Segmentation of Cerebral Tissues in Human Brain MRIs with uncertainty

Yacine BOUREGHDA
(supervisor: Nicolas BOUTRY)

Technical Report n^o202501-techrep-boureghda, Jan 2025
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. These methods will enable the computation of uncertainty by focusing on both epistemic uncertainty, related to model knowledge, and aleatoric uncertainty, arising from data noise.

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 d'effectuer 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 repose sur la génération de multiples prédictions en désactivant aléatoirement des neurones, tandis que les Deep Ensembles entraînent plusieurs réseaux avec des initialisations différentes. Ces méthodes permettront de quantifier l'incertitude en se concentrant à la fois sur l'incertitude épistémique, liée à la connaissance du modèle, et sur l'incertitude aléatoire, due au bruit des données.

Keywords
Segmentation, Medical Imaging, Neural Networks, MRI, Tumor Progression, Uncertainty Quantification, Monte Carlo Dropout, Deep Ensembles, Epistemic Uncertainty, Aleatoric Uncertainty, Prediction Reliability

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

Initially, medical image segmentation is used to extract organs, detect pathologies, and isolate anatomical structures from images such as X-rays, CT scans, or MRIs. With significant advancements in artificial intelligence, Deep Learning algorithms are now widely employed for these tasks, delivering remarkably accurate results. However, neural networks often act as "black boxes," making it difficult to understand the reasoning behind a given prediction. For small models, it is still possible to trace the decision path and interpret the results. However, with deeper architectures containing numerous layers and millions of parameters, interpretability becomes highly complex. The model receives an input, generates a prediction, but the exact reasoning behind the decision remains opaque. This challenge applies to both classification tasks, where the goal is to assign a global class to the image, and segmentation tasks, which involve pixel-wise classification.

In the medical domain, where decisions can directly impact patient health, it is crucial to better understand and interpret model outputs. Providing a precise prediction is not sufficient; it is equally important to know how confident the model is in its prediction. Ideally, the model should express high confidence when producing correct results and show uncertainty when it is less sure, particularly when dealing with noisy, out-of-distribution, or ambiguous data. However, neural networks often tend to be overly confident, even when their predictions are incorrect. This overconfidence can be problematic, as it may lead to excessive trust in erroneous results. Therefore, uncertainty quantification is essential, as it provides a measure of the model's confidence alongside its predictions, enabling better result analysis and more informed decision-making.

It is also important to distinguish between uncertainty metrics and traditional performance metrics used to evaluate Deep Learning models. Performance metrics, such as accuracy or F1-Score, assess whether a prediction is correct or incorrect. On the other hand, uncertainty metrics focus on the model's confidence level, regardless of whether the prediction is right or wrong. However, comparing uncertainty with prediction accuracy remains essential, as one would expect the model to be confident when making correct predictions and more hesitant when it makes mistakes. It is also interesting to observe the model's behavior when exposed to data it has never seen during training.

To quantify these uncertainties, distributions of predictions will be generated for the same input. Two state-of-the-art methods will be used in this project: Monte Carlo Dropout (MCD) and Deep Ensembles. By generating multiple stochastic predictions, these approaches estimate the variability in the model's outputs, thereby providing insights into its confidence levels.

A key objective of this work is also to explore a hybrid method combining Deep Ensembles and Monte Carlo Dropout to assess whether this combination improves uncertainty quantification or if it produces redundant results compared to using the methods separately.

This report will present the results obtained after implementing these three methods on two different tasks: an image classification task using the MNIST dataset and a brain MRI segmentation task using the iSEG-2017 dataset. For each task, the uncertainties obtained will be analyzed and compared to observe the differences between the methods and the potential benefit of combining MCD and Deep Ensembles. In the classification task, we will show that there are no significant differences between the MCD, DE, and Hybrid methods. However, in the segmentation task, the results demonstrate a clear advantage in uncertainty quantification with the Deep Ensembles and Hybrid methods, which perform significantly better compared to MCD.

Before concluding this introduction, I would like to thank Mr. Nicolas Boutry, my supervisor for this project over the past two semesters. His support and guidance have been essential in keeping me focused and making progress throughout the project. His ability to explain complex ideas clearly and provide helpful feedback has been a constant source of motivation, allowing me to tackle challenges and advance this work with confidence.

Chapter 2

Context and State of the Art

2.1 Monte Carlo Dropout

The Monte Carlo Dropout (MC Dropout) method is a powerful technique for quantifying uncertainty in Deep Learning models. Developed in 2016 by Yarin Gal and Zoubin Ghahramani, researchers at the University of Cambridge, it is based on the use of dropout, initially designed as a regularization method to prevent overfitting.

The main idea is to activate dropout not only during training but also during the evaluation phase. By doing so, multiple predictions are made with dropout activated. The resulting set of predictions forms a probability distribution, which allows differences between predictions to be studied and thus quantifies model uncertainty.

Yarin Gal et al. showed that complex calculations could be avoided, and the mean and variance of a network's predictive distribution could be directly obtained by following this procedure (for a regression task):

- Train a network with dropout activated.
- Evaluate the same input multiple times using the trained network, applying dropout each time.
- Compute the mean and variance of the resulting outputs.

