Segmentation of pathologies in Human Brain MRI's with uncertainty

Yacine BOUREGHDA
(supervisor: Nicolas BOUTRY)

Technical Report n^o202406-techrep-boureghda, June 2024
revision

Segmentation is a computer vision process used in medical imaging to support the diagnosis of various pathologies by healthcare teams. The purpose of my work is to develop a neural network that can perform segmentations on MRI images of human brains and provide prediction on the progression of the tumor through the potential contamination of different voxels. To improve the reliability of the model, we intend to develop 2 additional algorithms for quantifying uncertainty: Monte-Carlo Dropout and Deep Ensembles. Monte-Carlo Dropout is based on generating multiple predictions by randomly deactivating neurons, and Deep Ensembles train several networks with different initializations and architectures. These methods will enable the computation of uncertainty by measuring metrics such as standard deviation from the mean.

La segmentation est un processus de vision par ordinateur utilisé en imagerie médicale pour aider les équipes de soins de santé à diagnostiquer diverses pathologies. L'objectif de mon travail est de développer un réseau de neurones capable de réaliser des segmentations sur des images IRM de cerveaux humains et de fournir des prédictions sur la progression de la tumeur à travers la contamination potentielle de différents voxels. Pour améliorer la fiabilité du modèle, nous prévoyons de développer deux algorithmes supplémentaires pour quantifier l'incertitude : le Monte-Carlo Dropout et les Deep Ensembles. Le Monte-Carlo Dropout se base sur la génération de multiples prédictions en désactivant aléatoirement des neurones, tandis que les Deep Ensembles entraînent plusieurs réseaux avec des initialisations et des architectures différentes. Ces méthodes permettront de calculer l'incertitude en mesurant des métriques telles que l'écart type par rapport à la moyenne.

Keywords
Segmentation, Medical Imaging, Neural Networks, MRI, Tumor Progression, Uncertainty Quantification, Monte Carlo Dropout, Deep Ensembles, Standard Deviation, Prediction Reliability, Shannon Entropy

Laboratoire de Recherche de l'EPITA
14-16, rue Voltaire – FR-94276 Le Kremlin-Bicêtre CEDEX – France
Tél. +33 1 53 14 59 22 – Fax. +33 1 53 14 59 13
yacine.boureghda@lre.epita.fr – http://www.lre.epita.fr/

Copying this document

Contents

Chapter 1

Introduction

In the field of medical imaging, the use of deep learning algorithms for diagnostic assistance exhibits remarkable potential due to their ability to detect complex patterns and anomalies in data such as MRI scans. These models are increasingly favored for their efficiency and rapid learning capabilities, adapting to a wide array of application contexts. However, despite the often high accuracy of these algorithms, they present a major challenge related to uncertainty management: they tend to offer their predictions with excessively high confidence levels, without indicating the limits of their knowledge.

This issue of poorly calibrated confidence is clearly manifested when, for example, a classifier trained to recognize images within a defined set of categories encounters images outside of this set. The model might then provide erroneous predictions with great assurance. This phenomenon is particularly critical in the medical field where diagnostic stakes are high, as overconfidence in incorrect results can lead to diagnostic errors with severe consequences for patients.

The ability of a model to express "I don't know" is thus crucial, especially in medical applications where accuracy and safety are paramount. The transparency of these models is a necessity to build trust in their clinical use, highlighting the importance of reliably quantifying the level of certainty associated with their predictions.

My research project addresses this challenge. It aims to develop a convolutional neural network (CNN) capable of performing segmentations on brain MRIs to identify potential tumors. The goal is twofold: not only to enhance the precision of segmentation but also to integrate an uncertainty estimation to accompany each prediction with a confidence measure. Through this approach, the project seeks to enhance the reliability of AI-based diagnostics by providing not just a result, but also an evaluation of its certainty, thereby facilitating medical decision-making.

Chapter 2

Context and State of the Art

In the field of deep learning, quantifying uncertainty effectively is vital for decision-making, especially in areas such as medical imaging where safety and reliability are essential.

Segmenting an area on an image involves classifying each pixel of the image. To quantify the uncertainty in the predictions, we will construct for each input a set of predictions, each time different, in order to study the distribution of the membership probabilities of each pixel. The goal is then to construct a final prediction formed as the average of the results obtained.

During this semester, I focused my work on studying two basic but essential metrics. First, it is very interesting to study the standard deviation map for each pixel to get an overview of the distribution of the pixel membership probabilities.

How It Works : Standard deviation measures the variability or dispersion of predictions around the mean. It directly quantifies the spread of predicted probabilities for each input, making it useful for identifying areas of high and low uncertainty. The formula for the standard deviation is :

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^2$$

where:

- σ is the standard deviation.
- N is the total number of predictions.
- p_i is the probability of the i-th prediction.
- $\bar{p}$ is the mean of the probabilities of all predictions.

Next, I also calculated the Shannon entropy for each set of predictions. This metric allows us to quantify the degree of certainty the model has in its predictions.

How It Works: Shannon entropy measures the overall uncertainty of a probability distribution. It is particularly useful for classification tasks as it captures the spread of probabilities across all classes. The formula for Shannon entropy in binary classification is:

$$H(p) = -(p\log(p) + (1-p)\log(1-p))$$

where p is the predicted probability for the positive class. Higher entropy indicates greater uncertainty, meaning the predictions are more evenly spread across possible outcomes.

Utilizing Both Metrics: By using both Shannon entropy and standard deviation, I aim to capture a more comprehensive view of uncertainty:

- **Shannon Entropy** provides insights into the overall distributional uncertainty, indicating how spread out the probabilities are across different classes.
- **Standard Deviation** offers a measure of the localized variability in predictions, highlighting the consistency of the model's predictions around the mean.

