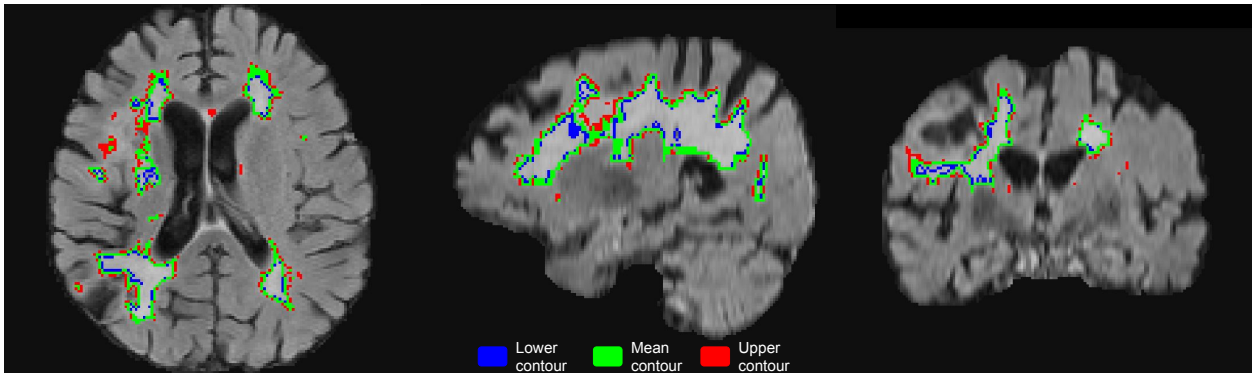


Figure V.2.3: Superposition of the lower, mean, and upper masks produced by TriadNet on a MS subject. Only the contours of the lesion delineations are presented. The restrictive mask (blue) is included within the mean mask (green), which is itself included in the permissive mask (red).

calibration is guaranteed by training a standard Dynamic U-Net with the $\mathcal{L}_3$ loss which has been introduced in the calibration benchmark (Section II.6.3).

Monte Carlo Dropout is used as a baseline to extract PI from a single model. This popular framework has been previously introduced in Section II.2.6. Here, it is implemented by performing T forward passes of the same input image with dropout activated to obtain 20 different estimations of volumes, from which the mean and standard deviation are extracted. A dedicated MC dropout Dynamic U-Net is trained to implement this approach, with a dropout rate of $p = 0.10$ after each convolution layer.

Test Time Augmentation, presented in Section II.2.10, is a popular way to generate various predictions for the same case, by generating variants of the input image. To implement the TTA baseline, T random augmentations for each input MRI are created using flipping, rotation, translation, and contrast augmentation with randomized parameters, implemented using the TorchIO Data Augmentation library [246]. Importantly, these augmentations do not modify the size of the target classes so that the estimation of volume is coherent from one augmentation to the other.

V.2.5 Evaluating the quality of predictive intervals

Evaluating the quality of predictive intervals requires the definition of dedicated metrics. First, PIs are constructed for a target level of coverage, set to 90% in these experiments. This means that empirically on the test set, 90% of the ground truth volumes should thus be contained in the intervals. Any deviation from this target level of coverage indicates a miscalibration of the intervals. A coverage **lower** than the target coverage (under-coverage) means that the PIs miss too many ground truth volumes and are thus probably too narrow. A coverage **larger** than the target coverage (over-coverage) means that PIs may be too large and thus, less informative. Second, PIs should be as narrow as possible to be informative. The second metric, W, is thus defined as the average distance between the lower and upper bounds of the intervals. An optimal PI should match the target level of coverage while being as narrow as possible. It should be noted that these two measures (coverage and with) are interdependent, as an increase in interval width leads to an increase in coverage. In practice, if the CP calibration is successful, all PI predictors will match the target coverage level. However, for poor PI predictors, the target coverage can only be reached using excessively large intervals to compensate for the poor bounds estimation. Thus, in CP studies, methods are generally ranked based on the size of the intervals [78]. Third, we report the Mean Average Error (MAE) between the estimated volume and the ground truth one. Finally, the average inference time to compute the PI is measured, to evaluate the compatibility of each method with industrial applications.

V.2.6 The importance of the size of the calibration dataset

A key element of the proposed PI framework is the conformal procedure that aims at calibrating the intervals so that they match the target coverage. In the literature, conformal calibration is generally carried out using the validation dataset. Yet, in 3D medical image applications, the validation dataset is usually small as the overall number of available annotated cases is scarce.

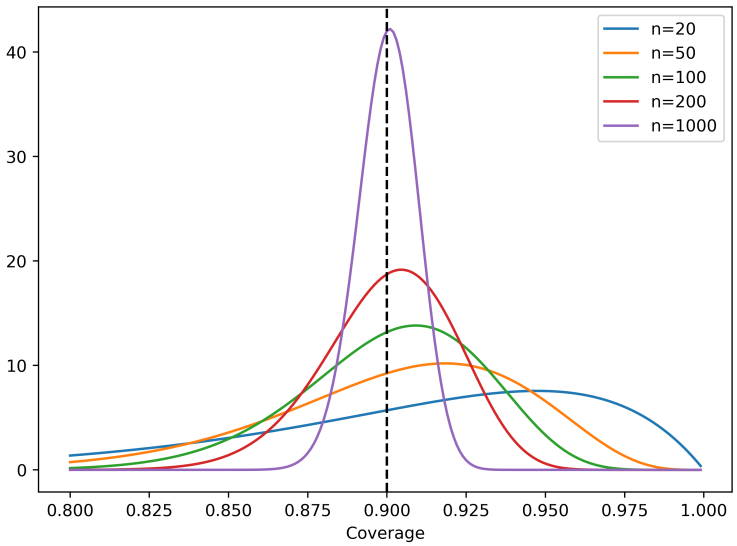


Figure V.2.4: Theoretical distribution of coverages for varying sizes of calibration datasets. The black vertical dashed line indicates the target coverage of 0.90.

For example, in the voxel-level experiments, $n = 21$ images composed the validation dataset for cross-sectional MS models and $n = 30$ for glioblastoma segmentation models. This is not in line with the standard conformal methodology that relies on calibration datasets that are several magnitudes larger. Typically, a size of $n = 1000$ is proposed as a guideline number for 2D image classification or regression tasks [76]. To analyze in more detail the importance of the calibration dataset size for conformal procedures, an analytic form of the distribution of coverages has been proposed by Vovk et al. [323]. It takes into account the user-defined error rate α and the size n of the calibration dataset $\{(X_i, Y_i)_{i=1}^n\}$. It is formulated as:

$$\mathcal{P}(Y_{test} \in \Gamma_\alpha(X_{test}) | \{(X_i, Y_i)_{i=1}^n\}) \sim \text{Beta}(n + 1 - l, l) \tag{V.2.10}$$

where $l = \lfloor (n + 1)\alpha \rfloor$. This theoretical distribution is plotted for a target coverage of 0.90 in Figure V.2.4, for increasing sizes of calibration sets: 20, 50, 100, 200, and 1000. It appears that for $n = 1000$, the usual guideline, the coverage is between 0.88 and 0.92, thus indeed extremely close to the target coverage of 0.90. However, for smaller calibration datasets, the distributions of coverage are more dispersed around the 0.90 value. For $n = 100$, a coverage between 0.85 and 0.95 can be expected, and the dispersion increases heavily with lower values of n. Additionally, as the analytic form follows a Beta distribution, the mode of the distribution can be computed, corresponding to the most likely value of the distribution (*i.e.* the peak in the probability distribution function). It is computed as $\frac{n-l}{n-1}$. The resulting values for each value of n are provided in Table V.1. For small calibration datasets ($n \leq 100$), the most likely value is superior from 0.90, and the deviation increases as the number of calibration samples shrinks. Thus, for small calibration datasets, it is expected that the actual coverage will be higher than the target one (0.90).

n	Mode
20	0.947
50	0.918
100	0.909
200	0.905
1000	0.901

Table V.1: Most likely coverage for different sizes of the calibration dataset. This corresponds to the mode of the analytic Beta distribution of coverages.

Thus, these theoretical results are not encouraging for our 3D medical image setting, with validation datasets typically containing a few dozen images. Using the original validation datasets would result in imprecise conformal calibration procedures. To alleviate this, a dedicated train/calibration/test stratification is adopted in this chapter to allow for an increase in the number of calibration data points. For the cross-sectional MS experiment, the set of 219 in-distribution images is distributed into 120 for training, 50 for validation, and 49 for a testing set (which is kept identical to the one used in the previous thesis experiments). For glioblastoma segmentation, the 1133 in-distribution images are distributed into 679 for training, 227 for calibration and 227 for testing.

Then, the recommended protocol to evaluate the quality of conformal procedures is adopted [76]. Recall that $N_{\text{test}} + N_{\text{val}}$ data points are available to calibrate and test the PIs. The coverage guarantee in Equation V.2.1 says that the coverage should be at least $1 - \alpha$ in **average**, for different realizations of the validation/test stratification. Thus, to robustly evaluate the conformal procedure, the experiment is reproduced for $R = 15000$ trials, with a random split of the $N_{\text{test}} + N_{\text{val}}$ datapoints into validation and test. This can be implemented efficiently by caching the predictions. This process allows to estimate the average coverage, interval width, MAE, as well as confidence estimates (Standard Error on the Mean, SEM) for these metrics.

V.2.7 Exchangeability of calibration and test datapoints

The conformal framework makes no assumptions about the model or the data. Yet, it is still based on a strong hypothesis that calibration and test data points are exchangeable. To follow the definitions used throughout this thesis, there should not be any **domain shifts** between the calibration and test datasets. This is because the conformal strategy is rooted in the principle that the errors on the calibration and test datasets have the same magnitude, which allows to achieve the desired coverage on the test dataset using the calibration samples. However, in most real-world applications and especially medical imaging, domain-shifts are extremely common. Thus, there is no guarantee that a conformal strategy calibrated on images from Hospital A will achieve the desired coverage on images from Hospital B. To analyze this phenomenon, domain-shift scenarios are explored. For MS lesions segmentation, we use the previously introduced MSLUB and 1.5 Tesla test datasets (Section II.1). For glioblastoma segmentation, we rely on the Sub-Saharan Africa test dataset (Section II.6.1). In this setting, the intervals are calibrated on in-distribution images, and the coverage is

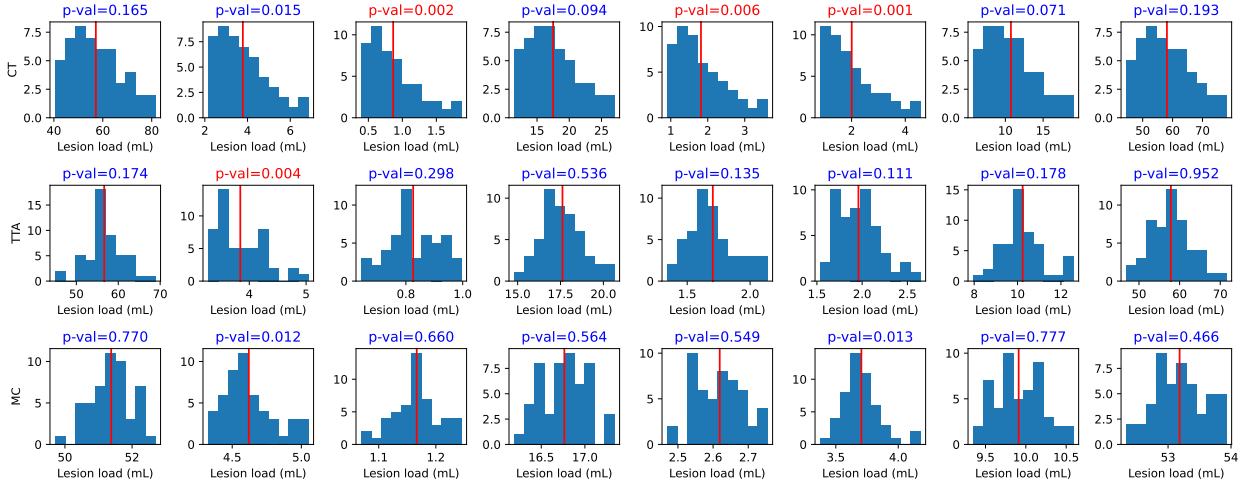


Figure V.2.5: Distribution of volumes for sampling-based approaches (CT, TTA, and MC) for the first 8 validation images. The red line indicates the average of the sampled volumes. In each case, a Shapiro-Wilk test is performed to test the normality hypothesis. The corresponding p-value is indicated on top of each plot. It is displayed in blue when the hypothesis cannot be rejected (p-value ≥ 0.01). Otherwise, it is indicated in red.

then estimated on a separate domain-shift dataset. The same evaluation metrics (coverage, average interval width, MAE) are then reported.

V.2.8 Application to lesion load estimation in MS patients

The first application investigated is the cross-sectional MS lesions segmentation task, which has already been investigated in detail for voxel-level, lesion-level, as well as quality control experiments. Here, the task is to compute the total lesion load of the patient, corresponding to the volume of identified white matter hyperintensities.

For sampling-based approaches (CT, TTA, MC), the distribution of the sampled volumes is presented for the first 8 validation images in Figure V.2.5. For each subject, the Shapiro-Wilk statistical test [324] is performed to verify if the distribution follows a normal distribution. More specifically, the null hypothesis is that the sampled volumes come from a normal distribution. Small p-values (e.g. ≤ 0.01) are evidence of departure from normality. It appears that in most cases, the normality hypothesis is indeed verified with a 0.01 level of significance. When it is not valid, it is associated with low lesion loads. In this setting, the distribution is indeed not Gaussian as the volume cannot be negative, leading to skewed distributions.

Then, the PI metrics (coverage, width, MAE) for each method are presented in Table V.2. Figure V.2.6 presents a visualization of the PIs for each method on the three test datasets. After calibration, all methods exhibit close coverages on in-distribution test data, with an average empiric coverage close to 0.92. Interestingly, this exactly corresponds to the theoretical mode of the analytic Beta distribution of coverages for a calibration dataset

of $n = 50$ samples. Thus, the experimental results perfectly match the expected results discussed in the previous section. All methods are thus equivalent in terms of coverage, which essentially validates the correctness of the conformal calibration procedure. The width of the intervals, however, is more heterogeneous across methods. More particularly, MC and TTA provide larger intervals as compared to CT and TriadNet. This can be observed in Figure V.2.6. This may be because the sampled volumes obtained via TTA or MC exhibit a low standard deviation. Thus, the $\hat{q}$ are generally much larger for MC and TTA than for CT, which produces more diverse sampled volumes. This phenomenon can be observed in Figure V.2.5. Then, intervals produced by TriadNet are slightly narrower than the ones produced by CT. In terms of MAE (estimation of the lesion load), TriadNet achieves the best result in the in-distribution test split. Finally, in terms of inference speed, the MC and TTA are largely slower than CT and TriadNet. For both sampling approaches, it takes roughly one minute to generate the interval. This is because $T = 50$ inferences are needed for MC and TTA. Contrarily, CT is very efficient as a single inference is needed. Sampling is then applied to the predicted probability map, which is computationally efficient. TriadNet, which does not require any sampling, is slightly faster (on average 1.35 seconds per MRI).

Regarding domain-shift datasets (MSLUB and 1.5 Tesla datasets), it can be noticed that there is a loss of coverage for each method. More particularly, the coverage approaches 100% for MSLUB for CT, MC, and TriadNet. This indicates that for MSLUB data, narrower intervals could be enough to encompass 90% of the ground truth volumes. For the 1.5T dataset, the coverages of CT and TTA fall around 85% on average, which shows the opposite tendency: larger intervals are needed for this dataset. This demonstrates the expected loss of guarantees due to the non-exchangeability of the data, which CP fails to tackle.

		Coverage (%)		W (ml)		MAE (ml)		Dice (%)		Time (s)
Method	**Dataset**	μ	SEM	μ	SEM	μ	SEM	μ	SEM	μ
CT	ID	92.03	5.29	14.92	2.18	3.55	0.43	77.69	0.95	1.42
	MSLUB	99.95	0.53	**12.50**	**1.66**	2.20	-	69.39	-	
	1.5T	86.32	6.97	**10.38**	**1.38**	3.69	-	67.58	-	
MC	ID	91.98	5.45	15.38	2.31	3.79	0.83	77.32	1.03	53.22
	MSLUB	100.00	0.00	23.16	2.33	1.82	-	**69.81**	-	
	1.5T	100.00	0.00	14.47	1.45	2.67	-	66.21	- X	
TTA	ID	92.01	5.34	17.64	3.46	3.12	0.42	77.54	0.94	62.93
	MSLUB	94.39	2.26	14.82	2.34	**1.44**	-	68.70	-	
	1.5T	84.18	9.10	10.56	1.67	3.04	-	**68.00**	-	
Triad	ID	92.06	5.34	**14.46**	**1.58**	**3.08**	**0.46**	**77.93**	0.97	**1.35**
	MSLUB	99.98	0.54	13.23	1.53	1.87	-	68.98	-	
	1.5T	96.04	1.48	11.48	1.48	**2.51**	-	67.89	-	

Table V.2: Quality of predictive intervals for cross-sectional MS lesions segmentation. Intervals are calibrated for a target coverage of 90%. Results are averaged over $R = 15000$ trials. W: average interval width. MAE: Mean Average Error on volume estimation. ml: milliliter. SEM: Standard Error on the Mean. ID: in-distribution test set. Note that the domain-shift test datasets (MSLUB and 1.5 Tesla) are fixed for each trial, hence the SEM is not estimated.

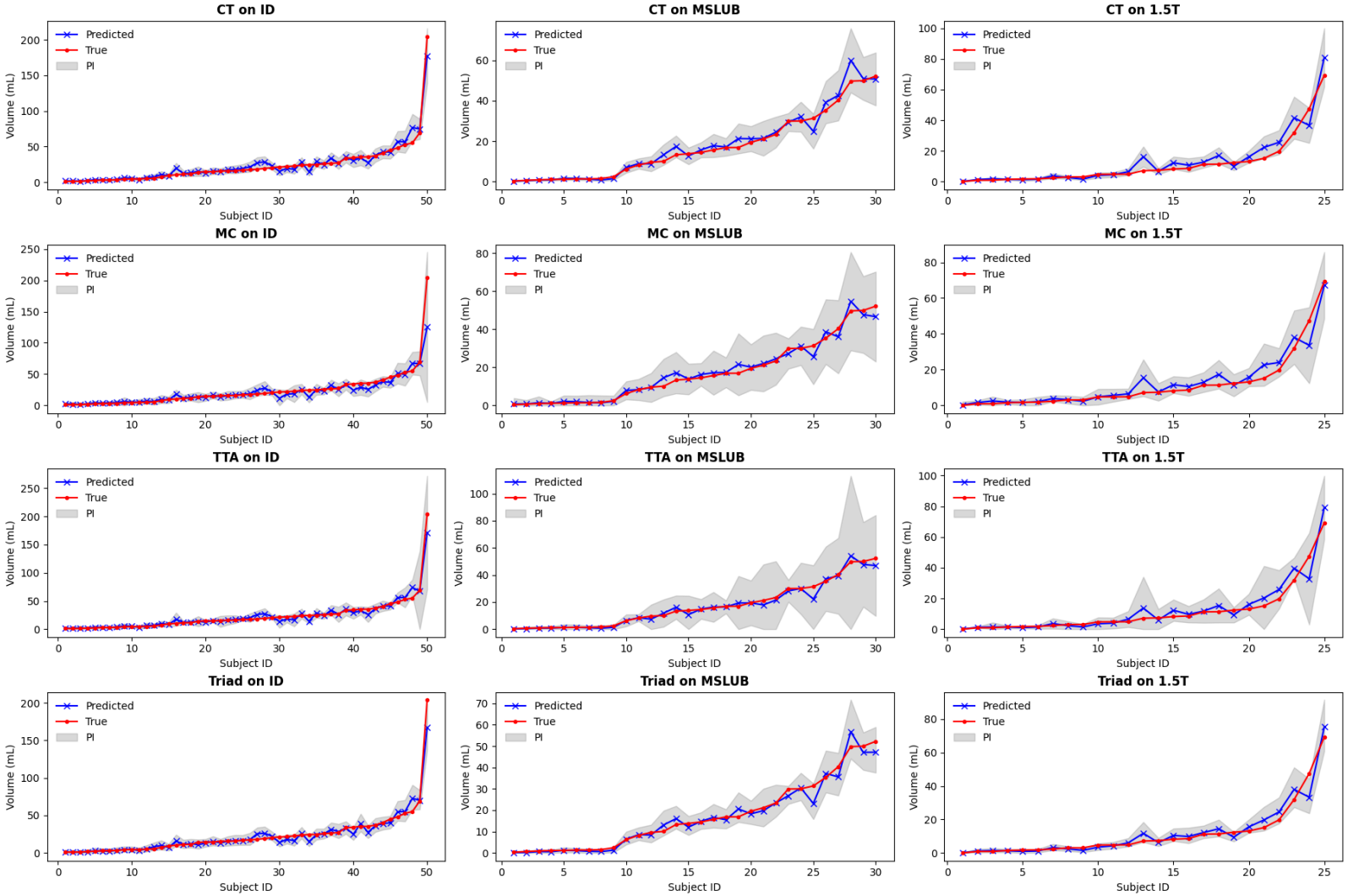


Figure V.2.6: Visualization of the predictive intervals for each method and on the three Multiple Sclerosis test datasets, for a randomly selected trial. Volumes are ordered from the smallest to largest for easier visualization. The red line indicates the ground truth volumes and the blue line indicates the predicted ones. The grey area indicates the predictive interval for each subject.

V.2.9 Application to brain tumor volume estimation

The second investigated application is the segmentation of glioblastoma in 3 tissue classes: Necrotic (label 1), Edematous (label 2), and GD-enhancing tumor (label 3). Intervals are estimated for each class independently. In practice, this implies that a distinct corrective value $\hat{q}_i$ is fitted for each class i on the calibration dataset (N=227 samples). Interval metrics are provided in Table V.3.

As for the cross-sectional MS experiment, the experimental results follow with great precision the theoretical results, with an average coverage between 90.7% and 90.8% for each class and method on the in-distribution dataset. The mode of the Beta distribution for a calibration dataset of $n = 200$ is 90.5 (Table V.1), thus very close to the empirical coverages. Now, in terms of interval width, MC and TTA also produce larger intervals than CT and TriadNet, which confirms the trend observed for MS lesions. The narrower intervals are provided by TriadNet for the Necrosis and GD-enhancing classes, while CT is the best approach for edematous. The same ranking of methods is observed for the MAE metric. On domain-shift data (SSA dataset), there is an important loss of coverage regarding the GD-enhancing tumor class, with coverages falling to around 50% for CT, TTA, and TriadNet. This phenomenon is clearly visible in Figure V.2.8, where it can be observed that the predicted volumes of GD-enhancing tumors (blue) are far from the true ones (red). This miscalibration is also observed for the necrosis and edematous classes (label 1-2) although more moderate. This is also linked to an increase in the MAE for each class and method. As the volumes are less accurately estimated, the quality of the PI degrades. Finally, in terms of inference speed, TriadNet produces the segmentation and associated intervals for the three classes in only 0.28 second. Note that this is faster than for the MS experiment as inference is carried in a fully 3D approach, while the MS models are based on patches. CT is slower (2.57 second on average) as the confidence thresholding is repeated for each class independently. Then, MC and TTA exhibit extended inference times. TTA is particularly slow for this brain tumor experiment as each MRI sequence (T1, T2, FLAIR, T1ce) has to be altered, for each inference.

			Coverage (%)		W (ml)		MAE (ml)		Dice (%)		Time (s)
Method	**Dataset**	**Class**	μ	SEM	μ	SEM	μ	SEM	μ	SEM	μ
CT	ID	1	90.78	2.75	11.37	1.35	3.12	0.46	79.63	1.21	
		2	90.82	2.73	**31.03**	**3.12**	7.54	0.52	**85.65**	0.59	
		3	90.78	2.71	8.06	0.76	1.86	0.20	**86.36**	0.89	
	SSA	1	81.97	2.54	44.57	4.50	14.51	-	54.20	-	2.57
		2	90.09	3.47	64.37	6.28	17.29	-	77.39	-	
		3	50.21	0.89	16.23	1.39	12.14	-	72.14	-	
MC	ID	1	90.79	2.68	13.70	1.49	3.56	0.47	78.22	1.24	
		2	90.82	2.69	38.94	3.27	9.75	0.61	84.93	0.63	
		3	90.80	2.71	8.04	0.42	2.15	0.19	85.01	0.94	
	SSA	1	95.94	1.62	47.47	2.77	10.39	-	58.97	-	
		2	92.57	2.13	80.22	6.06	21.72	-	78.05	-	12.76
		3	62.60	1.37	20.19	0.75	8.93	-	**79.0**	-	
TTA	ID	1	90.76	2.70	20.64	2.93	3.29	0.47	79.16	1.20	
		2	90.77	2.75	44.09	4.61	7.55	0.55	85.51	0.60	
		3	90.83	2.67	12.02	1.44	1.87	0.20	86.06	0.80	
	SSA	1	89.05	2.75	64.95	7.05	14.47	-	53.73	-	
		2	78.54	2.26	66.33	6.25	18.24	-	77.31	-	45.3
		3	50.94	3.87	13.54	1.47	12.24	-	72.08	-	
Triad	ID	1	90.78	2.71	**9.92**	**0.99**	3.10	0.46	**79.85**	1.18	
		2	90.76	2.70	32.36	2.85	8.22	0.57	85.19	0.60	
		3	90.79	2.71	**7.70**	**0.48**	1.73	0.19	86.02	0.89	
	SSA	1	77.01	2.32	22.03	1.08	10.57	-	**60.00**	-	
		2	84.15	2.26	51.58	3.00	19.81	-	**78.30**	-	**0.28**
		3	48.87	2.66	14.36	0.42	9.79	-	74.82	-	

Table V.3: Quality of predictive intervals for multi-class glioblastoma segmentation for a target coverage of 90%. Results are averaged over $R = 15000$ trials. W: Average interval width. MAE: Mean Average Error on volume estimation. ml: milliliter. SEM: Standard Error on the Mean. ID: in-distribution test dataset. SSA: Sub-Saharan Africa test dataset. Note that the domain-shift test dataset (SSA) is fixed for each trial, hence the SEM is not estimated.

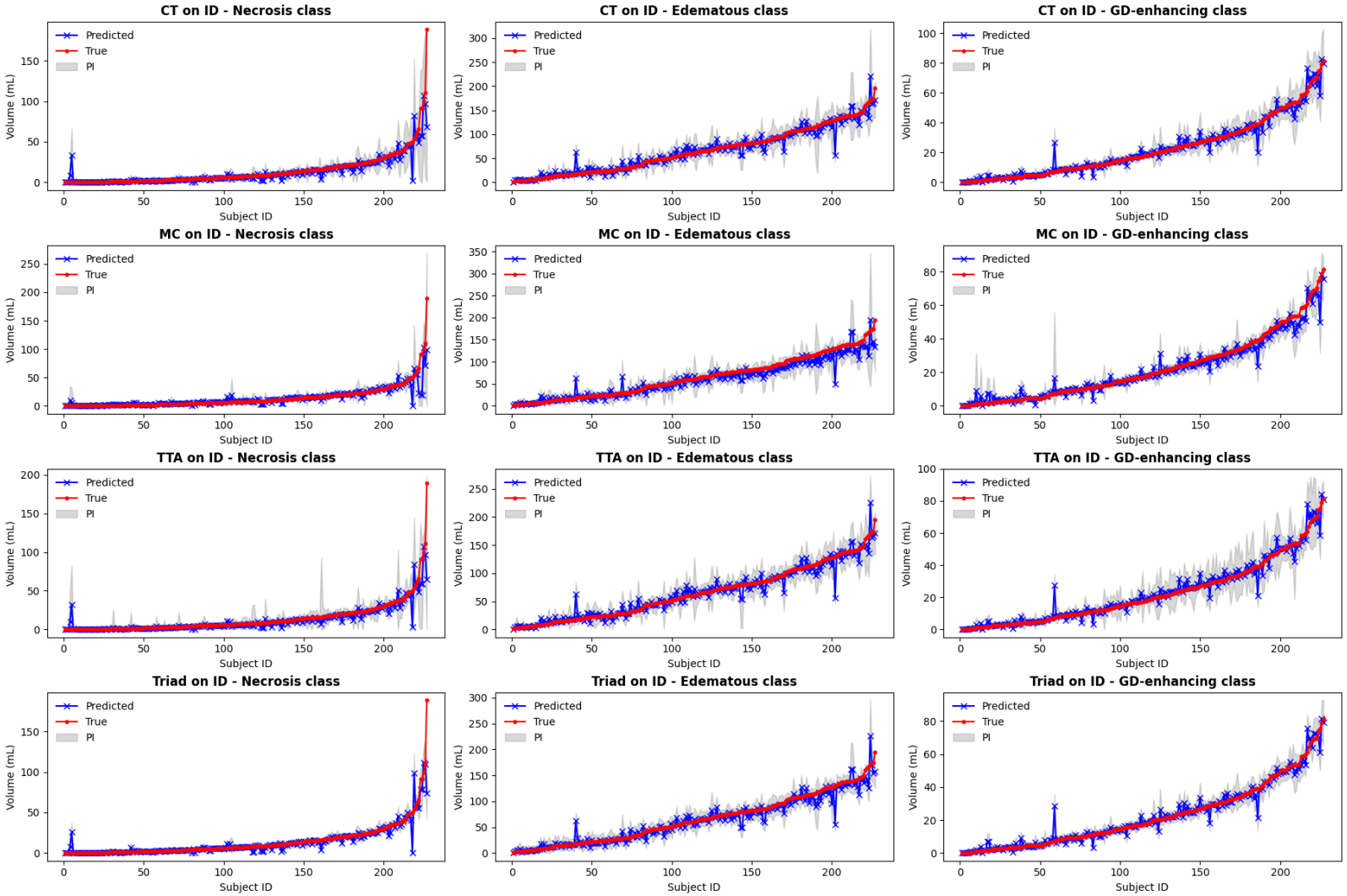


Figure V.2.7: Visualization of the predictive intervals for each method on the in-distribution brain tumor test dataset, for a randomly selected trial. Volumes are ordered from the smallest to largest for easier visualization. The red line indicates the ground truth volumes and the blue line indicates the predicted ones. The grey area indicates the predictive interval for each subject.

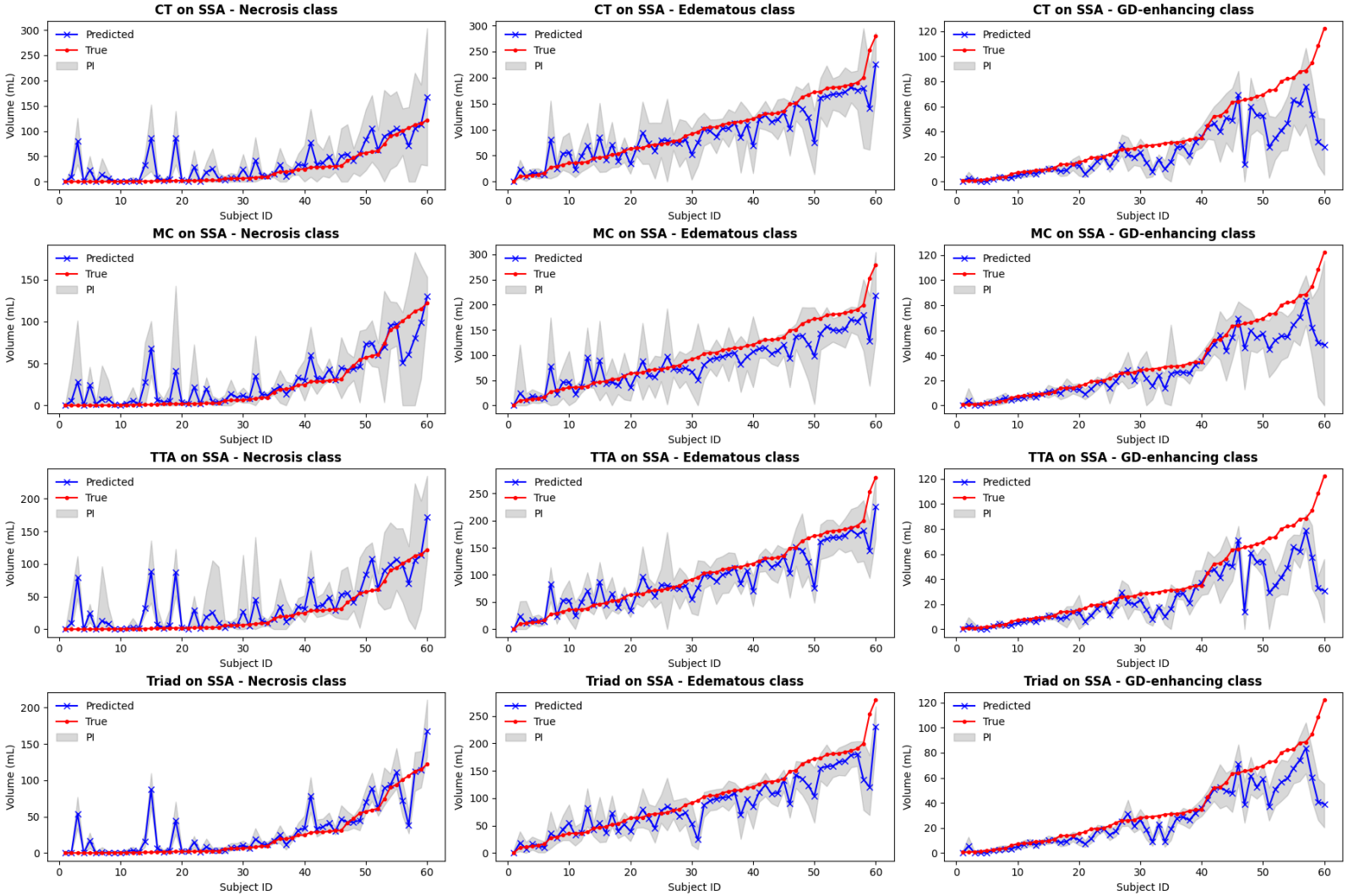


Figure V.2.8: Visualization of the predictive intervals for each method on the Sub-Saharan Africa (SSA) brain tumor test dataset, for a randomly selected trial. Volumes are ordered from the smallest to largest for easier visualization. The red line indicates the ground truth volumes and the blue line indicates the predicted ones. The grey area indicates the predictive interval for each subject.

V.2.10 Discussion on TriadNet

In this first part of the chapter, CP is investigated to calibrate PIs on lesion volumes. It can be applied to both sampling-based and direct PI estimation. In the latter direction, we propose a novel approach called TriadNet that directly predicts the three elements of the interval: lower bound, volume estimation, and upper bound. On two applications (MS lesions and glioblastoma), we show that CP performs an accurate calibration of the intervals to match the target coverage level on ID test data. Regarding inference time, TriadNet is extremely efficient as it does not require any sampling, contrary to the other proposed baselines. It could be argued however that the longer inference time of sampling-based approaches (especially MC and TTA) is due to the choice of T the number of sampling steps (set to $T = 50$ in these experiments). Indeed, it can be expected that the inference time of sampling-based approaches is linear with respect to T. However, smaller values of T would result in less precise estimations of the mean and standard deviation of the volumes.

Although very efficient on ID test data, our experiments show that when the test distribution is shifted (MSLUB, 1.5 Tesla, or SSA dataset), then there is a loss of the coverage guarantee. This appears as a major drawback for medical applications where this type of shift is common. It motivates the exploration of an enhancement of the CP framework, called weighted CP, explored in the rest of the chapter.

V.3 Perspectives on weighted conformal prediction to tackle domain shifts

V.3.1 Mathematical framework

As previously mentioned, the conformal framework is based on the hypothesis that calibration and test data points are exchangeable. As a consequence, its accuracy degrades when the test distribution shifts (MSLUB, 1.5 Tesla, and SSA datasets). This motivated the extension of the CP procedure in the setting of non-exchangeable data [76, 84, 325]. This framework is introduced below.