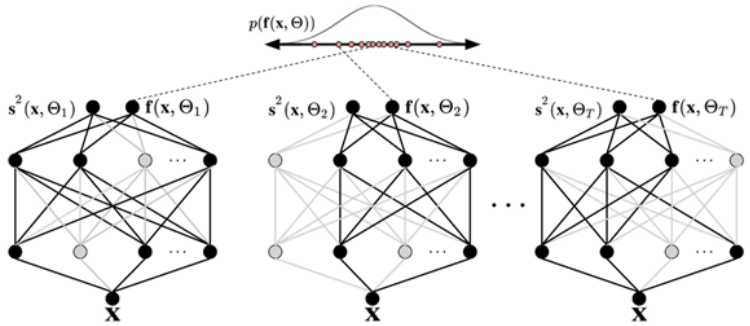
They demonstrated that variational inference could be approximated by dropout regularization, meaning the variational lower bound could resemble the dropout objective in a standard training scenario with certain assumptions about the prior and approximate posterior. Each version of the model at the evaluation stage can be seen as a sample from the posterior distribution $p(\theta|D)$, and each corresponding output can be considered a sample from the predictive distribution $p(y|x, D)$.

Thus, Gal et al. made the process of sampling from a network's posterior and predictive distributions computationally feasible. Given how effective Bayesian optimization is for non-neural network models, one might expect that MC Dropout, derived from Bayesian principles, would outperform other non-Bayesian uncertainty quantification approaches. However, subsequent works have shown this is not always the case.

The method is straightforward to implement since it requires only a single model and a single training run. However, it has certain limitations. The quality of uncertainty quantification depends heavily on the chosen dropout rate. A high dropout rate leads to significant differences in the prediction distribution but at the cost of quality loss, as the network becomes less effective in reasoning correctly. Conversely, a low dropout rate preserves prediction accuracy but significantly reduces variance in the distribution, thereby limiting uncertainty assessment.

In this research work on uncertainty quantification in Deep Learning, the Monte Carlo Dropout method is explored to generate probability distributions over predictions. Although simple and effective, the choice of the dropout rate is crucial and must be carefully adjusted to achieve a good balance between prediction quality and uncertainty measurement.

Figure 2.1: The same input, x, is fed to each, and the resulting outputs form a distribution we can use to estimate the uncertainty



2.2 Deep Ensembles

The Deep Ensembles method is a powerful and widely recognized approach for uncertainty quantification in Deep Learning. Developed in 2017 by Balaji Lakshminarayanan, a professor at Stanford University, this method is based on training multiple independent neural networks sharing the same architecture but with different random initializations. This variability in initializations allows exploration of diverse regions of the parameter space, improving the diversity of the predictions obtained.

The core idea is to form a collection of models, which can be trained either on distinct subsets of the training dataset or on the full dataset if data is limited. Each model in the ensemble generates its own predictions, which are then aggregated, often by averaging or majority voting, to form a global probability distribution. This distribution reflects the variability of results and thus quantifies the model's uncertainty on new data.

Deep Ensembles are currently considered one of the most effective methods for uncertainty quantification, providing accurate and robust results in numerous application contexts, including computer vision. Their ability to explore different regions of the parameter space helps improve the robustness and generalization of the model.

However, this method also presents certain limitations. It is complex to implement, requiring the training of multiple independent models, leading to high memory and computational costs. Moreover, multiple training runs can be challenging to manage on large datasets, limiting its applicability in resource-constrained environments.

Despite these challenges, Deep Ensembles remain a benchmark tool for uncertainty estimation in Deep Learning models, often used as a standard for comparing other uncertainty quantification techniques.

2.3 Uncertainty Metrics

Once prediction distributions have been constructed, it becomes possible to quantify uncertainty using various metrics. Initially, the mean prediction and standard deviation should be examined, as they provide a strong initial indication. However, more advanced metrics can be used for a more precise assessment of model uncertainty, which will be presented in this section.

It is crucial to understand that these metrics are distinct from performance metrics. The goal here is not to evaluate whether a prediction is correct or incorrect but rather to determine how confident the model was in its prediction.

Before exploring the uncertainty metrics, it is essential to understand that there are two primary types of uncertainties: aleatoric uncertainty and epistemic uncertainty.

2.3.1 Aleatoric Uncertainty

Derived from the Latin *aleator* (dice player), aleatoric uncertainty represents natural and random noise present in the data.

Source: Inherent noise in the data (e.g., blurry images, poor-quality sensors).

Characteristic: It cannot be reduced by adding more data but can be estimated or learned from the data.

Subtypes:

- **Homoscedastic Uncertainty:** Constant across all inputs and captures average uncertainty.
- **Heteroscedastic Uncertainty:** Depends on the inputs and varies based on the noise present in the observations.

2.3.2 Epistemic Uncertainty

Derived from the Greek *episteme* (knowledge), epistemic uncertainty is linked to the model itself, reflecting a lack of knowledge about the data-generating process.

Source: Poorly trained models or previously unseen data.

Characteristic: It can be reduced with more training data.

Manifestation: Appears when the model encounters data outside its training distribution.

2.3.3 Metrics to Quantify Uncertainty

Variation Ratios – Measures Epistemic Uncertainty

Definition:

$$VarRatio(x) = 1 - \frac{f_c}{T} \tag{2.1}$$

where:

- T is the total number of stochastic forward passes (e.g., through Monte Carlo Dropout).
- f_c is the frequency of the most frequent predicted class label.

Why does it measure epistemic uncertainty?

- If a model is uncertain, it will produce diverse predictions across multiple passes.
- If a model is confident, it will consistently predict the same label.

A high Variation Ratio indicates high epistemic uncertainty.

Predictive Entropy – Measures Aleatoric Uncertainty

Definition:

$$H[y|x, D_{train}] = -\sum_c \left(\frac{1}{T} \sum_{t=1}^{T} p(y = c|x, w_t) \right) \log \left(\frac{1}{T} \sum_{t=1}^{T} p(y = c|x, w_t) \right) \tag{2.2}$$

where:

- $p(y = c|x, w_t)$ is the predicted probability for class c during the t-th forward pass.

Why does it measure aleatoric uncertainty?