Utilizing Both Metrics: By using both Shannon entropy and standard deviation, I aim to capture a more comprehensive view of uncertainty :

- **Shannon Entropy** provides insights into the overall distributional uncertainty, indicating how spread out the probabilities are across different classes.
- **Standard Deviation** offers a measure of the localized variability in predictions, highlighting the consistency of the model's predictions around the mean.

This dual approach ensures that both the distributional and localized aspects of uncertainty are adequately captured, providing a deeper analysis of prediction uncertainty.

After determining the means to quantify uncertainties, the next step is to construct the set of predictions. One existing solution is the Bayesian neural network approach. This involves constructing a neural network where all parameters are random variables. By building such a model, it becomes possible to obtain a different result for each prediction. This solution works well in theory but is quite complex to implement, as training models with several million parameters following probability distributions can be very complicated and time-consuming.

However, there are currently several alternative methods for effectively quantifying uncertainties in Deep Learning. The two most renowned and effective methods are the Monte Carlo dropout method, developed by Yarin Gal and Zoubin Ghahramani, and the Deep Ensembles method, developed by Balaji Lakshminarayanan. A clear understanding of the mechanics of these methods is important to understand the direction and objectives of my research for this semester.

Firstly, the Monte Carlo dropout method was developed in 2016 by two Professors in Artificial Intelligence at the University of Cambridge, Yarin Gal and Zoubin Ghahramani. This method involves using the dropout technique, originally designed as a regularization method to prevent overfitting, during both the training and inference phases. It provides an approximation to Bayesian inference. The method works as described below :

Training with Dropout: During training, neurons in the network are randomly switched off with a given probability p. This ensures that the network is not overly dependent on certain neurons, thus improving its ability to generalize on unknown data.

Inference with Dropout Activated: In standard usage, the dropout is deactivated during inference. However, the interest of the Monte Carlo dropout method is to maintain this random neuronal deactivation during prediction. By running a certain number of predictions with the

dropout activated, we obtain a different result for each prediction, enabling us to model a probabilistic distribution.

By aggregating all predictions, it is then possible to calculate statistics such as the mean and standard deviation of predictions. This creates a predictive distribution, where the variance among outputs indicates the model's uncertainty level (higher variance suggests greater uncertainty). This makes it possible to estimate a probability distribution over the predictions, giving a measure of uncertainty.

The advantage of the Monte Carlo dropout method is its ease of implementation, as it does not require major modifications to the model architecture and can be easily integrated into a neural network.

Utilizing Both Metrics: By using both Shannon entropy and standard deviation, I aim to capture a more comprehensive view of uncertainty:

- **Shannon Entropy** provides insights into the overall distributional uncertainty, indicating how spread out the probabilities are across different classes.
- **Standard Deviation** offers a measure of the localized variability in predictions, highlighting the consistency of the model's predictions around the mean.

This dual approach ensures that both the distributional and localized aspects of uncertainty are adequately captured, providing a deeper analysis of prediction uncertainty.

After determining the means to quantify uncertainties, the next step is to construct the set of predictions. One existing solution is the Bayesian neural network approach. This involves constructing a neural network where all parameters are random variables. By building such a model, it becomes possible to obtain a different result for each prediction. This solution works well in theory but is quite complex to implement, as training models with several million parameters following probability distributions can be very complicated and time-consuming.

However, there are currently several alternative methods for effectively quantifying uncertainties in Deep Learning. The two most renowned and effective methods are the Monte Carlo dropout method, developed by Yarin Gal and Zoubin Ghahramani, and the Deep Ensembles method, developed by Balaji Lakshminarayanan. A clear understanding of the mechanics of these methods is important to understand the direction and objectives of my research for this semester.

Firstly, the Monte Carlo dropout method was developed in 2016 by two Professors in Artificial Intelligence at the University of Cambridge, Yarin Gal and Zoubin Ghahramani. This method involves using the dropout technique, originally designed as a regularization method to prevent overfitting, during both the training and inference phases. It provides an approximation to Bayesian inference. The method works as described below:

- **Training with Dropout:** During training, neurons in the network are randomly switched off with a given probability p. This ensures that the network is not overly dependent on certain neurons, thus improving its ability to generalize on unknown data.

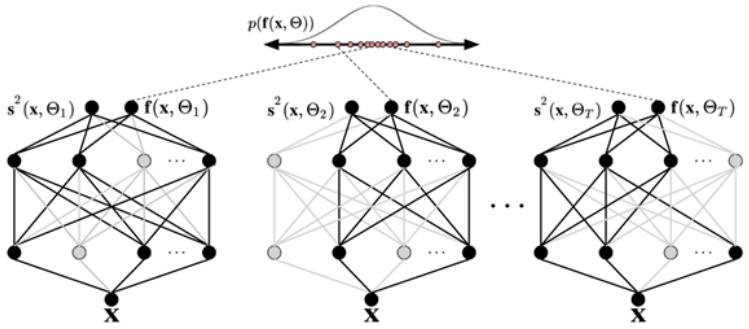
- **Inference with Dropout Activated:** In standard usage, the dropout is deactivated during inference. However, the interest of the Monte Carlo dropout method is to maintain this

random neuronal deactivation during prediction. By running a certain number of predictions with the dropout activated, we obtain a different result for each prediction, enabling us to model a probabilistic distribution.

- **Bayesian Approximation:** By aggregating all predictions, it is then possible to calculate statistics such as the mean and standard deviation of predictions. This creates a predictive distribution, where the variance among outputs indicates the model's uncertainty level (higher variance suggests greater uncertainty). This makes it possible to estimate a probability distribution over the predictions, giving a measure of uncertainty.