In the baseline formulation of CP, each calibration sample contributes equally to the overall conformal procedure. The core concept of weighted conformal prediction (WCP) is to reweight calibration conformal scores according to their likelihood under the observed test distribution. As an effect, we obtain a pseudo-calibration dataset that more accurately matches the target one. This is achieved by estimating the density ratio $w = dP_{\text{test}}/dP_{\text{train}}$ for each calibration and test sample. In practice, writing $(X_1, ..., X_n)$ the n calibration samples and x the fresh test point, importance weights are computed as:

$$p_i^w(x) = \frac{w(X_i)}{\sum_{i=1}^{N} w(X_j) + w(x)} \tag{V.3.1}$$

$$p_{\text{test}}^w(x) = \frac{w(x)}{\sum_{i=1}^{N} w(X_j) + w(x)} \tag{V.3.2}$$

Essentially, $p_i^w(x)$ is large when the calibration sample X_i is likely under the test distribution. Then, the corrective value $\hat{q}$ can be reframed as the $1-\alpha$ quantile of the weighted distribution:

$$\hat{q}(x) = \inf\left\{ s_j : \sum_{i=1}^{j} p_i^w(x)\mathbf{1}\{s_i \leq s_j\} \geq 1-\alpha \right\} \tag{V.3.3}$$

Note that when all weights are equal to $\frac{1}{n+1}$, the standard CP procedure is recovered and we end up choosing the $\lceil (n+1)(1-\alpha) \rceil$ quantile. It can also be noted that the baseline CP procedure produces an identical corrective value $\hat{q}$ for all test samples. Contrarily with WCP, there is now a dependence on $w(x)$, the density ratio estimated for the test sample x. However, this weighted formulation is not free. Indeed, it acts as an importance sampling protocol, alleviating more weights to calibration samples that look like test samples. As a downside, there is a reduction of the effective sample size [326] that can be estimated through the heuristic:

$$n_{eff} = \frac{[\sum_{i=1}^{N} |w(X_i)|]^2}{\sum_{i=1}^{N} |w(X_i)|^2} \tag{V.3.4}$$

As an effect, when weights deviate significantly from 1, the size of the calibration dataset virtually shrinks, which increases the variance of the CP procedure [325].

Importantly, this formulation was initially proposed to tackle covariate shifts. A covariate shift is intuitive when the input data correspond to covariate vectors. In this setting, the covariate shift manifests by one or several covariates that have different distributions in the calibration and test datasets (e.g. the sex or age of the patient). In our setting, the covariates are instead high-dimensional medical images and covariate shifts can be more ambiguous to interpret. In practice, shifts in the image space can occur because of variations in the image acquisition parameters or population demographics [327]. Importantly, the predictive task (e.g. segmentation of MS lesions or glioblastoma) must remain the same, meaning that there is no label shift between the calibration and test samples.

There are however several challenges in implementing WCP in our setting. Indeed, it supposes that i) the density ratio between the calibration and test distributions can be estimated and ii) P_{test} is absolutely continuous with respect to P_{train}, which is akin to say that the domain shift is not too important. Informally, when there are no samples in the calibration set that

are representative of the test dataset, then there is no hope of accounting for the covariate shift [328]. Second, unless in synthetic settings, the density ratio is never exactly known and must be estimated. A flourishing literature can be found for density ratio estimation. Popular approaches include training a classifier to distinguish between training and test distributions [329], moment matching [326], or ratio matching [330]. In the original proposal of WCP, Tibshirani et al. [325] propose to use an auxiliary classifier that only requires that unlabeled samples from the test distribution are available during the calibration step. The idea is to train a probabilistic classification model to classify samples between the training and test distributions. That is, writing $X_1, ..., X_n$ and $X_{n+1}, ..., X_{n+m}$ the training and test data points, one can form a classification dataset composed of the pairs $\{X_i, C_i\}$ where $C_i = 0$ for $i = 1, ..., n$ and $C_i = 1$ for $i = n + 1, ..., n + m$. Writing $\widehat{p}(x) = \mathcal{P}(C = 1|X = x)$ the probability predicted by a classifier model trained on the $\{X_i, C_i\}$ dataset that the input sample x belongs to the test distribution, the weight function can be expressed as:

$$\widehat{w}(x) = \frac{\widehat{p}(x)}{1 - \widehat{p}(x)} \tag{V.3.5}$$

However, this approach has several limitations. First, it requires access to a sufficient amount of calibration and test samples to allow for a supervised classification strategy. More particularly, having access to only one or several test images is not enough to allow for training the auxiliary classifier. Second, it heavily builds on the calibration of the auxiliary classifier as the predicted probabilities are used to compute the weights [328]. Moreover, training the classifier is efficient when the input data is a feature vector [325], but becomes cumbersome when dealing with high-dimensional medical images. In this setting, the dedicated classification approach would be the training of a deep learning CNN, requiring numerous examples of both classes (calibration and test). Moreover, recall that the training of the auxiliary classifier should be performed **during** the CP procedure to allow for the weight estimation. Incorporating the training of a CNN in the CP procedure is thus highly inefficient. Finally, if the domain shift is important, the weights estimated by the classifier will likely diverge, as can be easily stated from Equation V.3.1 when $\widehat{p}(x)$ converges to 1. As a conclusion, this classification task is computationally too costly when dealing with 3D medical images, and it may fail to provide accurate weights if the auxiliary classifier is miscalibrated or if the classification task is too easy, typically when the domain shift between calibration and test samples is high. Building on these limitations, we next investigate a more efficient approach making use of the latent representations extracted by the deployed segmentation model.

V.3.2 Investigation of an efficient approach to weight estimation in 3D MRI

As training the auxiliary classifier directly from the input images is too costly, more efficient approaches have to be investigated. In Chapter 3, it was shown that the latent representations extracted by trained segmentation models are efficient in detecting shifts in input images. One idea would be to train the classifier on these low-dimensional latent representations

directly, instead of on the high-dimensional 3D MRIs. To test this framework, we collect the activations of the penultimate convolution layer ϕ, which has 32 convolution kernels. The feature map has a shape of $32 \times H \times W \times D$, where H, W, and D are the spatial dimensions of the MRI. This feature map is reduced to a covariate vector $z(x)$ of dimension 32 by performing an averaging over the spatial dimensions, as done in Chapter 3 to compute the Mahalanobis distance:

$$z(x) = \frac{1}{H}\frac{1}{W}\frac{1}{D}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{d=1}^{D}\phi(x)(h,w,d) \tag{V.3.6}$$

This approach allows to perform classification on a compressed representation of the input MRI, which alleviates the curse of dimensionality. As a result, the classification can be performed efficiently during the WCP procedure and in practice should require fewer training examples than CNNs.

V.3.3 Proof-of-concept on a synthetic dataset with controlled covariate shift

To prove the relevancy of the proposed approach, we first rely on a synthetic setting allowing us to control covariate shift precisely. It allows us to have access to oracle values for the density ratio, as the shift between train and test distributions is closely controlled. Moreover, synthetic data allow the creation of large datasets, which is particularly interesting for the accuracy of the conformal procedure.

V.3.3.1 Synthetic Data description

The task that we propose here is the segmentation of spheres inside cubic volumes of shape $32 \times 32 \times 32$. The uncertainty task is to compute a predictive interval for the volume of each sphere. The covariate of interest is the signal-to-noise ratio (SNR) between the background of the image and the foreground spheres. For this experiment, the testing dataset will contain images with lower SNRs than the training images, which emulates a covariate shift setting. The SNR is defined here as:

$$\text{SNR} = \frac{\mu_{\text{spheres}}}{\sigma_{background}} \tag{V.3.7}$$

where μ_{spheres} is the average intensity of the foreground spheres, and $\sigma_{background}$ is the standard deviation of the background noise.

In each image, a sphere is generated by picking a random center and a diameter in the range $[6, 14]$ mm. To increase variability, the spheres are deformed using TorchIO's Elastic deformation function [246]. Then, we uniformly sample a random target SNR from the range $[1, 20]$ (low SNR corresponds to noisy images, in principle harder to segment). The next step is to convert the binary mask into an intensity image matching the predefined SNR.

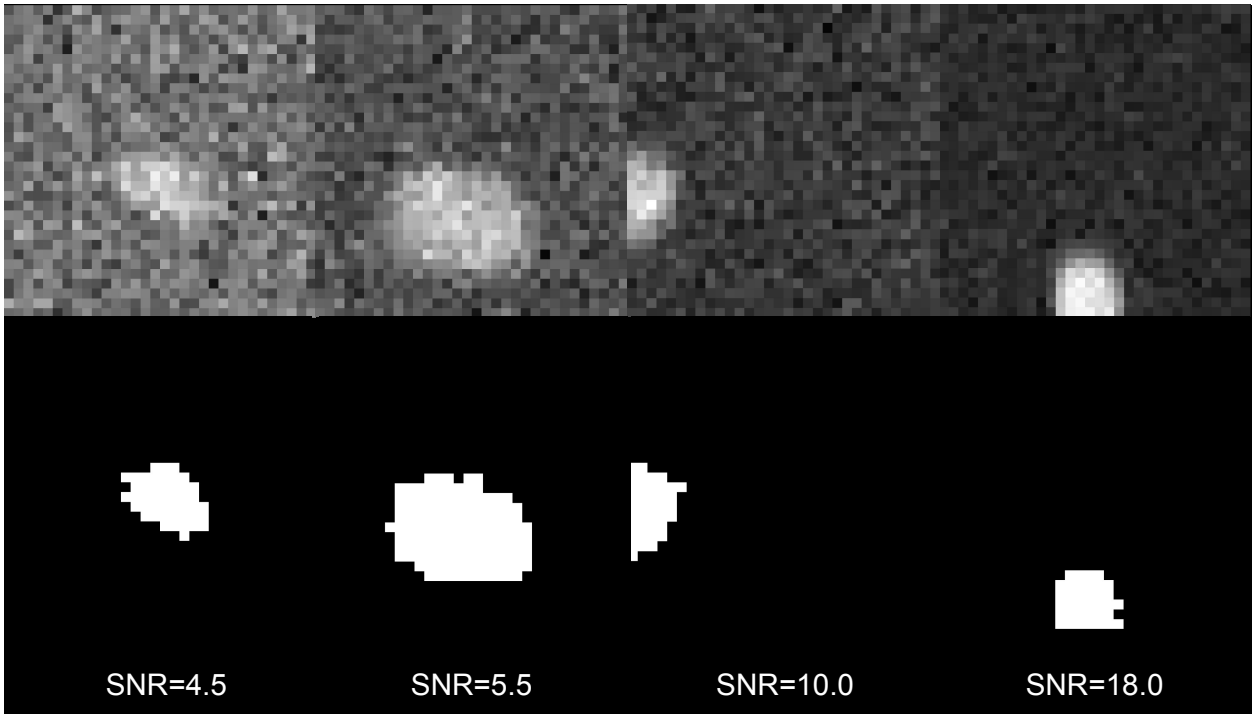


Figure V.3.1: Illustration of synthetic images (top row) with varying Signal-to-Noise ratios (SNRs) and associated ground truth (bottom row).

This is achieved by setting the background intensity to 0, the sphere intensity to 1, and then injecting an additive random Gaussian noise to the image following $\mathcal{N}(0, \frac{1}{SNR})$. As a result, the generated image has an SNR that matches the target one. Several examples of synthetic images with varying SNRs are presented in Figure V.3.1.

We generate a total of $N = 4000$ synthetic images with varying SNRs. We then split this dataset into an in-distribution split (3000 images) containing images with high SNRs, and a shifted dataset (1000 images) containing images with lower SNRs. To select the 1000 shifted test images, a sampling probability is assigned to each image X_i, proportional to :

$$w(X_i) = \frac{1}{\mathrm{SNR}(X_i)} \tag{V.3.8}$$

This allows the non-uniform sampling of images to favor low SNRs in the shifted test set. Because of this non-uniform sampling, we have $dP_{test}(X_i) \propto \frac{1}{\mathrm{SNR}(X_i)} dP_{train}(X_i)$. As a result, we can consider Equation V.3.8 as the oracle weights. Then, the in-distribution split corresponds to the remaining 3000 images. The densities of SNRs in both splits are presented in Figure V.3.2. The in-distribution dataset is further split into training, calibration, and in-distribution test parts, with 1000 images each.

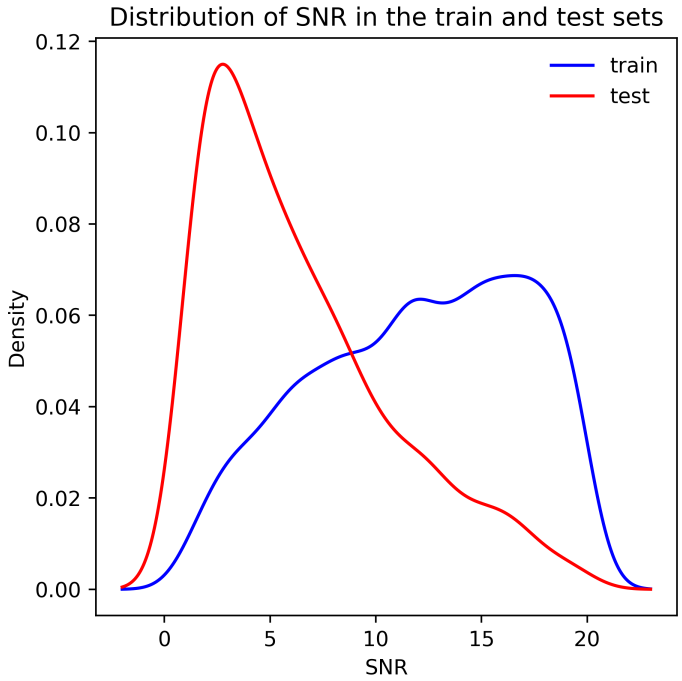


Figure V.3.2: Distribution of signal-to-noise ratios in the training and test synthetic datasets, allowing to emulate a covariate shift.

V.3.3.2 Experimental Setting

A TriadNet model with a Dynamic U-Net backbone is trained on the training split for 20 epochs, using the Triad loss and the ADAM optimizer [37] with a learning rate of 2×10^{-4}. Then, the WCP procedure is performed with a target coverage of 95%. We use here a higher coverage level (95% versus 90% for the real-world MRI tasks) because the model achieves high segmentation performance on the synthetic images, and the calibration dataset is larger (1000 images) allowing for higher coverage levels. Three ways of estimating the density ratio are investigated:

- **W-Oracle** uses the oracle weights $w(X) = \frac{1}{\text{SNR}(X)}$. Note that in our synthetic setting, we have access to Oracle weights as we control the data generation process. In real-world problems, we have no oracle knowledge of the density ratio.
- **W-Image** uses a CNN classifier to predict $\widehat{p}(x) = \mathcal{P}(C = 1|X = x)$ the probability that the input image x belongs to the shifted test distribution. To do so, a shallow CNN is used with three convolutional layers (with 2, 4, and 8 kernels, respectively), followed by a fully-connected layer. The model is trained on a separate dataset of 4000 images following the same generation process as the one previously presented. The Cross-entropy loss (Equation I.2.3) is used to perform training. To mitigate overfitting, a dropout rate of 50% is used after each layer. The predicted probabilities are clipped in the range $[0.01, 0.99]$ to avoid infinite weights.
- The **W-Latent** approach uses the latent representation of the image (Equation V.3.6),

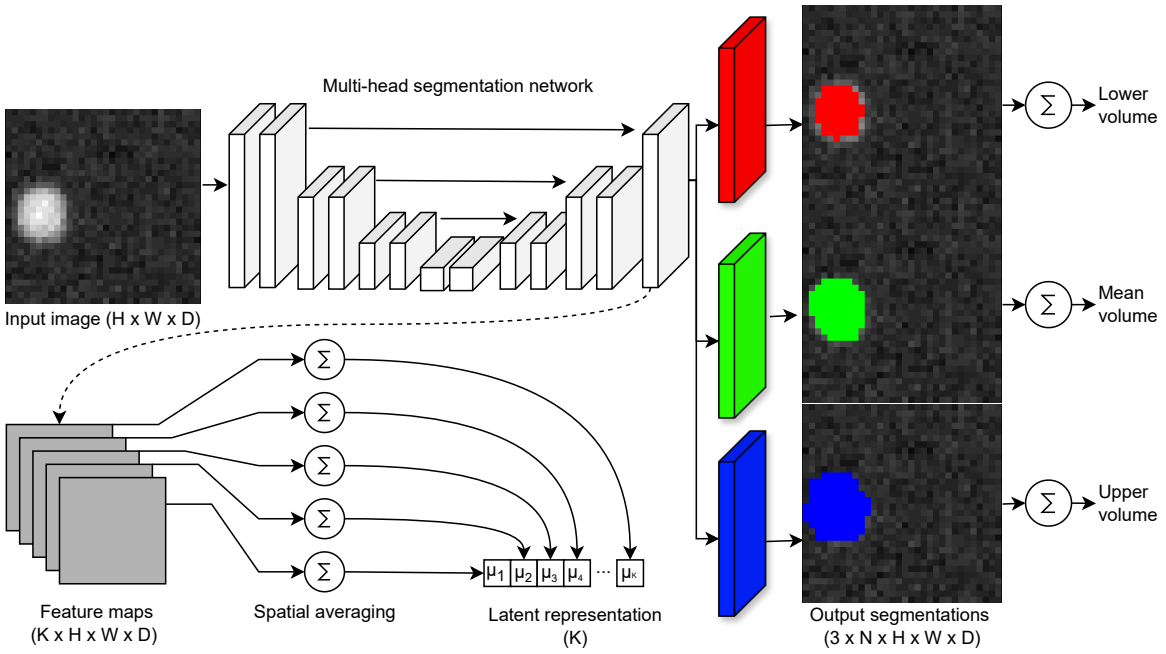


Figure V.3.3: The TriadNet framework enhanced for Weighted Conformal Prediction. Compressed latent representations are extracted at the penultimate convolutional layer, allowing an efficient estimation of the density ratio between the training and test distributions.

which has a dimension of 32. The overall framework is illustrated in Figure V.3.3. This low-dimensional feature vector is used to train an auxiliary Logistic Regression classification model to predict $\widehat{p}(x)$. It is trained using a 20-fold cross-validation paradigm. The predicted probabilities are clipped in the range $[0.01, 0.99]$ to avoid infinite weights.

We next report several metrics: the classification accuracy achieved by the auxiliary classification model on each dataset, the effective sample size (ESS, Equation V.3.4), the empirical coverage, and the average interval width. These metrics are estimated by running the experiments for $R = 250$ trials, by shuffling the in-distribution calibration and test samples. The shifted test set remains fixed for each trial.

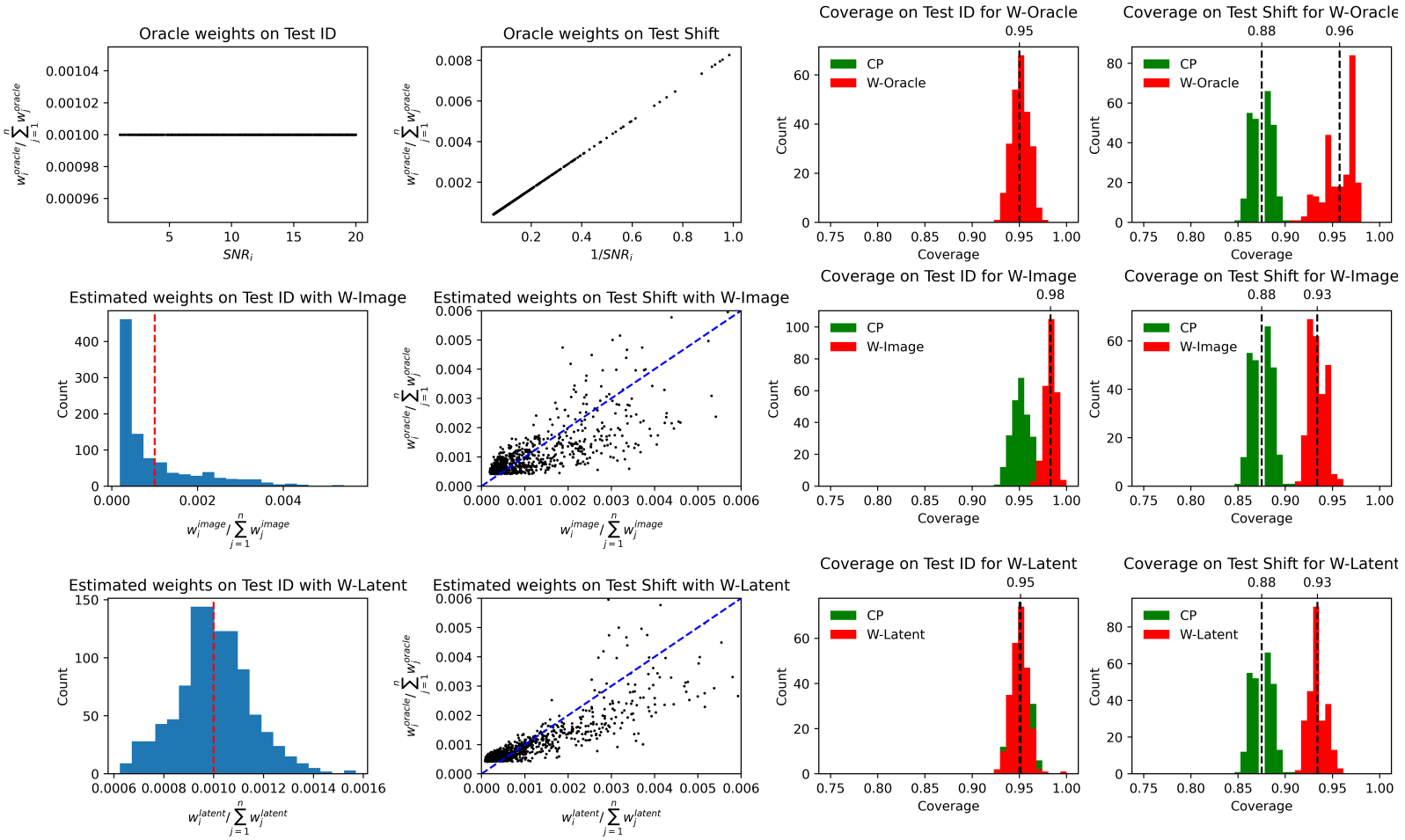


Figure V.3.4: Results of the simulation on synthetic data for standard and weighted conformal prediction. Top row (left to right): oracle calibration weight for Test ID and Test Shift, histograms of coverages on the Test ID dataset, and Shift Test. Middle row (left to right): distribution of estimated weights using W-Image on Test ID, comparison of estimated and oracle weights on Test Shift for W-Image, histograms of coverages on the Test ID and Shift datasets for W-Image. Bottom row (left to right): distribution of estimated weight using W-Latent on Test ID, comparison of estimated and oracle weights on Test Shift for W-Latent, and histograms of coverages on the Test Id and Shift datasets for W-Latent.

V.3.3.3 Results

Results of this experiment on synthetic data are presented in Table V.4 and Figure V.3.4. We start by commenting in detail Figure V.3.4.

- Top row, left to right: we start by presenting the oracle weights on ID data. When calibration and test data points are exchangeable, the density ratio is exactly one, so the p_i^w (Equation V.3.1) is set to $\frac{1}{n+1}$ with $n = 1000$ the size of the calibration dataset. When there is a covariate shift between calibration and test image, the density ratio is proportional to $\frac{1}{\text{SNR}}$, and thus the p_i^w increase proportionally to $\frac{1}{\text{SNR}}$. We then present the empirical distribution of coverages for standard CP (green) and W-Oracle (red) over the 250 trials. On Test ID, W-Oracle is strictly identical to standard CP, and in both cases, we have an average empirical coverage of exactly 0.95. In the presence of shift (Test Shift dataset), the standard CP undercover (average coverage of 0.88) while W-Oracle recovers the target coverage (0.96 on average).
- Middle row: we start by presenting the weights estimated on Test ID using W-CNN. The red line indicates the average. The expected behavior would be to have a distribution peaked around $\frac{1}{n+1}$, corresponding to equivalent weights for all calibration samples. However here it can be observed that the weights deviate from this value. As an effect the empirical coverage on Test ID is significantly above the target 95%, being 0.98 on average. It is also linked to an increase in the average interval width on Test ID and a reduction of the ESS (see Table V.4). This misestimation of the weights may be due to poor calibration of the CNN or due to overfitting. For Test Shift, we present the relationship between the weights estimated by W-Image and the oracle weights. The blue dashed line corresponds to the identity function. It can be observed that the weight estimation is not very accurate when taking the oracle weights as a reference, with the dispersion increasing with larger weights. Despite that, W-Image reduces the coverage gap on Test Shift, achieving an empirical coverage of around 0.93 versus 0.88 for standard CP.
- Bottom row is equivalent to the middle row but with W-Latent instead of W-Image. This time, the weights on Test ID follow the expected behavior with weights centered on the value $\frac{1}{n+1}$. The weights on Test Shift are also closer to the Oracle weights. On Test ID, W-Latent provides a distribution of coverages closely matching the standard CP procedure. On Test Shift, it reduces the coverage gap similarly to W-Image.

Overall, this controlled synthetic experiment shows that WCP recovers the target coverage level when oracle weights are known (W-Oracle). When oracle weights are unknown, two variants can be considered, namely W-Image and W-Latent. Our results show that W-Image is not fully satisfying for the weight estimation. While it efficiently reduces the coverage gap on Test Shift, we observed that the weight estimation in the absence of covariate shift (Test ID) is not reliable, with estimated density ratios deviating from the unit value. This phenomenon may thus be due to the miscalibration of the CNN, producing unreliable probabilities and thus poor estimations of the density ratios. Our proposed latent approach (W-Latent) is more satisfying in this regard, providing density ratios close to the unit on ID data.

Regarding the ESS, it can be noticed that all versions of WCP are associated with a reduced

CP version	Dataset	Accuracy	ESS	Coverage	Width (mm^3)	Dice
Standard	ID	-	1000 ± 0.0	95.11 ± 0.93	86.18 ± 3.00	$0.92\pm$ 0.07
W-Oracle		-	1000 ± 0.0	95.01 ± 1.01	86.06 ± 2.95	
W-Image		0.50 ± 0.01	505 ± 45.71	98.31 ± 0.61	118.94 ± 9.27	
W-Latent		0.50 ± 0.01	985.91 ± 21.30	95.03 ± 0.93	86.45 ± 3.79	
Standard	Shift	-	1000.0 ± 0.0	87.47 ± 1.08	94.76 ± 2.77	$0.89\pm$ 0.11
W-Oracle		-	98.20 ± 0.0	95.22 ± 1.57	153.28 ± 23.52	
W-Image		0.58 ± 0.01	162.09 ± 14.20	93.40 ± 0.83	128.56 ± 9.82	
W-Latent		0.72 ± 0.01	122.92 ± 13.92	93.39 ± 0.90	128.89 ± 11.34	

Table V.4: Comparison of standard and weighted Conformal Prediction on the synthetic task, for a target coverage of 95%. Results are averaged over $R = 250$ trials. CP=Conformal Prediction, ESS=Effective Sample Size. ID=In-Distribution.

ESS on the shifted test set. It is a drawback of the importance sampling that is operated. With W-Oracle, the ESS is reduced from 1000 to around 98. The reduction is less pronounced for W-Image and W-Latent. Note that there is an immediate link between the accuracy of the auxiliary classifiers (CNN for W-Image, Logistic Regression for W-Latent) and the ESS. If the accuracy is high, it means that the calibration and test distributions are far apart and thus the classification task is easy. In this setting, the predicted probabilities will likely diverge to extreme values (close to 0 or 1). As an effect, the estimated density ratios will be large (Equation V.3.1), and thus the ESS will shrink (Equation V.3.4). For instance, the auxiliary Logistic Regression reaches an accuracy of around 72% on Test Shift, showing that it can accurately distinguish the calibration and test data points. As a side effect, the ESS is reduced from 1000 to 123 approximately. The CNN only reaches an accuracy of 58% on Test Shift, linked with a smaller reduction of the ESS (around 162).

Lastly, it can be observed that the width of the intervals for the weighted version of CP is larger than for the standard CP, showing that in practice the reduction of the coverage gap is at the expense of the enlarging of the intervals.

To summarize, this experiment allows us to validate the relevancy of W-Latent on a controlled synthetic covariate shift. More particularly, we show that using the latent representations generated by the segmentation model allows a valid estimation of the density ratio. However, this framework can only work if there exist calibration samples that look like the test samples. In practice, if there were no calibration samples with low SNRs, there is no hope of accounting for the covariate shift. In the following, we investigate how our approach behaves on real MRI processing tasks.

V.3.3.4 Tackling unknown shifts on real-world MRI datasets

We now evaluate WCP in our two tasks (segmentation of MS lesions and glioblastoma) with the different domain-shift datasets (MSLUB and 1.5 Tesla datasets for MS lesions, SSA for glioblastomas). The target coverage is set to 90% (akin to set $\alpha = 0.10$). Intervals are estimated using the TriadNet technique. The W-Latent version of WCP is implemented as we do not have access here to Oracle weights, and W-Image proved to be unreliable. As

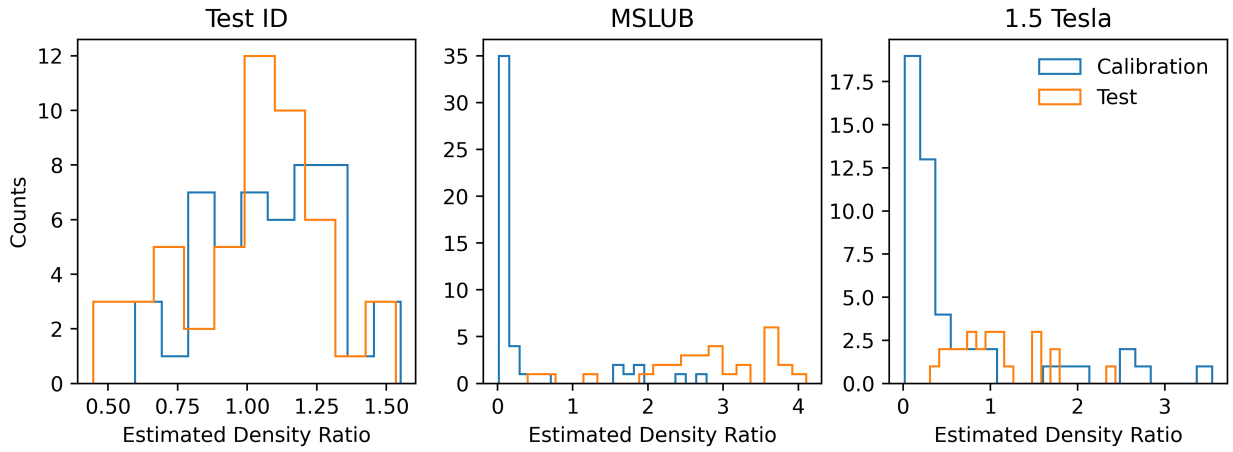


Figure V.3.5: Histograms of estimated density ratios for each Multiple Sclerosis dataset, using the logistic regression model.

CP version	**Test set**	**Accuracy** (%)	**ESS**	**Coverage** (%)	**Width** (ml)
Standard	ID	-	50.0 ± 0.0	92.06 ± 5.34	14.46 ± 1.58
Weighted-Latent	ID	0.50 ± 0.07	46.1 ± 9.8	90.97 ± 5.51	13.84 ± 1.97
Standard	MSLUB	-	50.0 ± 0.0	99.98 ± 0.54	13.23 ± 1.53
Weighted-Latent	MSLUB	0.90 ± 0.02	0.9 ± 0.3	100.00 ± 0.00	23.86 ± 0.64
Standard	1.5T	-	50.0 ± 0.0	96.04 ± 1.48	11.48 ± 1.48
Weighted-Latent	1.5T	0.74 ± 0.05	18.9 ± 8.0	93.44 ± 3.86	11.15 ± 4.15

Table V.5: Comparison of standard and weighted Conformal Prediction on MS lesion load estimation for a target coverage of 90%. Results are averaged over $R = 250$ trials. CP=Conformal Prediction, ESS=Effective Sample Size, MS=Multiple Sclerosis, ID=In-distribution. ml: milliliter. Metrics are estimated over 250 trials.

for the synthetic experiment, we use a standard Logistic Regression model to estimate the density ratio from the latent representations generated by the TriadNet model. We report the same metrics: the classification accuracy achieved by the classification model on each setting, the effective sample size, the empiric coverage, and the average interval width. These metrics are estimated by running the experiments for $R = 250$ trials, by shuffling the in-distribution calibration and test samples. They are presented in Table V.5 for the MS experiment and Table V.6 for glioblastoma. Figure V.3.5 presents the histograms of weights for each MS dataset (ID, MSLUB, and 1.5 Tesla).

First, for the MS experiment (Table V.5), it can be observed that when there is no domain shift (Test ID dataset), the accuracy of the classifier is 50% as expected. This means that calibration and test samples are indeed indistinguishable based on their latent representations. In this setting, the estimated density ratios are centered around 1 for each calibration and test sample (see Figure V.3.5, left). Both standard CP and W-Latent reach the target coverage level of 0.90, with standard CP being closer to the expected mode (0.92, see Table V.1) for a

calibration dataset of 50 samples.

Now, on the MSLUB dataset, it can be observed that the Logistic Regression model achieves a very high accuracy (90% on average) which means that calibration and test latent representations are extremely different. As a result, the weights diverge as $dP_{\text{test}} << dP_{\text{train}}$ for calibration samples, and $dP_{\text{test}} >> dP_{\text{train}}$ for test samples. It can be observed in Figure V.3.5, center, where most calibration samples are associated with a density ratio extremely close to 0. This has a dramatic effect on the Effective Sample Size which shrinks below the unit. In practice, we observed in this setting that most sets computed using Equation V.3.1 were empty: there exist no corrective values s_j such as the reweighted distribution is superior to $1 - \alpha$. For these cases, we set the corrective value to a large default value of 10mL, which translates to uninformative intervals. Thus, as a result, the interval width for the weighted CP algorithm jumps to an average of 23.86mL on the MSLUB dataset, as compared to the 13.23mL interval width achieved with the standard CP procedure.

A different tendency can be observed for the 1.5 Tesla test dataset. Here, the classifier reaches a moderate accuracy of 74% on average, showing that the latent representations are more alike between calibration and test samples. As a result, the ESS is more moderately reduced: roughly 19 on average (50 for the unweighted CP). Estimated density ratios are still close to 0 for calibration samples (Figure V.3.5, right) although this is less extreme than for the MSLUB dataset. Then, the coverage is closer to the target 90% than with the standard CP procedure, although the dispersion of coverages is higher. This is an expected downside of the reduced effective size, which makes the conformal procedure less precise.

Second, for the multi-class tumor segmentation task (Table V.6), the accuracy of the Logistic Regression model reaches 83% on average, which empirically divides by 4 the ESS (from 227 for standard CP to about 51 for weighted CP). On the SSA dataset, the coverage gap is reduced for all classes, yet the gain is minimal for class 2 (edematous). For classes 1 and 3 (necrosis and GD enhancing tumor), the gap is more efficiently reduced, which is linked to an increase in the interval widths.

To conclude, several insights can be noted from these experiments with weighted CP. First, estimating the weights from the latent representations (W-Latent) seems like an efficient and practical alternative to the density estimation on high-dimensional images (W-Image), which is computationally too costly. Moreover, the experiments on synthetic data showed that W-Image provided unreliable weights in the absence of covariate shifts, which may be due to a miscalibration of the probabilities and potential overfitting.

Importantly, WCP does not increase the coverage nor interval width when data are exchangeable (Test ID). When the shift is moderated (1.5 Tesla and SSA datasets), the coverage gap and interval width are reduced. This correspond to settings where the classifier does not perfectly distinguish between calibration and test samples. However, when the shift is more important (MSLUB dataset), then the estimated weights diverge because the classification task is too easy. In this setting the WCP procedure becomes highly unstable due to the reduced ESS. Then, the computed intervals are extremely large and uninformative. This is a downside of the weighted approach which assumes that the calibration and test distributions

Class	CP version	Test set	Accuracy %	ESS	Coverage (%)	Width (ml)
1	Standard	ID	-	227.0 ± 0.0	90.78 ± 2.71	9.92 ± 0.99
	W-Latent	ID	0.50 ± 0.04	219.9 ± 14.7	90.20 ± 2.65	9.46 ± 1.07
	Standard	SSA	-	227.0 ± 0.0	77.01 ± 2.32	22.03 ± 1.08
	W-Latent	SSA	0.83 ± 0.02	51.4 ± 8.1	82.92 ± 2.49	27.16 ± 2.77
2	Standard	ID	-	227.0 ± 0.0	90.76 ± 2.70	32.36 ± 2.85
	W-Latent	ID	0.50 ± 0.04	219.9 ± 14.7	90.03 ± 2.81	31.14 ± 2.17
	Standard	SSA	-	227.0 ± 0.0	84.15 ± 2.26	51.58 ± 3.00
	W-Latent	SSA	0.83 ± 0.02	51.4 ± 8.1	84.35 ± 2.24	52.58 ± 2.54
3	Standard	ID	-	227.0 ± 0.0	90.79 ± 2.71	7.70 ± 0.48
	W-Latent	ID	0.50 ± 0.04	219.9 ± 14.7	90.14 ± 2.66	7.52 ± 0.41
	Standard	SSA	-	227.0 ± 0.0	48.87 ± 2.66	14.36 ± 0.42
	W-Latent	SSA	0.83 ± 0.02	51.4 ± 8.1	56.76 ± 3.56	15.40 ± 0.47

Table V.6: Comparison of standard and weighted Conformal Prediction on multi-class tumor volume estimation for a target coverage of 90%. CP=Conformal Prediction, ESS=Effective Sample Size, ID=In-distribution. Class 1: Necrosis, Class 2: Edematous, Class 3: Gadolinium-enhancing tumor. ml: milliliter. Metrics are estimated over 250 trials.

are not too far, in which case the density ratio is not defined anymore. Importantly, we used a simple Logistic Regression model for these experiments. It can be anticipated that more efficient classifiers (e.g. Random Forest or Support Vector Machines) could achieve improved accuracies in distinguishing between calibration and shifted test samples. As an effect, the ESS will be reduced more heavily, which adds up to the instability of the weighted conformal procedure. This emphasizes the difficulty of implementing weighted CP in our medical image setting, as calibration datasets are rather small. Implementing the weighted CP is linked with an additional reduction effect of the calibration size, and then CP becomes intractable, as shown on MSLUB where the ESS converges to 0. In future work, improvement of the weight computation step could result in a more efficient reduction of the coverage gap on domain-shift settings. We note that there are recent studies tackling the problem of density ratio estimation when the two distributions are far apart [331, 332]. These methodological developments could partly alleviate the issues we face with the WCP procedure when the test data is significantly different from the training one.