- Aleatoric uncertainty arises from inherent data noise.
- If the data is ambiguous or noisy, the predictions will be naturally spread across classes.

A high predictive entropy indicates high aleatoric uncertainty, even for a well-trained model.

2.3.4 Summary of Uncertainty Types and Metrics

- **Aleatoric Uncertainty:**
  - Source: Noise in the data.
  - Key Metric: Predictive Entropy.
- **Epistemic Uncertainty:**
  - Source: Lack of model knowledge.
  - Key Metric: Variation Ratios.

Combining both metrics provides a thorough evaluation of uncertainty in Deep Learning models, making them particularly valuable for critical applications where reliable and well-calibrated predictions are essential. Building on this theoretical foundation, the next sections will present the results obtained from applying the Monte Carlo Dropout (MCD) method and Deep Ensembles, along with the aforementioned uncertainty metrics.

The experimental results will be structured into two main parts. The first part will focus on a classification task using the MNIST dataset, which serves as a benchmark for evaluating uncertainty estimation in a straightforward image classification scenario. The second part will address a more complex segmentation task using the iSeg dataset, where uncertainty quantification plays a crucial role in medical image analysis. By comparing both methods and metrics across these datasets, we aim to provide a comprehensive assessment of their performance and suitability for different types of tasks.

Chapter 3

Experimentations

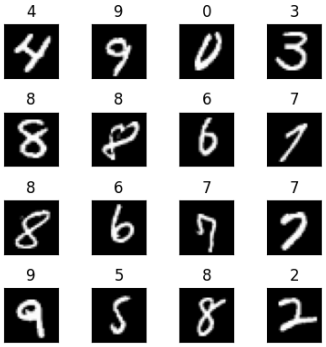
3.1 MNIST Classification

To validate the proposed uncertainty quantification techniques, we conducted experiments on two datasets: MNIST for classification and iSeg for segmentation tasks. The models utilized include Monte Carlo Dropout (MCD) and Deep Ensembles, both evaluated using Variation Ratios and Predictive Entropy as uncertainty metrics.

To begin my study on uncertainty quantification, I chose to experiment with these methods on a simpler task than segmentation: image classification. Segmentation can indeed be interpreted as pixel-wise classification. To better understand the behavior of uncertainty quantification methods in a controlled setting, I worked with the MNIST dataset, a standard benchmark in machine learning for classification tasks.

The MNIST dataset consists of grayscale images of handwritten digits (0 to 9) with a resolution of 28×28 pixels. Most digits are clear and easy to identify, but the dataset also includes slightly noisy or distorted digits, making them more challenging for a model to classify correctly. This diversity allowed me to study the behavior of uncertainty methods in both simple and ambiguous cases.

Figure 3.1: Examples of clear and noisy MNIST digits used for classification. The dataset includes both easily identifiable digits and challenging noisy samples



3.1.1 Model Architecture: AlexNet

For this classification task, I used a convolutional neural network (CNN) inspired by the AlexNet architecture, chosen for its simplicity and effectiveness in image classification tasks.

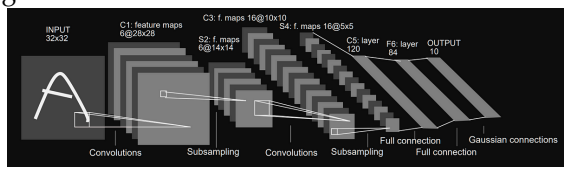
- Convolution with 6 filters, kernel size 5, padding 2, ReLU activation
- Max pooling, kernel size 2
- Convolution with 16 filters, kernel size 5, ReLU activation
- Max pooling, kernel size 2

Fully-Connected Layers:

- Dropout layer with $p = 0.25$
- Fully connected layer with 120 units and ReLU activation
- Dropout layer with $p = 0.5$
- Output layer with 10 units (one per digit class)

This architecture extracts progressively deeper features from the images while reducing the spatial dimension. Dropout is employed to regularize the learning process and introduce variability, which is essential for uncertainty estimation.

Figure 3.2: Diagram of the AlexNet architecture used for MNIST classification



3.1.2 Model Training and Usage

I trained a first model using the AlexNet architecture described above. This model served as the basis for implementing the Monte Carlo Dropout (MCD) uncertainty quantification method. Additionally, I trained five independently initialized models using the same architecture to implement the Deep Ensembles method. The randomness in initialization helps explore different regions of the parameter space and better estimate epistemic uncertainty.

Table 3.1: Performance metrics for the Monte Carlo Dropout (MCD) model.

Metric	Value
Accuracy	0.9920
Recall	0.9923
F1 Score	0.9922

Table 3.2: Performance metrics for the five models of the Deep Ensembles.

Model	Accuracy	Recall	F1-Score
Model 1	0.9890	0.9890	0.9890
Model 2	0.9902	0.9902	0.9902
Model 3	0.9909	0.9909	0.9909
Model 4	0.9745	0.9745	0.9746
Model 5	0.9862	0.9862	0.9862

3.1.3 Experiments and Results

To compare the different uncertainty quantification methods, I performed predictions on the MNIST test set using the following approaches:

- **Monte Carlo Dropout (MCD):** A single model with the dropout layer activated during inference, generating 100 stochastic predictions.
- **Deep Ensembles:** Five independently trained models, each making a single prediction.
- **Hybrid Approach (MCD + Deep Ensembles):** A combination of both methods where I generated 20 stochastic predictions per model, totaling 100 predictions.

This hybrid approach aims to evaluate whether combining Monte Carlo Dropout and Deep Ensembles provides better uncertainty quantification than using either method individually.