The advantage of the Monte Carlo dropout method is its ease of implementation, as it does not require major modifications to the model architecture and can be easily integrated into a neural network.

Figure 2.1: Illustration of the Monte Carlo Dropout Method



Secondly, another very famous method for estimating uncertainties in deep learning is the Deep Ensembles method. Developed by Stanford University professor Balaji Lakshminarayanan in 2017, this method involves training several independent neural network models with the same architecture but with different random initializations. The Deep Ensembles method is currently considered the most effective, offering highly accurate and powerful results. Here's how it works:

- **Training Independent Models:** Several neural network models are trained independently, each with different parameter initializations. This method involves training multiple instances of the same model, each starting with different initial weights and potentially exposed to different mini-batches of data during training. This diversity of training enables various representations of the parameter space to be captured, enriching the model's ability to generalize and avoiding overfitting specific to a single model configuration.
- **Combination of Predictions:** During inference, the predictions of the different models are combined to obtain a final estimate. This combination can be achieved in a number of ways, such as taking the average of each model's probabilistic predictions, or using majority voting for classifications.

The strength of Deep Ensembles lies in their collective ability to handle out-of-distribution data effectively, a typical shortfall of traditional single-model approaches. Consensus among the models signals high confidence, whereas significant variance indicates uncertainty, advising caution.

Chapter 3

Experimentations

3.1 iSEG-2017: Refining MRI Segmentation Algorithms for Comprehensive Analysis of Neonatal Brain Development

During this semester, my research has primarily centered on the iSEG-2017 dataset. The aim of this project is to enhance segmentation algorithms for MRI scans of newborn brains, facilitating a comprehensive analysis of early brain development.

3.1.1 Overview of Iseg

The iSEG-2017 dataset, released as part of the iSEG Grand Challenge at MICCAI 2017, is specifically designed for the segmentation of infant brain tissues, providing a significant tool for medical imaging researchers. This dataset is especially valuable because it targets a developmental stage where brain tissue undergoes rapid and profound changes. The high-resolution T1 and T2 weighted MRI scans included in iSEG offer a detailed view of the infant brain, facilitating a better understanding of its complex structures during critical growth periods.

3.1.2 Data Composition

The iSEG-2017 dataset includes MRI scans of 10 infants aged between 6 and 24 months, a period during which developmental changes in the brain are highly dynamic. Each scan includes expertly annotated ground-truth labels delineating three crucial brain tissue types :

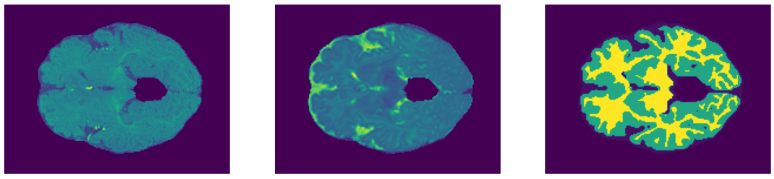
- **White Matter (WM):** Vital for the transmission of signals through the brain's neural networks. Segmentation of white matter enables us to study its development and integrity, which are essential for cognitive development and motricity.

- **Gray Matter (GM):** Important for the processing and interpretation of information flowing into the brain, understanding the development of gray matter provides a better understanding of the neurological basis of sensory processing, memory and decision-making.

- **Cerebrospinal Fluid (CSF):** Acts as a cushion and shock absorber for the brain, circulating nutrients and eliminating waste. CSF segmentation is crucial for assessing normal brain physiology and pathology in infant neurodevelopment.

The inclusion of these annotated tissue types enables accurate segmentation and the study of morphological changes during early brain development. The multimodal nature of the dataset, including both T1- and T2-weighted images, provides diverse contrasts, enhancing the ability to effectively differentiate between these tissue types.

Figure 3.1: Example of a 2D Slice for Patient 1



Here's an example of an average slice (slice number sz//2) for patient 1, showing three different views: T1-weighted, T2-weighted and VT (ventricular tissue) images. These slices offer a complete perspective of the infant brain, highlighting the different tissue contrasts and structures essential for segmentation and analysis.

- **T1-weighted image :** This image highlights the differences between white matter and gray matter, offering a clear view of the brain's overall structure.

- **T2-weighted image :** This image provides detailed contrast for cerebrospinal fluid and gray matter, complementing the information from the T1-weighted image.

- **Ground Truth Segmentation (GT Image):** This image represents the ventricular system, which is essential for identifying and analyzing brain abnormalities. It serves as the ground truth for evaluating the accuracy of segmentation algorithms.

Based on this dataset, the aim of my research work this semester was to propose a quantification of the uncertainties on the different segmentations of T1 and T2 slices by applying both the Deep Ensembles method and the Monte Carlo dropout in order to compare the results. To achieve this, I have used a famous convolutional neural network widely found in the field of medical imaging segmentation, the U-net.

3.2 U-net

3.2.1 Overview of U-net

The U-net is a type of convolutional neural network (CNN) designed specifically for image segmentation tasks. Developed by Olaf Ronneberger in 2015, the U-net has a U-shaped architecture, hence its name.

3.2.2 U-net architecture

- **Encoding Path :** This part of the network progressively reduces the spatial size of the image while increasing the number of features. It uses convolution layers followed by pooling layers

to extract important features from the image.

- **Decoding Path :** This part reconstructs the image at its original resolution while refining details. It uses up-scaling layers to increase the spatial size of the image and combine the features extracted by the contraction path.

- **Skip Connections :** These direct connections between the corresponding layers of the contraction path and the expansion path enable the preservation of high-resolution information lost during pooling operations.