V.4 Chapter conclusion

In this chapter, the building of predictive intervals for lesion volumes has been investigated. This application of uncertainty is currently overlooked in the medical DL literature, although being of crucial importance for real-world applications. At Pixyl, automated segmentations are used to generate reports displaying the total lesion volumes, and associating these estimations with proper intervals is important to build user trust.

Experiments carried out on MS lesions and brain tumor segmentation show that sampling-based approaches based on MC or TTA are time-consuming and thus not compatible with

industrial applications where the inference time is key. Moreover, the variability of the sampled volumes is low, which is compensated by large values of $\hat{q}$ and larger intervals.

This motivated the development of TriadNet, a framework allowing the direct estimation of intervals. Moreover, the framework is interpretable as the 3 masks can be provided to the user, allowing the visualization of the restrictive, balanced, and permissive masks (as shown in Figure V.2.3). Overlapping the three masks allows to visualize uncertain areas at the boundaries between classes, for instance, due to the partial volume effect in MRI. Finally, TriadNet is a simple framework that can be adapted with any segmentation backbone. It is associated with a marginal augmentation of the network parameters, which do not increase training or inference time.

This work also proposed the first investigation of conformal calibration for predictive intervals on lesion volumes in medical images. The framework provides interesting statistical guarantees and experimental results followed the theory with great precision when calibration and test data are exchangeable. However, two weak points of CP were highlighted. The first one is the requirement for a large calibration dataset, which is not possible in most 3D medical-image applications. The current guideline is to use 1000 data points for calibration [76], but our results demonstrate that useful intervals can be achieved with as low as 50 calibration datasets (MS experiment). In this setting, the average coverage is a few percent superior to 90%. With more calibration samples (tumor experiment), the coverages are centered at 90% with a smaller standard deviation. The second limitation is the assumption that calibration and test data points are exchangeable, which is rarely the case for industrial medical applications. Typically, if the distribution of lesion volumes in the test dataset changes or if the model produces poor segmentations on the test set, the coverage can diverge from the target, calibrated one. This could potentially be alleviated with a weighted formulation of CP, which has been investigated. This is based on the estimation of the density ratio between calibration and test distributions, for which we propose an efficient approach making use of the latent representations of the input images generated by the trained TriadNet model. This allows the reduction of the high-dimensional MRI to a small latent vector, which allows the use of a standard Logistic Regression model to estimate the weights. However, when the shift is too important, we show that the weights diverge which causes an important reduction of the effective sample size. It can even converge to 0 if the weights are too extreme. As a result, the weighted CP produces uninformative and excessively large intervals.

Despite this drawback, CP has proven to be a simple and practical tool for calibrating intervals without making any assumptions about the way they were generated or about the data distribution. In this chapter, CP has been investigated solely around the notion of coverage, which can be framed as a $0-1$ error: the ground truth volume is either contained or not in the PI. While being the main application of CP, the conformal framework is more general and can also be applied to control any monotone loss function, which has potentially many applications in the medical domain. The resulting framework, called Conformal Risk Control, can for instance be used to optimize decision thresholds of medical image segmentation networks. We propose an investigation of such a paradigm in Appendix A6.

General Conclusion

Deep Learning models have revolutionized the field of medical image analysis, but to be adopted and trusted in the clinic by healthcare professionals, confidence estimates should be provided alongside predictions. This thesis proposes a series of methodological developments intended to enhance the raw predictions with uncertainty estimates operating at different levels of radiological analysis.

At the voxel level, various popular uncertainty estimators were first compared on three different brain lesion segmentation tasks in MRI: MS lesions, glioblastoma, and strokes. A particular care was attributed to inference time, a crucial parameter for industrial applications. This series of experiments highlighted the relevancy of the Deep Ensemble technique to provide voxel-level uncertainty estimates, as well as boosting the segmentation accuracy in both in-distribution and out-of-distribution settings. Deep Ensemble is associated with an overhead of computation during the model development stage, however, it is very efficient at inference time as compared to other sampling-based techniques such as the popular Monte Carlo Dropout.

Building from voxel-level predictions, we then proposed to study the quantification of uncertainty at the lesion level. For applications revolving around the detection of multiple lesion instances per subject, as is typically the case for MS, instance-level confidence scores are preferred for a rapid overview of the automated prediction. Then, the clinician can directly review the uncertain lesions and discard them if judged inappropriate. To solve this challenge, we proposed a framework based on the training of an auxiliary classifier that predicts the status of the identified lesion (true positive or false positive) based on an extracted lesion representation. This representation choice is important, and three options were investigated: a Radiomics representation, a bounding-box representation, and finally graphs. The latter allows a flexible modelization of lesions and can naturally handle their heterogeneity while being frugal in terms of parameters. The interest of these approaches was demonstrated in three different lesion-oriented tasks: cross-sectional and longitudinal MS lesions detection, and lung nodules detection. One limit is the need for a sufficient amount of true and false positive lesions for training, which was found as a major drawback for the longitudinal experiment, for which lesion instances are particularly scarce.

The voxel and lesion-level experiments also allowed us to identify one important weakness of DL models, which is the robustness under domain shift. This has been investigated with two domain-shift scenarios for MS, one with data acquired in a different center (MSLUB), and one with a lower MRI quality (1.5 Tesla). For glioblastoma segmentation, it is investigated using a dataset comprising lower-quality MRI and more advanced stages of the disease (SSA dataset). In each case, the segmentation quality dropped. There is little hope of correcting this deficiency using standard data augmentation alone. Indeed, contrast and spatial transformations can not generate enough variability to account for all the differences between MRI scanners and acquisition protocols. Thus, being able to automatically flag

non-conform inputs is crucial to prevent suboptimal predictions. This has been investigated in detail in the third chapter focusing on out-of-distribution detection. Latent-space detectors based on the Mahalanobis distance were found particularly efficient, as well as being associated with low computational overhead. Indeed, it can be implemented in any trained segmentation model, only requiring access to the intermediate layer activation. Using a large benchmark of 24 OOD settings, we showed that this technique outperformed uncertainty and reconstruction-based approaches. However, this technique is highly dependent on the layer selection to gather the latent representations. This optimal layer is moreover dependent on the choice of the segmentation backbone. We show that the cumbersome layer selection can be alleviated by using a multi-layer aggregation strategy, which was found to perform consistently well across segmentation architectures.

OOD detection can be seen as a form of input-level QC, aiming at detecting images far from the training distribution. However, another definition of OOD is possible by taking into account the performance of the model on the OOD sample. This redefinition of OOD states that a sample is OOD if the associated segmentation is poor, regardless of the distance to the training database. This is akin to output-level QC, aiming at detecting segmentations that do not meet predefined standards. We explored how input and output level QC scores can be intertwined to provide a richer QC strategy. More specifically, we proposed to stratify the prediction space into 4 regions that present progressively degrading qualities of prediction. This unified QC strategy could be used to give insights to the user concerning the expected adequacy of the automated result.

Lastly, automated segmentations are used to extract high-level metrics such as lesion volumes, which are crucial bio-markers for many neurological diseases. Complementing these estimations with predictive intervals is a key to avoid misleading the reader. We proposed the first investigation of conformal prediction for lesion volume estimation in 3D brain MRI. Our approach called TriadNet is versatile and enables the construction of intervals without relying on sampling. Moreover, the approach is interpretable as three masks are provided to the user allowing them to visualize bounds in the form of restrictive and permissive delineations. The development of this module allowed us to emphasize several limitations of the conformal prediction framework in 3D medical image analysis pipelines. First, it requires a large calibration dataset that is usually unavailable due to the scarcity of medical-image datasets. Second it is based on the optimistic assumption that calibration and test data are exchangeable. A weighted version of conformal prediction can be adopted, but it is only efficient when the domain shift is not too important. In this direction, we propose an approach to estimate the gap between calibration and test samples relying on compressed latent representations.

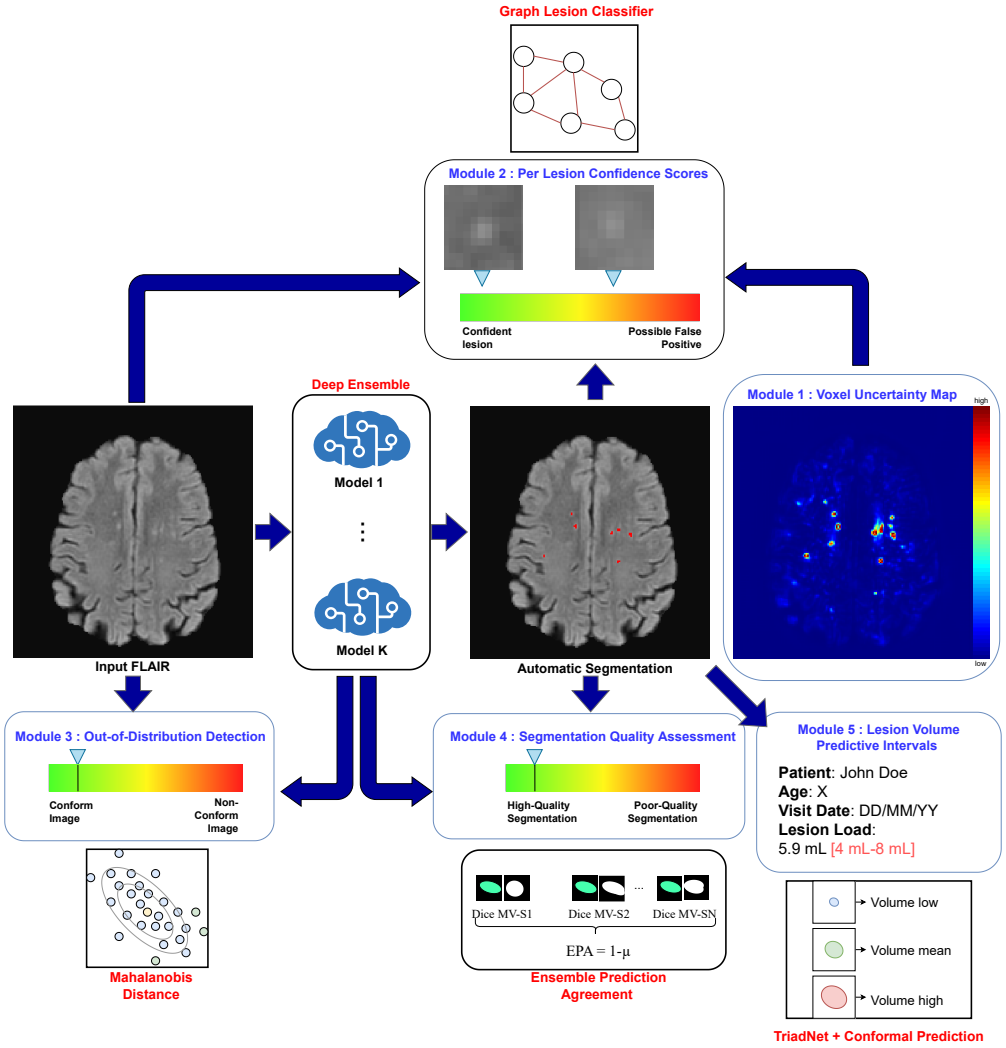


Figure A: Complete uncertainty framework incorporating the different modules developed during this thesis, allowing voxel, lesion, case, and volume-wise uncertainty quantification.

In conclusion, the methodologies developed during this Ph.D. project enable a complete estimation of uncertainties in medical image segmentation pipelines (see Figure **??**). It starts with voxel-level uncertainties that can be visualized next to the segmentation. Instance-level scores allow for confidence estimates aligned with the clinician's attention for lesion-based diseases such as MS. An overall appreciation concerning the conformity of the input image and quality of the output is provided to identify poor analyses. Finally, automatic reports displaying lesion volumes are enriched by predictive intervals for a more trustworthy estimation. Overall, particular attention was given to the development of versatile and robust tools. This is why each module has been evaluated on various pathologies and domain-shift settings when possible (in Appendix, Tables A.9.1 and A.9.2 present the complete list of databases used in the thesis). As inference time is key in industrial applications, the developed solutions are efficient and can be seamlessly incorporated into the inference pipeline. Incorporation of these tools in Pixyl workflow will ensure wiser decision-making and increase trust in the automated reports.

More generally, this Ph.D. project was the opportunity to investigate in detail exciting methodology developments in Deep Learning, such as Graph Neural Networks, Generative Adversarial Networks, and Conformal Predictions. The skills acquired in this thesis have also enabled us to take part in medical image segmentation challenges (Appendix A7-A8).

As a perspective, we note that the evaluation of uncertainty is generally limited to the detection of errors. Voxel uncertainty is used to detect incorrect voxels, lesion uncertainty is used to identify false positive lesions, and output QC aims at the detection of poor overall segmentations. This is because ground truth uncertainty labels are not available in most cases. Evaluating uncertainty estimates by their beneficial impact on decision-making would be an interesting future lead to validate the usefulness of the developed methods. Lastly, our different experiments stressed the limited generalization of DL models when the test distribution differs from the training one. Even if the drift seems moderate, such as the MSLUB or SSA datasets used throughout the thesis experiments, segmentations become suboptimal. This also manifested in the last part of this thesis during which the Conformal Prediction framework was investigated. While it provides provable statistical guarantees on in-distribution data, it no longer holds when the distribution shifts. For a software like Pixyl.Neuro processing hundreds of cases per day, it is crucial to guarantee the robustness of the algorithm to make sure that the same level of quality is met in each imaging center. Data Augmentation alone, used systematically to train the models in this thesis, is not sufficient to mimic the variability of the real world. This open the leads to exciting research on the amelioration of the generalization capacity of Deep Learning models.

Bibliography

[1] Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning* MIT Press, 2016.

[2] Christy JW Ledford, Dean A Seehusen, Alexander W Chessman and Navkiran K Shokar, How We Teach US Medical Students to Negotiate Uncertainty in Clinical: A CERA Study. Family Medicine **47** (2015), 31.

[3] Marieka A Helou, Deborah Diaz-Granados, Michael S Ryan and John W Cyrus, Uncertainty in decision-making in medicine: a scoping review and thematic analysis of conceptual models, Academic medicine: journal of the Association of American Medical Colleges **95** (2020), 157.

[4] George Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of control, signals and systems **2** (1989), 303.

[5] Chris Olah, Alexander Mordvintsev and Ludwig Schubert, *Feature Visualization*, 2017, URL: https://distill.pub/2017/feature-visualization.

[6] Fionn Murtagh, Multilayer perceptrons for classification and regression, Neurocomputing **2** (1991), 183.

[7] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM **60** (2017), 84.

[8] Adam Paszke et al., PyTorch: An Imperative Style, High-Performance Deep Learning Library, Advances in Neural Information Processing Systems **32** (2019), 8024.

[9] Martin Abadi et al., TensorFlow: a system for Large-Scale machine learning, 12th USENIX symposium on operating systems design and implementation (2016), 265.

[10] Zhou and Chellappa, Computation of optical flow using a neural network, IEEE 1988 international conference on neural networks (1988), 71.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten and Kilian Q Weinberger, Densely connected convolutional networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017), 4700.

[12] Mingxing Tan and Quoc Le, Efficientnet: Rethinking model scaling for convolutional neural networks, International Conference on Machine Learning (2019), 6105.

[13] Olaf Ronneberger, Philipp Fischer and Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention (2015), 234.

[14] Eric Kerfoot, James Clough, Ilkay Oksuz, Jack Lee, Andrew P King and Julia A Schnabel, Left-ventricle quantification using residual U-Net, International Workshop on Statistical Atlases and Computational Models of the Heart (2018), 371.

[15] Fausto Milletari, Nassir Navab and Seyed-Ahmad Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016 fourth international conference on 3D vision (2016), 565.

[16] Ozan Oktay et al., Attention u-net: Learning where to look for the pancreas, Medical Imaging with Deep Learning (2018).

[17] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen and Klaus H Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature methods **18** (2021), 203.

[18] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth and Daguang Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, Held in Conjunction with MICCAI 2021 (2022), 272.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin, Attention is all you need, Advances in neural information processing systems **30** (2017).

[20] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth and Daguang Xu, Unetr: Transformers for 3d medical image segmentation, Proceedings of the IEEE/CVF winter conference on applications of computer vision (2022), 574.

[21] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, et al., TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, arXiv preprint arXiv:2102.04306 (2021).

[22] Kelei He et al., Transformers in medical image analysis, Intelligent Medicine **3** (2023), 59.

[23] Sergey Ioffe and Christian Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, International conference on machine learning (2015), 448.

[24] Hongwei Yong, Jianqiang Huang, Deyu Meng, Xiansheng Hua and Lei Zhang, Momentum batch normalization for deep learning with small batch size, Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16 (2020), 224.

[25] Dmitry Ulyanov, Andrea Vedaldi and Victor Lempitsky, Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis, Proceedings of the IEEE conference on computer vision and pattern recognition (2017), 6924.

[26] Lei Jimmy Ba, Jamie Ryan Kiros and Geoffrey E. Hinton, Layer Normalization, CoRR **abs/1607.06450** (2016).

[27] Yuxin Wu and Kaiming He, Group normalization, Proceedings of the European Conference on Computer Vision (ECCV) (2018), 3.

[28] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang and Anne L Martel, Loss odyssey in medical image segmentation, Medical Image Analysis **71** (2021), 102035.

[29] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus and Ali Gholipour, Tversky loss function for image segmentation using 3D fully convolutional deep networks, International Workshop on Machine Learning in Medical Imaging (2017), 379.

[30] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin and M Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017 (2017), 240.

[31] Lena Maier-Hein, Bjoern Menze, et al., Metrics reloaded: Pitfalls and recommendations for image analysis validation, Nature Methods **2** (2022), 195.

[32] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz and Ismail Ben Ayed, Boundary loss for highly unbalanced segmentation, International conference on medical imaging with deep learning (2019), 285.

[33] Davood Karimi and Septimiu E Salcudean, Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks, IEEE Transactions on medical imaging **39** (2019), 499.

[34] Rosana EL Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina and Fahed Abdallah, A surprisingly effective perimeter-based loss for medical image segmentation, Medical Imaging with Deep Learning (2021), 158.

[35] Robert Hecht-Nielsen, Theory of the backpropagation neural network, Neural networks for perception (1992), 65.

[36] John C. Duchi, Elad Hazan and Yoram Singer, Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, J. Mach. Learn. Res. **12** (2011), 2121.

[37] Diederik P. Kingma and Jimmy Ba, Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings (2015).

[38] Yann LeCun, Léon Bottou, Genevieve B Orr and Klaus-Robert Müller, Efficient backprop, Neural networks: Tricks of the trade (2002), 9.

[39] Devansh Arpit et al., A closer look at memorization in deep networks, International conference on machine learning (2017), 233.

[40] Yarin Gal et al., Uncertainty in deep learning, PhD thesis, University of Cambridge, 2016.

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research **15** (2014), 1929.

[42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht and Oriol Vinyals, Understanding Deep Learning (Still) Requires Rethinking Generalization, Commun. ACM **64** (2021), 107–115.

[43] Richard Shaw, Carole H Sudre, Sebastien Ourselin, M Jorge Cardoso and Hugh G Pemberton, A Heteroscedastic Uncertainty Model for Decoupling Sources of MRI Image Quality, Machine Learning for Biomedical Imaging **1** (2021), 1.

[44] Jiheon Jeong, Ki Duk Kim, Yujin Nam, Kyungjin Cho, Jiseon Kang, Gil-Sun Hong and Namkug Kim, Generating High-Resolution 3D CT with 12-Bit Depth Using a Diffusion Model with Adjacent Slice and Intensity Calibration Network, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 366.

[45] Wei Peng, Ehsan Adeli, Tomas Bosschieter, Sang Hyun Park, Qingyu Zhao and Kilian M Pohl, Generating Realistic Brain MRIs via a Conditional Diffusion Probabilistic Model, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 14.

[46] Xinyi Yu, Guanbin Li, Wei Lou, Siqi Liu, Xiang Wan, Yan Chen and Haofeng Li, Diffusion-based data augmentation for nuclei image segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 592.

[47] Lingting Zhu, Zeyue Xue, Zhenchao Jin, Xian Liu, Jingzhen He, Ziwei Liu and Lequan Yu, Make-a-volume: Leveraging latent diffusion models for cross-modality

3d brain mri synthesis, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 592.

[48] Daniel P Huttenlocher, Gregory A. Klanderman and William J Rucklidge, Comparing images using the Hausdorff distance, IEEE Transactions on pattern analysis and machine intelligence **15** (1993), 850.

[49] Chuan Guo, Geoff Pleiss, Yu Sun and Kilian Q Weinberger, On calibration of modern neural networks, International Conference on Machine Learning (2017), 1321.

[50] Michael Yeung, Leonardo Rundo, Yang Nan, Evis Sala, Carola-Bibiane Schönlieb and Guang Yang, Calibrating the Dice loss to handle neural network overconfidence for biomedical image segmentation, Journal of Digital Imaging **36** (2023), 739.

[51] Anh Nguyen, Jason Yosinski and Jeff Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2015), 427.

[52] Roger Allan Ford, W Price and II Nicholson, Privacy and accountability in black-box medicine, Mich. Telecomm. Tech. L. Rev. **23** (2016), 1.

[53] Moloud Abdar et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion **76** (2021), 243.

[54] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden and Anna Goldenberg, What clinicians want: contextualizing explainable machine learning for clinical end use, Machine learning for healthcare conference (2019), 359.

[55] Eyke Hüllermeier and Willem Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Machine Learning **110** (2021), 457.

[56] Andrey Malinin, Uncertainty estimation in deep learning with application to spoken language assessment, PhD thesis, University of Cambridge, 2019.

[57] Alex Kendall and Yarin Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (2017), 5574.

[58] Zhiyun Xue, Feng Yang, Sivaramakrishnan Rajaraman, Ghada Zamzmi and Sameer Antani, Cross Dataset Analysis of Domain Shift in CXR Lung Region Detection, Diagnostics **13** (2023), 1068.

[59] Anton S Becker, Krishna Chaitanya, Khoschy Schawkat, Urs J Muehlematter, Andreas M Hötker, Ender Konukoglu and Olivio F Donati, Variability of manual segmentation of the prostate in axial T2-weighted MRI: A multi-reader study, European journal of radiology **121** (2019), 108716.

[60] Leo Joskowicz, D Cohen, N Caplan and Jacob Sosna, Inter-observer variability of manual contour delineation of structures in CT, European radiology **29** (2019), 1391.

[61] Jiachen Yao, Yikai Zhang, Songzhu Zheng, Mayank Goswami, Prateek Prasanna and Chao Chen, Learning to Segment from Noisy Annotations: A Spatial Correction Approach, International Conference on Learning Representations (2023).

[62] Miguel Angel Gonzalez Ballester, Andrew P Zisserman and Michael Brady, Estimation of the partial volume effect in MRI, Medical image analysis **6** (2002), 389.

[63] Olivier Commowick, Frédéric Cervenansky, François Cotton and Michel Dojat, MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure, MICCAI 2021-24th International

Conference on Medical Image Computing and Computer Assisted Intervention (2021), 126.

[64] Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene and Michel Dojat, Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis, Artificial Intelligence in Medicine (2024), 102830, ISSN: 0933-3657.

[65] Ananya Kumar, Percy Liang and Tengyu Ma, Verified Uncertainty Calibration, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (2019), 3787.

[66] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song and Peter Flach, Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration, Advances in neural information processing systems **32** (2019).

[67] Balamurali Murugesan, Bingyuan Liu, Adrian Galdran, Ismail Ben Ayed and Jose Dolz, Calibrating segmentation networks with margin-based label smoothing, Medical Image Analysis **87** (2023), 102826.

[68] Zhaoshuo Diao, Huiyan Jiang and Tianyu Shi, A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity, Knowledge-Based Systems **246** (2022), 108739.

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

[70] Terrance DeVries and Graham W Taylor, Leveraging uncertainty estimates for predicting segmentation quality, arXiv e-prints (2018).

[71] Christoph Berger, Magdalini Paschali, Ben Glocker and Konstantinos Kamnitsas, Confidence-based out-of-distribution detection: a comparative study and analysis, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis (2021), 122.

[72] Camila González, Karol Gotkowski, Moritz Fuchs, Andreas Bucher, Armin Dadras, Ricarda Fischbach, Isabel Jasmin Kaltenborn and Anirban Mukhopadhyay, Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation, Medical Image Analysis **82** (2022), 102596.

[73] Gustavo Carneiro, Leonardo Zorron Cheng Tao Pu, Rajvinder Singh and Alastair Burt, Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy, Medical Image Analysis **62** (2020), 101653.

[74] Gongbo Liang, Yu Zhang and Nathan Jacobs, Neural network calibration for medical imaging classification using dca regularization, International Conference on Machine Learning, Workshop on Uncertainty and Robustness in Deep Learning (2020).

[75] Vladimir Vovk, Alexander Gammerman and Glenn Shafer, Algorithmic learning in a random world, **29** (2005).

[76] Anastasios N. Angelopoulos and Stephen Bates, Conformal Prediction: A Gentle Introduction, Foundations and Trends in Machine Learning **16** (2023), 494.

[77] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik and Michael I Jordan, Uncertainty sets for image classifiers using conformal prediction, 9th International Conference on Learning Representations, ICLR 2021 (2021).

[78] Yaniv Romano, Evan Patterson and Emmanuel Candes, Conformalized quantile regression, Advances in neural information processing systems **32** (2019), 3538.

[79] Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder and Ola Spjuth, Predicting with confidence: using conformal prediction in drug discovery, Journal of Pharmaceutical Sciences **110** (2021), 42.

[80] Daniel Csillag, Lucas Monteiro Paes, Thiago Ramos, João Vitor Romano, Rodrigo Schuller, Roberto B Seixas, Roberto I Oliveira and Paulo Orenstein, AmnioML: amniotic fluid segmentation and volume prediction with uncertainty quantification, Proceedings of the AAAI Conference on Artificial Intelligence **37** (2023), 15494.

[81] Zach Eaton-Rosen, Thomas Varsavsky, Sebastien Ourselin and M Jorge Cardoso, As easy as 1, 2... 4? uncertainty in counting tasks for medical imaging, Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference (2019), 356.

[82] Yizhe Zhang, Shuo Wang, Yejia Zhang and Danny Z Chen, RR-CP: Reliable-Region-Based Conformal Prediction for Trustworthy Medical Image Classification, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 12.

[83] Hendrik Mehrtens, Tabea Bucher and Titus J Brinker, Pitfalls of Conformal Predictions for Medical Image Classification, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 198.

[84] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas and Ryan J Tibshirani, Conformal prediction beyond exchangeability, The Annals of Statistics **51** (2023), 816.

[85] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu and Daan Wierstra, Weight uncertainty in neural network, International Conference on Machine Learning (2015), 1613.

[86] Hao Wang and Dit-Yan Yeung, A survey on Bayesian deep learning, ACM Computing Surveys (CSUR) **53** (2020), 1.

[87] Tyler LaBonte, Carianne Martinez and Scott A Roberts, We know where we don't know: 3d bayesian cnns for credible geometric uncertainty, arXiv e-prints (2019).

[88] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine and Mohammed Bennamoun, Hands-on Bayesian neural networks—A tutorial for deep learning users, IEEE Computational Intelligence Magazine **17** (2022), 29.

[89] Kumar Shridhar, Felix Laumann and Marcus Liwicki, A comprehensive guide to bayesian convolutional neural network with variational inference, arXiv preprint arXiv:1901.02731 (2019).

[90] Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon and Yarin Gal, A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks, 4th Workshop on Bayesian Deep Learning (NeurIPS 2019) (2019).

[91] Hendrik A. Mehrtens, Alexander Kurz, Tabea-Clara Bucher and Titus J. Brinker, Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise, Medical Image Analysis **89** (2023), 102914.

[92] Pankaj Dhakal and Sashidhar Ram Joshi, Uncertainty Estimation in Detecting Knee Abnormalities on MRI using Bayesian Deep Learning, Proceedings of 10th IOE Graduate Conference **10** (2021).

[93] Haixing Li and Haibo Luo, Uncertainty Quantification in Medical Image Segmentation, 2020 IEEE 6th International Conference on Computer and Communications (ICCC) (2020), 1936.

[94] Moritz Fuchs, Camila Gonzalez and Anirban Mukhopadhyay, Practical uncertainty quantification for brain tumor segmentation, Medical Imaging with Deep Learning (2021).

[95] Jadie Adams and Shireen Y Elhabian, Benchmarking Scalable Epistemic Uncertainty Quantification in Organ Segmentation, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 53.

[96] Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan and Dustin Tran, Efficient and scalable bayesian neural nets with rank-1 factors, International conference on machine learning (2020), 2782.

[97] Yarin Gal and Zoubin Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, International Conference on Machine Learning (2016), 1050.

[98] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim and Myunghee Cho Paik, Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation, Computational Statisticsand Data Analysis **142** (2020), 106816.

[99] Alain Jungo et al., Towards uncertainty-assisted brain tumor segmentation and survival prediction, International MICCAI Brainlesion Workshop (2017), 474.

[100] José Ignacio Orlando, Philipp Seeböck, Hrvoje Bogunović, Sophie Klimscha, Christoph Grechenig, Sebastian Waldstein, Bianca S Gerendas and Ursula Schmidt-Erfurth, U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans, IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (2019), 1441.

[101] Robin Camarasa, Daniel Bos, Jeroen Hendrikse, Paul Nederkoorn, Eline Kooi, Aad van der Lugt and Marleen de Bruijne, Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020 (2020), 32.

[102] Biraja Ghoshal, Allan Tucker, Bal Sanghera and Wai Lup Wong, Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data, 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (2019), 318.

[103] Yarin Gal, Jiri Hron and Alex Kendall, Concrete dropout, Advances in neural information processing systems **30** (2017).

[104] Patrick McClure et al., Knowing what you know in brain segmentation using Bayesian deep neural networks, Frontiers in neuroinformatics **13** (2019), 67.

[105] Aryan Mobiny, Pengyu Yuan, Supratik K Moulik, Naveen Garg, Carol C Wu and Hien Van Nguyen, Dropconnect is effective in modeling uncertainty of bayesian deep networks, Scientific reports **11** (2021), 1.

[106] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun and Rob Fergus, Regularization of neural networks using dropconnect, International conference on machine learning (2013).

[107] Balaji Lakshminarayanan, Alexander Pritzel and Charles Blundell, Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, Annual Conference on Neural Information Processing Systems 2017 (2017), 6402.

[108] Andrey Malinin and Mark Gales, Uncertainty Estimation in Autoregressive Structured Prediction, International Conference on Learning Representations (2020).

[109] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi and Tina Kapur, Confidence calibration and predictive uncertainty estimation for deep medical image segmentation, IEEE Transactions on Medical Imaging **39** (2020), 3868.

[110] Jeppe Thagaard, Søren Hauberg, Bert van der Vegt, Thomas Ebstrup, Johan D Hansen and Anders B Dahl, Can you trust predictive uncertainty under real dataset shifts in digital pathology?, International Conference on Medical Image Computing and Computer-Assisted Intervention (2020), 824.

[111] Hamzeh Asgharnezhad, Afshar Shamsi, Roohallah Alizadehsani, Abbas Khosravi, Saeid Nahavandi, Zahra Alizadeh Sani, Dipti Srinivasan and Sheikh Mohammed Shariful Islam, Objective evaluation of deep uncertainty predictions for covid-19 detection, Scientific Reports **12** (2022), 1.

[112] Jasper Linmans, Jeroen van der Laak and Geert Litjens, Efficient Out-of-Distribution Detection in Digital Pathology Using Multi-Head Convolutional Neural Networks. Medical Imaging with Deep Learning (2020), 465.

[113] Kaisar Kushibar, Victor Campello, Lidia Garrucho, Akis Linardos, Petia Radeva and Karim Lekadir, Layer Ensembles: A Single-Pass Uncertainty Estimation in Deep Learning for Segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference (2022), 514.

[114] Lorena Qendro, Alexander Campbell, Pietro Lio and Cecilia Mascolo, Early exit ensembles for uncertainty quantification, Machine Learning for Health (2021), 181.

[115] Gengyan Zhao, Fang Liu, Jonathan A Oler, Mary E Meyerand, Ned H Kalin and Rasmus M Birn, Bayesian convolutional neural network based MRI brain extraction on nonhuman primates, Neuroimage **175** (2018), 32.

[116] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov and Andrew Gordon Wilson, A simple baseline for bayesian uncertainty in deep learning, Advances in neural information processing systems **32** (2019).

[117] Markus Lindén, Azat Garifullin and Lasse Lensu, Weight averaging impact on the uncertainty of retinal artery-venous segmentation, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020 (2020), 52.

[118] Yumin Liu, Claire Zhao and Jonathan Rubin, Uncertainty Quantification in Chest X-Ray Image Classification using Bayesian Deep Neural Networks. Knowledge Discovery in Healthcare Data, European Conference on Artificial Intelligence (2020), 19.

[119] Pablo M Granitto, Pablo F Verdes and H Alejandro Ceccatto, Neural network ensembles: evaluation of aggregation algorithms, Artificial Intelligence **163** (2005), 139.

[120] Ludmila I Kuncheva and Christopher J Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine learning **51** (2003), 181.

[121] Agostina J Larrazabal, César Martínez, Jose Dolz and Enzo Ferrante, Orthogonal ensemble networks for biomedical image segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference (2021), 594.

[122] Mariana-Iuliana Georgescu, Radu Tudor Ionescu and Andreea Iuliana Miron, Diversity-Promoting Ensemble for Medical Image Segmentation, The 38th ACM/SIGAPP Symposium On Applied Computing (2022).

[123] Moloud Abdar et al., Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning, Computers in biology and medicine **135** (2021), 104418.

[124] Moloud Abdar et al., UncertaintyFuseNet: Robust uncertainty-aware hierarchical feature fusion model with Ensemble Monte Carlo Dropout for COVID-19 detection, Information Fusion **90** (2023), 364.

[125] Wei Ji, Wenting Chen, Shuang Yu, Kai Ma, Li Cheng, Linlin Shen and Yefeng Zheng, Uncertainty quantification for medical image segmentation using dynamic label factor allocation among multiple raters, MICCAI on QUBIQ Workshop (2020).

[126] Sabri Can Cetindag, Mert Yergin, Deniz Alis and Ilkay Oksuz, Meta-learning for Medical Image Segmentation Uncertainty Quantification, International MICCAI Brainlesion Workshop (2022), 578.

[127] Yanwu Yang, Xutao Guo, Yiwei Pan, Pengcheng Shi, Haiyan Lv and Ting Ma, Uncertainty Quantification in Medical Image Segmentation with Multi-decoder U-Net, International MICCAI Brainlesion Workshop (2022), 570.

[128] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman and Max Welling, Supervised uncertainty quantification for segmentation with multiple annotations, International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), 137.

[129] Tanya Nair, Doina Precup, Douglas L Arnold and Tal Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, Medical image analysis **59** (2020), 101557.

[130] Zach Eaton-Rosen, Felix Bragman, Sotirios Bisdas, Sébastien Ourselin and M Jorge Cardoso, Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions, International Conference on Medical Image Computing and Computer-Assisted Intervention (2018), 691.

[131] Mark S Graham, Carole H Sudre, Thomas Varsavsky, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin and M Jorge Cardoso, Hierarchical brain parcellation with uncertainty, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020 (2020), 23.

[132] Thierry Judge, Olivier Bernard, Woo-Jin Cho Kim, Alberto Gomez, Agisilaos Chartsias and Pierre-Marc Jodoin, Asymmetric Contour Uncertainty Estimation for Medical Image Segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 210.