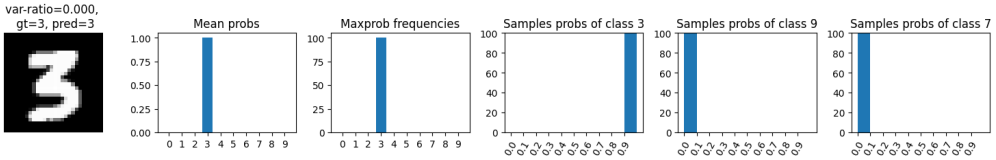
To represent uncertainty, I primarily used the Variation Ratio (Var-Ratio), a metric particularly suited for analyzing epistemic uncertainty, reflecting the model's confidence in its predictions.

3.1.4 Results Presentation

To illustrate the model's behavior in different scenarios, I selected three examples of predictions using the Monte Carlo Dropout method :

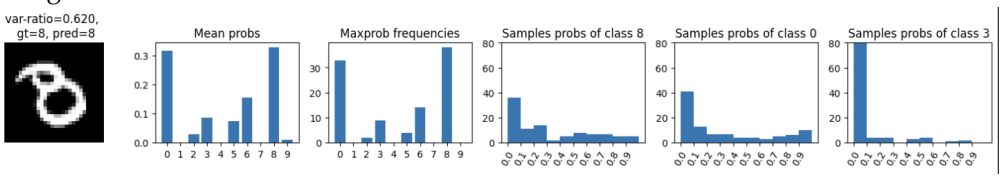
Well-Classified Digit without Uncertainty: A digit 3 correctly classified by the model. Across 100 stochastic predictions, the model predicted the correct class every time with a Var-Ratio of 0, indicating total confidence.

Figure 3.3: A digit 3 correctly classified with total confidence using the Monte Carlo Dropout method. The model predicted the same class across all 100 stochastic passes, resulting in a Var-Ratio of 0.



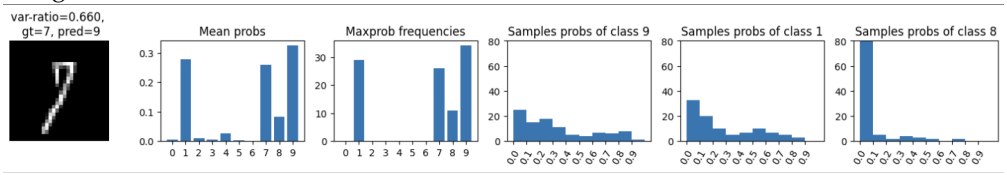
Noisy Digit with Moderate Uncertainty: A slightly distorted 8, correctly classified by the model but with significant uncertainty (Var-Ratio of 0.620). The predictions show the model hesitated between classes 8 and 0, explaining the increased uncertainty.

Figure 3.4: A noisy digit 8 correctly classified using the Monte Carlo Dropout method. Despite the correct prediction, the model exhibited moderate uncertainty, with a Var-Ratio of 0.620, indicating hesitations between classes 8 and 0



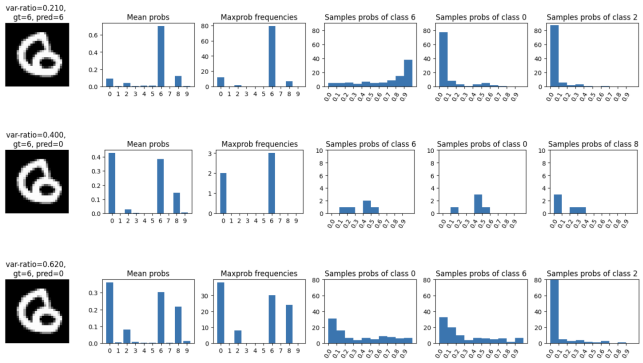
Misclassified Digit with High Uncertainty: A digit 7 misclassified as a 9. The uncertainty is high, reflecting confusion between multiple classes, mainly 7, 9, and 1.

Figure 3.5: A noisy digit 8 correctly classified using the Monte Carlo Dropout method. Despite the correct prediction, the model exhibited moderate uncertainty, with a Var-Ratio of 0.620, indicating hesitations between classes 8 and 0



To further explore uncertainty estimation, I tested a highly distorted digit 6 using all three methods: MCD, Deep Ensembles, and the Hybrid Approach.

Figure 3.6: A distorted digit 6 analyzed using three uncertainty quantification methods: Monte Carlo Dropout, Deep Ensembles, and a Hybrid Approach. Each method demonstrates varying levels of epistemic uncertainty, with the Hybrid Approach showing the highest Var-Ratio due to increased stochastic sampling



All three methods misclassified the 6 as a 0, but the key observation here lies in the variation of epistemic uncertainty across the methods:

- **Monte Carlo Dropout (MCD):** Var-Ratio of 0.210
- **Deep Ensembles (DE):** Var-Ratio of 0.400

- **Hybrid Approach:** Var-Ratio of 0.620

The uncertainty is higher with Deep Ensembles compared to Monte Carlo Dropout. This can be explained by the fact that Deep Ensembles leverage multiple independently trained models with varying initializations, leading to a broader spread in predictions and thus capturing more epistemic uncertainty.

Finally, the Hybrid Approach yields the highest uncertainty level. This results from combining both model diversity and the stochastic nature of dropout during inference. With 20 stochastic predictions per model, the overall distribution becomes more spread, leading to higher uncertainty.

Having explored how these methods perform on a classification task, I will now apply the same metrics and techniques to a more complex image segmentation task. This will help assess how Monte Carlo Dropout, Deep Ensembles, and the Hybrid Approach capture uncertainty when segmenting medical images, where predictions need to be made for each pixel.

3.2 Iseg Segmentation

3.2.1 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.

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.