3.2.3 Applications in Medical Imaging

U-net is widely used in medical imaging for segmentation tasks, where it is crucial to distinguish different anatomical structures in complex images such as MRI, CT scans and ultrasound. Here are some specific applications:

- **Segmentation of brain tissue :** Identify and segment different brain regions (gray matter, white matter, ventricles) in MRI images.

- **Tumor detection :** Locate and segment tumors or other abnormalities in various imaging modalities.

- **Organ analysis :** Segmentation of internal organs for detailed anatomical studies and radiotherapy treatment plans.

3.2.4 Using U-net in the context of my work

For my work, I used U-net with the following specifications:

- **Number of layers :** 23
- **Parameters :** Several million parameters (the exact number depends on the input and output dimensions)
- **Dropout :** 0.5 for deep layers

In order to focus my efforts on quantifying uncertainty rather than on the challenges of producing an accurate segmentation of complex structures, the first part of my semester was devoted to running a binary segmentation on brain slices. The initial objective was to distinguish the brain structure from the rest of the image. This brain structure includes cerebrospinal fluid, white matter and gray matter.

3.3 Initiating with Binary Segmentation of Brain Slices

Separating the brain area from the background of the input image requires that we perform a binary segmentation to categorize the pixels of the brain from those of the background. Input images are initially composed of pixels divided into four categories:

1. Background
2. Cerebrospinal fluid
3. White matter
4. Gray matter

To separate the background from the brain region, therefore, the data must first be labeled so that all background pixels are set to 0 and all other pixels are set to 1. For the median slice (sz//2) of patient 0, this labeling yields this result :

Figure 3.2: Comparison of Unlabeled and Labeled Brain Slices

- **Left :** Unlabeled image showing the initial classification of pixels into four categories: background (purple), cerebrospinal fluid (green), white matter (blue), and gray matter (yellow).
- **Right :** Labeled image where background pixels are set to 0 (purple) and all other pixels (cerebrospinal fluid, white matter, and gray matter) are set to 1 (yellow), demonstrating the binary segmentation approach.

3.3.1 Results Using the Monte Carlo Dropout Method

Using these labeled images, it was therefore possible to train the U network to perform the segmentations. To quantify uncertainty using the Monte Carlo dropout method described by Yarin Gal, it is recommended to make between 30 and 100 predictions for each input in order to obtain a realistic and usable distribution.

In my work, I have chosen to make 100 different predictions for each input to ensure an accurate assessment of uncertainty. This approach captures the variability of the predictions and provides a reliable estimate of the uncertainty associated with each segmentation.

The aim was then to construct an average image from the 100 samples, where each pixel in the image is calculated as the average value of the corresponding 100 pixels in the samples. The figure below shows the result obtained for a slice of patient 1.

Figure 3.3: Segmentation Process for Mid-Slice (sz//2) of Patient 1 : T1 - T2 - Predictions - Ground Truth

From left to right :

1. **T1-weighted Image:** Shows the T1-weighted MRI scan
2. **T2-weighted Image:** Shows the T2-weighted MRI scan
3. **Average Prediction:** Displays the average segmentation prediction using the Monte Carlo dropout method with 100 predictions
4. **Ground Truth (GT):** Represents the ventricular tissue as labeled in the ground truth

The above result shows that the average prediction is very close, if not almost identical, to the ground truth. This indicates that the network has learned very well and that the 100 predictions are highly consistent with each other and with reality. Additionally, the reliability of the result can be observed by examining the standard deviation map, which models the standard deviation of each pixel relative to its average.

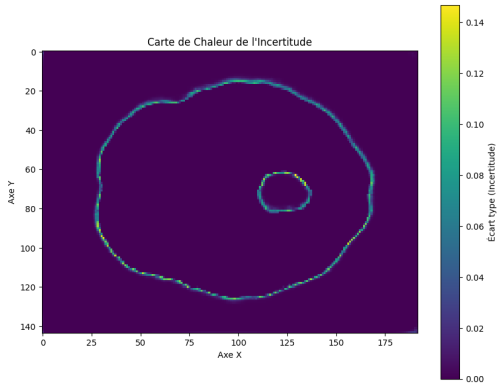


Figure 3.4: Standard Deviation Map for Monte Carlo Dropout

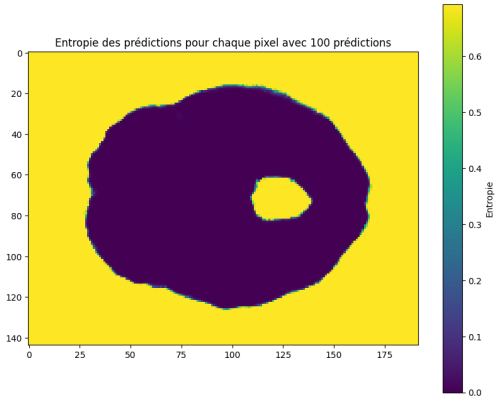


Figure 3.5: Shannon Entropy for Monte Carlo Dropout

By examining the uncertainty map, we observe that the area representing the boundary between the brain region and the background is clearly identifiable as the zone with the highest standard deviation. This result indicates ambiguity in the contours, showing that even for a binary classification task, which is supposed to be relatively easy, the five networks do not necessarily predict the same classes consistently.

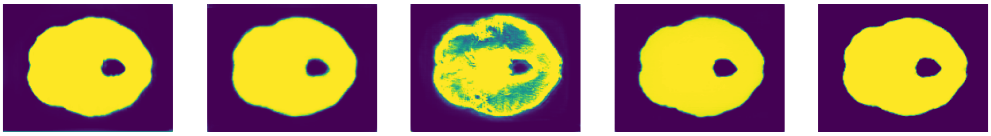
The results obtained using the Monte Carlo dropout method are consistent with our theoretical expectations. Indeed, we observe that the 100 predictions are effective in identifying the background of the image and the brain region, with almost no differences between the 100 predictions. However, there is noticeable ambiguity near the segmentation boundaries. This outcome confirms the intuition to conduct additional checks when approaching the contours.