[133] Richard McKinley, Micheal Rebsamen, Katrin Daetwyler, Raphael Meier, Piotr Radojewski and Roland Wiest, Uncertainty-driven refinement of tumor-core segmentation using 3D-to-2D networks with label uncertainty, International MICCAI Brainlesion Workshop (2020), 401.

[134] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency and Masahito Ueda, Deep gamblers: Learning to abstain with portfolio theory, Advances in Neural Information Processing Systems **32** (2019).

[135] Till J Bungert, Levin Kobelke and Paul F Jäger, Understanding Silent Failures in Medical Image Classification, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 400.

[136] Simon Kohl et al., A Probabilistic U-Net for Segmentation of Ambiguous Images, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (2018), 6965.

[137] Simon AA Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman and Olaf Ronneberger, A hierarchical probabilistic u-net for modeling multi-scale ambiguities, arXiv e-prints (2019).

[138] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati and Ender Konukoglu, Phiseg: Capturing uncertainty in medical image segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), 119.

[139] Marc Gantenbein, Ertunc Erdil and Ender Konukoglu, Revphiseg: A memory-efficient neural network for uncertainty quantification in medical image segmentation, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis (2020), 13.

[140] MM Amaan Valiuddin, Christiaan GA Viviers, Ruud JG van Sloun, Peter HN de With and Fons van der Sommen, Improving Aleatoric Uncertainty Quantification in Multi-annotated Medical Image Segmentation with Normalizing Flows, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021 (2021), 75.

[141] Christiaan GA Viviers, Amaan MM Valiuddin, Peter HN de With and Fons van der Sommen, Probabilistic 3D segmentation for aleatoric uncertainty quantification in full 3D medical data, Medical Imaging 2023: Computer-Aided Diagnosis **12465** (2023), 343.

[142] Raghavendra Selvan, Frederik Faye, Jon Middleton and Akshay Pai, Uncertainty quantification in medical image segmentation with normalizing flows, International Workshop on Machine Learning in Medical Imaging (2020), 80.

[143] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk and Ben Glocker, Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty, Advances in Neural Information Processing Systems **33** (2020), 12756.

[144] Tomer Amit, Shmuel Shichrur, Tal Shaharabany and Lior Wolf, Annotator Consensus Prediction for Medical Image Segmentation with Diffusion Models, Medical Image

Computing and Computer Assisted Intervention – MICCAI 2023 (2023), ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood and Russell Taylor, 544.

[145] Murat Seckin Ayhan and Philipp Berens, Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks, International conference on Medical Imaging with Deep Learning (2018).

[146] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin and Tom Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing **338** (2019), 34.

[147] Laura Mora Ballestar and Veronica Vilaplana, MRI brain tumor segmentation and uncertainty estimation using 3D-UNet architectures, Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020 (2021), 376.

[148] Golara Javadi et al., Towards targeted ultrasound-guided prostate biopsy by incorporating model and label uncertainty in cancer detection, International Journal of Computer Assisted Radiology and Surgery **17** (2022), 121.

[149] Janis Postels, Mattia Segu, Tao Sun, Luc Van Gool, Fisher Yu and Federico Tombari, On the practicality of deterministic epistemic uncertainty, International Conference on Machine Learning (2021).

[150] Erdi Calli, Bram Van Ginneken, Ecem Sogancioglu and Keelin Murphy, FRODO: An In-Depth Analysis of a System to Reject Outlier Samples From a Trained Neural Network, IEEE Transactions on Medical Imaging **42** (2022), 971.

[151] Harry Anthony and Konstantinos Kamnitsas, On the use of Mahalanobis distance for out-of-distribution detection with neural networks for medical imaging, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 136.

[152] Mickael Tardy, Bruno Scheffer and Diana Mateus, Uncertainty measurements for the reliable classification of mammograms, International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), 495.

[153] McKell Woodland, Nihil Patel, Mais Al Taie, Joshua P. Yung, Tucker J. Netherton, Ankit B. Patel and Kristy K. Brock, Dimensionality Reduction for Improving Out-of-Distribution Detection in Medical Image Segmentation, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 147.

[154] Davood Karimi and Ali Gholipour, Improving calibration and out-of-distribution detection in deep models for medical image segmentation, IEEE Transactions on Artificial Intelligence (2022).

[155] Benjamin Lambert, Florence Forbes, Senan Doyle and Michel Dojat, Multi-layer Aggregation as a key to feature-based OOD detection, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging - 5th International Workshop, UNSURE 2023, Held in Conjunction with MICCAI 2023, Lecture Notes in Computer Science **14291** (2023), 104.

[156] Arthur P Dempster, A generalization of Bayesian inference, Journal of the Royal Statistical Society: Series B (Methodological) **30** (1968), 205.

[157] Ke Zou, Xuedong Yuan, Xiaojing Shen, Meng Wang and Huazhu Fu, TBraTS: Trusted Brain Tumor Segmentation, International Conference on Medical Image Computing

and Computer-Assisted Intervention, Lecture Notes in Computer Science **13438** (2022), 503.

[158] Murat Sensoy, Lance M. Kaplan and Melih Kandemir, Evidential Deep Learning to Quantify Classification Uncertainty, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (2018), 3183.

[159] Wei Fu, Yufei Chen, Wei Liu, Xiaodong Yue and Chao Ma, Evidence Reconciled Neural Network for Out-of-Distribution Detection in Medical Images, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 305.

[160] Ling Huang, Su Ruan, Pierre Decazes and Thierry Denoeux, Evidential segmentation of 3D PET/CT images, International Conference on Belief Functions (2021), 159.

[161] Ling Huang, Su Ruan and Thierry Denoeux, Belief function-based semi-supervised learning for brain tumor segmentation, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 160.

[162] Florin C Ghesu, Bogdan Georgescu, Eli Gibson, Sebastian Guendel, Mannudeep K Kalra, Ramandeep Singh, Subba R Digumarthy, Sasa Grbic and Dorin Comaniciu, Quantifying and leveraging classification uncertainty for chest radiograph assessment, International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), 676.

[163] Florin C Ghesu et al., Quantifying and leveraging predictive uncertainty for medical image assessment, Medical Image Analysis **68** (2021), 101855.

[164] Tareen Dawood, Emily Chan, Reza Razavi, Andrew P King and Esther Puyol-Anton, Addressing deep learning model calibration using evidential neural networks and uncertainty-aware training, 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (2023).

[165] Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen and Martin Aastrup Olsen, Improving uncertainty estimation in convolutional neural networks using inter-rater agreement, Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22 (2019), 540.

[166] Santiago Toledo-Cortés, Melissa De La Pava, Oscar Perdómo and Fabio A González, Hybrid deep learning gaussian process for diabetic retinopathy diagnosis and uncertainty quantification, International Workshop on Ophthalmic Medical Image Analysis (2020), 206.

[167] Lin Wang et al., Medical matting: a new perspective on medical segmentation with uncertainty, Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference (2021), 573.

[168] Nataliia Molchanova, Vatsal Raina, Andrey Malinin, Francesco La Rosa, Henning Muller, Mark Gales, Cristina Granziera, Mara Graziani and Meritxell Bach Cuadra, Novel structural-scale uncertainty measures and error retention curves: application to multiple sclerosis, 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (2022).

[169] Ishaan Bhat, Hugo J Kuijf, Veronika Cheplygina and Josien PW Pluim, Using uncertainty estimation to reduce false positives in liver lesion detection, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 663.

[170] Ishaan Bhat, Josien PW Pluim, Max A Viergerver and Hugo J Kuijf, Influence of uncertainty estimation techniques on false-positive reduction in liver lesion detection, Machine Learning for Biomedical Imaging **1** (2022), 1.

[171] Joost JM Van Griethuysen et al., Computational radiomics system to decode the radiographic phenotype, Cancer research **77** (2017), e104.

[172] Onur Ozdemir, Rebecca L Russell and Andrew A Berlin, A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans, IEEE Transactions on Medical Imaging **39** (2019), 1419.

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

[174] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang and Nasir Rajpoot, MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images, Medical image analysis **52** (2019), 199.

[175] Yuta Hiasa, Yoshito Otake, Masaki Takao, Takeshi Ogawa, Nobuhiko Sugano and Yoshinobu Sato, Automated muscle segmentation from clinical CT using Bayesian U-net for personalized musculoskeletal modeling, IEEE Transactions on Medical Imaging **39** (2019), 1030.

[176] Xiaobin Hu, Rui Guo, Jieneng Chen, Hongwei Li, Diana Waldmannstetter, Yu Zhao, Biao Li, Kuangyu Shi and Bjoern Menze, Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images, IEEE journal of biomedical and health informatics **24** (2020), 2599.

[177] Sarahi Rosas-Gonzalez, Taibou Birgui-Sekou, Moncef Hidane, Ilyess Zemmoura and Clovis Tauber, Asymmetric Ensemble of Asymmetric U-Net Models for Brain Tumor Segmentation With Uncertainty Estimation, Frontiers in Neurology (2021), 1421.

[178] Guotai Wang, Wenqi Li, Sébastien Ourselin and Tom Vercauteren, Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation, Frontiers in computational neuroscience **13** (2019), 56.

[179] Anjali Balagopal et al., A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy, Medical image analysis **72** (2021), 102101.

[180] Alain Jungo, Raphael Meier, Ekin Ermis, Evelyn Herrmann and Mauricio Reyes, Uncertainty-driven Sanity Check: Application to Postoperative Brain Tumor Cavity Segmentation, Medical Imaging with Deep Learning (2022).

[181] Sambuddha Ghosal, Audrey Xie and Pratik Shah, Uncertainty quantified deep learning for predicting dice coefficient of digital histopathology image segmentation, arXiv e-prints (2021).

[182] Evan Hann, Ricardo A Gonzales, Iulia A Popescu, Qiang Zhang, Vanessa M Ferreira and Stefan K Piechnik, Ensemble of deep convolutional neural networks with monte carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets, Annual Conference on Medical Image Understanding and Analysis (2021), 280.

[183] Tewodros Weldebirhan Arega, Stéphanie Bricq, François Legrand, Alexis Jacquier, Alain Lalande and Fabrice Meriaudeau, Automatic uncertainty-based quality controlled

T1 mapping and ECV analysis from native and post-contrast cardiac T1 mapping images using Bayesian vision transformer, Medical image analysis **86** (2023), 102773.

[184] Ziyi Huang, Yu Gan, Theresa Lye, Haofeng Zhang, Andrew Laine, Elsa D Angelini and Christine Hendon, Heterogeneity measurement of cardiac tissues leveraging uncertainty information from image segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2020), 782.

[185] Kristoffer Wickstrøm, Michael Kampffmeyer and Robert Jenssen, Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps, Medical image analysis **60** (2020), 101619.

[186] Parth Natekar, Avinash Kori and Ganapathy Krishnamurthi, Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis, Frontiers in computational neuroscience **14** (2020), 6.

[187] Bernhard Föllmer, Federico Biavati, Christian Wald, Sebastian Stober, Jackie Ma, Marc Dewey and Wojciech Samek, Active multi-task learning with uncertainty weighted loss for coronary calcium scoring, Medical Physics (2022).

[188] Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A González Ballester and Gemma Piella, Curriculum learning for improved femur fracture classification: Scheduling data with prior knowledge and uncertainty, Medical Image Analysis **75** (2022), 102273.

[189] Lie Ju, Xin Wang, Lin Wang, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, Tongliang Liu and Zongyuan Ge, Improving Medical Images Classification With Label Noise Using Dual-Uncertainty Estimation, IEEE Trans. Medical Imaging **41** (2022), 1533.

[190] Chaoyi Li, Meng Li, Can Peng and Brian C Lovell, Dynamic Curriculum Learning via In-Domain Uncertainty for Medical Image Classification, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 747.

[191] Soufiane Belharbi, Jérôme Rony, Jose Dolz, Ismail Ben Ayed, Luke McCaffrey and Eric Granger, Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty, IEEE Transactions on Medical Imaging **41** (2021), 702.

[192] Jinyi Xiang, Peng Qiu and Yang Yang, FUSSNet: Fusing Two Sources of Uncertainty for Semi-supervised Medical Image Segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference (2022), 481.

[193] Suman Sedai, Bhavna Antony, Ravneet Rai, Katie Jones, Hiroshi Ishikawa, Joel Schuman, Wollstein Gadi and Rahil Garnavi, Uncertainty guided semi-supervised segmentation of retinal layers in OCT images, International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), 282.

[194] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu and Pheng-Ann Heng, Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2019), 605.

[195] Xuyang Cao, Houjin Chen, Yanfeng Li, Yahui Peng, Shu Wang and Lin Cheng, Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation, IEEE transactions on medical imaging **40** (2020), 431.

[196] Wenjing Lu, Jiahao Lei, Peng Qiu, Rui Sheng, Jinhua Zhou, Xinwu Lu and Yang Yang, UPCoL: Uncertainty-Informed Prototype Consistency Learning for Semi-supervised Medical Image Segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 662.

[197] Raghav Mehta, Thomas Christinck, Tanya Nair, Paul Lemaitre, Douglas Arnold and Tal Arbel, Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures (2019), 23.

[198] Leonhard F Feiner, Martin J Menten, Kerstin Hammernik, Paul Hager, Wenqi Huang, Daniel Rueckert, Rickmer F Braren and Georgios Kaissis, Propagation and Attribution of Uncertainty in Medical Imaging Pipelines, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 1.

[199] Roger D Soberanis-Mukul, Nassir Navab and Shadi Albarqouni, Uncertainty-based graph convolutional networks for organ segmentation refinement, Medical Imaging with Deep Learning (2020), 755.

[200] Meng Wang, Lianyu Wang, Xinxing Xu, Ke Zou, Yiming Qian, Rick Siow Mong Goh, Yong Liu and Huazhu Fu, Federated Uncertainty-Aware Aggregation for Fundus Diabetic Retinopathy Staging, Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (2023), ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood and Russell Taylor, 222.

[201] Jiayi Zhu, Bart Bolsterlee, Brian VY Chow, Yang Song and Erik Meijering, Uncertainty and Shape-Aware Continual Test-Time Adaptation for Cross-Domain Segmentation of Medical Images, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 659.

[202] Alireza Norouzi, Ali Emami, Kayvan Najarian, Nader Karimi, SM Reza Soroushmehr, et al., Exploiting uncertainty of deep neural networks for improving segmentation accuracy in MRI images, ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019), 2322.

[203] Glenn W Brier, Verification of forecasts expressed in terms of probability, Monthly weather review **78** (1950), 1.

[204] Sora Iwamoto, Bisser Raytchev, Toru Tamaki and Kazufumi Kaneda, Improving the Reliability of Semantic Segmentation of Medical Images by Uncertainty Modeling with Bayesian Deep Networks and Curriculum Learning, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis (2021), 34.

[205] Rajeev Kumar Singh, Rohan Gorantla, Sai Giridhar Rao Allada and Pratap Narra, SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability, Plos one **17** (2022), e0276836.

[206] Thierry Judge, Olivier Bernard, Mihaela Porumb, Agisilaos Chartsias, Arian Beqiri and Pierre-Marc Jodoin, CRISP-Reliable Uncertainty Estimation for Medical Image Segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference (2022), 492.

[207] Biraja Ghoshal and Allan Tucker, Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection, arXiv e-prints (2020).

[208] Saul Calderon-Ramirez et al., Improving uncertainty estimation with semi-supervised deep learning for covid-19 detection using chest x-ray images, IEEE Access **9** (2021), 85442.

[209] Ruotao Zhang, Constantine Gatsonis and Jon Arni Steingrimsson, Role of calibration in uncertainty-based referral for deep learning, Statistical Methods in Medical Research **32** (2023), 927.

[210] Jörg Sander, Bob D de Vos, Jelmer M Wolterink and Ivana Išgum, Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI, Medical imaging 2019: image Processing **10949** (2019), 324.

[211] Lisa Herzog, Elvis Murina, Oliver Dürr, Susanne Wegener and Beate Sick, Integrating uncertainty in deep neural networks for MRI based stroke analysis, Medical Image Analysis **65** (2020), 101790.

[212] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens and Siegfried Wahl, Leveraging uncertainty information from deep neural networks for disease detection, Scientific reports **7** (2017), 1.

[213] Raghav Mehta et al., QU-BraTS: MICCAI BraTS 2020 Challenge on QuantifyingUncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results, Journal of Machine Learning for Biomedical Imaging **1** (2022).

[214] Andrey Malinin et al., Shifts 2.0: Extending The Dataset of Real Distributional Shifts, arXiv preprint arXiv:2206.15407 (2022).

[215] Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy and Veronica Vilaplana, Uncertainty estimation in deep neural networks for dermoscopic image classification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2020), 744.

[216] Katharina Hoebel, Vincent Andrearczyk, Andrew Beers, Jay Patel, Ken Chang, Adrien Depeursinge, Henning Müller and Jayashree Kalpathy-Cramer, An exploration of uncertainty information for segmentation quality assessment, Medical Imaging 2020: Image Processing **11313** (2020), 381.

[217] Vatsal Raina, Nataliia Molchanova, Mara Graziani, Andrey Malinin, Henning Muller, Meritxell Bach Cuadra and Mark Gales, Tackling Bias in the Dice Similarity Coefficient: Introducing nDSC for White Matter Lesion Segmentation, 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (2022).

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

[219] Parinaz Roshanzamir, Hassan Rivaz, Joshua Ahn, Hamza Mirza, Neda Naghdi, Meagan Anstruther, Michele C Battié, Maryse Fortin and Yiming Xiao, How inter-rater variability relates to aleatoric and epistemic uncertainty: a case study with deep learning-based paraspinal muscle segmentation, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 74.

[220] Yoav Wald, Amir Feder, Daniel Greenfeld and Uri Shalit, On calibration and out-of-domain generalization, Advances in neural information processing systems **34** (2021), 2215.

[221] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz and Mohsen Ali, Towards improving calibration in object detection under domain shift, Advances in Neural Information Processing Systems **35** (2022), 38706.

[222] Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers and Florian Buettner, Post-hoc uncertainty calibration for domain drift scenarios, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), 10124.

[223] Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran and Melinda Gervasio, Confidence calibration for domain generalization under covariate shift, Proceedings of the IEEE/CVF International Conference on Computer Vision (2021), 8958.

[224] Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz Khan, Anirudh Jain, Runa Eschenhagen, Richard E. Turner and Rio Yokota, Practical Deep Learning with Bayesian Principles, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (2019), 4289.

[225] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov and Dmitry P. Vetrov, Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning, 8th International Conference on Learning Representations, ICLR 2020 (2020).

[226] Bjoern H Menze et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Transactions on Medical Imaging **34** (2014), 1993.

[227] Clare Walton et al., Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, Multiple Sclerosis Journal **26** (2020), 1816.

[228] Anthony L Traboulsee and DK Li, The role of MRI in the diagnosis of multiple sclerosis. Advances in neurology **98** (2006), 125.

[229] Olivier Commowick et al., Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset, Neuroimage **244** (2021), 118589.

[230] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L. Cuzzocreo, et al., Longitudinal multiple sclerosis lesion segmentation: Resource and challenge, NeuroImage **148** (2017), 77.

[231] Žiga Lesjak, Alfiia Galimzianova, Aleš Koren, Matej Lukin, Franjo Pernuš, Boštjan Likar and Žiga Špiclin, A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus, Neuroinformatics **16** (2018), 51.

[232] Fabian Isensee et al., Automated brain extraction of multisequence MRI using artificial neural networks, Human brain mapping **40** (2019), 4952.

[233] Kathy Rock, O McArdle, P Forde, M Dunne, D Fitzpatrick, B O'Neill and C Faul, A clinical review of treatment outcomes in glioblastoma multiforme—the validation in a non-trial population of the results of a randomised Phase III clinical trial: has a more radical approach improved survival?, The British journal of radiology **85** (2012), e729.

[234] Neil Grech, Theresia Dalli, Sean Mizzi, Lara Meilak, Neville Calleja and Antoine Zrinzo, Rising incidence of glioblastoma multiforme in a well-defined population, Cureus **12** (2020).

[235] Farina Hanif, Kanza Muzaffar, Kahkashan Perveen, Saima M Malhi and Shabana U Simjee, Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment, Asian Pacific journal of cancer prevention: APJCP **18** (2017), 3.

[236] Gaurav Shukla, Gregory S Alexander, Spyridon Bakas, Rahul Nikam, Kiran Talekar, Joshua D Palmer and Wenyin Shi, Advanced magnetic resonance imaging in glioblastoma: a review, Chin Clin Oncol **6** (2017), 40.

[237] Ujjwal Baid et al., The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification, arXiv preprint arXiv:2107.02314 (2021).

[238] Maruf Adewole et al., The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa), ArXiv (2023).

[239] Valery L Feigin et al., Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, The Lancet Neurology **20** (2021), 795.

[240] Jong S Kim, tPA helpers in the treatment of acute ischemic stroke: are they ready for clinical use?, Journal of Stroke **21** (2019), 160.

[241] Keith W Muir, Alastair Buchan, Rudiger von Kummer, Joachim Rother and Jean-Claude Baron, Imaging of acute stroke, The Lancet Neurology **5** (2006), 755.

[242] Sook-Lei Liew et al., A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms, Scientific data **9** (2022), 320.

[243] Vladimir S Fonov, Alan C Evans, Robert C McKinstry, C Robert Almli and DL Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage **47** (2009), S102.

[244] The MONAI Consortium, *Project MONAI*, Dec. 2020, DOI: 10.5281/zenodo.4323059, URL: https://doi.org/10.5281/zenodo.4323059.

[245] Michał Futrega, Alexandre Milesi, Michał Marcinkiewicz and Pablo Ribalta, Optimized U-Net for brain tumor segmentation, International MICCAI Brainlesion Workshop (2021), 15.

[246] Fernando Pérez-García, Rachel Sparks and Sébastien Ourselin, TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning, Computer Methods and Programs in Biomedicine **208** (2021), 106236.

[247] Silvia Seidlitz et al., Robust deep learning-based semantic organ segmentation in hyperspectral images, Medical Image Analysis **80** (2022), 102488.

[248] Dan Hendrycks and Kevin Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, arXiv preprint arXiv:1610.02136 (2016).

[249] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He and Piotr Dollár, Focal Loss for Dense Object Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence **42** (2020), 318.

[250] Nataliia Molchanova et al., Structural-Based Uncertainty in Deep Learning Across Anatomical Scales: Analysis in White Matter Lesion Segmentation, arXiv preprint arXiv:2311.08931 (2023).

[251] Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Beyond Voxel Prediction Uncertainty: Identifying brain lesions you can trust, Interpretability of Machine Intelligence in Medical Image Computing: 5th International Workshop, iMIMIC 2022, Held in Conjunction with MICCAI 2022, Lecture Notes in Computer Science **13611** (2022), 61.

[252] Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht and Hanno Gottschalk, Prediction error meta classification in semantic

segmentation: Detection via aggregated dispersion measures of softmax probabilities, 2020 International Joint Conference on Neural Networks (IJCNN) (2020), 1.
[253] F. Pedregosa et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research **12** (2011), 2825.
[254] Thomas Lewiner, Hélio Lopes, Antônio Wilson Vieira and Geovan Tavares, Efficient implementation of marching cubes' cases with topological guarantees, Journal of graphics tools **8** (2003), 1.
[255] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals and George E Dahl, Message passing neural networks, Machine learning meets quantum physics (2020), 199.
[256] Xiyuan Wang and Muhan Zhang, How Powerful are Spectral Graph Neural Networks, International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, Proceedings of Machine Learning Research **162** (2022), 23341.
[257] Byung-Hoon Kim and Jong Chul Ye, Understanding graph isomorphism network for rs-fMRI functional connectivity analysis, Frontiers in neuroscience **14** (2020), 630.
[258] Si Zhang, Hanghang Tong, Jiejun Xu and Ross Maciejewski, Graph convolutional networks: a comprehensive review, Computational Social Networks **6** (2019), 1.
[259] Minjie Yu Wang, Deep graph library: Towards efficient and scalable deep learning on graphs, ICLR workshop on representation learning on graphs and manifolds (2019).
[260] Rosana El Jurdi and Olivier Colliot, How Precise are Performance Estimates for Typical Medical Image Segmentation Tasks? (2023), 1.
[261] Rebecca Siegel, Jiemin Ma, Zhaohui Zou and Ahmedin Jemal, Cancer statistics, 2014. CA: a cancer journal for clinicians **64** (2014), 9.
[262] Edwin JR van Beek, Saeed Mirsadraee and John T Murchison, Lung cancer screening: Computed tomography or chest radiographs?, World journal of radiology **7** (2015), 189.
[263] Konstantinos Loverdos, Andreas Fotiadis, Chrysoula Kontogianni, Marianthi Iliopoulou and Mina Gaga, Lung nodules: a comprehensive review on current approach and management, Annals of thoracic medicine **14** (2019), 226.
[264] Peter J Mazzone and Louis Lam, Evaluating the patient with a pulmonary nodule: a review, Jama **327** (2022), 264.
[265] Denise R Aberle et al., Results of the two incidence screenings in the National Lung Screening Trial, New England Journal of Medicine **369** (2013), 920.
[266] Geoffrey D Rubin, Lung nodule and cancer detection in CT screening, Journal of thoracic imaging **30** (2015), 130.
[267] Jinglun Liang, Guoliang Ye, Jianwen Guo, Qifan Huang and Shaohui Zhang, Reducing false-positives in lung nodules detection using balanced datasets, Frontiers in Public Health **9** (2021), 671070.
[268] Samuel G Armato III et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Medical physics **38** (2011), 915.
[269] Hui Yu, Jinqiu Li, Lixin Zhang, Yuzhen Cao, Xuyao Yu and Jinglai Sun, Design of lung nodules segmentation and recognition algorithm based on deep learning, BMC bioinformatics **22** (2021), 1.
[270] Saurabh Jain et al., Two time point MS lesion segmentation in brain MRI: an expectation-maximization framework, Frontiers in neuroscience **10** (2016), 576.

[271] José V Manjón, José E Romero, Roberto Vivo-Hernando, Gregorio Rubio, Fernando Aparici, Maria de La Iglesia-Vaya, Thomas Tourdias and Pierrick Coupé, Blind MRI brain lesion inpainting using deep learning, Simulation and Synthesis in Medical Imaging: 5th International Workshop, SASHIMI 2020, Held in Conjunction with MICCAI 2020 (2020), 41.

[272] Reda Abdellah Kamraoui, Boris Mansencal, José V Manjon and Pierrick Coupé, Longitudinal detection of new MS lesions using deep learning, Frontiers in Neuroimaging **1** (2022), 948235.

[273] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, Generative adversarial networks, Communications of the ACM **63** (2020), 139.

[274] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A Efros, Image-to-image translation with conditional adversarial networks, Proceedings of the IEEE conference on computer vision and pattern recognition (2017), 1125.

[275] Lin Zhang, Lei Zhang, Xuanqin Mou and David Zhang, A comprehensive evaluation of full reference image quality assessment algorithms, 2012 19th IEEE International Conference on Image Processing (2012), 1477.

[276] Zhou Wang, Alan C Bovik, Hamid R Sheikh and Eero P Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing **13** (2004), 600.

[277] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang and Stephen Paul Smolley, Least squares generative adversarial networks, Proceedings of the IEEE international conference on computer vision (2017), 2794.

[278] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz and Bryan Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, Proceedings of the IEEE conference on computer vision and pattern recognition (2018), 8798.

[279] Walter HL Pinaya et al., Generative AI for medical imaging: extending the MONAI framework, arXiv preprint arXiv:2307.15208 (2023).

[280] Taesung Park, Ming-Yu Liu, Ting-Chun Wang and Jun-Yan Zhu, Semantic image synthesis with spatially-adaptive normalization, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019), 2337.

[281] *The BraTS 2023 Challenge website* URL: https://www.synapse.org/#!Synapse:syn51156910/wiki/622461.

[282] Yannick Suter, Urspeter Knecht, Waldo Valenzuela, Michelle Notter, Ekkehard Hewer, Philippe Schucht, Roland Wiest and Mauricio Reyes, The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation, Scientific data **9** (2022), 768.

[283] Richard Shaw, Carole Sudre, Sebastien Ourselin and M. Jorge Cardoso, MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty, Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research **102** (2019), 427.

[284] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen and Paul Suetens, Automated model-based tissue classification of MR images of the brain, IEEE transactions on medical imaging **18** (1999), 897.

[285] Benjamin Billot, Eleanor Robinson, Adrian V Dalca and Juan Eugenio Iglesias, Partial volume segmentation of brain MRI scans of any resolution and contrast, Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference (2020), 177.

[286] Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan and Adolf Pfefferbaum, The SRI24 multichannel atlas of normal adult human brain structure, Human brain mapping **31** (2010), 798.

[287] Ian J Goodfellow, Jonathon Shlens and Christian Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572 (2014).

[288] Anahita Fathi Kazerooni et al., The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics, ArXiv (2023).

[289] Ahmed W Moawad et al., The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri, ArXiv (2023).

[290] Dominic LaBella et al., The ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2023: Intracranial Meningioma, arXiv preprint arXiv:2305.07642 (2023).

[291] Mark Oliver Gueld, Michael Kohnen, et al., Quality of DICOM header information for image categorization, Medical imaging 2002: PACS and integrated medical information systems: design and evaluation **4685** (2002), 280.

[292] *The Head CT CQ500 dataset* URL: http://headctstudy.qure.ai/dataset.

[293] Hugo J Kuijf et al., Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge, IEEE transactions on medical imaging **38** (2019), 2556.

[294] Fernando Pérez-García, Roman Rodionov, et al., Simulation of brain resection for cavity segmentation using self-supervised and semi-supervised learning, Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020 (2020), 115.

[295] *The Information eXtraction from Images (IXI) dataset* URL: https://brain-development.org/ixi-dataset.

[296] A Emre Kavur et al., CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation, Medical Image Analysis **69** (2021), 101950.

[297] Friska Natalia, Hira Meidia, et al., Development of Ground Truth Data for Automatic Lumbar Spine MRI Image Segmentation, HPCC/SmartCity/DSS 2018 (2018), 1449.

[298] Dennis Ulmer and Giovanni Cinà, Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection, Uncertainty in Artificial Intelligence (2021), 1766.

[299] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh and Anton van den Hengel, Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection, Proceedings of the IEEE/CVF International Conference on Computer Vision (2019), 1705.

[300] Hyunjong Park, Jongyoun Noh and Bumsub Ham, Learning memory-guided normality for anomaly detection, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020), 14372.

[301] Camila Gonzalez, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn and Anirban Mukhopadhyay, Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference (2021), 304.

[302] Richard G Brereton, The Mahalanobis distance and its relationship to principal component scores, Journal of Chemometrics **29** (2015), 143.

[303] Haoliang Wang, Chen Zhao, Xujiang Zhao and Feng Chen, Layer Adaptive Deep Neural Networks for Out-of-Distribution Detection, Pacific-Asia Conference on Knowledge Discovery and Data Mining (2022), 526.

[304] Anton Vasiliuk, Daria Frolova, Mikhail Belyaev and Boris Shirokikh, Redesigning Out-of-Distribution Detection on 3D Medical Images, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 126.

[305] Jonathan Lennartz and Thomas Schultz, Segmentation Distortion: Quantifying Segmentation Uncertainty Under Domain Shift via the Effects of Anomalous Activations, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 316.

[306] Evan Hann et al., Quality control-driven image segmentation towards reliable automatic image analysis in large-scale cardiovascular magnetic resonance aortic cine imaging, Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22 (2019), 750.

[307] Khaled ELKarazle, Valliappan Raman, Patrick Then and Caslon Chua, Detection of colorectal polyps from colonoscopy using machine learning: A survey on modern techniques, Sensors **23** (2023), 1225.

[308] Konstantin Pogorelov et al., Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, Proceedings of the 8th ACM on Multimedia Systems Conference (2017), 164.

[309] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray and Bertrand Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, International journal of computer assisted radiology and surgery **9** (2014), 283.

[310] Jorge Bernal, Javier Sánchez and Fernando Vilarino, Towards automatic polyp detection with a polyp appearance model, Pattern Recognition **45** (2012), 3166.

[311] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez and Fernando Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Computerized medical imaging and graphics **43** (2015), 99.

[312] Sharib Ali et al., A multi-centre polyp detection and segmentation dataset for generalisability assessment, Scientific Data **10** (2023), 75.

[313] P Roca, A Attye, et al., Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI, Diagnostic and Interventional Imaging **101** (2020), 795.

[314] Ahmed Ghoneem, Michael T Osborne, et al., Association of socioeconomic status and infarct volume with functional outcome in patients with ischemic stroke, JAMA Network Open **5** (2022), e229178.

[315] Mustafa Mahmut Baris, Ahmet Orhan Celik, et al., Role of mass effect, tumor volume and peritumoral edema volume in the differential diagnosis of primary brain tumor and metastasis, Clinical neurology and neurosurgery **148** (2016), 67.

[316] José Contador, Agnès Pérez-Millán, et al., Longitudinal brain atrophy and CSF biomarkers in early-onset Alzheimer's disease, NeuroImage: Clinical **32** (2021), 102804.

[317] Jakob Wasserthal et al., Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images, Radiology: Artificial Intelligence **5** (2023).
[318] Danijel Kivaranovic, Kory D Johnson and Hannes Leeb, Adaptive, distribution-free prediction intervals for deep networks, International Conference on Artificial Intelligence and Statistics (2020), 4346.
[319] Tim Pearce, Alexandra Brintrup, et al., High-quality prediction intervals for deep learning: A distribution-free, ensembled approach, International Conference on Machine Learning (2018), 4075.
[320] Natasa Tagasovska and David Lopez-Paz, Single-model uncertainties for deep learning, Advances in Neural Information Processing Systems **32** (2019).
[321] Youngseog Chung, Willie Neiswanger, Ian Char and Jeff Schneider, Beyond pinball loss: Quantile methods for calibrated uncertainty quantification, Advances in Neural Information Processing Systems **34** (2021), 10971.
[322] Benjamin Lambert, Florence Forbes, Senan Doyle and Michel Dojat, TriadNet: Sampling-Free Predictive Intervals for Lesional Volume in 3D Brain MR Images, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging - 5th International Workshop, UNSURE 2023, Held in Conjunction with MICCAI 2023, Lecture Notes in Computer Science **14291** (2023), 32.
[323] Vladimir Vovk, Conditional validity of inductive conformal predictors, Asian conference on machine learning (2012), 475.
[324] Samuel Sanford Shapiro and Martin B Wilk, An analysis of variance test for normality (complete samples), Biometrika **52** (1965), 591.
[325] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes and Aaditya Ramdas, Conformal prediction under covariate shift, Advances in neural information processing systems **32** (2019).
[326] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, Bernhard Schölkopf, et al., Covariate shift by kernel mean matching, Dataset shift in machine learning **3** (2009), 5.
[327] Sotirios Panagiotis Chytas, Vishnu Suresh Lokhande, Peiran Li and Vikas Singh, Pooling Image Datasets With Multiple Covariate Shift and Imbalance, arXiv preprint arXiv:2403.02598 (2024).
[328] Jérôme Dockès, Gaël Varoquaux and Jean-Baptiste Poline, Preventing dataset shift from breaking machine-learning biomarkers, GigaScience **10** (2021), giab055.
[329] Steffen Bickel, Michael Brückner and Tobias Scheffer, Discriminative learning for differing training and test distributions, Proceedings of the 24th international conference on Machine learning (2007), 81.
[330] Takafumi Kanamori, Shohei Hido and Masashi Sugiyama, A least-squares approach to direct importance estimation, The Journal of Machine Learning Research **10** (2009), 1391.
[331] Benjamin Rhodes, Kai Xu and Michael U Gutmann, Telescoping density-ratio estimation, Advances in neural information processing systems **33** (2020), 4905.
[332] Kristy Choi, Chenlin Meng, Yang Song and Stefano Ermon, Density ratio estimation via infinitesimal classification, International Conference on Artificial Intelligence and Statistics (2022), 2552.
[333] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei and Tal Schuster, Conformal risk control, arXiv preprint arXiv:2208.02814 (2022).

[334] António Farinhas, Chrysoula Zerva, Dennis Ulmer and André FT Martins, Non-exchangeable conformal risk control, arXiv preprint arXiv:2310.01262 (2023).

[335] Félix Quinton et al., A Tumour and Liver Automatic Segmentation (ATLAS) Dataset on Contrast-Enhanced Magnetic Resonance Imaging for Hepatocellular Carcinoma, Data **8** (2023), 79.

[336] Benjamin Lambert, Florence Forbes, Senan Doyle and Michel Dojat, Anisotropic Hybrid Networks for liver tumor segmentation with uncertainty quantification, Resource-Efficient Medical Image Analysis - 2nd International Workshop, REMIA 2023, Held in Conjunction with MICCAI 2023, Lecture Notes in Computer Science **14394** (2023).