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 :

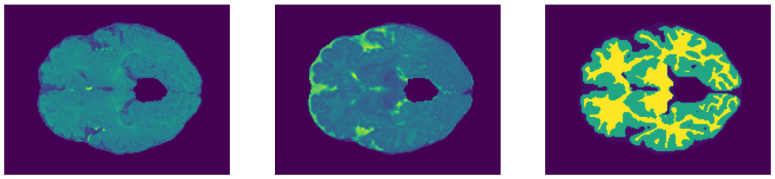
- **White Matter (WM):** Vital for the transmission of signals through the brain's neural networks. Segmentation of white matter enables us to study its development and integrity, which are essential for cognitive development and motricity.

- **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 3.7: 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.

3.2.2 U-net

Overview of U-Net

The U-Net is a convolutional neural network architecture designed for image segmentation tasks, particularly effective in biomedical imaging. It consists of a symmetrical architecture where the input image undergoes a series of transformations to extract features and later reconstructs the segmented output with high accuracy.

U-Net Architecture

The U-Net architecture is composed of three key elements: the encoder path, the decoder path, and the skip connections.

The encoder path extracts essential features while progressively reducing the spatial resolution of the image. It consists of a series of convolutional blocks, each including two 3x3 convolutions followed by ReLU activations, with a max-pooling operation (2x2) halving spatial dimensions. Simultaneously, the number of feature channels doubles to capture more complex patterns.

The decoder path reconstructs the segmented image back to its original resolution. Each stage begins with a transposed convolution (up-sampling) that doubles the spatial resolution. This is followed by concatenation with the corresponding features from the encoder path via skip connections. Two 3x3 convolutions with ReLU activations refine the predictions further.

Skip connections play a crucial role in preserving high-resolution information lost during pooling. They transfer high-level features from the encoder directly to the decoder, improving segmentation boundary accuracy and enabling finer detail recovery.

Applications in Medical Imaging

The U-Net has widespread applications in medical imaging due to its precise segmentation capabilities, particularly in differentiating complex anatomical structures. Some typical use cases include:

- **Brain tissue segmentation:** Identifying gray matter, white matter, and ventricles in MRI scans.
- **Tumor detection:** Localizing and segmenting tumors in medical scans such as MRI and CT images.
- **Organ analysis:** Segmenting internal organs for anatomical studies and radiotherapy planning.
- **Microscopic cell analysis:** Segmenting cells in microscopy for cellular biology applications.

Using U-Net in the Context of My Work

In my work, I employed a U-Net configured with the following parameters:

- **Number of layers:** 23.

- **Number of parameters:** Several million (depending on input and output dimensions).
- **Dropout:** 0.5 in deeper layers to regularize training.

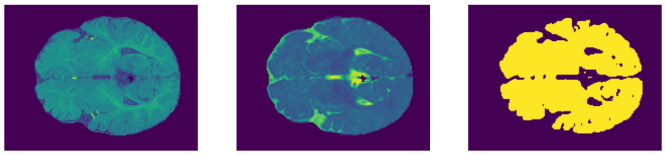
This model was applied to the segmentation of 2D slices from the iSeg dataset, enabling precise delineation of structures of interest in medical images. This segmentation step lays the foundation for the next phase, focused on evaluating the results and analyzing the model's performance.

3.3 Experimental Results on Iseg

The focus now shifts to the segmentation task using the iSeg dataset. The goal is to compare the uncertainty quantification methods of Deep Ensembles and Monte Carlo Dropout on 2D slices from the dataset. For this, binary segmentations were performed to distinguish white matter and gray matter from other structures present in the MRI images.

The model was trained to segment these regions using T1 and T2 MRI images, with ground truth generated through automatic thresholding of the reference image. This segmentation analysis aims to further investigate the performance of the uncertainty metrics and methods previously explored in the classification task, now in a more complex pixel-wise classification scenario.

Figure 3.8: Example of T1 and T2 MRI images used as inputs along with the corresponding ground truth segmentation. These images serve as the input data for the segmentation task, with the ground truth providing a reference for model evaluation.



3.3.1 Uncertainty Quantification with Monte Carlo Dropout

To evaluate the model's uncertainty, a U-Net was trained with the following configuration:

- **Number of layers:** 23
- **Number of parameters:** Several million
- **Dropout:** 0.5 in the deeper layers

Using this model, both aleatoric and epistemic uncertainties were estimated via the Monte Carlo Dropout method. This approach generates a distribution of predictions for a single input by keeping the Dropout layer active even during inference.

Specifically, for a given image, 100 stochastic predictions were generated with the Dropout layer enabled.

Table 3.3: Summary of the model's performance after training.

Metric	Value
Loss	0.095
Accuracy	0.9886
Dice Score	0.9249
Precision	0.9121
Recall	0.9390

Figure 3.9: Mean prediction obtained after 100 stochastic forward passes using the Monte Carlo Dropout method. The image represents the average predicted segmentation map, providing insight into the model's confidence across multiple inferences.

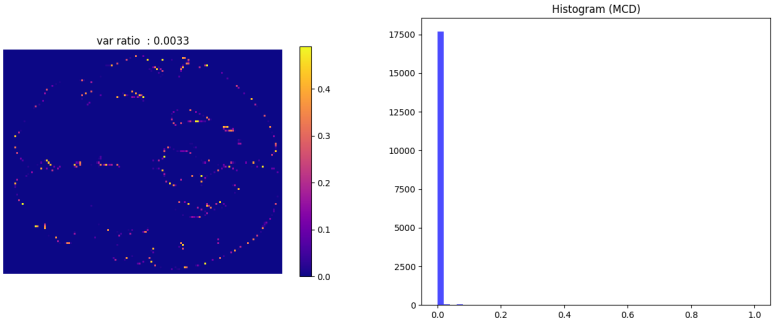
3.3.2 Uncertainty Maps Calculation

Based on the distribution of generated predictions, two key uncertainty metrics were computed:

Variation Ratio (Var-Ratio):

- Measures epistemic uncertainty based on the frequency of the majority class among the predictions.
- Higher variation ratio values indicate greater uncertainty.