In contrast, the Shannon entropy shows an inconsistency. Specifically, we observe high entropy for the background of the image (yellow color), which is higher than expected. We would expect lower entropy in these areas, similar to the brain region within the segmentation. We will now compare these results with those obtained using the Deep Ensembles method to determine which is more effective. This comparison will help us identify any notable differences between the two methods or confirm if their results are similar.

3.3.2 Results Using the Deep Ensembles Method

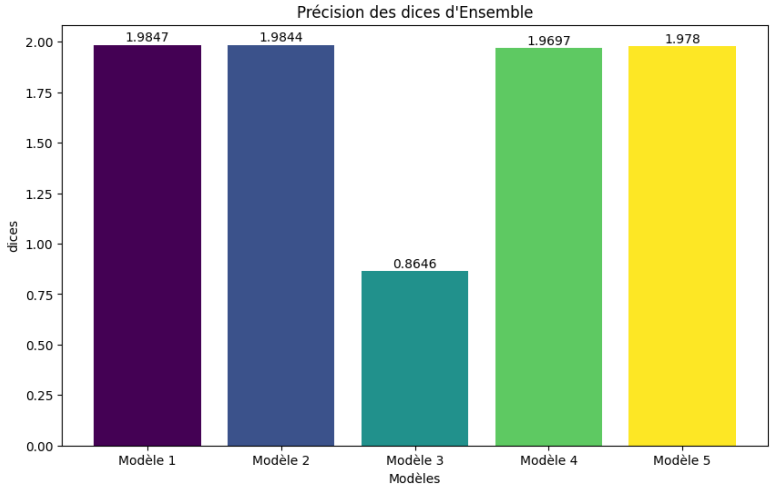
To perform the segmentation using the Deep Ensembles method as described by Balaji Lakshminarayanan, I trained 5 models with different initializations. The figure below shows the results of the 5 models for the prediction of the mid-slice (sz//2) of patient 1.

Figure 3.6: Predictions of Mid-Slice (sz//2) of Patient 1 Using Deep Ensembles



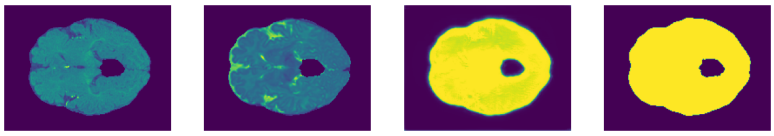
We observe that the results are quite similar except for the third model. This prediction indicates that there was an error in the training of this network, leading to a segmentation failure. This conclusion is further supported by examining the Dice scores of the 5 predictions compared to the ground truth :

Figure 3.7: Dice Scores of Predictions Compared to Ground Truth



We can see that the Dice score for Model 3 is significantly lower. This result will therefore influence the construction of the average prediction and the standard deviation map.

Figure 3.8: Average Prediction for Mid-Slice (sz//2) of Patient 1 Using Deep Ensembles



From left to right:

1. **T1-weighted Image:** Shows the T1-weighted MRI scan
2. **T2-weighted Image:** Shows the T2-weighted MRI scan
3. **Average Prediction:** Displays the average segmentation prediction using the Deep Ensembles method
4. **Ground Truth (GT):** Represents the ventricular tissue as labeled in the ground truth

In the average prediction, we observe blue traces resulting from the poorly trained network's prediction. However, the overall result remains coherent, as it is mitigated by the four other predictions from the networks that were trained correctly.

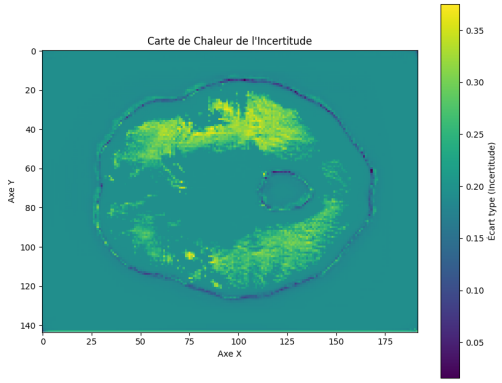


Figure 3.9: Standard Deviation Map for Deep Ensembles

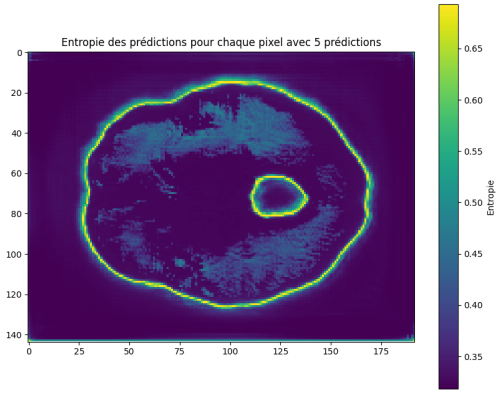


Figure 3.10: Shannon Entropy for Deep Ensembles

Here are several observations on this map: First, the yellow traces present in the average prediction correspond to the areas with the highest standard deviation (around 0.30). This is expected, as the poorly trained model's results differ significantly from others, leading to a high standard deviation. Next, we see the brain region's contour with a mix of pixels marked by both high and low standard deviations. This result is difficult to interpret because both the inside and outside areas of the brain region also show non-zero standard deviations. In regions far from the contours, we expect stable and similar predictions across models. However, we observe a standard deviation close to 0.3 in these areas, which is counterintuitive and indicates poor network training.