[337] Maria Reig et al., BCLC strategy for prognosis prediction and treatment recommendation: The 2022 update, Journal of Hepatology **76** (Mar. 2022), 681, (visited on 04/11/2022).

[338] Maarten LJ Smits, Mattijs Elschot, Daniel Y Sze, Yung H Kao, Johannes FW Nijsen, Andre H Iagaru, Hugo WAM de Jong, Maurice AAJ van den Bosch and Marnix GEH Lam, Radioembolization dosimetry: the road ahead, Cardiovascular and interventional radiology **38** (2015), 261.

[339] Siqi Liu et al., 3D anisotropic hybrid network: Transferring convolutional features from 2D images to 3D anisotropic volumes, Medical Image Computing and Computer Assisted Intervention: 21st International Conference (2018), 851.

[340] Pedro Furtado, Loss, post-processing and standard architecture improvements of liver deep learning segmentation from Computed Tomography and magnetic resonance, Informatics in Medicine Unlocked **24** (2021), 100585.

[341] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, et al., SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods **17** (2020), 261.

[342] Ziyan Huang et al., STU-Net: Scalable and Transferable Medical Image Segmentation Models Empowered by Large-Scale Supervised Pre-training, arXiv preprint arXiv:2304.06716 (2023).

[343] Patrick Bilic et al., The liver tumor segmentation benchmark (lits), Medical Image Analysis **84** (2023), 102680.

[344] Kamrul SM Hasan and Cristian A Linte, A Multi-Task Cross-Task Learning Architecture for Ad Hoc Uncertainty Estimation in 3D Cardiac MRI Image Segmentation, 2021 Computing in Cardiology (CinC) **48** (2021), 1.

[345] Muhammad Ahtazaz Ahsan, Adnan Qayyum, Adeel Razi and Junaid Qadir, An active learning method for diabetic retinopathy classification with uncertainty quantification, Medical and Biological Engineering and Computing (2022), 1.

[346] Łukasz Rączkowski, Marcin Możejko, Joanna Zambonelli and Ewa Szczurek, ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning, Scientific reports **9** (2019), 1.

[347] Bofan Song et al., Bayesian deep learning for reliable oral cancer image classification, Biomedical Optics Express **12** (2021), 6422.

[348] SM Kamrul Hasan and Cristian A Linte, Calibration of cine MRI segmentation probability for uncertainty estimation using a multi-task cross-task learning architecture, Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling **12034** (2022), 174.

[349] Parisa Mojiri Forooshani et al., Deep Bayesian networks for uncertainty estimation and adversarial resistance of white matter hyperintensity segmentation, Human brain mapping **43** (2022), 2089.

[350] Xiaohong Gou, Xuenong He, et al., Deep learning-based detection and diagnosis of subarachnoid hemorrhage, Journal of Healthcare Engineering **2021** (2021).

[351] Xuyang Cao, Houjin Chen, Yanfeng Li, Yahui Peng, Shu Wang and Lin Cheng, Dilated densely connected U-Net with uncertainty focus loss for 3D ABUS mass segmentation, Computer Methods and Programs in Biomedicine **209** (2021), 106313.

[352] Yidong Zhao, Changchun Yang, Artur Schweidtmann and Qian Tao, Efficient Bayesian Uncertainty Estimation for nnU-Net, Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference (2022), 535.

[353] Ge Zhang, Hao Dang and Yulong Xu, Epistemic and aleatoric uncertainties reduction with rotation variation for medical image segmentation with ConvNets, SN Applied Sciences **4** (2022), 1.

[354] Biraja Ghoshal, Allan Tucker, Bal Sanghera and Wai Lup Wong, Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection, Computational Intelligence **37** (2021), 701.

[355] Jiawei Yang, Yuan Liang, Yao Zhang, Weinan Song, Kun Wang and Lei He, Exploring instance-level uncertainty for medical detection, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 448.

[356] Yongkai Liu, Guang Yang, Melina Hosseiny, Afshin Azadikhah, Sohrab Afshari Mirak, Qi Miao, Steven S Raman and Kyunghyun Sung, Exploring uncertainty measures in Bayesian deep attentive neural networks for prostate zonal segmentation, IEEE Access **8** (2020), 151817.

[357] Ishaan Bhat and Hugo J Kuijf, Extending Probabilistic U-Net Using MC-Dropout to Quantify Data and Model Uncertainty, International MICCAI Brainlesion Workshop (2022), 555.

[358] Milda Pocevičiūtė, Gabriel Eilertsen, Sofia Jarkman and Claes Lundström, Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology, Scientific Reports **12** (2022), 1.

[359] Saul Calderon-Ramirez, Diego Murillo-Hernandez, Kevin Rojas-Salazar, Luis-Alexander Calvo-Valverd, Shengxiang Yang, Armaghan Moemeni, David Elizondo, Ezequiel Lopez-Rubio and Miguel A Molina-Cabello, Improving uncertainty estimations for mammogram classification using semi-supervised learning, 2021 International Joint Conference on Neural Networks (IJCNN) (2021), 1.

[360] Dwarikanath Mahapatra, Alexander Poellinger, Ling Shao and Mauricio Reyes, Interpretability-driven sample selection using self supervised learning for disease classification and segmentation, IEEE Transactions on Medical Imaging **40** (2021), 2548.

[361] Zakaria Senousy, Mohammed M Abdelsamea, Mohamed Medhat Gaber, Moloud Abdar, U Rajendra Acharya, Abbas Khosravi and Saeid Nahavandi, MCUa: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification, IEEE Transactions on Biomedical Engineering **69** (2021), 818.

[362] Joohyun Lee, Dongmyung Shin, Se-Hong Oh and Haejin Kim, Method to Minimize the Errors of AI: Quantifying and Exploiting Uncertainty of Deep Learning in Brain Tumor Segmentation, Sensors **22** (2022), 2406.

[363] Axel-Jan Rousseau, Thijs Becker, Jeroen Bertels, Matthew B Blaschko and Dirk Valkenborg, Post training uncertainty calibration of deep networks for medical image segmentation, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 1052.

[364] Adrian Tousignant, Paul Lemaître, Doina Precup, Douglas L Arnold and Tal Arbel, Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data, International conference on medical imaging with deep learning (2019), 483.

[365] Onur Ozdemir, Benjamin Woodward and Andrew A Berlin, Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection, 2nd Workshop on Bayesian Deep Learning (NeurIPS 2017) (2017).

[366] Huitong Pan, Yushan Feng, Quan Chen, Craig Meyer and Xue Feng, Prostate segmentation from 3d mri using a two-stage model and variable-input based uncertainty measure, 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (2019), 468.

[367] Pieter Van Molle, Tim Verbelen, Cedric De Boom, Bert Vankeirsbilck, Jonas De Vylder, Bart Diricx, Tom Kimpe, Pieter Simoens and Bart Dhoedt, Quantifying uncertainty of deep neural networks in skin lesion classification, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures (2019), 52.

[368] Aryan Mobiny, Aditi Singh and Hien Van Nguyen, Risk-aware machine learning classifier for skin lesion diagnosis, Journal of clinical medicine **8** (2019), 1241.

[369] Zain Ul Abideen, Mubeen Ghafoor, Kamran Munir, Madeeha Saqib, Ata Ullah, Tehseen Zia, Syed Ali Tariq, Ghufran Ahmed and Asma Zahra, Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks, IEEE Access **8** (2020), 22812.

[370] Raghav Mehta, Angelos Filos, Yarin Gal and Tal Arbel, Uncertainty evaluation metric for brain tumour segmentation, Medical Imaging with Deep Learning (2020).

[371] Sidi Yang and Thomas Fevens, Uncertainty Quantification and Estimation in Medical Image Classification, International Conference on Artificial Neural Networks (2021), 671.

[372] Max-Heinrich Laves, Sontje Ihler and Tobias Ortmaier, Uncertainty Quantification in Computer-Aided Diagnosis: Make Your Model say" I don't know" for Ambiguous Cases, Medical Imaging with Deep Learning (2019).

[373] Sivaramakrishnan Rajaraman, Ghada Zamzmi, Feng Yang, Zhiyun Xue, Stefan Jaeger and Sameer K Antani, Uncertainty Quantification in Segmenting Tuberculosis-Consistent Findings in Frontal Chest X-rays, Biomedicines **10** (2022), 1323.

[374] Yingda Xia et al., Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation, Medical Image Analysis **65** (2020), 101766.

[375] Ekaterina Redekop and Alexey Chernyavskiy, Uncertainty-based method for improving poorly labeled segmentation datasets, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 1831.

[376] Alain Jungo, Raphael Meier, Ekin Ermis, Evelyn Herrmann and Mauricio Reyes, Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation, Medical Imaging with Deep Learning (2018).

[377] James M Dolezal et al., Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology, Nature communications **13** (2022), 6572.

[378] Yanan Ruan et al., Mt-UcGAN: Multi-task uncertainty-constrained GAN for joint segmentation, quantification and uncertainty estimation of renal tumors on CT, Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference (2020), 439.

[379] Lin Hu, Jiaxin Li, Xingchen Peng, Jianghong Xiao, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou and Yan Wang, Semi-supervised NPC segmentation with uncertainty and attention guided consistency, Knowledge-Based Systems **239** (2022), 108021.

[380] Ling Huang, Su Ruan, Pierre Decazes and Thierry Denœux, Lymphoma segmentation from 3D PET-CT images using a deep evidential network, International Journal of Approximate Reasoning **149** (2022), 39.

[381] Anton Vasiliuk, Daria Frolova, Mikhail Belyaev and Boris Shirokikh, Limitations of Out-of-Distribution Detection in 3D Medical Image Segmentation, Journal of Imaging **9** (2023).

[382] Qiao Lin, Xin Chen, Chao Chen and Jonathan M. Garibaldi, A Novel Quality Control Algorithm for Medical Image Segmentation Based on Fuzzy Uncertainty, IEEE Transactions on Fuzzy Systems **31** (2023), 2532.

[383] Yinglin Zhang, Ruiling Xi, Huazhu Fu, Dave Towey, RuiBin Bai, Risa Higashita and Jiang Liu, Elongated Physiological Structure Segmentation via Spatial andScale Uncertainty-Aware Network, Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (2023), 323.

[384] Shishuai Wang, Johan Nuyts and Marina Filipovic, Uncertainty Estimation in Liver Tumor Segmentation Using the Posterior Bootstrap, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 188.

[385] Elina Thibeau-Sutre, Dieuwertje Alblas, Sophie Buurman, Christoph Brune and Jelmer M Wolterink, Uncertainty-based quality assurance of carotid artery wall segmentation in black-blood MRI, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 95.

[386] Thomas Buddenkotte, Lorena Escudero Sanchez, Mireia Crispin-Ortuzar, Ramona Woitek, Cathal McCague, James D Brenton, Ozan Öktem, Evis Sala and Leonardo Rundo, Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation, Computers in Biology and Medicine (2023), 107096.

[387] Jasper Linmans, Stefan Elfwing, Jeroen van der Laak and Geert Litjens, Predictive uncertainty estimation for out-of-distribution detection in digital pathology, Medical Image Analysis **83** (2023), 102655.

[388] Hao Li, Yang Nan, Javier Del Ser and Guang Yang, Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation, Neural Computing and Applications **35** (2023), 22071.

[389] Afshar Shamsi, Hamzeh Asgharnezhad, Shirin Shamsi Jokandan, Abbas Khosravi, Parham M Kebria, Darius Nahavandi, Saeid Nahavandi and Dipti Srinivasan, An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis, IEEE Transactions on neural networks and learning systems **32** (2021), 1408.

[390] Fumin Guo, Matthew Ng, Grey Kuling and Graham Wright, Cardiac MRI Segmentation With Sparse Annotations: Ensembling Deep Learning Uncertainty and Shape Priors, Medical Image Analysis (2022), 102532.

[391] Murat Seçkin Ayhan, Laura Kühlewein, Gulnar Aliyeva, Werner Inhoffen, Focke Ziemssen and Philipp Berens, Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection, Medical Image Analysis **64** (2020), 101724.

[392] Jimut Bahan Pal, Holistic Network for Quantifying Uncertainties in Medical Images, International MICCAI Brainlesion Workshop (2022), 560.

[393] Minh H Vu, Tufve Nyholm and Tommy Löfstedt, Multi-decoder networks with multi-denoising inputs for tumor segmentation, International MICCAI Brainlesion Workshop (2020), 412.

[394] Alireza Mehrtash, Tina Kapur, Clare M Tempany, Purang Abolmaesumi and William M Wells, Prostate Cancer Diagnosis With Sparse Biopsy Data And In Presence Of Location Uncertainty, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 443.

[395] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang and Danny Z Chen, Suggestive annotation: A deep active learning framework for biomedical image segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2017), 399.

[396] Xi Wang, Fangyao Tang, Hao Chen, Luyang Luo, Ziqi Tang, An-Ran Ran, Carol Y Cheung and Pheng-Ann Heng, UD-MIL: uncertainty-driven deep multiple instance learning for OCT image classification, IEEE journal of biomedical and health informatics **24** (2020), 3431.

[397] Haochen Mei, Wenhui Lei, Ran Gu, Shan Ye, Zhengwentai Sun, Shichuan Zhang and Guotai Wang, Automatic segmentation of gross target volume of nasopharynx cancer using ensemble of multiscale deep neural networks with spatial attention, Neurocomputing **438** (2021), 211.

[398] Wenjing Lu, Jiahao Lei, Peng Qiu, Rui Sheng, Jinhua Zhou, Xinwu Lu and Yang Yang, UPCoL: Uncertainty-Informed Prototype Consistency Learning for Semi-supervised Medical Image Segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 662.

[399] Adrian Galdran, Johan W Verjans, Gustavo Carneiro and Miguel A González Ballester, Multi-Head Multi-Loss Model Calibration, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 108.

[400] Evan Hann, Iulia A Popescu, Qiang Zhang, Ricardo A Gonzales, Ahmet Barutçu, Stefan Neubauer, Vanessa M Ferreira and Stefan K Piechnik, Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping, Medical image analysis **71** (2021), 102029.

[401] Natália Alves, Joeran S Bosma, Kiran V Venkadesh, Colin Jacobs, Zaigham Saghir, Maarten de Rooij, John Hermans and Henkjan Huisman, Prediction variability to identify reduced AI performance in cancer diagnosis at MRI and CT, Radiology **308** (2023), e230275.

[402] Xiaoyan Zhang et al., Generalizability and quality control of deep learning-based 2D echocardiography segmentation models in a large clinical dataset, The International Journal of Cardiovascular Imaging **38** (2022), 1685.

[403] Charles Lu, Anastasios N Angelopoulos and Stuart Pomerantz, Improving Trustworthiness of AI Disease Severity Rating in Medical Imaging with Ordinal Conformal Prediction Sets, Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference (2022), 545.

[404] Guotai Wang et al., Interactive medical image segmentation using deep learning with image-specific fine tuning, IEEE Transactions on Medical Imaging **37** (2018), 1562.

[405] Pedram Mojabi, Vahab Khoshdel and Joe Lovetri, Tissue-type classification with uncertainty quantification of microwave and ultrasound breast imaging: A deep learning approach, IEEE Access **8** (2020), 182092.

[406] João Lourenço-Silva and Arlindo L Oliveira, Using Soft Labels to Model Uncertainty in Medical Image Segmentation, International MICCAI Brainlesion Workshop (2022), 585.

[407] Haoneng Lin, Zongshang Li, Zefan Yang and Yi Wang, Variance-aware attention U-Net for multi-organ segmentation, Medical Physics **48** (2021), 7864.

[408] Håkan Wieslander, Philip J Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan Karlsson, Ola Spjuth and Carolina Wählby, Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images, IEEE journal of biomedical and health informatics **25** (2020), 371.

[409] Jayaraman J Thiagarajan, Bindya Venkatesh, Deepta Rajan and Prasanna Sattigeri, Improving reliability of clinical models using prediction calibration, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020 (2020), 71.

[410] Agostina J Larrazabal, César Martínez, Jose Dolz and Enzo Ferrante, Maximum Entropy on Erroneous Predictions: Improving Model Calibration for Medical Image Segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 273.

[411] Xiaojiao Xiao, Qinmin Vivian Hu and Guanghui Wang, Edge-Aware Multi-task Network for Integrating Quantification Segmentation and Uncertainty Prediction of Liver Tumor on Multi-modality Non-contrast MRI, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 652.

[412] Xiangyu Zhao, Zhenrong Shen, Dongdong Chen, Sheng Wang, Zixu Zhuang, Qian Wang and Lichi Zhang, One-Shot Traumatic Brain Segmentation with Adversarial Training and Uncertainty Rectification, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 120.

[413] Jiayi Zhu, Bart Bolsterlee, Brian VY Chow, Yang Song and Erik Meijering, Uncertainty and Shape-Aware Continual Test-Time Adaptation for Cross-Domain Segmentation of Medical Images, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 659.

[414] Balamurali Murugesan, Sukesh Adiga Vasudeva, Bingyuan Liu, Hervé Lombaert, Ismail Ben Ayed and Jose Dolz, Trust your neighbours: Penalty-based constraints for model calibration, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 572.

[415] Neerav Karani, Neel Dey and Polina Golland, Boundary-weighted logit consistency improves calibration of segmentation networks, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 367.

[416] Changjian Shui, Justin Szeto, Raghav Mehta, Douglas L Arnold and Tal Arbel, Mitigating calibration bias without fixed attribute grouping for improved fairness in medical imaging analysis, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 189.

[417] Ben Philps, Maria del C Valdes Hernandez and Miguel Bernabeu Llinares, Proper Scoring Loss Functions Are Simple and Effective for Uncertainty Quantification of White Matter Hyperintensities, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 208.

[418] Richard McKinley et al., Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence, NeuroImage: Clinical **25** (2020), 102104.

[419] Richard McKinley, Raphael Meier and Roland Wiest, Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation, International MICCAI Brainlesion Workshop (2018), 456.

[420] Suman Sedai, Bhavna Antony, Dwarikanath Mahapatra and Rahil Garnavi, Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using Bayesian deep learning, Computational Pathology and Ophthalmic Medical Image Analysis (2018), 219.

[421] Suraj Mishra, Danny Z Chen and X Sharon Hu, Objective-dependent uncertainty driven retinal vessel segmentation, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 453.

[422] Richard McKinley, Michael Rebsamen, Raphael Meier and Roland Wiest, Triplanar ensemble of 3D-to-2D CNNs with label-uncertainty for brain tumor segmentation, International MICCAI Brainlesion workshop (2019), 379.

[423] Teresa Araújo, Guilherme Aresta, Luís Mendonça, Susana Penas, Carolina Maia, Ângela Carneiro, Ana Maria Mendonça and Aurélio Campilho, DR| GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images, Medical Image Analysis **63** (2020), 101715.

[424] Hooman Vaseli, Ang Nan Gu, S Neda Ahmadi Amiri, Michael Y Tsang, Andrea Fung, Nima Kondori, Armin Saadat, Purang Abolmaesumi and Teresa SM Tsang, ProtoAS-Net: Dynamic Prototypes for Inherently Interpretable and Uncertainty-Aware Aortic Stenosis Classification in Echocardiography, International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), 368.

[425] Wufeng Xue, Tingting Guo and Dong Ni, Left ventricle quantification with sample-level confidence estimation via Bayesian neural network, Computerized Medical Imaging and Graphics **84** (2020), 101753.

[426] Qiao Lin, Xin Chen, Chao Chen and Jonathan M Garibaldi, Quality quantification in deep convolutional neural networks for skin lesion segmentation using fuzzy uncertainty measurement, 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (2022), 1.

[427] Esther Puyol-Antón, Bram Ruijsink, Christian F Baumgartner, Pier-Giorgio Masci, Matthew Sinclair, Ender Konukoglu, Reza Razavi and Andrew P King, Automated quantification of myocardial tissue characteristics from native T 1 mapping using neural networks with uncertainty-based quality-control, Journal of Cardiovascular Magnetic Resonance **22** (2020), 1.

[428] Kai Ren, Ke Zou, Xianjie Liu, Yidi Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang and Huazhu Fu, Uncertainty-Informed Mutual Learning for Joint Medical Image Classification and Segmentation, Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (2023), ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood and Russell Taylor, 35.

[429] Craig K Jones, Guoqing Wang, Vivek Yedavalli and Haris Sair, Direct quantification of epistemic and aleatoric uncertainty in 3D U-net segmentation, Journal of Medical Imaging **9** (2022), 034002.

[430] Juan E Arco, Andres Ortiz, Javier Ramirez, Francisco J Martinez-Murcia, Yu-Dong Zhang and Juan M Gorriz, Uncertainty-driven ensembles of multi-scale deep architectures for image classification, Information Fusion **89** (2023), 53.

[431] Yan Li, Xiaoyi Chen, Li Quan and Ni Zhang, Uncertainty-Guided Robust Training For Medical Image Segmentation, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 1471.

[432] Eric W Prince, Debashis Ghosh, Carsten Görg and Todd C Hankinson, Uncertainty-Aware Deep Learning Classification of Adamantinomatous Craniopharyngioma from Preoperative MRI, Diagnostics **13** (2023), 1132.

[433] Yizhe Zhang, Shuo Wang, Yejia Zhang and Danny Z Chen, RR-CP: Reliable-Region-Based Conformal Prediction for Trustworthy Medical Image Classification, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 12.

[434] Hendrik Mehrtens, Tabea Bucher and Titus J Brinker, Pitfalls of Conformal Predictions for Medical Image Classification, International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (2023), 198.

[435] Benjamin Lambert, Maxime Louis, Senan Doyle, Florence Forbes, Michel Dojat and Alan Tucholka, Leveraging 3D information in unsupervised brain MRI segmentation, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 187.

[436] Benjamin Lambert, Florence Forbes, Alan Tucholka, Senan Doyle and Michel Dojat, Multi-Scale Evaluation of Uncertainty Quantification Techniques for Deep Learning based MRI Segmentation, ISMRM-ESMRMB & ISMRT 2022-31st Joint Annual Meeting International Society for Magnetic Resonance in Medecine (2022), 1.

[437] Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Uncertainty-based Quality Control for Subcortical Structures Segmentation in T1-weighted Brain MRI, ISMRM-ESMRMB & ISMRT 2023-32nd Joint Annual Meeting International Society for Magnetic Resonance in Medecine (2023).

[438] Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Fast Uncertainty Quantification for Deep Learning-based MR Brain Segmentation, EGC 2022-Conference francophone pour l'Extraction et la Gestion des Connaissances (2022), 1.

[439] Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Intervalles de confiance pour l'estimation de superficies à partir d'images satellitaires, GRETSI 2023-XXIXème Colloque Francophone de Traitement du Signal et des Images (2023), 1.

[440] Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Safety-Net: Identification automatique des erreurs de segmentation des lésions de

la Sclérose-en-Plaques, Société Française de Résonance Magnétique en Biologie et Médecine (2023).

Appendix

A1 Lesion matching edge cases

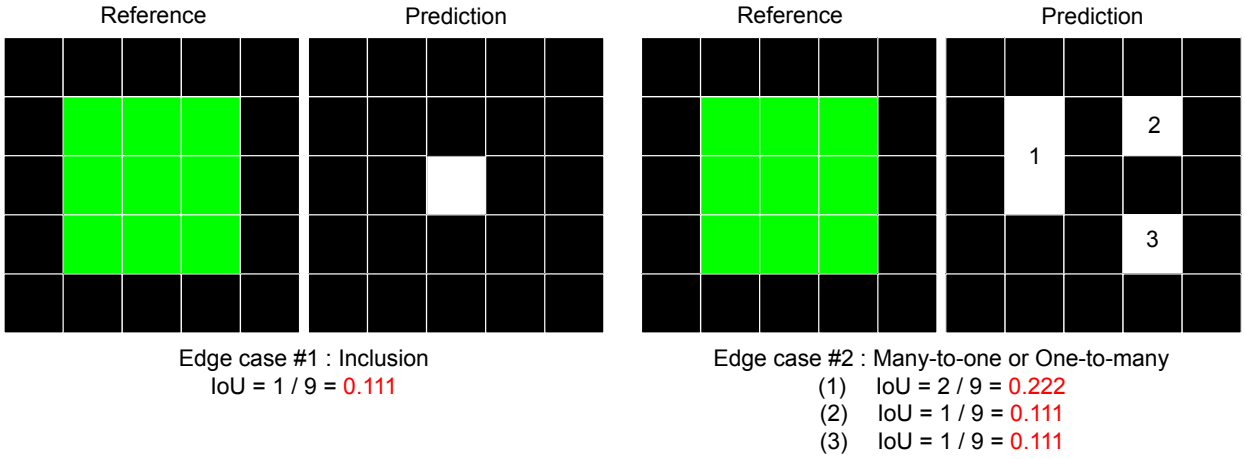


Figure A.1.1: Illustration of two edge cases that can occur when matching predicted and reference lesions. In these cases, using a rigorous IoU threshold of 25% or above would consider all predicted lesions as FP_{les}, although the detection is correct.

A2 Hyper-parameters of lesion classifiers

Model	Parameters	Tested values
Logistic Regression	C	1, 2, 5, 10
Random Forest	N trees	10, 20, 50, 100, 150
SVC	Kernel	Linear, RBF, Poly
	C	1, 2, 5, 10
	Degree	3, 4, 5

Table A.2.1: Hyper-parameters tested during the grid-search cross-validation for the lesion ML classifiers.

A3 Feature importance of Machine Learning classifiers

Feature Name	Category	Coefficient
glrlm RunEntropy	FLAIR	1.82
Average Interior Entropy	Entropy	1.79
Maximum2DDiameterRow	Shape	1.45
Maximum2DDiameterSlice	Shape	1.20
Average Contour Entropy	Entropy	1.18
glcm MaximumProbability	FLAIR	1.06
Maximum3DDiameter	Shape	0.91
glrlm RunLengthNonUniformityNormalized	FLAIR	0.86
glrlm RunPercentage	FLAIR	0.80

Table A.3.1: Weight of the top 10 features in the Logistic Regression model trained on cross-sectional MS lesions.

Feature Name	Category	Coefficient
gldm DependenceVariance	CT	1.06
Surface Area	Shape	1.06
glszm SizeZoneNonUniformityNormalized	CT	0.93
Least Axis Length	Shape	0.83
ngtdm Complexity	CT	0.79
Maximum2DDiameter Slice	CT	0.73
Average Interior Entropy	Entropy	0.72
glcm ClusterProminence	CT	0.69
Maximum2DDiameterColumn	CT	0.64

Table A.3.2: Weight of the top 10 features in the Logistic Regression model trained on lung nodules.

Feature Name	Category	Coefficient
First order TotalEnerg	FLAIR V1	1.27
First order Energy	FLAIR V1	1.27
glcm Imc2	FLAIR V0	1.21
First order RootMeanSquared	FLAIR V1	1.17
First order 90Percentile	FLAIR V1	0.91
First order Median	FLAIR V0	0.90
Average Interior Entropy	Entropy	0.86
glszm ZonePercentage	FLAIR V0	0.86
glcm DifferenceVariance	FLAIR V0	0.81

Table A.3.3: Weight of the top 10 features in the Logistic Regression model trained on new MS lesions. FLAIR V0 designates the prior FLAIR while FLAIR V1 designates the posterior FLAIR.

Samples from BraTS 2023 datasets

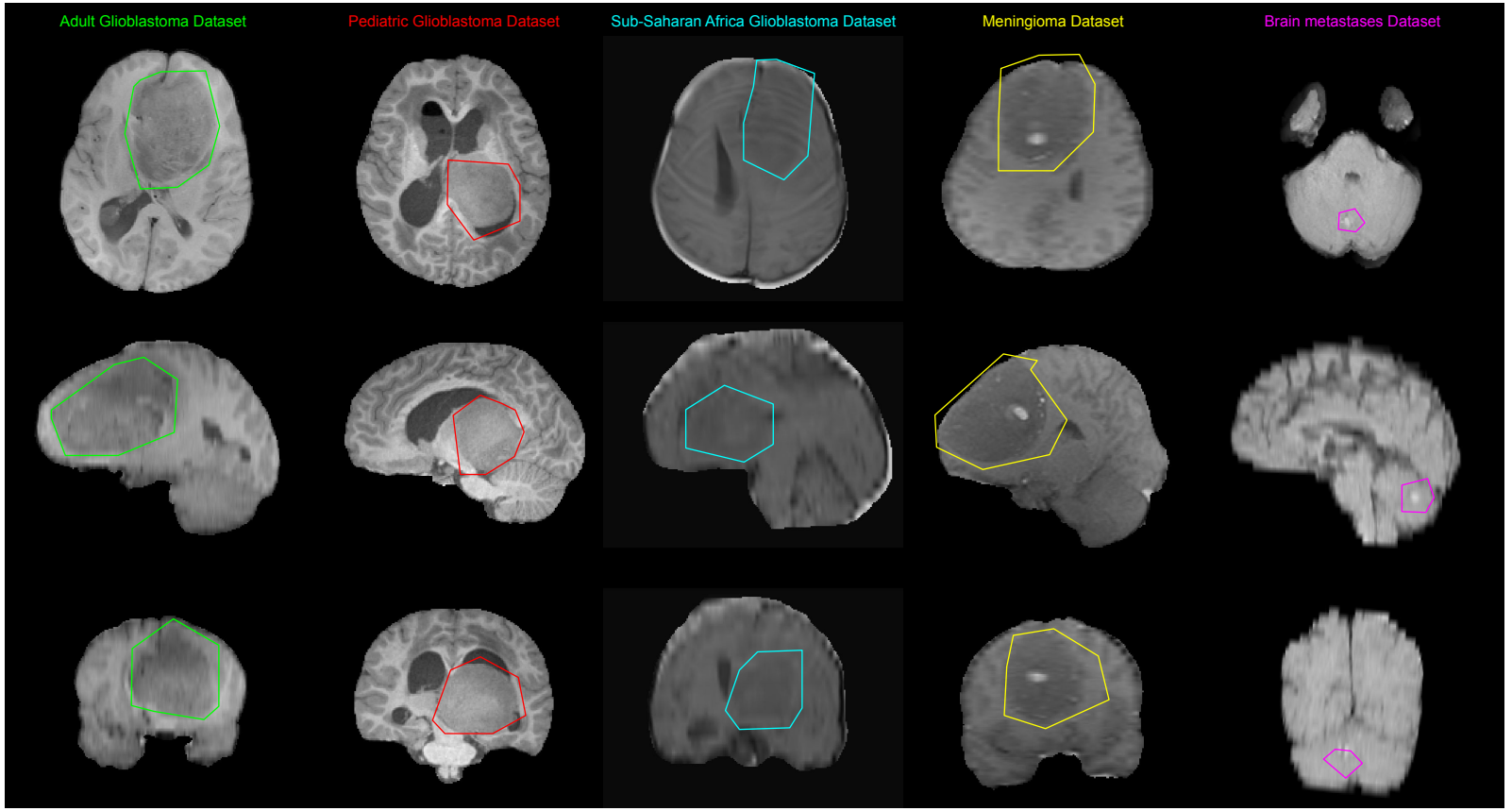


Figure A.3.1: Samples from BraTS 2023 datasets, exhibiting the variability in tumor appearance, location, and size according to the dataset. Pediatric brain MR exhibits important differences in appearance as compared to adult brain MR, for instance, due to the overabundance of gray matter in early childhood. Images from the sub-saharan dataset have a lower resolution and present artifacts. Meningiomas are challenging due to the difference in their location and appearance as compared to glioblastoma. Metastases exhibit heterogeneous sizes and lesions can be particularly small.

A4 Additional OOD benchmark results

This section presents the results of the OOD benchmark with the Attention U-Net, V-Net and Residual U-Net ensembles, respectively. More particularly, the segmentation performance of each ensemble on each OOD dataset is presented in Figures A.4.1, A.4.2, A.4.3.

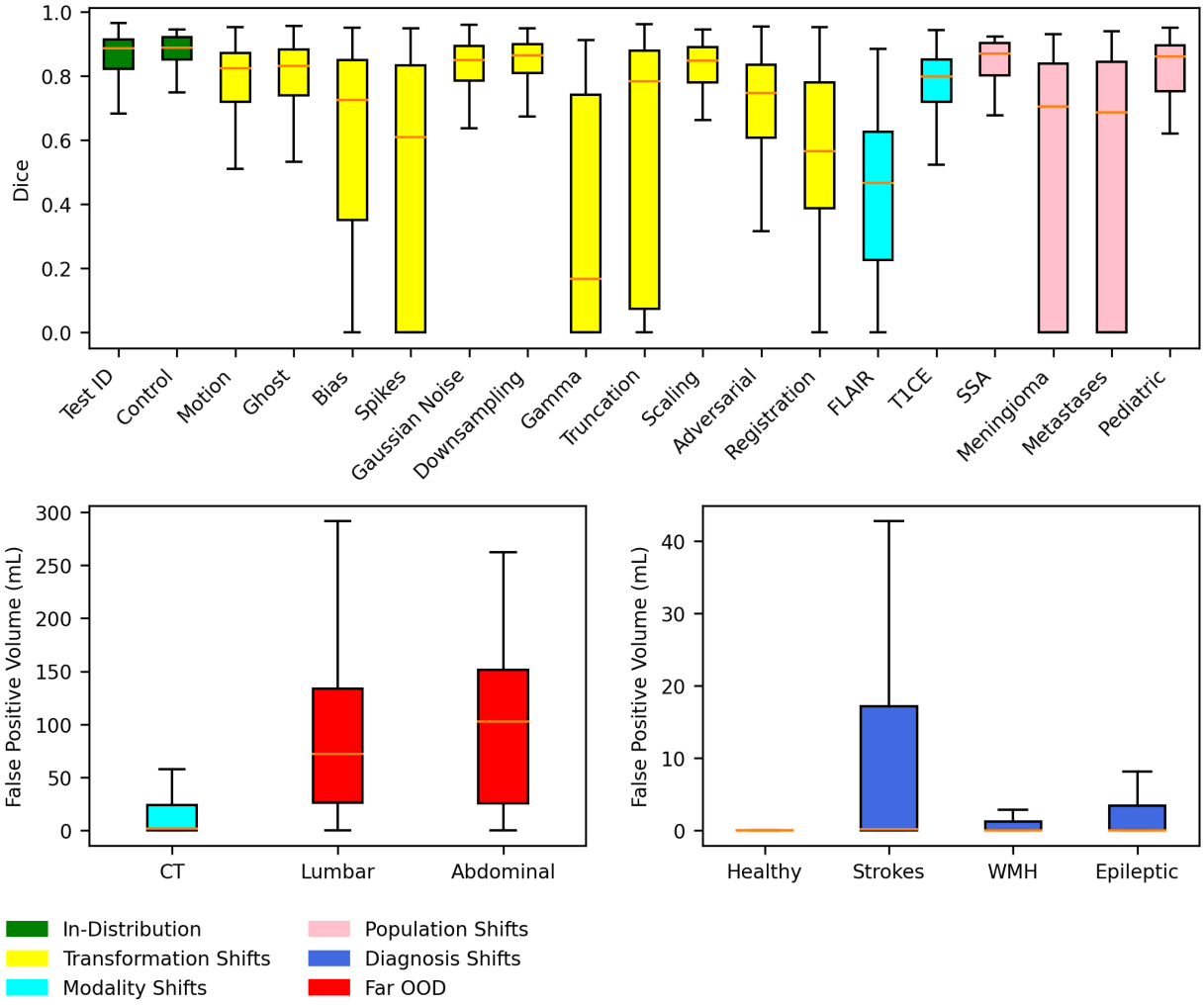


Figure A.4.1: Segmentation performance of the V-Net Ensemble on the different datasets used in the OOD experiments. The average Dice is presented for datasets where the ground truth delineation of the whole tumor is available. For the rest of the datasets, we present the average False Positive (FP) volume per subject, in milliliter.