Figure 3.10: Variation Ratio map obtained after 100 stochastic forward passes using the Monte Carlo Dropout method. The image highlights regions of high uncertainty, with the accompanying histogram representing the distribution of uncertainty values across the segmented image

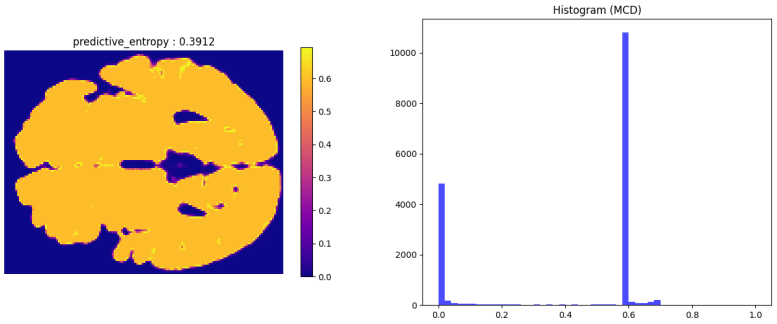


Predictive Entropy:

- Measures aleatoric uncertainty by analyzing the overall dispersion of predictive probabilities.

- High entropy values indicate significant uncertainty due to data noise or ambiguity.

Figure 3.11: Predictive Entropy map obtained after 100 stochastic forward passes using the Monte Carlo Dropout method. The image highlights regions of high uncertainty, with the accompanying histogram representing the distribution of uncertainty values across the segmented image.



3.3.3 Uncertainty Quantification with Deep Ensembles

Next, the Deep Ensembles method was applied to compare how it quantifies uncertainty in the segmentation task. This approach involves training multiple independent models with the same architecture but different random initializations, allowing for a broader exploration of uncertainty.

Table 3.4: Performance metrics for the five models trained for segmentation.

Model	Dice Score	IoU	Precision	Recall
Model 1	0.9254	0.8944	0.9243	0.9275
Model 2	0.9233	0.8911	0.9035	0.9459
Model 3	0.9221	0.8886	0.9241	0.9212
Model 4	0.9176	0.8813	0.8967	0.9422
Model 5	0.9019	0.8545	0.8601	0.9529

3.3.4 Uncertainty Maps Calculation

As with the previously described Monte Carlo Dropout method, uncertainty maps for Variation Ratio and Predictive Entropy were estimated from the aggregated predictions of the five models.

Figure 3.12: Variation Ratio map obtained after 100 stochastic forward passes using the Monte Carlo Dropout method. The map highlights regions of high uncertainty, while the accompanying histogram displays the distribution of variation ratio values across the segmented image.

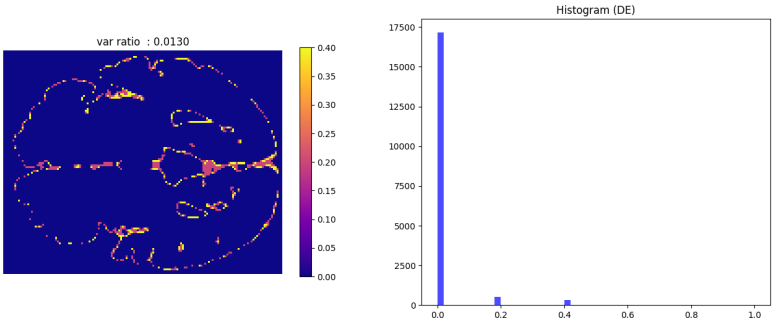
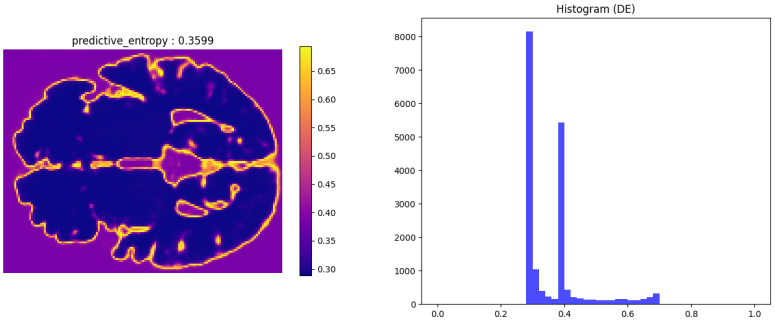


Figure 3.13: Predictive Entropy map obtained after 100 stochastic forward passes using the Monte Carlo Dropout method. The map emphasizes regions of high uncertainty, with the histogram showing the spread of predictive entropy values across the segmented image.



The Deep Ensembles method captures a broader range of prediction variability, both for epistemic and aleatoric uncertainty. This behavior is expected since multiple independently trained models with different initializations result in significant variations between predictions.

In contrast, the Monte Carlo Dropout (MCD) method relies on a single model where the only source of variability comes from stochastic dropout activation. This structural modification alone is often insufficient to introduce substantial differences between predictions.

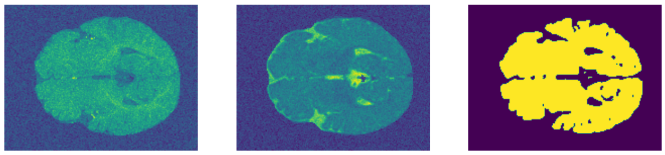
However, it is essential to note that increasing the dropout rate excessively can artificially increase variability among predictions. A very high dropout rate disrupts the model's structure, reducing its ability to predict correctly and leading to degraded results.