In contrast, the entropy results are more coherent. The probabilities decrease as we approach the contours, which aligns with our expectations.

3.4 Segmentation of White and Gray Matter in the Brain Using a Cascaded Neural Network Method

In the second phase of my work, I focused on segmenting the area containing white matter and gray matter. To do this, I used a cascade segmentation method, which exploits several neu-

ral network models in sequence in order to achieve progressive refinement.

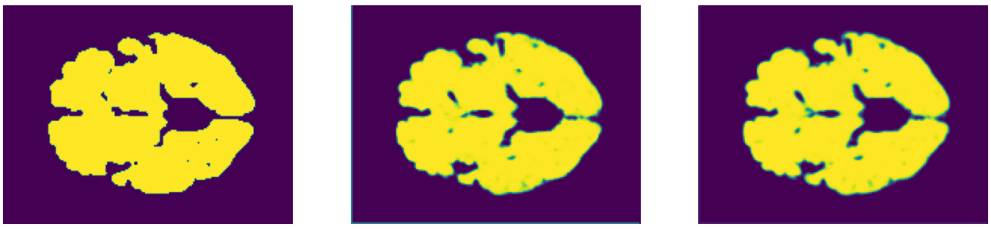
The cascade segmentation method uses an initial network to perform a coarse segmentation, capturing the main structures and providing a basic pattern. Subsequent networks then take this preliminary segmentation as input and refine it, focusing on more complex or ambiguous regions to improve overall accuracy.

In the context of my work, I constructed a second neural network that takes as input the T1 and T2 images, as well as the complete brain region encompassing white matter, gray matter, and cerebrospinal fluid. This time, the ground truths are the areas containing only the white and gray matter.
From left to right:

1. **T1-weighted Image:** Shows the T1-weighted MRI scan
2. **T2-weighted Image:** Shows the T2-weighted MRI scan
3. **Brain Region Mask:** Represents the complete brain region including white matter, gray matter, and cerebrospinal fluid.
4. **Ground Truth (White and Gray Matter):** Represents the segmentation ground truth for white and gray matter only.

Using this method, we obtain the following average predictions for the Deep Ensembles and Monte Carlo Dropout methods:

Figure 3.11: Comparison of Segmentation Methods for White and Gray Matter



From left to right:

1. **Ground Truth (White and Gray Matter):** Represents the ground truth segmentation for white and gray matter
2. **Monte Carlo Dropout Mean Prediction:** Displays the average prediction using the Monte Carlo dropout method
3. **Deep Ensembles Mean Prediction:** Displays the average prediction using the Deep Ensembles method

From these two results, we can make the following observations :

Accuracy and Consistency :

- Both the Monte Carlo Dropout and Deep Ensembles methods produce segmentations that are largely consistent with the ground truth, successfully identifying the regions of white and gray matter.

Boundary Precision :

- Monte Carlo Dropout : The mean prediction shows some variability around the boundaries of the segmented regions, indicating slightly higher uncertainty in these areas.
- Deep Ensembles : The mean prediction is smoother and exhibits less variability, suggesting higher confidence and more precise boundary definitions. This is likely because each model in the ensemble captures different aspects of the data, and the averaging process helps to mitigate individual model errors.

Segmentation Quality :

- Monte Carlo Dropout: Minor discrepancies near the edges of the white and gray matter regions indicate some sensitivity to boundary areas, but the overall segmentation remains robust.
- Deep Ensembles: The segmentation is more stable and uniform across the brain region, with fewer boundary discrepancies, indicating a more reliable performance in distinguishing tissue types.

Uncertainty Levels :

- Monte Carlo Dropout: Displays lower overall uncertainty, making it suitable for applications requiring high confidence in predictions.
- Deep Ensembles: Shows higher overall variability, providing valuable insights into regions of uncertainty, which can guide targeted review and refinement.

We also observe greater uncertainty quantified by the Deep Ensembles method when comparing the uncertainty maps of the two methods :

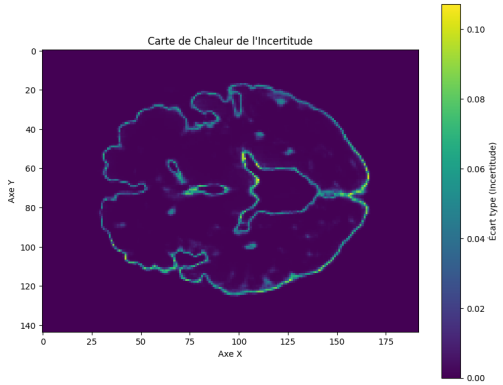


Figure 3.12: Standard Deviation Map for Monte Carlo Dropout

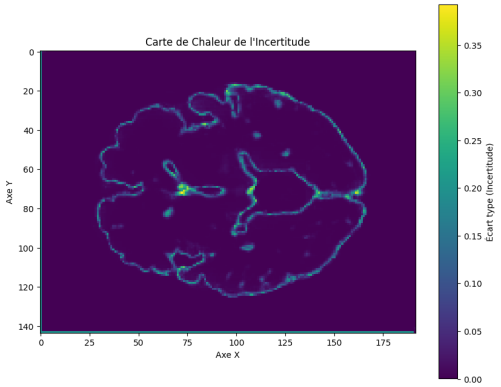


Figure 3.13: Standard Deviation Map for Deep Ensembles

- **Left (Monte Carlo Dropout):** Minor discrepancies near the edges of the white and gray matter regions indicate some sensitivity to boundary areas, but the overall segmentation remains robust. The standard deviation decreases as we move away from the contours, showing more consistent results in the interior regions.
- **Right (Deep Ensembles):** The uncertainty map for the Deep Ensembles method displays higher standard deviation values, reaching up to 0.35. There is a noticeable increase in uncertainty within the brain region, particularly around the boundaries and more complex areas. However, the standard deviation also decreases as we move away from the contours, indicating more consistent predictions in the inner regions.