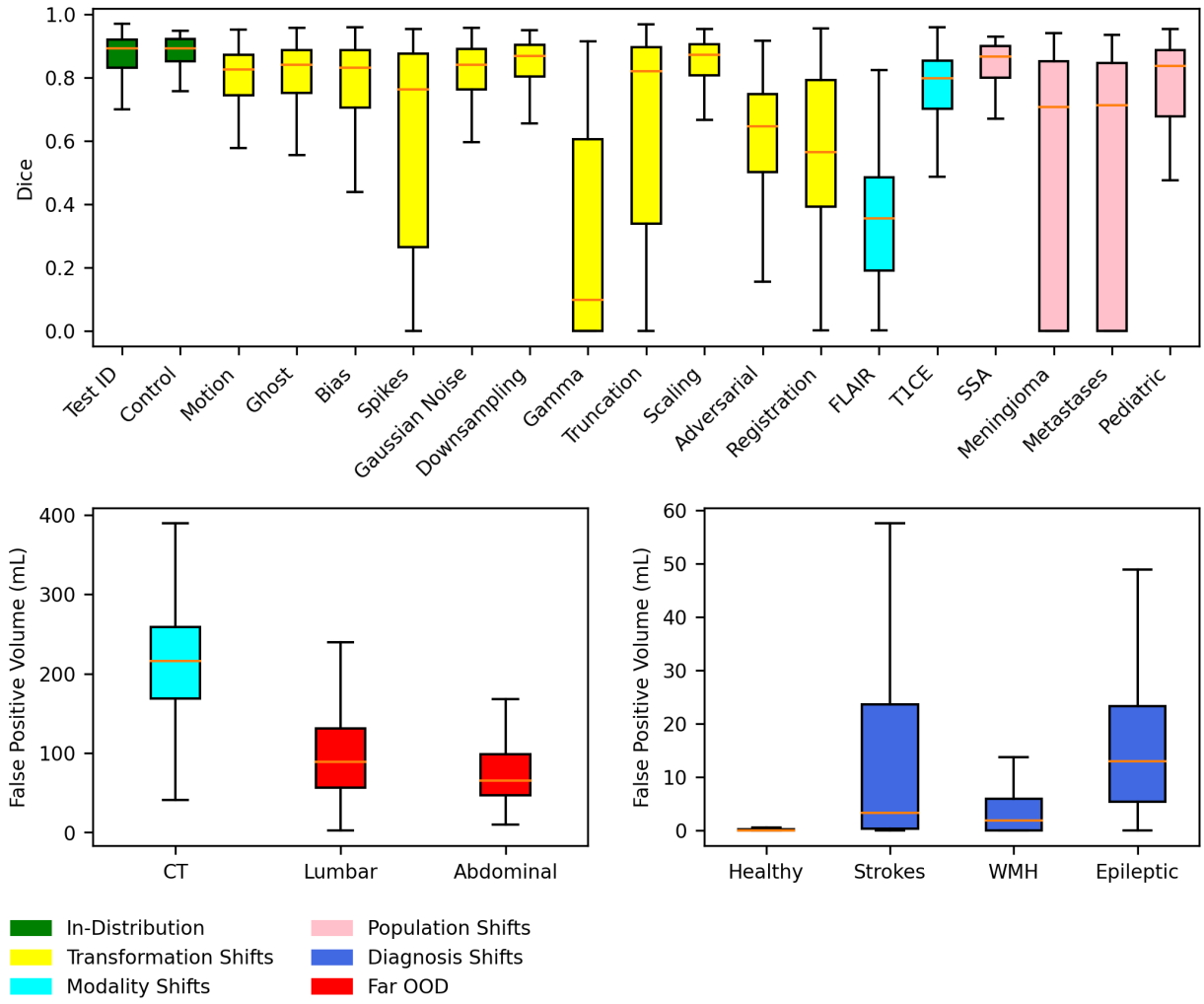


Figure A.4.2: Segmentation performance of the Attention U-Net Ensemble on the different datasets used in the OOD experiments. The average Dice is presented for datasets where the ground truth delineation of the whole tumor is available. For the rest of the datasets, we present the average False Positive (FP) volume per subject, in milliliter.

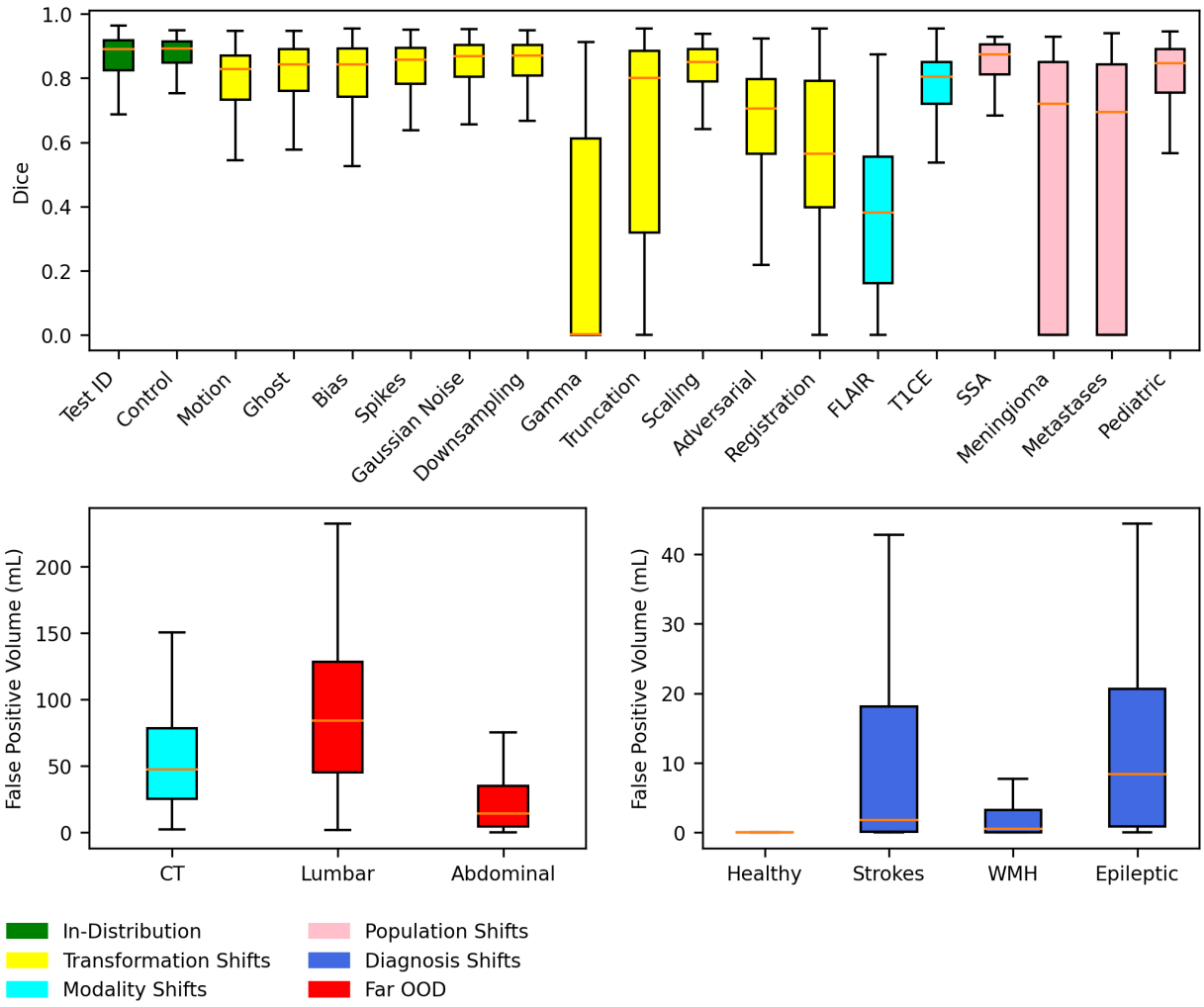


Figure A.4.3: Segmentation performance of the Residual U-Net Ensemble on the different datasets used in the OOD experiments. The average Dice is presented for datasets where the ground truth delineation of the whole tumor is available. For the rest of the datasets, we present the average False Positive (FP) volume per subject, in milliliter.

Type	ID	Transformation											Modality			Diagnosis				Population				Far		Avg
Dataset	Control	Motion	Ghost	Bias	Spikes	Noise	Downsample	Gamma	Truncation	FGSM	Registration	Scaling	FLAIR	T1CE	CT	Healthy	Strokes	WMH	Epilepsy	SSA	Meningioma	Metastases	Pediatric	Abdominal	Lumbar	Average
Deep Ensemble	52	58	69	97	95	91	64	100	52	98	53	64	100	83	100	29	46	65	42	80	53	56	55	100	100	72
MNAD	78	71	99	98	100	100	29	99	85	87	65	98	88	85	99	53	76	71	56	61	47	43	70	100	100	78
MD Conv 1	29	54	73	98	88	79	45	97	95	52	61	98	71	67	62	46	63	84	56	43	59	37	58	100	100	69
MD Conv 2	34	71	77	100	100	100	88	99	94	75	71	97	88	83	94	37	70	95	62	61	62	68	62	100	100	80
MD Conv 3	51	84	88	100	100	100	84	100	96	72	60	98	99	92	99	87	88	97	87	79	79	81	83	100	100	88
MD Conv 4	41	82	91	100	100	100	85	100	93	79	81	94	92	90	96	36	70	94	62	75	64	70	68	100	100	83
MD Conv 5	50	85	92	100	100	100	83	100	91	73	89	94	93	91	96	49	70	93	74	81	65	73	73	100	100	85
MD Conv 6	56	82	85	100	100	100	79	100	92	67	71	95	98	95	99	78	83	95	76	85	78	81	81	100	100	87
MD Conv 7	55	87	92	100	100	100	87	100	91	70	88	93	95	94	97	42	68	92	76	85	68	76	80	100	100	85
MD Conv 8	56	85	93	100	100	100	88	100	90	73	83	93	96	93	97	39	68	92	76	85	69	74	81	99	99	85
MD Conv 9	50	87	92	100	100	100	88	100	91	72	75	94	97	94	97	41	79	93	66	83	71	73	84	99	99	85
MD Conv 10	59	85	92	100	100	100	88	100	89	70	81	93	95	94	96	37	69	92	74	86	69	73	81	99	99	85
MD Conv 11	56	85	92	100	100	100	89	100	87	69	80	93	97	93	97	41	69	90	79	87	69	73	83	100	100	85
MD Conv 12	50	88	94	100	100	100	83	99	94	75	73	95	98	91	98	23	59	84	56	82	58	61	82	100	100	82
MD Conv 13	54	84	93	100	100	100	87	100	92	68	80	94	97	92	97	33	62	84	67	85	66	69	85	100	100	84
MD Conv 14	52	86	93	100	100	100	84	100	92	69	77	93	97	93	97	28	59	82	63	87	64	67	82	100	100	83
MD Mean	52	86	92	100	100	100	87	100	93	74	82	95	97	93	97	55	77	94	75	84	73	78	81	100	100	87
MD Max	54	83	90	100	100	100	83	100	93	72	80	96	97	93	98	83	86	96	84	82	78	81	83	99	100	88

Table A.4.1: OOD detection performance (AUROC, expressed in percentage) for each OOD detector and dataset, for the Attention U-Net ensemble.

Type	ID	Transformation											Modality			Diagnosis				Population				Far		Avg
Dataset	Control	Motion	Ghost	Bias	Spikes	Noise	Downsample	Gamma	Truncation	FGSM	Registration	Scaling	FLAIR	T1CE	CT	Healthy	Strokes	WMH	Epilepsy	SSA	Meningioma	Metastases	Pediatric	Abdominal	Lumbar	Average
Deep Ensemble	25	56	66	96	94	90	61	100	52	98	53	71	100	83	100	39	55	64	38	53	55	55	44	100	100	70
MNAD	56	71	99	99	100	100	39	99	89	80	70	98	90	86	99	53	75	60	48	31	53	49	56	100	100	76
MD Conv 1	17	52	68	94	90	76	45	97	93	52	59	98	67	63	60	46	60	68	45	18	59	45	44	100	100	65
MD Conv 2	19	67	77	100	100	100	77	99	88	68	66	94	87	81	91	45	76	93	61	26	64	63	38	100	100	75
MD Conv 3	26	79	84	100	100	100	75	100	92	66	57	95	98	89	98	75	86	93	77	41	76	75	63	100	100	82
MD Conv 4	21	76	87	100	100	100	74	100	86	71	73	88	90	86	93	43	75	92	61	34	66	62	42	100	99	77
MD Conv 5	24	77	87	100	100	100	72	100	83	66	82	88	90	86	92	48	73	89	65	40	65	66	45	100	100	77
MD Conv 6	27	76	80	100	100	100	68	100	86	61	64	90	96	91	95	67	80	89	64	49	75	75	58	99	100	80
MD Conv 7	26	79	86	100	100	100	76	100	83	64	81	86	92	89	93	44	71	88	66	45	66	68	54	99	99	78
MD Conv 8	27	77	87	100	100	100	76	100	82	66	74	85	92	89	92	44	72	88	66	45	67	67	56	99	95	78
MD Conv 9	24	80	86	98	100	100	78	100	83	67	67	87	94	89	93	44	77	88	58	45	67	66	61	98	95	78
MD Conv 10	29	77	87	99	100	100	76	100	80	64	73	86	92	89	92	43	71	87	64	47	66	66	56	98	95	77
MD Conv 11	28	77	86	99	100	100	77	100	78	63	73	86	93	89	93	46	70	84	67	49	67	67	60	99	99	78
MD Conv 12	25	82	88	99	100	100	73	99	87	69	67	89	95	86	94	38	64	79	52	47	60	58	60	100	99	76
MD Conv 13	27	78	88	99	100	100	76	100	84	64	73	88	93	86	92	43	65	79	58	46	65	64	62	100	99	77
MD Conv 14	26	79	88	100	100	100	73	99	85	64	71	86	93	88	93	41	64	79	55	49	64	63	61	100	99	77
MD Mean	24	78	86	100	100	100	75	100	86	67	73	89	94	88	93	51	76	89	65	44	69	69	56	100	100	79
MD Max	26	76	83	100	100	100	72	100	87	65	72	89	95	88	94	70	83	90	72	42	74	73	59	99	99	80

Table A.4.2: OOD detection performance (AUPR, expressed in percentage) for each OOD detector and dataset, for the Attention U-Net ensemble.

Type	ID	Transformation											Modality			Diagnosis				Population				Far		Avg
Dataset	Control	Motion	Ghost	Bias	Spikes	Noise	Downsample	Gamma	Truncation	FGSM	Registration	Scaling	FLAIR	T1CE	CT	Healthy	Strokes	WMH	Epilepsy	SSA	Meningioma	Metastases	Pediatric	Abdominal	Lumbar	Average
Deep Ensemble	49	57	69	91	83	70	57	92	45	97	55	59	100	83	99	0	17	11	12	81	30	35	53	100	100	62
MNAD	78	71	99	98	100	100	29	99	85	87	65	98	88	85	99	53	76	71	56	61	47	43	70	100	100	78
MD Conv 1	33	53	69	98	88	77	44	97	99	51	64	100	70	65	61	38	57	80	50	59	60	42	60	99	100	69
MD Conv 2	41	68	70	100	100	94	68	99	99	55	86	99	94	95	99	38	62	77	41	80	58	56	61	100	100	78
MD Conv 3	41	74	74	100	100	99	76	100	97	52	92	99	98	96	100	52	72	83	76	80	63	62	67	100	100	82
MD Conv 4	54	71	70	100	100	100	82	100	99	58	93	97	97	96	100	44	71	90	81	85	64	70	73	100	100	84
MD Conv 5	59	82	78	100	100	99	89	100	97	56	93	99	98	97	100	47	71	90	84	88	66	74	80	100	100	86
MD Conv 6	60	85	83	100	100	100	91	100	94	54	86	97	98	97	100	41	66	86	75	89	66	75	82	100	100	85
MD Conv 7	55	87	83	100	100	100	90	100	97	54	90	98	99	97	100	37	67	88	70	90	65	75	81	100	100	85
MD Conv 8	57	85	86	100	99	99	90	100	96	54	88	98	99	97	100	41	67	86	72	89	65	72	84	100	100	85
MD Conv 9	55	84	87	100	99	98	82	100	95	52	82	98	98	95	100	44	70	91	69	84	65	67	78	100	100	84
MD Conv 10	45	82	86	100	98	97	78	100	96	53	76	98	98	95	99	45	68	92	63	81	66	63	78	100	100	82
MD Conv 11	52	81	83	100	99	99	83	100	99	55	84	98	99	97	100	33	59	82	60	88	62	63	83	100	100	82
MD Conv 12	43	81	87	100	99	98	80	100	98	55	84	99	99	97	100	31	60	85	58	87	61	60	83	100	100	82
MD Conv 13	44	83	89	100	98	97	77	100	97	54	75	98	99	93	99	28	49	73	48	82	56	53	79	100	100	79
MD Conv 14	51	87	92	100	99	99	91	100	96	54	82	98	99	97	100	60	75	92	78	87	67	70	86	100	100	86
MD Conv 15	51	88	91	100	99	99	90	100	96	53	75	97	99	95	100	59	76	92	70	87	69	70	88	100	100	86
MD Conv 16	53	89	91	100	100	100	88	100	97	57	85	98	99	96	100	57	78	94	78	88	68	73	87	100	100	87
MD Conv 17	51	88	91	100	99	99	83	100	97	58	80	98	99	96	100	51	74	92	73	87	70	71	89	100	100	86
MD Conv 18	50	87	91	100	100	99	82	100	98	59	82	99	100	97	100	57	75	91	71	88	72	72	87	100	100	86
MD Conv 19	41	84	89	100	99	97	78	100	99	58	69	100	99	91	100	61	71	92	67	81	73	72	82	100	100	84
MD Mean	49	84	86	100	100	99	86	100	99	56	88	99	99	97	100	45	69	90	70	87	67	70	82	100	100	85
MD Max	49	83	86	100	100	99	81	100	99	56	87	99	99	96	100	49	69	89	71	84	67	69	83	100	100	85

Table A.4.3: OOD detection performance (AUROC, expressed in percentage) for each OOD detector and dataset, for the V-Net ensemble.

Type	ID	Transformation											Modality			Diagnosis				Population				Far		Avg
Dataset	Control	Motion	Ghost	Bias	Spikes	Noise	Downsample	Gamma	Truncation	FGSM	Registration	Scaling	FLAIR	T1CE	CT	Healthy	Strokes	WMH	Epilepsy	SSA	Meningioma	Metastases	Pediatric	Abdominal	Lumbar	Average
Deep Ensemble	23	55	66	90	86	67	55	93	49	97	55	68	100	80	99	33	37	27	26	58	43	45	34	100	100	63
MNAD	56	71	99	99	100	100	39	99	89	80	70	98	90	86	99	53	75	60	48	31	53	49	56	100	100	76
MD Conv 1	18	51	63	94	89	73	45	97	98	51	63	99	65	61	62	42	55	63	41	28	60	46	47	99	100	64
MD Conv 2	20	61	65	100	100	92	57	99	98	53	83	98	92	93	97	42	59	63	39	46	53	50	37	100	100	72
MD Conv 3	23	68	70	100	100	98	66	100	94	53	89	97	97	95	99	49	66	70	64	47	61	55	49	100	100	76
MD Conv 4	29	66	69	100	100	99	73	100	99	56	91	95	96	95	98	45	71	86	71	57	65	64	52	100	100	79
MD Conv 5	31	78	76	100	100	99	83	100	95	54	92	97	97	95	98	46	70	84	74	65	65	69	61	100	100	81
MD Conv 6	33	79	79	100	100	99	84	100	92	52	82	95	97	96	99	44	66	79	64	64	63	70	63	100	100	80
MD Conv 7	30	83	81	100	100	99	84	100	95	52	87	96	98	96	98	42	69	83	62	63	65	71	62	100	100	81
MD Conv 8	29	81	83	100	99	99	83	100	93	53	86	97	98	95	98	44	69	81	64	61	64	68	63	100	100	80
MD Conv 9	26	80	84	100	99	97	73	100	92	51	78	96	97	93	98	46	73	88	62	50	65	61	52	100	100	79
MD Conv 10	22	78	83	100	98	96	69	100	92	52	71	96	97	93	98	45	73	89	58	48	66	58	51	100	100	77
MD Conv 11	28	79	81	100	99	98	76	100	97	54	82	96	99	96	99	41	65	80	57	62	63	60	67	100	100	79
MD Conv 12	23	79	84	100	99	98	72	100	96	54	83	97	99	95	99	41	67	83	56	58	63	57	66	100	100	79
MD Conv 13	22	81	86	100	99	97	70	100	94	53	72	96	98	91	99	39	57	69	48	51	57	51	57	100	100	75
MD Conv 14	27	83	87	100	99	99	85	100	94	53	80	96	99	96	98	54	75	88	69	59	66	65	70	100	100	82
MD Conv 15	26	85	88	100	99	98	83	100	93	52	72	95	99	94	98	54	76	88	63	59	67	65	72	100	100	81
MD Conv 16	28	87	88	100	100	99	80	100	95	55	83	96	99	95	98	52	77	90	69	58	67	67	70	100	100	82
MD Conv 17	26	86	88	100	99	99	76	100	95	55	77	97	99	95	98	49	75	88	66	58	68	65	72	100	100	81
MD Conv 18	25	85	89	100	100	99	75	100	96	56	78	97	99	95	98	52	75	87	63	62	69	64	69	100	100	81
MD Conv 19	20	81	86	100	99	97	70	100	97	57	67	98	98	90	98	54	72	87	60	49	70	64	59	100	100	79
MD Mean	26	80	83	100	100	99	77	100	97	54	85	97	98	96	98	45	70	85	62	58	65	63	62	100	100	80
MD Max	25	80	82	100	100	99	71	100	98	54	85	97	98	95	98	47	69	83	61	52	64	62	60	100	100	79

Table A.4.4: OOD detection performance (AUPR, expressed in percentage) for each OOD detector and dataset, for the V-Net ensemble. The highest score for each dataset is indicated in **bold**.

Type	ID	Transformation											Modality			Diagnosis				Population				Far		Avg
Dataset	Control	Motion	Ghost	Bias	Spikes	Noise	Downsample	Gamma	Truncation	FGSM	Registration	Scaling	FLAIR	T1CE	CT	Healthy	Strokes	WMH	Epilepsy	SSA	Meningioma	Metastases	Pediatric	Abdominal	Lumbar	Average
MNAD	78	71	99	98	100	100	29	99	85	87	65	98	88	85	99	53	76	71	56	61	47	43	70	100	100	78
MD Conv 1	29	51	71	91	83	80	39	93	100	98	60	99	65	63	57	26	46	66	38	60	55	38	60	100	100	67
MD Conv 2	48	97	99	100	100	100	96	99	100	100	81	100	93	93	100	42	62	86	68	85	61	75	64	100	100	86
MD Conv 3	45	95	99	100	100	100	94	100	99	93	71	100	95	95	99	62	65	89	74	91	63	77	70	100	100	87
MD Conv 4	25	52	71	96	86	81	39	96	100	97	58	99	68	66	59	32	52	75	44	55	56	36	59	100	100	68
MD Conv 5	47	93	100	100	100	100	90	100	100	98	71	100	97	95	99	57	67	85	70	86	67	75	75	100	100	87
MD Conv 6	50	96	99	100	100	100	90	100	99	97	74	100	97	95	99	68	72	91	75	84	71	76	80	100	100	89
MD Conv 7	54	98	100	100	100	100	86	100	99	99	69	100	99	94	99	73	75	91	73	81	72	78	85	100	100	89
MD Conv 8	54	95	100	100	100	100	91	100	100	99	76	100	97	96	99	61	68	87	75	89	68	78	77	100	100	88
MD Conv 9	53	97	100	100	100	100	90	100	100	99	73	100	99	95	100	64	72	90	69	88	71	79	84	100	100	89
MD Conv 10	50	97	100	100	100	100	88	100	100	98	79	100	99	96	100	63	73	91	70	87	68	75	84	100	100	89
MD Conv 11	51	97	100	100	100	100	80	100	100	96	77	100	100	95	100	56	66	87	67	87	68	72	86	100	100	87
MD Conv 12	52	97	100	100	100	100	89	100	100	99	73	100	99	96	100	60	70	88	70	88	71	78	83	100	100	88
MD Conv 13	42	94	100	100	100	100	78	100	100	99	75	100	99	96	100	36	60	82	50	86	65	69	84	100	100	85
MD Conv 14	42	96	100	100	100	100	81	100	100	98	85	100	99	95	100	48	62	85	56	90	66	69	84	100	100	86
MD Conv 15	39	89	99	100	100	99	69	100	100	92	84	100	99	93	99	34	46	71	51	87	61	58	83	100	100	82
MD Conv 16	41	92	100	100	100	100	77	100	100	99	75	100	99	96	100	35	61	83	50	86	64	69	84	100	100	84
MD Conv 17	41	94	100	100	100	100	83	100	100	98	81	100	99	94	100	40	59	85	52	90	69	72	84	100	100	86
MD Conv 18	37	89	100	100	100	99	69	100	100	90	86	100	100	93	100	34	49	77	48	90	64	63	84	100	100	83
MD Conv 19	44	81	98	100	97	90	57	100	100	81	78	100	100	85	100	34	38	57	38	87	63	60	81	100	100	79
MD Conv 20	42	91	100	100	100	100	78	100	100	99	75	100	99	96	99	38	61	82	49	86	66	70	84	100	100	85
MD Conv 21	42	84	99	100	100	99	68	100	100	84	78	100	99	94	100	59	64	86	59	82	62	64	85	100	100	84
MD Conv 22	51	89	99	100	100	99	79	100	100	89	71	100	100	95	100	86	84	96	78	83	75	79	88	100	100	90
MD Conv 23	51	86	98	100	100	99	76	100	99	85	66	100	99	93	100	85	84	94	81	81	78	80	89	100	100	89
MD Mean	44	95	100	100	100	100	88	100	100	99	75	100	99	95	100	59	67	89	68	87	69	76	81	100	100	88
MD Max	49	95	99	100	100	100	92	100	100	99	72	100	99	95	99	80	80	92	77	89	76	82	86	100	100	90

Table A.4.5: OOD detection performance (AUROC, expressed in percentage) for each OOD detector and dataset, for the Residual U-Net ensemble. The highest score for each dataset is indicated in **bold**.

Type	ID	Transformation											Modality			Diagnosis				Population				Far		Avg
Dataset	Control	Motion	Ghost	Bias	Spikes	Noise	Downsample	Gamma	Truncation	FGSM	Registration	Scaling	FLAIR	T1CE	CT	Healthy	Strokes	WMH	Epilepsy	SSA	Meningioma	Metastases	Pediatric	Abdominal	Lumbar	Average
MNAD	56	71	99	99	100	100	39	99	89	80	70	98	90	86	99	53	75	60	48	31	53	49	56	100	100	76
MD Conv 1	17	48	65	80	78	68	41	92	100	98	61	99	56	57	55	38	46	48	33	24	52	43	40	100	100	62
MD Conv 2	22	95	97	100	100	100	91	98	100	99	78	100	91	92	99	45	62	77	54	54	58	71	40	100	100	81
MD Conv 3	22	91	98	100	100	100	88	99	99	89	68	100	92	93	98	55	64	78	56	63	59	73	45	100	100	81
MD Conv 4	16	48	64	89	84	72	42	95	99	97	57	99	60	60	56	40	51	56	37	23	54	43	41	100	100	63
MD Conv 5	23	90	98	99	100	99	83	100	99	95	68	100	95	94	98	52	66	72	56	52	62	67	53	100	100	81
MD Conv 6	23	93	98	100	100	99	81	100	99	94	69	100	96	92	97	59	71	84	62	48	66	68	58	100	100	82
MD Conv 7	25	96	98	100	100	98	76	100	99	97	67	100	98	93	98	62	72	83	61	46	67	70	65	100	100	83
MD Conv 8	26	93	98	99	100	100	85	100	100	97	73	100	95	94	98	54	66	75	59	57	63	70	56	100	100	82
MD Conv 9	26	95	98	99	100	100	83	100	100	97	71	100	98	94	98	57	72	83	58	56	67	72	65	100	100	83
MD Conv 10	24	95	98	99	100	99	79	100	99	96	77	100	98	94	98	56	74	85	59	58	65	69	65	100	100	83
MD Conv 11	25	95	98	99	99	98	73	100	99	93	75	99	99	94	98	52	70	83	58	59	66	67	69	100	100	83
MD Conv 12	26	94	98	99	100	100	82	100	100	98	72	100	98	94	98	54	70	81	58	56	67	72	65	100	100	83
MD Conv 13	22	92	99	100	100	100	72	100	100	97	74	100	98	95	98	42	63	75	45	58	63	64	68	100	100	81
MD Conv 14	21	94	99	100	100	99	75	100	99	96	84	100	98	94	99	47	66	81	50	66	65	64	67	100	100	83
MD Conv 15	19	87	98	99	99	97	64	100	98	87	83	99	98	91	98	41	52	60	44	62	60	55	66	100	100	78
MD Conv 16	22	90	98	100	100	100	72	100	100	97	73	100	98	95	98	41	64	76	44	57	62	64	69	100	100	81
MD Conv 17	22	91	99	100	100	100	76	100	100	95	79	100	98	93	99	43	64	79	47	64	66	66	67	100	100	82
MD Conv 18	19	88	100	100	100	98	64	100	100	87	85	100	99	91	99	41	56	67	42	68	63	58	69	100	100	80
MD Conv 19	21	80	97	100	97	88	55	100	100	78	77	100	99	83	99	41	44	45	34	58	60	57	63	100	100	75
MD Conv 20	23	89	98	100	100	100	73	100	100	97	73	100	98	95	98	42	64	75	44	57	64	65	68	100	100	81
MD Conv 21	23	83	98	100	100	97	64	100	100	82	76	100	99	93	99	55	67	80	53	57	64	61	71	100	100	81
MD Conv 22	25	87	97	99	99	98	72	100	99	85	69	99	98	94	98	76	81	91	67	55	72	74	74	100	100	84
MD Conv 23	23	82	96	99	100	98	68	100	98	79	63	99	98	91	98	73	78	85	65	50	73	71	73	100	100	82
MD Mean	21	93	98	100	100	100	80	100	100	97	73	100	97	94	98	53	68	81	56	56	64	68	61	100	100	82
MD Max	23	91	98	100	100	100	84	100	99	98	70	100	98	93	98	67	73	81	59	58	69	74	65	100	100	84

Table A.4.6: OOD detection performance (AUPR, expressed in percentage) for each OOD detector and dataset, for the Residual U-Net ensemble. The highest score for each dataset is indicated in **bold**.

A5 Additional notes on conformal score functions

In this Appendix section, we motivate the choice of the score functions used to find the corrective values $\hat{q}$ on the predictive intervals (PIs). We also clarify the link between the choice of the quantile $\hat{q}$ and the target $(1-\alpha)$ coverage level.

Sampling-based Intervals

For sampling-based intervals, a PI on the volume X has the following form:

$$\Gamma_\alpha(X) = [\mu(X) - z\sigma(X), \mu(X) + z\sigma(X)]$$

where μ_X and σ_X are the mean and the standard deviation estimated by sampling, respectively, and z is the number of standard deviations to match the target confidence level (e.g. $z = 1.65$ for 90% PIs). The ground truth volume Y is contained within the interval if:

$$\mu_X - \sigma(X) \times z \leq Y \leq \mu(X) + \sigma(X) \times z$$

which can be written equivalently as:

$$|Y - \mu(X)| \leq \sigma(X) \times z \Rightarrow \frac{|Y - \mu(X)|}{\sigma(X)} \leq z$$

which is exactly the score function used to calibrate sampling-based PIs:

$$s(X, Y) = \frac{|Y - \mu(X)|}{\sigma(X)}$$

Writing $(X_i, Y_i)_{i=1,...,n}$ the calibration dataset, the multiplicative corrective value $\hat{q}$ is computed as $\hat{q} = \text{Quantile}(s_1, s_2, ..., s_n, \frac{(n+1)(1-\alpha)}{n})$. The conformalized PIs are further obtained by replacing z by $\hat{q}$:

$$[\mu_i - \sigma_i\hat{q}, \mu_i + \sigma_i\hat{q}]$$

It makes sure that at least $1 - \alpha$ of the PIs will encompass the ground truth values on the calibration dataset. Now, writing (X_{test}, Y_{test}) a fresh test datapoint and $\Gamma_\alpha(X_{test})$ the corresponding PI, we have the following result:

$$P(Y_{test} \in \Gamma_\alpha(X_{test})) \geq 1-\alpha \Rightarrow P(\frac{|Y_{test} - \mu(test)|}{\sigma(test)} \leq \hat{q}) \geq 1-\alpha \Rightarrow P(s(X_{test} \leq \hat{q}) \geq 1-\alpha$$

Direct PI estimation

In sampling-free PI estimation, a $(1-\alpha)\%$ PI on the volume X has the following form:

$$\Gamma_\alpha(X) = [\hat{t}_{\alpha/2}(X), \hat{t}_{1-\alpha/2}(X)]$$

where $\hat{t}_{\alpha/2}$ and $\hat{t}_{1-\alpha/2}$ are the estimated quantiles, allowing to get $(1-\alpha\%)$ coverages. For 90% PI, the ground truth volume Y is supposed to land below $\hat{t}_{0.05}(X)$ with 5% probability and above $\hat{t}_{0.95}(X)$ with 5% probability. The Y is contained within the interval if $\hat{t}_{\alpha/2}(X) \leq Y_i \leq \hat{t}_{1-\alpha/2}(X)$. The score function is defined as the difference between Y_i and its nearest quantile:

$$s(X,Y) = \max\{\hat{t}_{\alpha/2}(X) - Y, Y - \hat{t}_{1-\alpha/2}(X)\}$$

The interpretation of the score is as follows. When Y is below the lower bound, the magnitude of the error is $|Y - \hat{t}_{\alpha/2}(X)|$. Alternatively, if Y is superior to the upper bound, the magnitude of the error is $|Y - \hat{t}_{1-\alpha/2}(X)|$. If Y is correctly bounded by the interval, then the score corresponds to the larger of the two negative numbers $\{\hat{t}_{\alpha/2}(X) - Y, Y - \hat{t}_{1-\alpha/2}(X)\}$. This formulation allows to correct for both potential under and over-coverages.

As for sampling-based PI, the corrective value is taken as $\hat{q} = \text{Quantile}(s_1, s_2, ..., s_n, \frac{(n+1)(1-\alpha)}{n})$ and the conformalized PIs are further obtained as :

$$\Gamma_\alpha(X) = [\hat{t}_{\alpha/2}(X) - \hat{q}, \hat{t}_{1-\alpha/2}(X) + \hat{q}]$$

Now, we have the following equivalence [78]:

$$\{Y_{test} \in \Gamma_\alpha(X_{test})\} \Leftrightarrow \{s_{test} \leq \hat{q}\}$$

and thus:

$$P(Y_{test} \in \Gamma_\alpha(X_{test})) \geq 1-\alpha \Rightarrow P(s_{test} \leq \hat{q} \geq 1-\alpha)$$

A6 Disgression on conformal risk control for thresholds tuning

Conformal Risk Control (CRC) has been proposed as a generalization of the standard Conformal Prediction framework to control any monotone loss function [333]. In this appendix, we propose to investigate how that can be used in practice to control the decision thresholds of segmentation models.

Motivations

One limit of our investigation of CP for lesion volumes (Chapter V) is that we do not consider the prediction adequacy at the voxel level, but only at the image level. Let's consider a simple setting where a ground truth mask contains 10 mL of lesions. If a segmentation model makes 10 mL of FP predictions and 10 mL of FN predictions, it will end up predicting the correct lesion load of 10 mL, even though there is no intersection between the predicted and reference masks. Moreover, coverage can be seen as a $0-1$ loss, as the ground truth (e.g. volume) is either contained in the predictive interval or not. Yet, in many real-world problems including image segmentation, the notion of error is continuous, with different mistakes having different costs (e.g. FNR, FDR). For example, in the context of medical image segmentation, we may want to control the proportion of false positive or negative voxels in a segmentation. These quantities are continuous and depend on the decision threshold of the neural network, generally set to 0.5 for binary problems. Reducing the threshold with increase the number of FP and decrease the number of FN. In contrast, increasing the threshold results in more conservative predictions with higher FNs but lower FPs. This concept is illustrated in Figure A.6.1. To solve this challenge, Conformal Risk Control (CRC) [333] has been proposed as a generalization of split conformal prediction to control the expectation of any monotone loss function (here, FDR or FNR). CRC acts as a quality-assurance policy, providing statistical guarantees that the loss on unseen test data will be, in expectation, equal or inferior to a user-defined threshold (e.g. 5%, 10%).

Mathematical Framework

In this section, the concept of Conformal Risk Control is introduced, as presented in [333], and we discuss how it can be implemented in the setting of medical image segmentation. We focus on simplicity in binary segmentation tasks, although the described procedure could be applied with multi-class segmentation networks. Additionally, we introduce our method for 2D images, although the process is strictly identical for 3D images, which have an extra spatial dimension d.