Therefore, balancing the dropout rate is crucial to maximize prediction quality while capturing meaningful uncertainty.

3.3.5 Behavior of the Three Methods on a Noisy Image

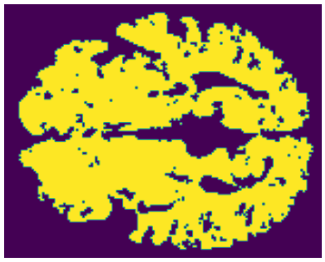
After evaluating how the three uncertainty quantification methods behave on standard inputs, the next step involves assessing their performance on a noisy input. To introduce variability and increase the complexity of the task, Gaussian noise was added to the T1 and T2 MRI images. The same prediction protocols were applied to investigate how uncertainty is quantified in this altered context.

Figure 3.14: Noisy T1 and T2 MRI images used as inputs for segmentation. Gaussian noise was applied to evaluate the behavior of uncertainty quantification methods under challenging conditions.



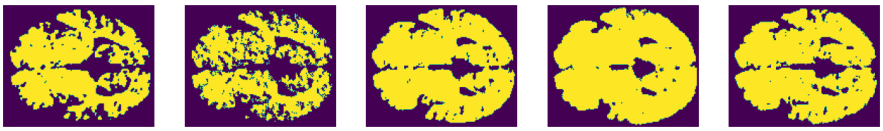
Given the increased noise in the inputs, the resulting segmentations were visibly distorted, as shown in the mean prediction below using the Monte Carlo Dropout method:

Figure 3.15: Mean segmentation prediction using Monte Carlo Dropout on noisy MRI data. The segmentation output is visibly degraded due to the added noise.



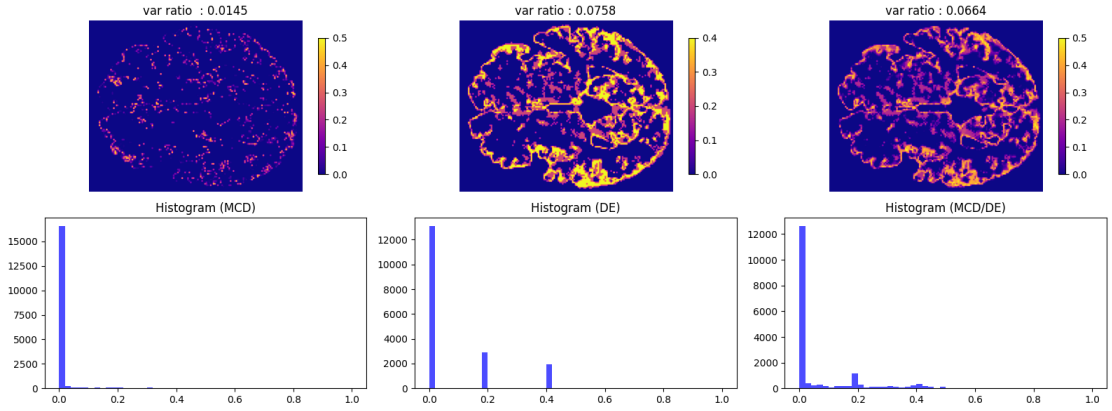
Similarly, the individual predictions obtained using the five models of the Deep Ensembles method reveal significant variation:

Figure 3.16: Segmentations obtained from the five independently trained models of the Deep Ensembles method on the noisy MRI images.



To further analyze the uncertainty, the Variation Ratio (Var-Ratio) maps were computed for the three methods: Monte Carlo Dropout, Deep Ensembles, and the Hybrid Approach. The results are shown below:

Figure 3.17: Variation Ratio maps for the three methods (MCD, DE, and Hybrid) on the noisy input. The maps highlight regions of increased epistemic uncertainty due to noise.



From these results, it can be observed that the Monte Carlo Dropout method fails to capture a significant increase in uncertainty despite the noisy input, as the uncertainty values remain relatively low across the image. On the other hand, the Deep Ensembles method demonstrates a clear increase in epistemic uncertainty, which is consistent with the noisy nature of the input. A similar increase in uncertainty is also visible with the Hybrid method, combining both approaches.

To refine this analysis further, uncertainty maps were calculated exclusively on the regions where the model's prediction differed from the ground truth. The results are shown below :

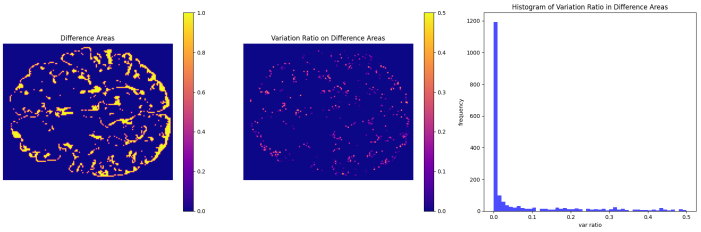


Figure 3.18: Variation Ratio map for Monte Carlo Dropout (MCD) focused on error regions.

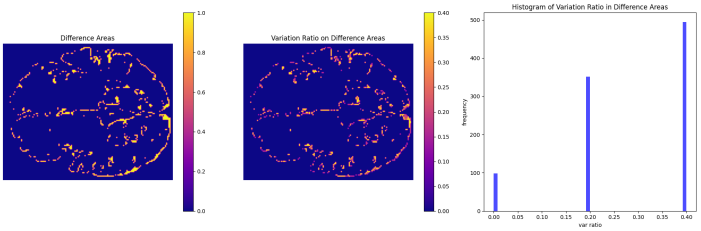


Figure 3.19: Variation Ratio map for Deep Ensembles (DE) focused on error regions.