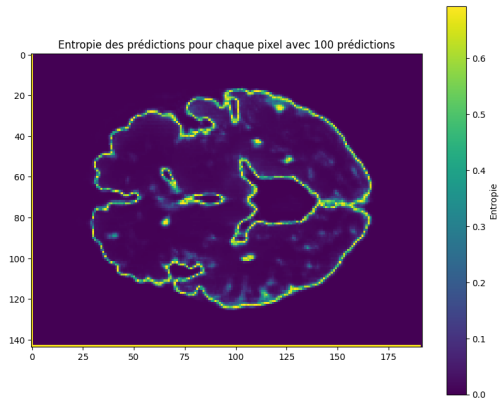


Figure 3.14: Shannon Entropy for Monte Carlo Dropout

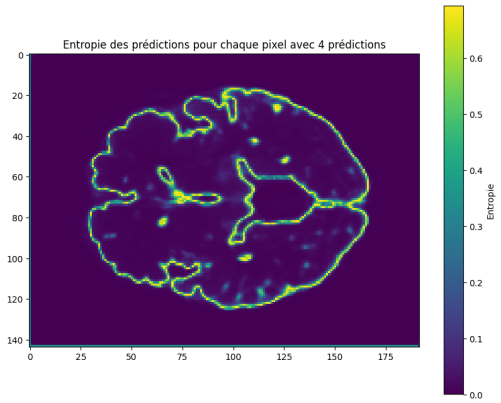


Figure 3.15: Shannon Entropy for Deep Ensembles

By examining the entropy maps, we observe consistent results. Entropy increases as we approach the contours, indicating that the membership probabilities are moving closer to 0.5 and

away from the extremes (0 and 1). This suggests higher uncertainty near the boundaries.

Moreover, the appearance of yellow points corresponds to uncertain areas not visible in the ground truth. This result is particularly interesting as it challenges the ground truth and suggests that there may be potential errors in it. The network might be correcting these errors based on the T1 and T2 information. The entropy results show that while the predictions are stable and certain in most regions, the areas near the boundaries and some isolated points are flagged as uncertain. This highlights the capability of the entropy metric to identify and provide insights into regions where the model's predictions may be less reliable.

Overall, the entropy metric not only confirms the expected increase in uncertainty near the segmentation boundaries but also reveals additional areas of potential uncertainty. This allows for a more nuanced analysis and potentially aids in refining the ground truth by identifying areas where the model's predictions differ from it.

3.5 Exclusive Segmentation of White Matter in the Brain

In the last part of my work, I further complexified my segmentations to tackle more challenging structures. Still using the principle of cascade segmentation, I tried this time to segment only the white matter region. I provided the model with T1, T2, the brain region, and the white matter/gray matter region as inputs. This allowed me to obtain new results on more complex structures, thereby enabling a more interesting study of uncertainty metrics.

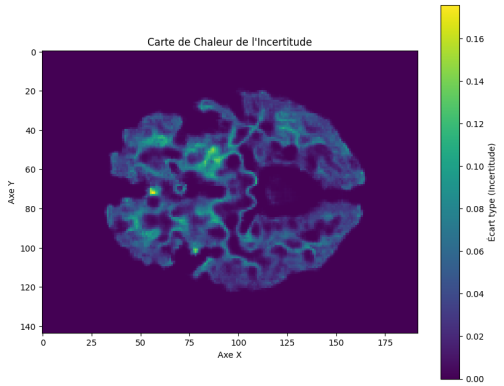


Figure 3.16: Standard Deviation for Monte Carlo Dropout

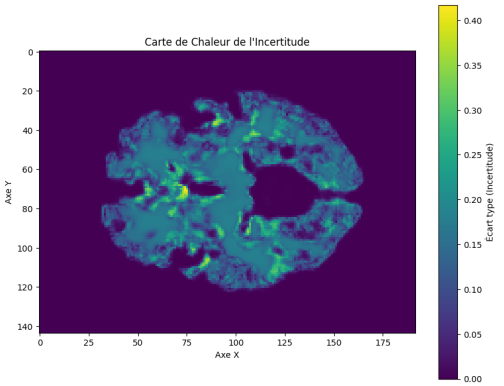


Figure 3.17: Standard Deviation for Deep Ensembles

The results obtained from the standard deviation maps remain satisfactory and consistent. An observed increase in standard deviation within the segmentation area, which aligns with the ground truth data, confirms that the segmentation is precise and the networks have learned effectively. Notably, the standard deviation map produced by the Deep Ensembles method shows a greater deviation in this area compared to that obtained by the Monte Carlo dropout method. This can be explained by the fact that Deep Ensembles train multiple models independently, each learning differently, which tends to diversify predictions more significantly. In contrast, Monte Carlo dropout, by only modifying a portion of the neurons, has a slightly less pronounced impact on the results. Increasing the dropout rate could diversify the set of

predictions even further, but too high an increase could impair the model's performance and compromise the segmentation quality, as there would not be enough active neurons for effective processing.