We consider a trained segmentation model f that maps input images $x \in \mathcal{X}$ to a probability map $f : \mathcal{X} \rightarrow [0,1]^{h \times w}$. Using a calibration dataset $\{X_i, Y_i\}_{i=1}^{n}$ composed of pairs of images and associated ground truth segmentations, we aim at building a *predictive region* $\mathcal{C}$ by post-processing the probability map predicted by f such as:

$$\mathbb{E}[\ell(\mathcal{C}_\lambda(X_{n+1}), Y_{n+1})] \leq \alpha \tag{.6.1}$$

where $\ell \in [-\infty, B]$ is a bounded loss function, and λ is the parameter that we want to optimize, controlling the size of the predictive region. In the following we will write $L_i(\lambda) = \ell(\mathcal{C}_\lambda(X_i), Y_i)$ for simplicity. The CRC algorithm will find the optimal λ to control the risk ℓ at a user-defined threshold α, using the n calibration data points, by solving:

$$\begin{aligned} \hat{\lambda} = \inf \left\{ \lambda : \frac{n}{n+1}\widehat{R_n}(\lambda) + \frac{B}{n+1} \leq \alpha \right\} \\ \text{where } \widehat{R_n} = \frac{1}{n}\sum_{i=1}^{n} L_i(\lambda) \end{aligned} \tag{.6.2}$$

Concretely, for segmentation models, λ corresponds to the decision threshold used to binarize the probability map. We will apply CRC to control two antagonist losses separately, namely the FNR and FDR, which are meaningful for medical image analysis. The CRC procedure takes the following forms for FNR and FDR control:

$$\begin{aligned} \ell_{FNR}(\mathcal{C}_\lambda(X_i), Y_i) = 1 - \frac{|Y_i \cap \mathcal{C}_\lambda(X_i)|}{|Y_i|} = \frac{FN}{FN+TP} \\ \text{with } \mathcal{C}_\lambda(X_i) = \{y : f(X_i) \geq 1 - \lambda\} \end{aligned} \tag{.6.3}$$

$$\begin{aligned} \ell_{FDR}(\mathcal{C'}_\lambda(X_i), Y_i) = \frac{|\overline{Y_i} \cap \mathcal{C'}_\lambda(X_i)|}{|\mathcal{C'}_\lambda(X_i)|} = \frac{FP}{FP+TP} \\ \text{with } \mathcal{C'}_\lambda(X_i) = \{y : f(X_i) \geq \lambda\} \end{aligned} \tag{.6.4}$$

where FP, TP, and FN correspond to False Positive, True Positive, and False Negative predictions respectively. Note that this procedure is extremely similar to the usual CP procedure performed on predictive intervals. The principal difference is that the $0-1$ miscoverage loss $\mathbf{1}_{\{Y_{test} \notin \mathcal{C}(X_{test})\}}$ is replaced by the monotone risk $\ell(\mathcal{C}_\lambda(X_{test}), Y_{test})$.

Experimental setting

We illustrate the CRC procedure on a task of polyp segmentation in 2D colonoscopy images. The datasets have been presented in Chapter 3 IV.4.4 and a summary of the polyp datasets is presented in Table A.9.1. Briefly, a dataset composed of 1312 images is used to train a 2D segmentation DynU-Net, trained with the Dice++ loss II.6.3. After training, CRC is used to control the FDR and FNR at a level of $\alpha = 0.10$ using the calibration split composed of 438

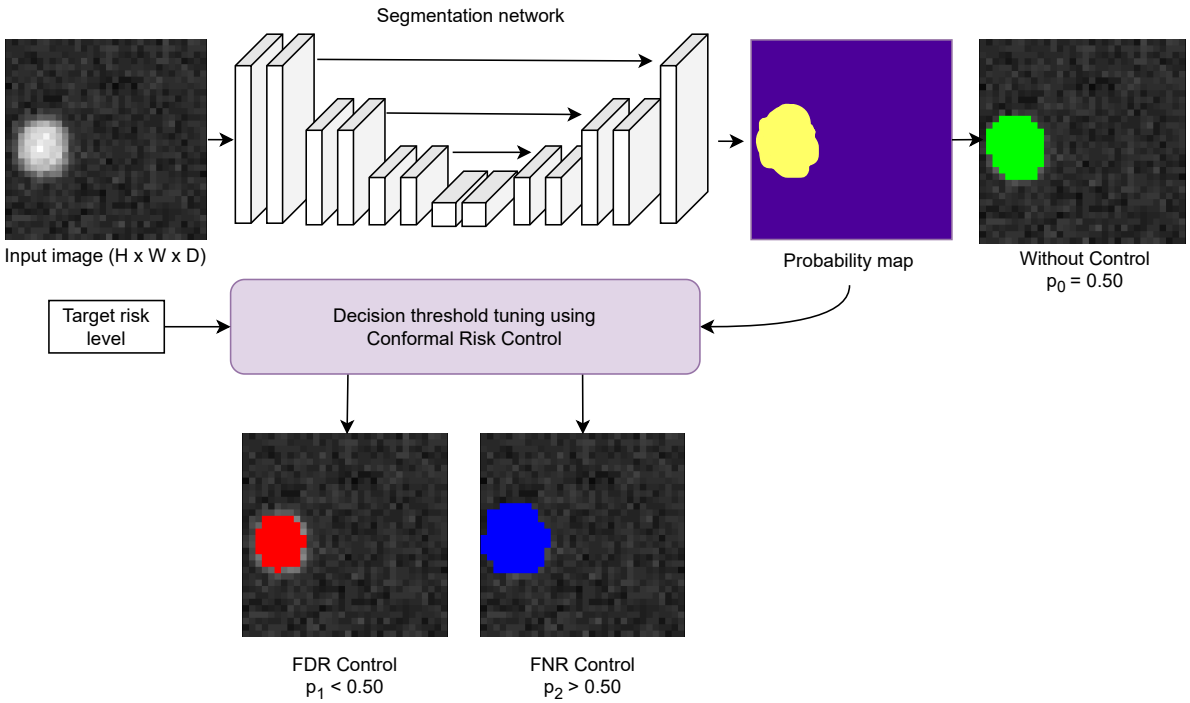


Figure A.6.1: Illustration of threshold tuning using Conformal Risk Control (CRC). In a standard segmentation network, the probability map is binarized using a default threshold of $p_0 = 0.50$ for binary problems. Alternatively, the decision threshold can be tuned to control the False Negative Rate (FNR) or False Discovery Rate (FDR), based on a user-defined risk level. Controlling the FDR yields to a decision threshold $p_1 \leq p_0$, while controlling the FNR yields to a decision threshold $p_2 \geq p_0$.

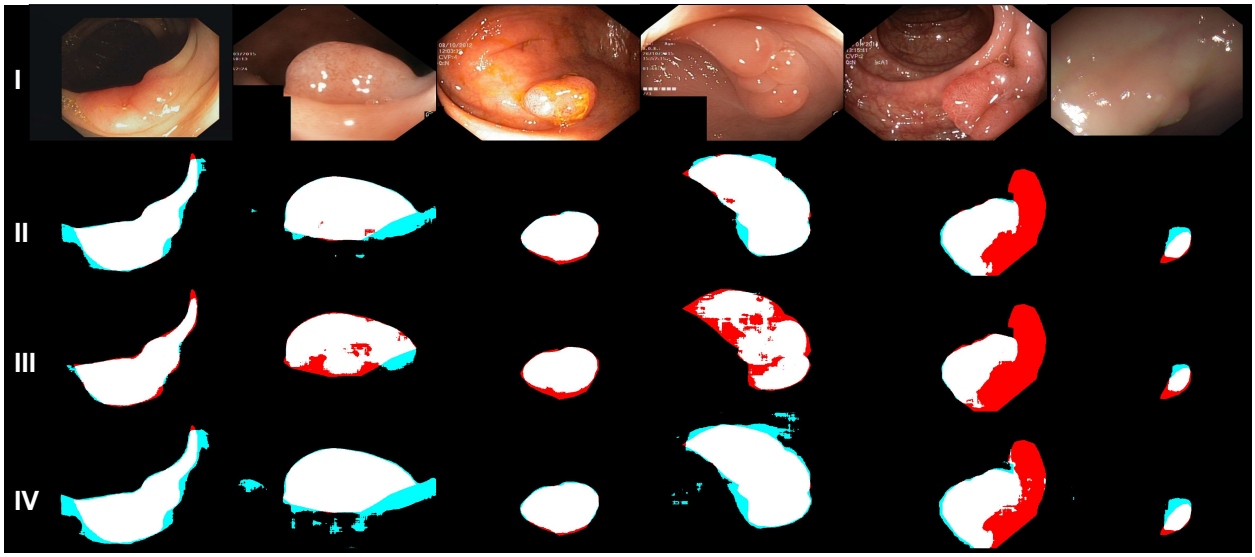


Figure A.6.2: Illustration of Conformal Risk Control in the context of polyp segmentation. I: input images, II: baseline segmentation with TP in white, FP in cyan, and FN in red, III: segmentation obtained by controlling the FDR rate at 10%, IV: segmentation obtained by controlling the FNR rate at 10%.

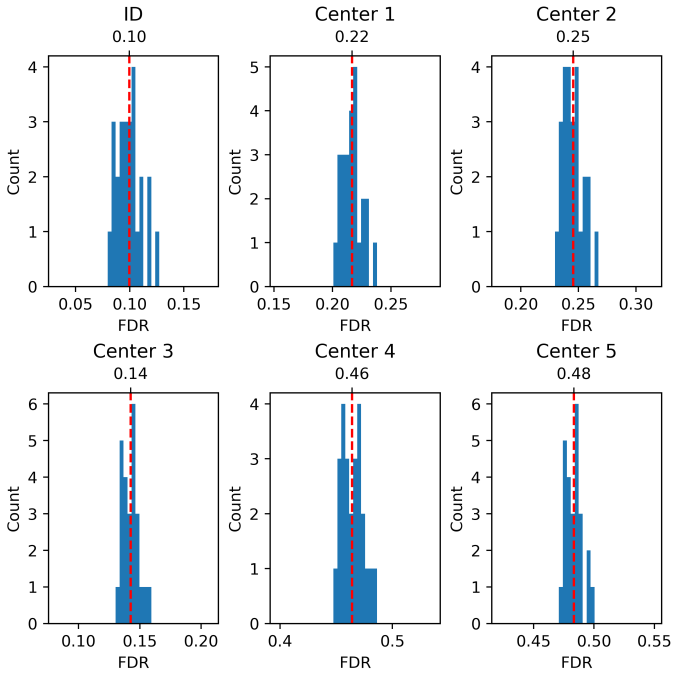
Dataset	Mean Dice (%)
In-distribution	89.22
PolypGen - Center 1	78.42
PolypGen - Center 2	73.51
PolypGen - Center 3	84.50
PolypGen - Center 4	55.70
PolypGen - Center 5	53.12

Table A.6.1: Average segmentation performance on the polyp test datasets.

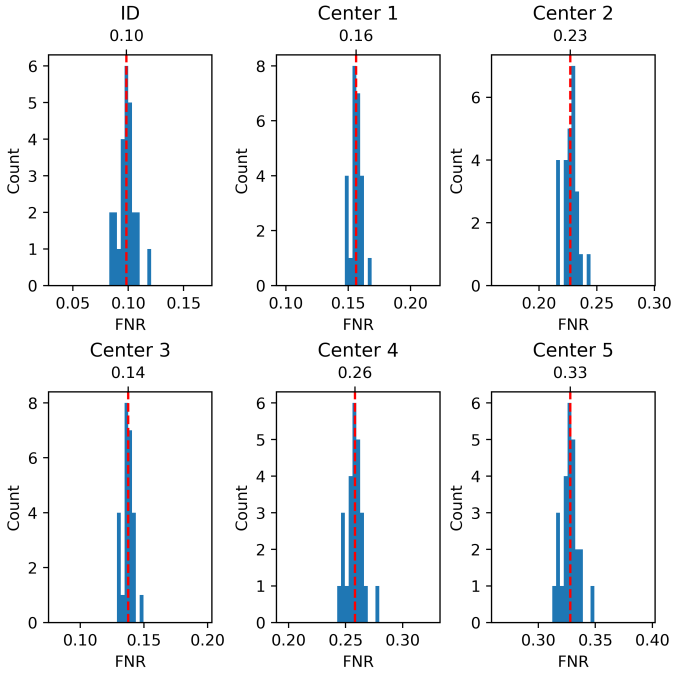
images. The accuracy of the conformal procedure is tested on an in-distribution test split of 438 images. Finally, we evaluate the robustness of the approach to domain-shift settings using the PolypGen datasets comprising images from 6 different imaging centers, exhibiting important variability. Figure A.6.2 presents a visualization of the different decision thresholds obtained with and without CRC. The experiment is reproduced for 25 trials by shuffling the in-distribution calibration and test splits.

Results

The average segmentation performance of the model is presented in Table A.6.1 for each test dataset. The model reaches top-quality segmentation on the in-distribution test split, however the performance is very variable on the PolypGen datasets. While images from Center 3 are well-segmented, Centers 4 and 5 are much more challenging. Regarding the CRC procedure, the empirical FDR and FNR are exactly 0.10 on the ID test set, showing that the



a



b

Figure A.6.3: False Discovery Rate (a, FDR) and False Negative Rate (b, FNR) control on the polyp test datasets. The graphs present the histograms (blue) of empirical risks over the 25 trials. The red dashed line indicates the average risk over the trials. The numerical value is indicated above the line. ID: in-distribution test dataset. Centers 1 to 6 correspond to the PolypGen dataset, used for domain-shift robustness evaluation.

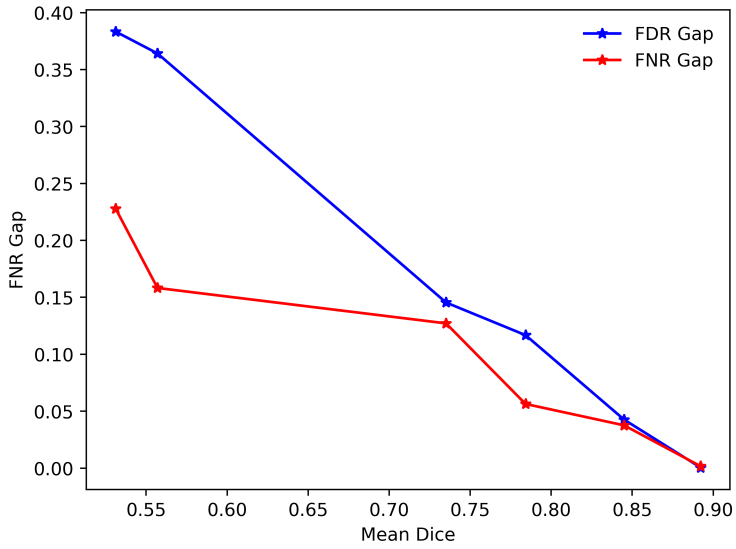


Figure A.6.4: Relationship between the risk gap (distance between the empirical risk and the target risk $\alpha = 0.10$) and the average segmentation performance (Dice). As segmentation quality drops, the gap to the target risk increases.

conformal decision threshold tuning is highly precise. However, the precision is degraded on the PolypGen datasets. The shift is proportional to the drop in segmentation accuracy: in Centers 4 and 5, the empirical FDR and FNR are very distant from the target level of $\alpha = 0.10$. The gap is less pronounced for Centers 1 and 3, for which the segmentation model is more robust. Overall, conclusions similar to Chapter V can be drawn here. When the test data is exchangeable with calibration data, the conformal procedure is highly efficient. However, when the exchangeability hypothesis is violated, the CP procedure precision highly degrades. For CRC, we can further show that this gap is proportional to the drop in segmentation quality (see Figure A.6.4).

Discussion

Conformal Risk Control is an interesting generalization of the conformal framework to control any monotone loss function. Here, we illustrate how it can be used to tune the decision threshold of segmentation models to control the rate of FP and FN predictions, respectively (note that the exact same procedure can be used for medical image classification models). As a contribution, we test the robustness of the procedure to domain-shift settings, and our experiments show that the risk gap increases as the segmentation quality drops. Note that a weighted formulation of conformal risk control can be framed [333, 334], similar to the WCP framework presented in this thesis (Section V.3).

A7 Participation in the SHIFT Challenge on WMH segmentation uncertainty

The Shift 2.0 Challenge[1] [214] has taken place between September 2022 and March 2023. The scope of this challenge is to evaluate the robustness and quality of uncertainty estimates on real-world problems. For this second edition, the challenge included a task of White Matter Hyperintensities (WMH) segmentation in brain T2 FLAIR MRI, thus fitting perfectly into the scope of this thesis. In this section, we present our proposed contribution which ranked in second position in the leaderboard.

Challenge Objectives

The challenge aims at the development of models that 1) are robust under various domain shifts, and 2) provide useful uncertainty estimates. For the WMH segmentation task, the training dataset proposed in this challenge is a subset of the data used in the thesis cross-sectional MS experiments. It relies on the ISBI 2015 dataset [230], MSSEG [229], MSLUB [231], and a private MS dataset acquired at the Swiss universities of Lausanne and Basel. The main difference with the MS dataset used in this thesis is that the WMH 2017 dataset is not included, which reduces the number of available scans to 33 for training, 7 for validation, and 33 for in-distribution test. MSLUB is used for out-of-distribution test (N=25) for phase I of the challenge, while the private Lausanne dataset (N=74) is used for out-of-distribution test for phase II. Note that the challenge does not evaluate models on in-distribution data.

The participants had to provide a Docker producing a segmentation of the WMH and a voxel-level uncertainty map. To evaluate the quality of uncertainty estimates, a metric similar to the R-AUC used in Chapter 2 of this thesis is used. The only difference is that challenge organizers used the normalized Dice score (nDSC) [217] instead of the Dice to construct the performance versus retention curve. This metric is named nDSC R-AUC.

Proposed Algorithm

The proposed algorithm is based on an ensemble of 5 individually trained Attention UNets [16]. The particularity is that we do not use directly the entropy derived from the predicted probabilities of each member, as done in the thesis. Instead, this challenge was the occasion to experiment with a learning paradigm for uncertainty. More particularly, each model has two outputs: one for the segmentation and one for a predicted uncertainty map. This is akin to the learned uncertainty framework introduced as a baseline in the thesis (II.2.7). More specifically, models are trained using the Focal Labelflip loss II.6.4. As a result, each member of the ensemble produces a segmentation paired with its associated uncertainty map. Then, the ensemble uncertainty map is obtained by averaging the individual maps. Data Augmentation is used extensively to account for the small size of the training dataset.

[1] https://shifts.grand-challenge.org/

Algorithm	Ranking	nDSC R-AUC ↓	nDSC ↑
Ours	1	0.0102 ± 0.0075	0.5944
martakaczmarska	2	0.0126 ± 0.0109	0.6057
umaimarahman.ai	3	0.0160 ± 0.0152	0.7084

Table A.7.1: Performance of the top 3 algorithms during Phase I. The test dataset is MSLUB.

Algorithm	Ranking	nDSC R-AUC ↓	nDSC ↑
agaldran	1	0.0128 ± 0.0169	0.5110
Ours	2	0.0134 ± 0.0147	0.5832
martakaczmarska	3	0.0136 ± 0.0138	0.6611

Table A.7.2: Performance of the top 3 algorithms during Phase II. The test dataset is the private Lausane dataset.

Implementation Details

The framework is implemented in PyTorch [8]. Models are trained with the usual paradigm used throughout this thesis: the ADAM optimizer is used with a fixed learning rate of 2×10^{-4} until the validation Dice seizes to improve to 60 epochs. The Attention U-Nets operate on patches extracted from the input FLAIR MRI, with a fixed size of $128 \times 128 \times 128$. Data Augmentation is implemented using the TorchIO library [246].

Ranking and Discussion

The different algorithms were benchmarked by the challenge organizers with respect to the nDSC R-AUC metric that quantifies the quality of uncertainty. For Phase I, the test data corresponds to the MSLUB dataset which is provided to the participants. Thus participants could optimize their algorithms so that the performance on MSLUB is optimized. Then, during Phase II, the test dataset is the private Lausanne dataset, not provided to participants. It was thus not possible to tune the algorithm directly for this test set. The final retained ranking is the one of Phase II. The performances of the top 3 teams in each phase are presented in Tables A.7.1 and A.7.2, respectively.

It can be noticed that segmentation performance (estimated using the normalized Dice) is quite poor for each algorithm. This is because only 33 images for training were provided, and moreover, the test data were not in-distribution samples. Thus the predictive task was particularly challenging. Finally, it was possible to optimize the quality of uncertainty (nDSC R-AUC) while having poor segmentation accuracy. As a result, our algorithm achieved the best uncertainty quality in Phase I while proving the poorest nDSC. Similarly, the top algorithm in Phase 2 with respect to uncertainty was also the worst in terms of nDSC. This can be seen as a limitation of the challenge, which rewards poor segmentation models that generate a lot of errors. When these incorrect voxels are associated with high uncertainties, the nDSC R-AUC metric can be minimized. This echoes the observations presented in the

voxel-level uncertainty benchmark (Chapter 2), where we argued that evaluation uncertainty separately from segmentation accuracy was cumbersome.

A8 Participation in the ATLAS Challenge on liver tumor segmentation

As a side project to this thesis, a contribution was developed for the ATLAS Challenge [335] organized in the context of MICCAI 2023. The challenges focus on the segmentation of liver tumors in MRI. This participation was the subject of a publication at the Resource-Efficient Medical Image Analysis (REMIA) workshop at MICCAI 2023 [336]. Here, we present a summary of the proposed algorithm. This work has been carried out in collaboration with Dr. Pauline Roca (Pixyl).

Challenge Objective

Liver cancer ranks as the sixth most prevalent form of cancer globally and is the fourth leading cause of cancer-related mortality. More particularly, hepatocellular carcinoma (HCC) is the primary type affecting adults. When the tumor cannot be surgically removed, the treatment involves transarterial radioembolization (TARE), inducing tumor necrosis through radiation-induced DNA damage and cell death [337]. To calculate dosimetry and plan the intervention, the volume and location of the tumor need to be precisely estimated [338]. This can be done from contrast-enhanced magnetic resonance imaging (CE-MRI) with four phases (precontrast, arterial, portal venous, delayed phases). However, manual delineation is time-consuming and error-prone and could thus benefit from automatic tools.

The goal of the ATLAS challenge is to develop accurate segmentation algorithms that should carry the simultaneous segmentation of the liver and the tumors. For each patient, one of the 4 MRI phases is provided. In this setting, the challenges are multiple. First, the training dataset is limited for model development (60 cases for training and validation) and there is a lot of variability between the MRI phase (precontrast, arterial, portal venous, delayed phases). This variability concerns the image contrast, tumors appearance, and also the image resolution (see Figure A.8.1). More precisely, training images exhibit a resolution ranging from 0.6841.4mm in the XY plane, and from 24.6mm in the Z-axis. Then, there is also a large variability in terms of MRI acquisition device [335]. The last challenge is that the hidden test dataset (30 subjects) was acquired more recently, and as an effect the overall quality of MRIs is superior to the one of the training images.

Proposed Pipelines

For this challenge, we propose 2 different pipelines: a multi-class model that performs the simultaneous segmentation of the liver and the tumors. It operates in a patch-based approach, with a patch size set to $256 \times 256 \times 64$. The second pipeline is composed of two models, one segmenting the overall liver, and one the tumors. The motivation is that segmenting tumors is much more challenging than segmenting the liver, thus potential benefit could be gained by disentangling both tasks. For the binary liver model, the patch size of $256 \times 256 \times 64$ is kept.

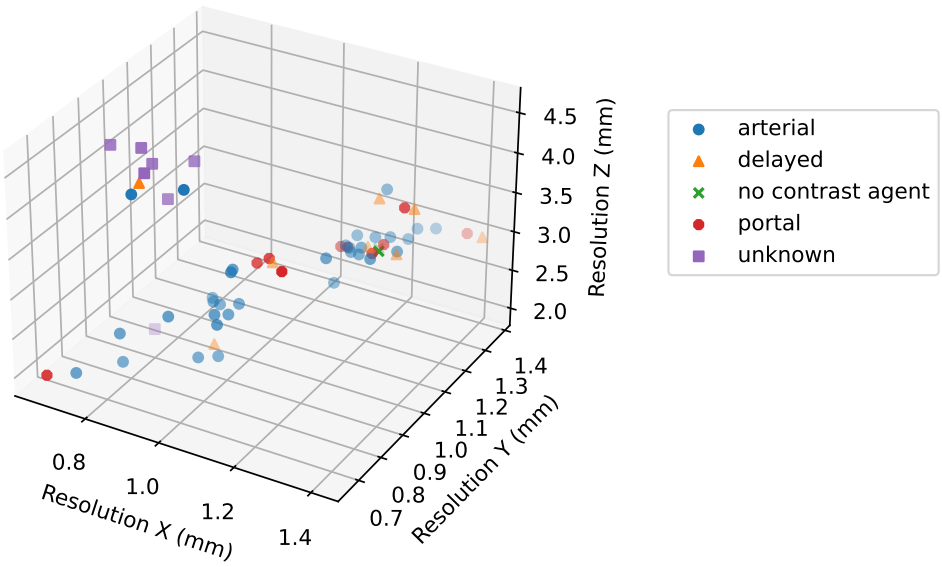


Figure A.8.1: Variability in voxel resolution based on the MRI phase for the 60 training subjects.

For the binary tumor model, a smaller patch size of $128 \times 128 \times 64$ is employed as it allows to reduce the imbalance between background and tumor voxels. At inference, the input MRI is processed by each binary model. The two resulting binary masks are then aggregated to reconstitute the final 3-class segmentation.

For each pipeline, we use an Anisotropic Hybrid U-Net (AHUNet) [339] as the segmentation backbone to tackle the anisotropy of the data, with the Z resolution being up to 4 times that in the plane. This alleviates the need for resampling to a uniform resolution, which inevitably increases interpolation blur in the images. This model is composed of a pre-trained 2D convolutional encoder that ignores between-slice information. It is followed by a 3D convolutional decoder that incorporates the 3D context.

Post-processing module

Post-processing is a crucial step for the automated segmentation of the liver in CT and MRI [340]. The raw predictions of our algorithms are post-processed using a 3-steps procedure:

- First, only the largest connected component for the liver is kept. Then, eventual holes in the liver mask are filled [341].
- Tumor lesions outside the liver mask are discarded.
- Binary closing is applied to remaining lesions to improve the smoothness of their borders.

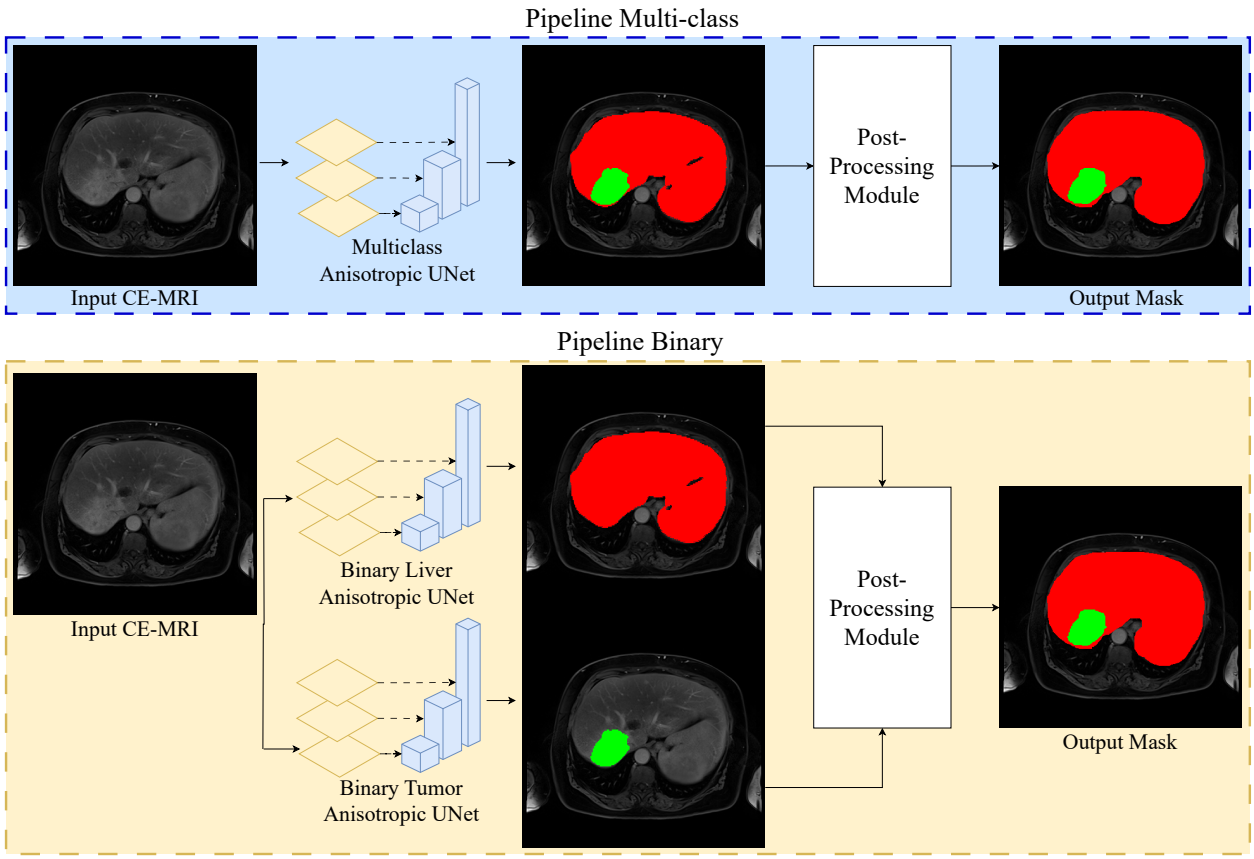


Figure A.8.2: Illustration of our two proposed pipelines, namely multi-class and binary.

Lesion uncertainty quantification

As the predictive task involves the detection of lesions, the lesion-uncertainty paradigm developed for this thesis can be explored. However, the number of lesions is rather low in this dataset, and the methodology using auxiliary classifiers to quantify lesion uncertainty is ill-adapted in these conditions, as shown in Chapter 3 of this thesis. Thus we opt for a simple baseline to compute lesion uncertainty. More particularly, we collect the tumor probabilities $p_{i,tumor}$ ($i \in [1, N]$) for each of the N voxels of the lesion, as produced by the Ensemble of Anisotropic UNets. The *lesion-wise* uncertainty score is then taken as:

$$L_{unc} = 1 - \frac{1}{N} \sum_{i=1}^{N} p_{i,tumor}$$

Ranking and discussion

The challenge algorithms were evaluated based on a panel of segmentation metrics: the Dice, the 5mm surface Dice, the symmetric surface distance, and the Hausdorff distance, which were calculated for both the liver and tumor classes. The Root Mean Square Error on tumor burden was also calculated to assess the precision of the tumor volume estimation. The performances of our pipelines in the private test dataset are presented in Table A.8.1, along with the performance of the challenge winner. Figure A.8.3 displays the histogram of uncertainty scores for true positives and false positives lesions.

Regarding segmentation metrics, our proposed pipelines and the winner algorithm produce similar results on the liver class, for each metric. However, ours produce significantly lower results on the tumor class. There are several design choices in the winning algorithm that explain this success:

- Data are resampled to a uniform voxel spacing. It appears that other teams followed this approach. Thus our design choice to use an anisotropic model on the raw data may not be optimal.
- They used a model called Scalable and Transferable U-Net (STU-Net) [342] which is a large pre-trained segmentation backbone. More precisely the model is pretrained on the TotalSeg dataset [317] that contains 1204 CT images with the manual segmentation of 104 anatomical structures, including 27 organs.
- They extend the training dataset with the LiTs 2017 dataset [343] which contains CT scans of patients diagnosed with liver tumors.

In summary, a net gain in tumor segmentation accuracy could be gained by extending the ATLAS training dataset with images from another modality (CT). Pretraining was according to the authors an important feature of their winning algorithm. In future work, we will include these concepts in our liver tumor segmentation pipelines to see if they translate well with our Anisotropic UNet models.

Regarding the lesion-wise uncertainty scores (Figure A.8.3, left), it can be observed that true

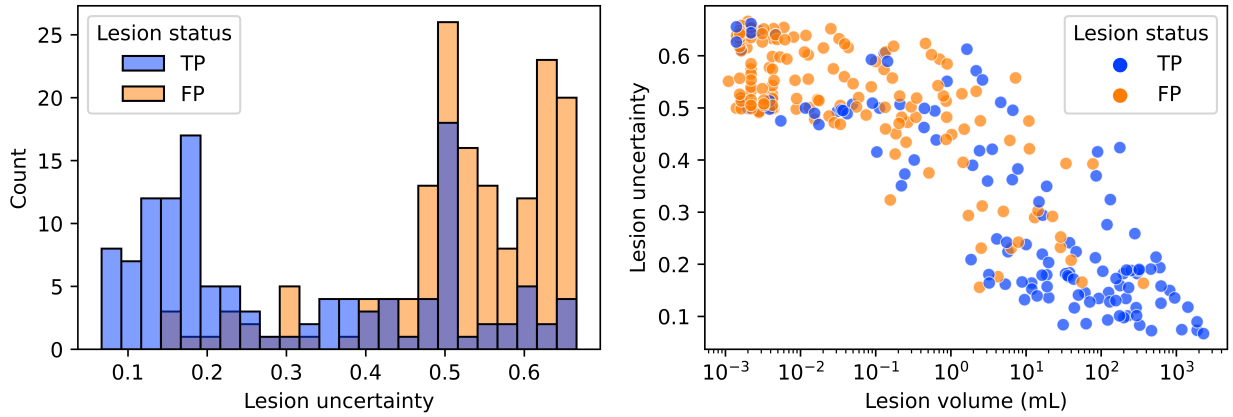


Figure A.8.3: Qualitative evaluation of the tumor lesion uncertainty obtained with the multi-class pipeline. Left: histogram of uncertainty estimates with respect to the lesion status (True Positive, TP or False Positive, FP). Right: lesion uncertainty with respect to the lesion volume (in log-scale).

positive lesions are associated with lower uncertainty scores than false positive lesions, which follows the same observations presented in this thesis for MS lesions and lung nodules. A strong correlation with the lesion size can be noted (Figure A.8.3, right).

		Multi-class	Dual binary	Winner (Jin Ye et al.)
Liver	ASD ↓	1.7	1.5	1.5
	Dice ↑	0.95	0.95	0.96
	HD ($\times 10^1$)	2.6	2.5	2.4
	SD ↑	0.95	0.95	0.96
Tumor	ASD ($\times 10^1$) ↓	3.0	4.0	0.7
	Dice ↑	0.60	0.59	0.75
	HD ($\times 10^1$) ↓	7.8	9.3	0.4
	SD ↑	57.4	55.3	75.4
	RMSE ($\times 10^{-2}$) ↓	0.4	0.6	0.2

Table A.8.1: Performance on the hidden test dataset for both pipelines, as reported in the public leaderboard. ASD = Asymmetric Surface Distance, HD: Haussdorf Distance, SD: Surface Dice.

A9 Datasets summary

In this section, the different datasets used throughout this thesis are summarised (Tables A.9.1 and A.9.2).

Pathology	Usage in Thesis	Dataset	Modalities	Size	Task	# Centers	# Annotators
Multiple Sclerosis	Calibration benchmark II.6.3 Voxel uncertainty II.5 Lesion uncertainty III Quality Control IV.4.2 Predictive Intervals V.2.8 Data synthesis III.8.1	MSSEG 2016 [229]	3D T1-w 3D T2-w 3D FLAIR 2D PD-T2-w	53	Cross-sectional WMH segmentation	4	7
		ISBI 2015 [230]	3D T1-w 3D T2-w 3D FLAIR 2D PD-T2-w	21	Cross-sectional WMH segmentation	1	2
		Ljubljana [231]	2D T1-w 2D T2-w 3D FLAIR	30	Cross-sectional WMH segmentation	1	3
		WMH 2017 [293]	3D T1-w 3D FLAIR	170	Cross-sectional WMH segmentation	3	2
	Lesion uncertainty III.8	MSSEG-2 [63]	3D FLAIR (2 visits)	100	Longitudinal WMH segmentation	15	4
Lung cancer	Lesion uncertainty III.7	LIDC-IDRI [268]	Helical thoracic CT	1018	Nodules segmentation	Multi	4
Strokes	Calibration benchmark II.6.1.3 Voxel uncertainty II.7 Out-of-distribution IV	ATLAS-2 [242]	T1-w	655	Stroke lesion segmentation	11	2
Polyps	Quality Control IV.4.4 Conformal Risk Control A6	Kvasir [308]	2D RGB images	1000	Polyp segmentation	4	≥ 1
		ETIS-LaribPolyp [309]	2D RGB images	196	Polyp segmentation	1	N/A
		CVC-ColonDB [310]	2D RGB images	380	Polyp segmentation	1	N/A
		CVC-ClinicDB [311]	2D RGB images	196	Polyp segmentation	1	N/A
		PolypGen [312]	2D RGB images	1412	Polyp segmentation	6	6
Healthy	Out-of-distribution IV	IXI [295]	3D brain T1-w	600	N/A	3	N/A
		CHAOS [296]	Abdominal T1-w MRI	80	Organ segmentation	1	3

Table A.9.1: Summary of the datasets used in this thesis (Part I). WMH: White-Matter Hyperintensities, PD: Proton Density.