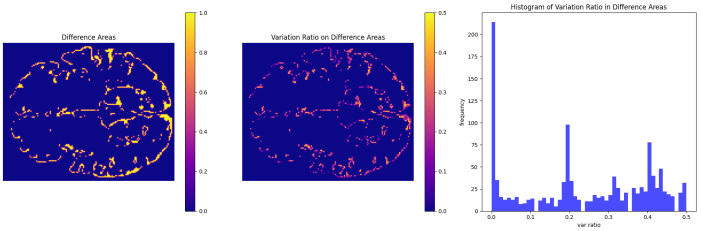


Figure 3.20: Variation Ratio map for Hybrid approach focused on error regions.

Figure 3.21: Comparison of Variation Ratio maps on error regions for the three methods: MCD, DE, and Hybrid.

Examining the error regions reveals that the Monte Carlo Dropout method fails to significantly raise the uncertainty in areas where the model made incorrect predictions. However, the Deep Ensembles method effectively captures increased epistemic uncertainty in these regions, which is also reflected in the distribution histograms of uncertainty values.

These findings highlight that the Monte Carlo Dropout method struggles to accurately quantify uncertainty for segmentation tasks, as it fails to capture a sufficiently broad distribution using a single model with stochastic dropout. In contrast, Deep Ensembles provide a more reliable estimation of uncertainty by leveraging multiple independently trained models with diverse initializations.

Finally, the Hybrid approach, which combines both the Monte Carlo Dropout and Deep Ensembles methods, produces results that are nearly indistinguishable from those obtained using

the Deep Ensembles method alone. This lack of significant difference is expected, as the Hybrid method involves performing stochastic dropout on each model within the ensemble. However, as previously observed, Monte Carlo Dropout alone often fails to introduce sufficient variability between predictions, as the stochastic nature of the dropout layers does not significantly alter the model's decision boundaries in most cases. Consequently, when applied in the Hybrid approach, the additional stochastic sampling does not create enough diversity among the predictions beyond what is already captured by the independently initialized models in the Deep Ensembles. Therefore, while the Hybrid method does combine both techniques, it does not appear to provide a substantial improvement in uncertainty quantification compared to Deep Ensembles alone. To confirm this observation, further experimentation on more challenging datasets with higher variability and noise levels would be necessary to determine whether the Hybrid approach could demonstrate clearer benefits in scenarios where greater model diversity might be required.

Chapter 4

Discussion and Conclusion

4.1 Discussion

4.1.1 Related Work

During the previous semester, my work focused on binary segmentation of multiple brain tissues using the iSEG dataset, with uncertainty quantification mainly based on the observation of the mean and standard deviation of the predictions for each pixel. This semester, I further developed this approach by implementing two more specific metrics : the Variation Ratio (Var-Ratio) to quantify epistemic uncertainty and Predictive Entropy to quantify aleatoric uncertainty, allowing for a more precise evaluation of model confidence.

Additionally, I expanded the application of these methods by first testing them on simpler classification tasks before returning to segmentation. A significant advancement this semester was the implementation of a hybrid method combining Monte Carlo Dropout (MCD) and Deep Ensembles (DE). This new approach was tested on both classification and segmentation tasks to assess its effectiveness and relevance compared to the individual methods.

4.1.2 Future Work

The main avenue for future exploration would be to continue comparing the hybrid method with standard MCD and DE techniques for uncertainty quantification. At this stage, it remains unclear whether the hybrid approach provides a significative advantage or if it is redundant compared to using Deep Ensembles alone. Further investigation would help clarify the specific benefits of combining both methods.

Another interesting direction would be to test these techniques on more complex and diverse datasets. For classification tasks, using a dataset like ImageNet could provide more significant results. For segmentation tasks, the BraTS dataset, which focuses on brain tumor detection, could serve as an excellent benchmark to evaluate the robustness of the proposed methods.

4.2 Conclusion

This report has explored multiple methods for uncertainty quantification applied to both classification and segmentation tasks in Deep Learning. By implementing two prominent state-of-the-art techniques, Monte Carlo Dropout and Deep Ensembles, it has been demonstrated that prediction uncertainty can be effectively measured and analyzed. These methods operate by generating multiple stochastic predictions on the same input data, leading to the construction of probability distributions. From these distributions, key metrics such as Variation Ratio (Var-Ratio) for epistemic uncertainty and predictive entropy for aleatoric uncertainty can be extracted, offering deeper insights into model confidence and reliability.

Additionally, a hybrid approach combining MCD and DE was developed and evaluated in this work, aiming to investigate whether the fusion of these techniques could enhance uncertainty quantification. Preliminary results showed promise, suggesting that the hybrid approach could indeed capture a broader range of uncertainties. However, further research is required to validate these findings and determine if the hybrid method consistently outperforms the standalone approaches.

This study opens several avenues for future work. One potential direction would be to extend the evaluation of the hybrid method to more complex datasets and medical imaging scenarios, such as the BraTS dataset for brain tumor segmentation or ImageNet for large-scale classification tasks. Furthermore, a comparative analysis with other advanced uncertainty quantification techniques could provide a clearer understanding of the strengths and limitations of the methods explored here. Exploring the influence of model architectures and hyperparameter tuning could also offer additional insights.

In conclusion, this report provides a foundation for further research in uncertainty quantification in Deep Learning, particularly within the context of medical imaging. The methods and insights presented here aim to contribute towards developing more reliable and interpretable models for critical applications where uncertainty estimation plays a important role in decision-making.

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] Sorbonne Université, Computer Science Master Données, Apprentissage et Connaissances (DAC) Bayesian Deep Learning, Nicolas Thome https://cord.isir.upmc.fr/teaching-rdfia/

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

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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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