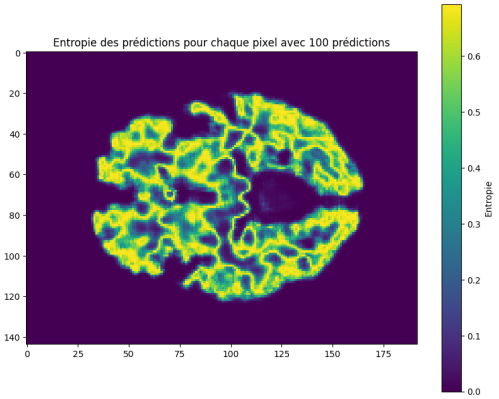


Figure 3.18: Shannon Entropy for Monte Carlo Dropout

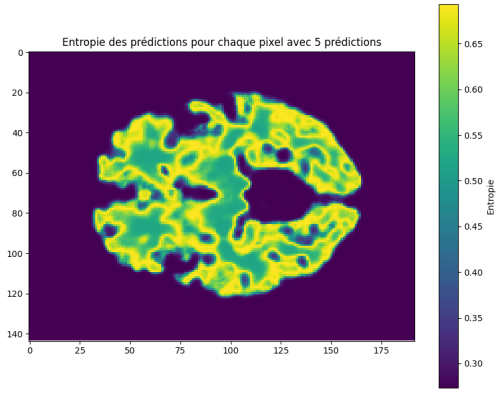


Figure 3.19: Shannon Entropy for Deep Ensembles

As for the entropy figures obtained with the Deep Ensembles and Monte Carlo dropout methods, they show similarities in certain respects. Entropy always increases as one moves away from the contours. However, this increase is now observed across the entire segmentation area, not just around the contours. This phenomenon is evident in both the results from the Deep Ensembles and the Monte Carlo dropout, with even higher entropy in the case of Deep Ensembles.

Chapter 4

Discussion and Conclusion

4.1 Discussion

4.1.1 Related Work

During this semester, my research focused on implementing two methods for quantifying uncertainties in basic segmentation tasks using the I-Seg dataset. I also performed precise uncertainty measures on my predictions using fundamental yet essential metrics: standard deviation and Shannon entropy.

4.1.2 Future Work

To deepen my work on uncertainty estimation, I intend to identify sources of uncertainty over the next semester. There are two main types of uncertainty: random uncertainty, which is related to the data, and epistemic uncertainty, which is related to the model. Successfully identifying the source of uncertainties will lead to much more accurate estimates and more realistic, convincing results. In addition, I intend to work with the MRBrains dataset to perform further segmentations on even more complex structures.

4.2 Conclusion

In conclusion, this report presents an approach to quantify uncertainty in deep learning algorithms, utilizing two techniques that currently represent the state of the art in this field. By constructing a probability distribution of each pixel's belonging, the results demonstrate the feasibility of quantifying the uncertainty of predictions based on basic yet essential metrics: the standard deviation relative to the mean and Shannon entropy.

These two metrics capture different types of uncertainty. The standard deviation examines the disagreement among predictions, thus providing a measure of how slight parameter modifications influence the outcomes. A high standard deviation indicates a certain fragility in the prediction, representing a form of uncertainty. On the other hand, Shannon entropy assesses the models' confidence in their predictions, regardless of whether the results are correct. An area with high entropy is thus a zone where the models' predicted pixel membership probabilities are close to 0.5, indicating significant ambiguity.

To further this study, it would be pertinent to quantify the sources of uncertainties, particularly by identifying the degree of epistemic uncertainty, which is related to the model's knowledge itself. Improving this understanding could help reduce uncertainty and enhance model performance. The next step in my work will involve applying these methods to another dataset for segmenting tumors in more complex structures, still using a U-Net architecture.

Additionally, my future work will include a detailed study of the implementation of Monte Carlo dropout and Deep Ensembles techniques. The intelligent placement of dropout layers in the U-Net is crucial to achieve a well-balanced and evenly distributed outcome while maintaining good results.

Chapter 5

Bibliography

[1] Mathématiques et imagerie, Bibliothèque Tangente n°77, Edition Pôle, 2022. https://infinimath.com/librairie/pdf/BIB77_sommaire.pdf

[2] Mobarakol Islam, Vibashan Jeya Maria Jose, Navodini Wijethilake, Uppal Utkarsh, Hongliang, "Brain Tumor Segmentation and Survival Prediction using 3D Attention UNet", Published on ResearchGate.

[3] Steven Czolbe, Kasra Arnavaz, Oswin Krause, Aasa Feragen, "Is segmentation uncertainty useful?", https://arxiv.org/abs/2103.16265.

[4] Balaji Lakshminarayanan, Alexander Pritzel, Charles Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles", https://arxiv.org/abs/1612.01474.

[5] Alireza Mehrtash, William M.Wells III, "Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation", https://arxiv.org/abs/1911.13273.

[6] Yarin Gal, Zoubin Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", University of Cambridge, https://arxiv.org/abs/1506.02142.

[7] Josiah Davis, Jason Zhu, PhD, Jeremy Oldfather, Samual MacDonald, Maciej Trzaskowski, PhD, "Quantifying Uncertainty in Deep Learning Systems", AWS Prescriptive Guidance, August 2020. https://d1.awsstatic.com/APG/quantifying-uncertainty-in-deep-learning-systems.pdf

[8] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", CoRR, abs/1505.04597, 2015. http://arxiv.org/abs/1505.04597

[9] Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, Ender Konukoglu, "PHiSeg: Capturing Uncertainty in Medical Image Segmentation", Computer Vision Lab, ETH Zürich, Switzerland, University Hospital Zürich, Switzerland, Memorial Sloan Kettering Cancer Center, New York, USA, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, USA. https://arxiv.org/pdf/1906.04045

[10] Zoubin Ghahramani, "Probabilistic machine learning and artificial intelligence", Nature, vol. 521, no. 7553, 2015. https://www.repository.cam.ac.uk/items/ec26a18e-5e5b-4426-8d5b-828a2371efa7

[11] José Miguel Hernández-Lobato, Ryan P. Adams, "Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks", ICML-15, 2015. https://arxiv.org/pdf/1502.05336