POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Calibration and Uncertainty Visualisation in Medical Image Segmentation

Master's Thesis in
Computer Science and Engineering
Ingegneria Informatica

Author: **Lorenzo Reitani**

Advisor: Prof. Loiacono Daniele
Co-advisors: Ricardo Brioso
Academic Year: 2025-26

Abstract

Deep learning models for medical image segmentation must provide not only accurate predictions but also calibrated confidence and clinically meaningful uncertainty estimates. In full-body CT segmentation of small and low-contrast structures such as the lymphatic system, predictive probabilities are often miscalibrated and voxel-wise uncertainty maps are dominated by trivial boundary effects. This thesis investigates calibration and uncertainty estimation within the nnU-Net v2 framework on 45 annotated CT volumes. Three configurations are compared: a single baseline model, a deep ensemble (5-fold cross-validation), and a checkpoint ensemble based on cyclical learning rates. Confidence calibration is performed via post-hoc temperature scaling, with the temperature parameter estimated by minimizing negative log-likelihood (NLL) on a held-out region of interest. On the test set, temperature scaling consistently reduces miscalibration across all models, with relative improvements up to 41% in NLL and 56% in Expected Calibration Error, without altering segmentation decisions. Voxel-wise uncertainty maps derived from single or ensemble predictions (entropy, variance, mutual information, disagreement and anti-confidence metrics) are quantitatively evaluated as error localization tools through threshold calibration based on Dice overlap with segmentation errors. Within a 15 mm ROI, ensemble-based measures achieve the highest overall error recall (recall_FP $\sim$0.81 and recall_total $\sim$ 0.64), with F1-scores up to 0.44. When excluding a 2 mm boundary band, standard maps show a recall reduction of up to 37%, whereas distance-aware entropy reaches recall_total $\sim$0.56, outperforming conventional measures in non-boundary regions, especially Distance Expected Uncertainty (DEU). Exceedance-Based Contextual Uncertainty (EBCU) further suppresses trivial boundary uncertainty and increases relative recall in the no-border setting by more than 40% compared to the standard ROI. An interactive visualization tool integrates calibrated predictions and thresholded uncertainty maps to support structured qualitative inspection. Overall, the results show that post-hoc calibration improves the quantitative alignment between uncertainty and actual segmentation errors. Voxel-wise uncertainty maps can be used as error detection tool enhancing clinical interpretability without architectural modifications or substantial computational overhead.

Keywords: uncertainty estimation, confidence calibration, nnU-Net, ensemble, temper-

ature scaling

Abstract in lingua italiana

I modelli di deep learning per la segmentazione di immagini mediche devono fornire non solo predizioni accurate, ma anche stime di confidenza calibrate e misure di incertezza clinicamente significative. Nella segmentazione full-body in TC di strutture piccole e a basso contrasto, come il sistema linfatico, le probabilità predittive risultano spesso mal calibrate e le mappe di incertezza voxel-wise sono dominate da effetti banali di bordo. Questa tesi indaga la calibrazione e la stima dell'incertezza all'interno del framework nnU-Net v2 su 45 volumi TC annotati. Vengono confrontate tre configurazioni: un modello base singolo, un deep ensemble (5-fold cross-validation) e un checkpoint ensemble basato su learning rate ciclici. La calibrazione della confidenza è effettuata tramite temperature scaling post-hoc, con il parametro di temperatura stimato minimizzando la negative log-likelihood (NLL) su una regione di interesse separata. Sul test set, il temperature scaling riduce in modo consistente la miscalibrazione in tutti i modelli, con miglioramenti relativi fino al 41% in termini di NLL e fino al 56% in termini di Expected Calibration Error, senza alterare le decisioni di segmentazione. Le mappe di incertezza voxel-wise derivate da predizioni singole o ensemble (entropia, varianza, mutual information, disaccordo e metriche di anti-confidence) sono valutate quantitativamente come strumenti di localizzazione dell'errore attraverso una calibrazione della soglia basata sull'overlap di Dice con gli errori di segmentazione. All'interno di una ROI di 15 mm, le misure basate su ensemble raggiungono il valore più elevato di richiamo complessivo dell'errore (recall_FP circa 0,81 e recall_total circa 0,64), con F1-score fino a 0,44. Escludendo una banda di 2 mm dal bordo, le mappe standard mostrano una riduzione del recall fino al 37%, mentre l'entropia distance-aware raggiunge un recall_total di circa 0,56, superando le misure convenzionali nelle regioni non di bordo, in particolare la Distance Expected Uncertainty (DEU). La Exceedance-Based Contextual Uncertainty (EBCU) riduce ulteriormente l'incertezza banale di bordo e aumenta il richiamo relativo nel setting senza bordo di oltre il 40% rispetto alla ROI standard. Uno strumento di visualizzazione interattivo integra predizioni calibrate e mappe di incertezza sogliate per supportare un'ispezione qualitativa strutturata. Nel complesso, i risultati mostrano che la calibrazione post-hoc migliora l'allineamento quantitativo tra incertezza e reali errori di segmentazione. Le

mappe di incertezza voxel-wise possono essere utilizzate come strumenti di rilevazione dell'errore, migliorando l'interpretabilità clinica senza modifiche architetturali o un significativo aumento del costo computazionale.