Pathology	Usage in Thesis	Dataset	Modalities	Size	Task	# Centers	# Annotators
Brain Tumors	Calibration benchmark II.6.3 Voxel uncertainty II.6 Out-of-distribution IV Quality Control IV.4.3 Predictive Intervals V.3	BraTS Adult Gliomas [226]	T1-w T2-w FLAIR T1-ce	1133	Multi-class Glioblastoma segmentation	Multi	Multi
		BraTS SSA [238]	T1-w T2-w FLAIR T1-ce	60	Multi-class Glioblastoma segmentation	Multi	Multi
	Quality Control IV.4.3 Predictive Intervals V.3 Out-of-distribution IV	BraTS Meningioma [290]	T1-w T2-w FLAIR T1-ce	944	Multi-class Glioblastoma segmentation	Multi	Multi
	Quality Control IV.4.3 Out-of-distribution IV	BraTS Metastases [289]	T1-w T2-w FLAIR T1-ce	238	Multi-class Glioblastoma segmentation	Multi	Multi
	Quality Control IV.4.3 Out-of-distribution IV	BraTS Pediatric [288]	T1-w T2-w FLAIR T1-ce	99	Multi-class Glioblastoma segmentation	Multi	Multi
	Out-of-distribution IV	LUMIERE [282]	T1-w T2-w FLAIR T1-ce	74	Multi-class Glioblastoma segmentation	Multi	Multi
Epilepsy	Out-of-distribution IV	EPISURG [294]	162	T1-w	Resection cavity segmentation	1	3
Critical Findings	Out-of-distribution IV	CQ-500 [292]	491	Head CT	Detection of bleeds, fractures, and mass effects	1	3
Lumbar stenosis	Out-of-distribution IV	Lumbar MRI dataset [297]	Lumbar T1-w MRI	568	Lumbar spine segmentation	> 1	N/

Table A.9.2: Summary of the datasets used in this thesis (Part II).

List of Papers included in the Litterature Review

Uncertainty Frameworks	Count	Studies
Monte Carlo Dropout	118	[172] [179] [344] [90] [68] [146] [345] [216] [69] [346] [175] [347] [173] [251] [348] [110] [176] [109] [71] [206] [188] [349] [73] [350] [186] [351] [352] [112] [353] [207] [354] [355] [356] [347] [357] [358] [184] [154] [189] [204] [165] [208] [359] [211] [360] [104] [70] [212] [361] [362] [111] [218] [363] [364] [197] [365] [366] [367] [368] [205] [157] [210] [130] [148] [99] [100] [185] [369] [215] [92] [370] [193] [152] [371] [372] [373] [123] [98] [181] [374] [194] [199] [375] [376] [169] [195] [94] [72] [147] [101] [105] [102] [114] [377] [378] [379] [380] [170] [301] [81] [381] [382] [132] [383] [135] [209] [304] [95] [384] [385] [322] [219] [183] [182] [386] [387] [91] [388]
Deep Ensemble	55	[90] [216] [389] [69] [177] [110] [390] [109] [71] [352] [112] [391] [192] [358] [392] [165] [113] [393] [168] [111] [394] [162] [163] [395] [157] [396] [370] [371] [127] [123] [375] [397] [117] [118] [121] [114] [170] [81] [381] [398] [399] [209] [304] [95] [384] [219] [400] [401] [306] [182] [402] [386] [387] [91] [388]
Softmax	44	[90] [68] [216] [69] [67] [71] [206] [191] [73] [352] [391] [403] [208] [359] [404] [70] [74] [363] [405] [370] [406] [407] [408] [72] [409] [170] [410] [301] [81] [381] [411] [412] [413] [414] [415] [416] [135] [417] [304] [384] [322] [336] [387] [50]
Learned Uncertainty	28	[43] [68] [187] [164] [69] [418] [206] [419] [355] [129] [420] [70] [126] [421] [128] [130] [422] [125] [372] [93] [133] [423] [131] [132] [424] [135] [198] [425]
Test Time Augmentation	24	[68] [146] [178] [391] [202] [358] [165] [366] [205] [145] [148] [215] [375] [72] [147] [174] [170] [426] [382] [132] [385] [322] [219] [91]
Generative Models	15	[136] [137] [68] [357] [138] [139] [143] [157] [206] [142] [140] [94] [141] [427] [144]
Features	14	[68] [71] [150] [154] [208] [152] [72] [305] [301] [381] [151] [153] [304] [155]
Evidential Deep Learning	14	[164] [160] [162] [163] [157] [152] [161] [380] [190] [200] [428] [159] [429] [388]
Dropout Ensemble	11	[90] [354] [111] [370] [371] [123] [124] [209] [182] [430] [91]
Bayesian Neural Networks	7	[90] [92] [431] [94] [432] [95] [91]
Conformal Prediction	5	[81] [80] [433] [434] [322]
Other	4	[69] [166] [165] [167]
Total	**338**	

Table A.9.3: Resume of the papers included in the literature review of this thesis, classified according to the uncertainty framework. The same study can be present in different rows, if several uncertainty approaches have been compared.

Evaluation Frameworks	Count	Studies
Error Detection - Referral	70	[172] [90] [68] [146] [345] [69] [346] [177] [347] [251] [110] [206] [73] [353] [207] [354] [391] [355] [347] [129] [358] [166] [204] [208] [359] [211] [104] [212] [393] [168] [111] [364] [162] [163] [367] [368] [205] [395] [157] [145] [210] [215] [370] [152] [371] [372] [373] [123] [133] [124] [423] [131] [397] [147] [101] [105] [432] [118] [377] [174] [170] [132] [424] [209] [417] [95] [384] [91] [388]
Qualitative Assessment	61	[344] [187] [389] [418] [348] [390] [176] [188] [349] [191] [186] [351] [419] [160] [202] [356] [192] [184] [189] [404] [360] [420] [361] [362] [421] [197] [394] [405] [130] [99] [422] [396] [185] [92] [193] [125] [98] [374] [194] [199] [375] [431] [169] [407] [195] [102] [378] [379] [161] [190] [411] [383] [200] [412] [413] [428] [398] [198] [336] [430] [429]
Quality Control	32	[43] [179] [146] [216] [69] [177] [175] [178] [173] [109] [206] [104] [113] [70] [366] [100] [181] [376] [305] [426] [382] [132] [385] [400] [183] [401] [306] [427] [182] [402] [386] [425]
Calibration	44	[172] [164] [69] [67] [110] [109] [71] [206] [73] [350] [352] [354] [391] [154] [403] [165] [211] [113] [74] [111] [363] [365] [157] [210] [148] [408] [94] [117] [409] [121] [114] [380] [410] [132] [414] [415] [416] [399] [417] [384] [386] [50] [91] [388]
OOD Detection	20	[68] [110] [71] [112] [150] [154] [215] [152] [72] [114] [305] [301] [381] [159] [135] [151] [153] [304] [155] [387]
Label Distribution	19	[136] [137] [357] [392] [126] [218] [138] [139] [143] [128] [93] [127] [142] [406] [140] [167] [141] [219] [144]
Coverage Error	5	[81] [80] [433] [322] [83]

Table A.9.4: Resume of the papers included in the literature review of this thesis, classified according to the uncertainty evaluation paradigm. The same study can be present in different rows, if several evaluation strategies have been employed.

Published work

Journal Papers

- Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene and Michel Dojat, Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis, Artificial Intelligence in Medicine (2024), 102830, ISSN: 0933-3657

International Conference Papers

- Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Beyond Voxel Prediction Uncertainty: Identifying brain lesions you can trust, Interpretability of Machine Intelligence in Medical Image Computing: 5th International Workshop, iMIMIC 2022, Held in Conjunction with MICCAI 2022, Lecture Notes in Computer Science **13611** (2022), 61
- Benjamin Lambert, Florence Forbes, Senan Doyle and Michel Dojat, Multi-layer Aggregation as a key to feature-based OOD detection, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging - 5th International Workshop, UNSURE 2023, Held in Conjunction with MICCAI 2023, Lecture Notes in Computer Science **14291** (2023), 104
- Benjamin Lambert, Florence Forbes, Senan Doyle and Michel Dojat, TriadNet: Sampling-Free Predictive Intervals for Lesional Volume in 3D Brain MR Images, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging - 5th International Workshop, UNSURE 2023, Held in Conjunction with MICCAI 2023, Lecture Notes in Computer Science **14291** (2023), 32
- Benjamin Lambert, Maxime Louis, Senan Doyle, Florence Forbes, Michel Dojat and Alan Tucholka, Leveraging 3D information in unsupervised brain MRI segmentation, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021), 187

International Conference Abstracts

- Benjamin Lambert, Florence Forbes, Alan Tucholka, Senan Doyle and Michel Dojat, Multi-Scale Evaluation of Uncertainty Quantification Techniques for Deep Learning based MRI Segmentation, ISMRM-ESMRMB & ISMRT 2022-31st Joint Annual Meeting International Society for Magnetic Resonance in Medecine (2022), 1 - Poster presentation.
- Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Uncertainty-based Quality Control for Subcortical Structures Segmentation in T1-weighted Brain MRI, ISMRM-ESMRMB & ISMRT 2023-32nd Joint Annual Meeting International Society for Magnetic Resonance in Medecine (2023) - Poster presentation.

French Conference Papers

- Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Fast Uncertainty Quantification for Deep Learning-based MR Brain Segmentation, EGC 2022-Conference francophone pour l'Extraction et la Gestion des Connaissances (2022), 1
- Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Intervalles de confiance pour l'estimation de superficies à partir d'images satellitaires, GRETSI 2023-XXIXème Colloque Francophone de Traitement du Signal et des Images (2023), 1

French Conference Abstracts

- Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka and Michel Dojat, Safety-Net: Identification automatique des erreurs de segmentation des lésions de la Sclérose-en-Plaques, Société Française de Résonance Magnétique en Biologie et Médecine (2023) - Power pitch

Résumé français - French Summary

Introduction

L'établissement d'un diagnostic médical est par nature sujet à l'incertitude. Le manque de données, leur incomplétude ou les informations conflictuelles peuvent mener le docteur à douter et à être incertain concernant la cause des symptômes observés. Confrontés à une pathologie rare, différents experts peuvent être en désaccord concernant le traitement adéquat. En imagerie médicale également, une image faiblement résolue ou présentant un artefact peut rendre son analyse ambiguë. Il est donc communément admis que l'incertitude fait partie du quotidien du corps médical, et leur formation implique d'apprendre à prendre des décisions éclairées en prenant en compte cette incertitude.

Il serait donc attendu d'un algorithme analysant automatiquement les images médicales de pouvoir raisonner avec l'incertitude d'une manière similaire, afin d'éviter d'induire en erreur les utilisateurs du logiciel. Cependant, les modèles prédictifs basés sur les réseaux de neurones profonds sont typiquement incapables d'exprimer le doute. En général, toute prédiction est effectuée avec un niveau de confiance absolu, ce qui met en question leur fiabilité. Dans le cadre d'applications critiques comme l'analyse d'images médicales, il est donc crucial d'améliorer ces algorithmes complexes afin qu'ils puissent avertir l'utilisateur quand le résultat automatique est incertain. Cela est nécessaire afin d'éviter d'induire en erreur le praticien, ce qui pourrait avoir des conséquences négatives pour le patient, comme une prise en charge retardée de la pathologie.

La quantification de l'incertitude d'un réseau de neurones profonds est une tâche complexe. En effet, ces modèles sont généralement composés de millions de paramètres dont l'interprétation par l'humain est laborieuse. Par ailleurs, leur mode d'entraînement implique l'apprentissage automatique de caractéristiques à partir des données brutes, sans supervision humaine concernant le choix de cesdites caractéristiques. Par conséquent, le processus de décision d'un réseau de neurones est opaque pour le développeur et l'utilisateur du logiciel. Cet effet boîte noire peut rendre l'utilisateur réticent à l'utilisation de l'outil.

L'objectif de cette thèse est le développement de méthodologies permettant de quantifier l'incertitude liée aux analyses automatiques d'images médicales par réseaux de neurones profonds. Concernant la segmentation d'images médicales 3D, l'incertitude est utile à divers niveaux pour quantifier de manière complète la confiance du résultat automatique:

- À l'**échelle du voxel**, des cartes d'incertitudes peuvent être construites pour associer chaque voxel de l'image à un score de confiance. Ce score correspond au degré de

confiance que l'on peut attribuer au modèle concernant le label associé au voxel dans la segmentation (ex: lésion ou sain). Ces cartes peuvent être superposées à l'image d'entrée pour visualiser les zones qui ont suscité l'hésitation du modèle.

- Pour des pathologies cérébrales comme la Sclérose-en-Plaques, la segmentation automatique du cerveau permet généralement de mettre en évidence plusieurs dizaines de lésions individuelles. Dans ce cadre, des scores d'incertitude à l'**échelle de la lésion** sont désirés. Ces scores permettraient à l'utilisateur de contrôler directement les lésions les plus incertaines afin de rejeter les potentiels faux positifs. Ce niveau de quantification de l'incertain est aligné avec l'attention du clinicien, qui est à l'échelle de la lésion dans le cadre de la SEP.
- En guise de troisième niveau, l'incertitude peut également être quantifiée à l'**échelle du cas**. Cela peut être considéré comme une forme de contrôle qualité. Plus précisément, ce contrôle qualité peut être effectuée sur l'image d'entrée, avec pour objectif d'identifier les images qui ne conforment pas à ce pour quoi le modèle a été entraîné. Par exemple, un modèle entraîné à partir de séquences d'IRM pondérées T1 sera a priori incompétent sur des séquences d'IRM pondérées T2. Dans cette situation, il est attendu que l'incertitude du modèle soit élevée, permettant de détecter ce cas anormal et d'alerter l'utilisateur. De plus, le contrôle qualité peut également être appliqué sur la segmentation produite par le modèle. L'idée est de détecter automatiquement une segmentation qui n'atteint pas le degré attendu de qualité afin d'alerter l'utilisateur.
- Enfin, la segmentation est généralement utilisée afin d'extraire des métriques de haut-niveau comme le volume des lésions identifiées. Ces informations sont présentées à l'utilisateur sous la forme d'une rapport textuel. À l'heure actuelle, ces estimations ne sont pas complémentées par des intervalles prédictifs. La dernière échelle envisageable est donc l'**échelle des volumes estimés**. L'objectif est d'associé à ces estimations des intervalles prédictifs, construits de manière à contenir le vrai volume avec un degré de confiance donné (par exemple 90%).

Organisation de la thèse

Cette thèse contient 5 chapitres. Le premier chapitre permet d'introduire plusieurs notions clés du Deep Learning en se concentrant sur les techniques utilisées dans l'analyse d'images médicales. Le second chapitre permet d'introduire l'étude bibliographique qui a été effectuée dans le cadre de cette thèse, permettant une revue systématique des méthodes proposées pour quantifier l'incertitude des modèles d'analyse d'images médicales. Ce chapitre est également l'occasion d'effectuer deux études comparatives différentes. La première est dédiée à l'étude de la calibration des probabilités des modèles de segmentation. Cette propriété désirée permet de s'assurer que les probabilités en sortie du modèle de segmentation soient bien représentatives de la confiance du modèle. La seconde étude s'intéresse à la comparaison de diverses techniques pour quantifier l'incertitude à l'échelle du voxel. Cette expérience permet de mettre en avant l'intérêt de l'assemblage de modèle, qui consiste à entraîner plusieurs instances d'un même réseau de neurones puis d'agréger leurs prédictions. Les chapitres 3, 4 et 5 présentent ensuite les différentes méthodologies développées pour quantifier l'incertitude des lésions, des sujets et des volumes estimés, respectivement. Ces contributions sont détaillées plus en avant ci-dessous.

Chapitre 3 : Quantification de l'incertitude des lésions à partir de classificateurs auxiliaires

La première étape à la quantification de l'incertitude à l'échelle des lésions est d'effectuer une analyse des composantes connexes de la segmentation, permettant d'identifier chaque instance unique de lésion. Ensuite, une approche simple et directe consiste à moyenner l'incertitude des voxels qui composent la lésion. Néanmoins, cette approche standard suppose que chaque voxel contribue uniformément au niveau global d'incertitude de la lésion. Dans ce chapitre, nous cherchons à déterminer si une approche plus sophistiquée permet d'améliorer l'estimation de la confiance de chaque lésion.

L'angle proposé pour quantifier l'incertitude à l'échelle des lésions est l'utilisation de classificateurs auxiliaires. Plus précisément, un modèle est entraîné à prédire la probabilité que la lésion détectée soit un faux positif. Ce score est utilisé comme marqueur d'incertitude. La motivation est qu'il est souhaité que les fausses détections soient plus incertaines que les vrais positifs. Cela requiert de développer des classificateurs opérant à l'échelle des lésions. Trois variantes sont proposées ici. La première consiste à l'extraction de caractéristiques à partir de l'image d'entrée, de la segmentation et de la carte d'incertitude. Cela permet de construire des vecteurs de variables représentatives de la lésion, permettant ensuite d'entraîner un modèle de régression logistique pour prédire le statut de la lésion (vrai positif ou faux positif). La deuxième technique testée est l'utilisation d'un réseau de neurones convolutif qui travaille à partir de boîtes englobantes centrées sur la lésion. Enfin, une technique faisant usage de graphe est présentée. L'idée est de représenter chaque lésion par un graphe, ce qui permet de prendre en compte l'hétérogénéité des formes de lésion. Le graphe est composé de nœuds qui correspondent aux voxels composant la lésion. Une série de caractéristiques est ensuite définie pour chaque nœud à partir de l'intensité de l'image d'entrée et du niveau d'incertitude des voxels et de la géométrie de la lésion. Enfin, un réseau de neurones opérant sur les graphes prédit la probabilité que le graphe corresponde à une lésion fausse positive.

Ces différentes techniques sont évaluées sur 3 tâches impliquant la détection de lésions multiples : la segmentation de lésions SEP en IRM transversale (une image à la fois) et longitudinale (comparaison de deux images du même patient séparées dans le temps), ainsi que la détection de nodules pulmonaires en scanner.

Les scores d'incertitude à l'échelle de la lésion sont évalués par leur capacité à séparer les vrais positifs des faux. Nous montrons que ces approches basées sur la classification des lésions sont efficaces quand une base de données suffisamment grande peut être construite, contenant suffisamment d'exemples de lésions correctes et incorrectes. Par ailleurs, ces scores sont corrélées aux scores d'incertitude définies par les experts (score de subtilité et variabilité inter-expert). Cependant, quand peu de lésions sont disponibles, comme dans le cas de l'analyse longitudinale de patients SEP, ces méthodes sont mises en défaut. Cela vient du fait qu'elles reposent sur un apprentissage supervisé qui est sous-optimal quand peu d'exemples sont disponibles.

Chapitre 4 : Détection des images hors-distribution et Contrôle qualité des segmentations

Dans ce chapitre, l'objectif est de développer un module capable de détecter automatiquement les images qui ne correspondent pas à ce qui a été vu pendant l'apprentissage du modèle. En imagerie médicale, de nombreuses anomalies peuvent se glisser dans l'image d'entrée et rendre son analyse sous-optimale. Plus particulièrement, nous avons identifié 5 catégories différentes. Tout d'abord, la présence d'artefact (mouvement, biais) dans l'image est commune en IRM, ce qui peut rendre la lecture de l'image ambiguë. Ensuite, les modèles d'IA sont entraînés à partir de séquences IRM bien spécifiées (par exemple T2 FLAIR pour les modèles SEP ou T1 pour les modèles segmentant le cerveau en régions anatomiques). Il est donc important de pouvoir détecter quand la séquence en entrée ne correspond pas à ce qui est attendu. En troisième place, la présence d'un type de lésion cérébrale non vue pendant l'apprentissage cause généralement des erreurs dans la segmentation. Ainsi un modèle de tumeur risque de segmenter des lésions AVC car ces lésions sont typiquement absentes de la base d'apprentissage. Ce type de cas est donc évalué dans notre banc d'essai. Une situation plus subtile ou la pathologie reste la même (glioblastome par exemple) mais que la population change (par exemple des sujets mineurs pour un modèle entraîné à partir de patients adultes) est étudié. Enfin, des cas extrêmes ou l'image ne présente pas l'organe attendu sont inclus dans le banc d'essai (par exemple, des IRMs abdominales pour un modèle opérant sur le cerveau).

Ces divers cas sont étudiés sur une tâche portant sur la segmentation de tumeur dans des IRMs T1 du cerveau. Le choix est fait de se concentrer sur des IRMs T1 car elles sont facilement accessibles et permettent d'étudier de nombreux cas d'images hors-distribution. Nous commençons par évaluer l'approche standard qui dérive un score pour l'image à partir de la carte d'incertitude à l'échelle du voxel. Cette méthode se montrant limitée, une technique plus avancée faisant usage des représentations latentes du modèle de segmentation est explorée. Cette dernière s'avère particulièrement robuste, tout en étant peu coûteuse en termes de calcul. Néanmoins nous montrons que la performance de l'approche dépend fortement du choix de la couche de convolution permettant d'extraire les représentations latentes. Ce choix est particulièrement sensible et dépend de l'architecture de segmentation sélectionnée. Nous montrons donc qu'une solution évitant la sélection de la couche de convolution est l'agrégation des scores des différentes couches. Cette solution permet d'obtenir des résultats performants peu importe l'architecture de segmentation utilisée.

Ensuite, notre étude s'intéresse à une seconde définition des images hors-distribution, qui consiste à définir une image comme hors-distribution si la segmentation correspondante n'est pas de bonne qualité. Cette définition est intéressante car elle permet de prendre en compte la capacité de généralisation du modèle qui peut segmenter correctement une image présentant un artefact. En général, la détection des segmentations de mauvaise qualité se fait avec des techniques différentes que celles utilisées pour détecter les images hors-distribution. Ici, nous proposons d'utiliser un cadre unifié faisant appel à deux estimées de qualité : une portant sur l'image d'entrée, une portant sur la segmentation de sortie. Cela permet de stratifier l'espace des prédictions en quatre régimes de fonctionnement: optimal, robuste, dysfonctionnel ou

divergent. Nous montrons à travers différents exemples que ces régimes permettent une estimation graduée de la qualité de la segmentation.

Chapitre 5 : Construction d'intervalles prédictifs pour l'estimation de volumes lésionnels

La segmentation automatique peut être utilisée pour générer des rapports d'analyse reportant les volumes des lésions identifiées ainsi que le volume de régions anatomiques. Cela peut être vu comme une tâche de régression des volumes à partir des segmentations. Néanmoins il n'existe pour le moment pas de méthode fiable pour équiper ces estimations avec des intervalles prédictifs, cruciaux pour éviter d'induire en erreur l'utilisateur du logiciel. Dans ce chapitre, nous proposons une nouvelle méthodologie pour estimer des intervalles de confiance à 90% pour l'estimation de volumes lésionnel. Ce modèle nommé TriadNet permet la construction d'intervalles sans besoin d'échantillonner les prédictions, permettant donc une nette réduction du temps d'inférence. L'intérêt de ce modèle est démontré à travers deux tâches: la segmentation des lésions SEP, et la segmentation multi-classes des glioblastomes.

Pour garantir que les intervalles construits contiennent bien la fraction désirée des vrais volumes (ex: 90%), nous utilisons le concept des prédictions conformes. Proposé dans les années 80, ce cadre mathématique regagne en popularité et est revisité principalement dans des problématiques de classification d'images 2D, ou en régression à uni-dimension. Nous proposons la première investigation de ce principe pour l'analyse d'images médicales 3D. Le concept est le suivant : les intervalles sont calculés sur une base de calibration, et leur couverture est mesurée. Sans calibration, il y a très peu de chance que la couverture atteigne la cible de 90%. Une calibration est possible avec les prédictions conformes, en calculant un facteur correctif $\hat{q}$ qui vise à corriger les intervalles pour qu'ils atteignent, une fois calibrés, la couverture désirée. Un challenge immédiat en imagerie 3D et la taille de la base de calibration, qui doit être suffisamment grande pour atteindre avec précision les 90% de couverture souhaités. Dans nos applications, il est impossible d'atteindre ce chiffre. Nous montrons cependant que des intervalles informatifs peuvent être obtenus avec aussi peu que 50 images de calibration.

Enfin, les prédictions conformes reposent sur l'hypothèse que les données de calibration et de test sont échangeables. Hors, en imagerie médicale, les données de test sont souvent d'une distribution différente (différent appareil d'acquisition, différente population imagée). Dans ce cadre, nous montrons que la couverture des intervalles se dégrade quand les données ne sont pas échangeables. Nous proposons enfin une investigation d'une version pondérée des prédictions conformes, qui repose sur l'estimation du ratio des densités entre distribution de calibration et de test. Comme l'estimation de ce ratio directement sur les images est insoluble en raison de leur dimension, nous proposons une alternative plus frugale basée sur les représentations latentes des images. Cela permet une estimation du ratio des densités rapide et à faible coût en calculs. Quand la distance entre la base de calibration et de test est modérée (faible déplacement des variables), la méthode est efficace pour diminuer l'erreur de couverture. Néanmoins quand la différence est plus marquée, le ratio ne peut plus être estimé correctement et les intervalles deviennent trop larges et non informatifs. Cela nous permet de

souligner deux limitations des prédictions conformes qui sont un frein à leur utilisation dans un contexte d'imagerie médicale. Tout d'abord, le besoin pour de grandes bases d'images pour la calibration n'est pas compatible avec la réalité des datasets d'images médicales 3D, qui sont souvent très limités en taille. Secondement, il est impossible de garantir que les données de calibration et de test seront strictement échangeables, ce qui compromet la pertinence des intervalles prédictifs.

Conclusion

Les modèles de Deep Learning ont révolutionné l'analyse d'images médicales, mais leur adoption reste encore à être obtenue. Pour cela, il est essentiel que les algorithmes soient capables d'exprimer le doute afin d'alerter leur utilisateur quand le cas à analyser est incertain. Dans cette thèse, nous avons proposé plusieurs développements méthodologiques visant à compléter les prédictions brutes des modèles avec des marqueurs d'incertitudes opérants à différents niveaux.

Au niveau du voxel, nous avons pu démontrer l'intérêt de l'assemblage de modèles, qui permet à la fois d'améliorer la qualité de la segmentation et de produire des estimations d'incertitude de qualité. Pour cela, nous nous sommes appuyés sur un banc d'essai comportant 3 tâches de segmentation d'anomalie en IRM cérébrale : la segmentation de lésions SEP, d'AVC, et de tumeurs. Cette technique d'assemblage est associée à un surcoût de calcul lors de l'entraînement des modèles, néanmoins la procédure est efficace au moment de l'inférence.

À partir de ces estimées à l'échelle du voxel, nous avons ensuite proposé d'estimer l'incertitude à l'échelle des instances de lésion. Cela est pertinent pour des applications telles que la Sclérose-en-Plaques, ou un seul cerveau peut contenir plusieurs dizaines de lésions individuelles. Ces scores à l'échelle des lésions permettent à l'utilisateur de directement contrôler les lésions les plus incertaines qui ont des chances d'être des faux positifs. Pour cela, nous proposons un paradigme construit autour de l'utilisation de classificateurs auxiliaires qui opèrent directement sur les lésions. Ces modèles prédisent pour chaque lésion la probabilité qu'elle soit un faux positif, que nous utilisons comme une estimation de l'incertitude. Trois variantes sont explorées : un modèle de Régression Logistique faisant usage de caractéristiques extraites de l'image d'entrée, de l'incertitude des voxels et du masque de la lésion. Un modèle de classification convolutif est également testé, travaillant à partir de boîtes englobantes centrées sur les lésions. Enfin, pour adresser l'hétérogénéité des lésions, nous mettons à l'essai un modèle de graph. Plus particulièrement, un graphe représentant chaque lésion est construit, permettant d'utiliser un modèle de classification des graphes. Ce dernier modèle est intéressant car il permet une modélisation flexible des lésions tout en étant frugal en termes de calcul. Ces techniques sont mises à l'essai sur 3 tâches : la segmentation transversale et longitudinale des lésions SEPs, ainsi que la détection des nodules pulmonaires en CT-scan. Une limite de ce paradigme est cependant le besoin de suffisamment d'exemples de lésions vraies positives et fausses positives pour entraîner les classificateurs.

Ces expériences nous ont également permis d'identifier une limite connue des réseaux de neurones profonds, qui est leur manque de robustesse quand l'image de test ne correspond

pas parfaitement à ce qui a été vu pendant l'apprentissage. Nous avons donc travaillé sur le développement d'un module permettant d'estimer la conformité de l'image d'entrée. Pour cela, une approche se basant sur l'espace latent des modèles de segmentation entrainés s'est montrée particulièrement adéquate sur un banc d'essai comportant 24 types d'images hors-distribution. Cependant, la technique dépend du choix de la couche sélectionnée pour générer les représentations latentes. Pour éviter d'avoir à sélectionner pour chaque réseau la couche optimale, nous proposons d'agréger les scores des différentes couches, ce qui fonctionne bien en pratique pour les différentes architectures testées. Pour finir, nous proposons de compléter ce score avec une estimation de la qualité de la segmentation. En combinant ces deux scores, nous montrons qu'il est possible d'estimer la qualité de l'analyse de manière automatique, permettant d'avertir l'utilisateur en cas de prédictions sous-optimales.

Pour terminer, nous avons étudié la construction d'intervalles prédictifs pour l'estimation du volume des lésions cérébrales. En mêlant une architecture multi-tête et les prédictions conformes, des intervalles de 90% sont obtenus. Cela permet de compléter les rapports automatiques d'analyses avec des barres d'erreur permettant de mieux quantifier l'erreur possible sur l'estimation des volumes.

En conclusion, les méthodologies développées au cours de cette thèse permettent une estimation complète de l'incertitude dans les pipelines d'analyse d'images médicales. Cela commence par une visualisation des zones incertaines à l'aide des cartes d'incertitude opérant au niveau du voxel. Ensuite, des scores à l'échelle des instances de lésion permettent d'aligner l'incertitude avec l'attention du clinicien pour des pathologies comme la Sclérose-en-Plaques. Enfin, une estimation globale sur la conformité de l'image d'entrée et sur la qualité de la segmentation associée permettent de détecter les analyses ratées. Pour finir, les rapports automatiques qui rapportent les volumes totaux de lésions sont enrichis avec des intervalles prédictifs pour une analyse plus sûre. Au cours de cette thèse, un soin particulier a été apporté au développement de solutions versatiles et robustes. C'est pourquoi les différentes solutions ont été évaluées sur différentes pathologies et avec des images hors distribution dès que possible. L'incorporation de ces divers outils dans la suite de logicielle développée par Pixyl permettra d'améliorer la prise de décision assistée par AI et augmenter la confiance accordée dans les rapports automatisés. En guise de perspective, il peut être noté que l'évaluation de l'incertitude demeure un problème délicat. En effet, l'incertitude est généralement évaluée par sa corrélation avec les erreurs du modèle. Il pourrait être intéressant de considérer l'utilité de l'incertitude en routine clinique, par exemple par son impact bénéfique sur la prise de décision. Enfin, les différentes expériences ont démontré que les réseaux de neurones souffraient d'une baisse de performance sur des images éloignées de la distribution d'apprentissage. Pour des applications industrielles comme celles développées par Pixyl, un soin particulier doit être donné à la capacité de généralisation des algorithmes en dehors de la distribution pour laquelle ils ont été explicitement entraînés, afin de garantir que des performances optimales soient atteintes dans chaque site où le logiciel est déployé. L'augmentation des données, une procédure devenue standard lors du développement des modèles d'analyses d'images médicales, est néanmoins insuffisante pour reproduire la variabilité observée dans le monde réel. Cela ouvre le champ à des recherches innovantes pour l'amélioration de la généralisation des modèles de Deep Learning.

Abstract

Quantifying the inherent uncertainty in an automated medical image analysis is crucial to guarantee the safe deployment of deep learning models. However, these models, often referred to as black boxes, are known to produce errors with high confidence, potentially leading to misinformed conclusions.

The operative goal of Pixyl, Grenoble Institute of Neurosciences, and Inria for this Ph.D. is to develop flexible uncertainty quantification tools that would address the opacity of deep learning models. In medical image analysis, uncertainty estimates are useful at different levels: the voxel, the lesion, the subject (input image or output segmentation), and the estimated volumes.

At the voxel level, there is a multitude of solutions proposed in the literature. Through three different brain lesion segmentation tasks, we show the adequacy of the Deep Ensemble framework to detect incorrectly classified voxels. The resulting uncertainty map can be overlayed on the input image to identify the ambiguous regions. Second, specifically tailored for diseases involving the detection of multiple lesions such as Multiple Sclerosis, we propose a lesion-level uncertainty module that associates each identified brain lesion with a confidence score. This allows to draw the clinician's attention to these uncertain lesion instances, which may correspond to false positive detections.
Then, as deep learning models lack robustness when the test image is not represented in the training dataset, we build two different subject-level quality control tools. Non-conform inputs are detected using a compressed latent representation, allowing to efficiently compute its distance to the training distribution in a low-dimensional manifold. Yet, this approach focuses on the conformity of the input image and thus is unhelpful in detecting poor-quality segmentations. To alleviate this, a second quality check focusing on segmentation quality is implemented, allowing to enrich the informativeness of the case-level uncertainty quantification.
Lastly, we leverage the conformal prediction framework to equip lesional volume estimations with robust predictive intervals. The proposed framework, TriadNet, computes the segmentation and associated predictive intervals in a second, thus being ideal for industrial software.

The tools developed during this PhD will allow to enhance medical image analysis software with useful uncertainty estimates, allowing increased trust in the automated results and enabling informed decision-making.

Keywords: Uncertainty, Segmentation, Quality Control, Brain, MRI, Robustness, Predictive Intervals

Résumé

La quantification de l'incertitude inhérente à une analyse d'image médicale automatisée est cruciale pour garantir le déploiement sécurisé des logiciels basés sur les réseaux de neurones profonds. Cependant, ces modèles, souvent considérés comme des boîtes noires, sont connus pour produire des erreurs avec une grande confiance, pouvant potentiellement conduire à des conclusions erronées.

L'objectif opérationnel de Pixyl, de l'Institut des Neurosciences de Grenoble et de l'Inria pour cette thèse est de développer des outils de quantification de l'incertitude flexibles qui permettraient de pallier l'opacité des réseaux de neurones profonds. Pour l'analyse automatique d'images médicales, les estimations d'incertitude sont utiles à différents niveaux pour les radiologues: le voxel, la lésion, le sujet (image d'entrée ou segmentation de sortie), et les volumes estimés.

Au niveau du voxel, une multitude de méthodes ont été proposées dans la littérature. À travers un banc d'essai comportant 3 tâches différentes de segmentation de lésions cérébrales, nous montrons l'adéquation de l'assemblage de modèles pour détecter les voxels incorrects. La carte d'incertitude résultante peut être superposée sur l'image d'entrée pour identifier les régions ambiguës. Ensuite, spécifiquement conçu pour les maladies impliquant la détection de lésions multiples telles que la Sclérose-en-Plaques, nous proposons un module d'incertitude au niveau de la lésion qui associe chaque lésion cérébrale identifiée à un score de confiance. Cela permet d'attirer l'attention du clinicien sur ces instances de lésions douteuses, qui peuvent correspondre à des détections de faux positifs.
Ensuite, comme les réseaux de neurones profonds manquent de robustesse lorsque l'image de test n'est pas représentée dans l'ensemble de données d'entraînement, nous construisons deux outils de contrôle qualité au niveau du sujet. Les entrées non conformes sont détectées en utilisant une représentation latente compressée de l'image, permettant de calculer efficacement sa distance à la distribution d'entraînement dans un espace de dimension réduite. Cependant, cette approche se concentre sur la conformité de l'image d'entrée et est donc peu utile pour détecter les segmentations de mauvaise qualité. Pour remédier à cela, un deuxième contrôle qualité axé sur la qualité de la segmentation est mis en œuvre, permettant d'enrichir l'utilité de la quantification de l'incertitude à l'échelle du sujet.
Enfin, nous explorons le cadre des prédictions conformes pour doter les estimations de volume lésionnel d'intervalles prédictifs robustes. Le modèle proposé, TriadNet, calcule la segmentation et les intervalles prédictifs correspondant en une seconde, ce qui en fait un algorithme idéal pour les logiciels industriels.

Les outils développés au cours de cette thèse permettront d'améliorer les logiciels d'analyse d'images médicales avec des estimations utiles de l'incertitude, permettant ainsi une plus grande confiance dans les résultats automatisés tout en promouvant la prise de décision éclairée.

Mots-clés : Incertitude, Segmentation, Contrôle Qualité, Cerveau, IRM, Robustesse, Intervalles Prédictifs