Parole chiave: stima dell'incertezza, calibrazione della confidenza, nnU-Net, deep ensemble, checkpoint ensemble, temperature scaling

Contents

1 | Introduction

Deep learning has significantly advanced medical image analysis, enabling accurate automated segmentation of complex anatomical structures [34]. However, most models provide deterministic predictions without indicating how reliable these predictions are, which is problematic in safety-critical domains such as healthcare. At the same time, clinicians face an increasing workload, with radiologists and radiation oncologists required to inspect large volumes of imaging data under strict time constraints. In this context, tools that can both automate part of the segmentation process and highlight potentially unreliable regions are essential to support expert decision-making rather than replace it.

This thesis focuses on uncertainty-aware medical image segmentation as a way to make deep learning models more transparent and clinically useful, with the aim of easing clinicians' workload while maintaining human control over final decisions.

1.1. Motivation and Clinical Context

The lymphatic system is clinically critical for cancer staging and immune function assessment, yet its segmentation from full-body CT scans is challenging due to small lymph nodes with poor contrast and high inter-observer variability. Moreover, clinicians face growing workload pressures, spending excessive time on manual delineations that compete with other diagnostic tasks.

This thesis is motivated by the need for uncertainty-aware segmentation tools that reduce clinicians' burden. By providing confidence maps and interactive visualisations, the proposed approach enables experts to focus on potentially unreliable regions rather than inspecting entire volumes, thus supporting efficient clinical workflows.

1.2. Research Objectives

The overarching goal of this thesis is to investigate, implement, and evaluate techniques for uncertainty estimation and confidence calibration in deep learning-based medical image

segmentation, with a particular focus on making uncertainty maps clinically actionable. Specifically, the work pursues the following objectives:

1. **Confidence calibration via temperature scaling.** A region-of-interest-aware temperature scaling procedure is implemented to correct the systematic overconfidence of neural network predictions. The optimal temperature is estimated by minimising the Negative Log-Likelihood (NLL) on a held-out calibration set restricted to a clinically relevant region around the segmentation boundary.

2. **Voxel-wise uncertainty estimation.** Several complementary strategies are integrated within the nnU-Net v2 pipeline, including Deep Ensembles, Checkpoint Ensembles with cyclical learning rate schedules. A comprehensive set of uncertainty maps is derived from base model and ensemble predictions, including both base measures (entropy, ensemble variance, mutual information, boundary disagreement, anti-confidence) and novel distance-aware measures (DEU and EBCU) that explicitly account for the expected relationship between uncertainty and distance from the segmentation boundary.

3. **Threshold calibration for error localisation.** A dedicated calibration procedure is developed to convert continuous uncertainty maps into local maps that optimally localise actual segmentation errors, bridging the gap between abstract uncertainty values and clinical decision support.

4. **Evaluation on standard and no-border ROI settings.** The quality of uncertainty maps as error detection tools is assessed within clinically meaningful regions of interest, with an additional no-border evaluation that excludes the trivially uncertain boundary band to provide a more stringent assessment of distance-aware measures.

5. **Interactive visualisation tool.** A Streamlit-based interactive viewer is developed to allow real-time inspection of segmentation masks, uncertainty heatmaps, and ground truth overlays, supporting slice-by-slice navigation and customisable uncertainty filtering.

1.3. Structure of the Thesis

The thesis is organised as follows:

Chapter 2 – Background introduces the theoretical foundations, including medical image segmentation, the U-Net and nnU-Net architectures, the concepts of aleatoric

and epistemic uncertainty, model calibration, and the main methods for uncertainty quantification in deep learning (Deep Ensembles, Monte Carlo Dropout, Checkpoint Ensembles, Test-Time Augmentation).

Chapter 3 – Dataset and Models describes the private dataset of 45 full-body CT scans used in the experiments, including volume characteristics, voxel spacing heterogeneity, and class imbalance analysis. It also details the training configurations for the Deep Ensemble and Checkpoint Ensemble within the nnU-Net v2 framework.

Chapter 4 – Temperature Scaling for Confidence Calibration presents the region-of-interest-aware temperature scaling procedure, including the calibration set design, the two-stage grid search optimisation, and the evaluation of calibration improvements across all model configurations.

Chapter 5 – Uncertainty Estimation and Visualisation covers the computation of voxel-wise uncertainty maps, introduces the distance-aware uncertainty measures (DEU and EBCU), describes the threshold calibration procedure, and presents the evaluation of uncertainty maps as error detection tools under both standard and no-border ROI settings.

Chapter 6 – Interactive Viewer describes the Streamlit-based application developed for real-time visualisation and exploration of segmentation results and uncertainty maps.

Chapter 7 – Conclusions and Future Developments summarises the main findings, discusses the limitations of the current study, and outlines possible directions for future research.

2 | Background

This chapter provides the theoretical foundation needed to understand the methods and contributions presented in the following chapters. It first introduces medical image segmentation[29], emphasizing its clinical relevance and the practical challenges posed by anatomical variability, limited annotations, and imaging artifacts. It then reviews the U-Net architecture, a widely adopted backbone in biomedical segmentation thanks to its encoder–decoder design and skip connections [30].

In the second part of the chapter, we focus on the notion of predictive confidence and model calibration in deep learning[35]. We discuss why modern neural networks can produce accurate yet miscalibrated probability estimates, and we introduce the main tools used to assess calibration.

We then turn to uncertainty in machine learning[9], distinguishing between epistemic uncertainty (stemming from limited knowledge or model uncertainty) and aleatoric uncertainty (arising from irreducible noise in the data). We will see how uncertainty can be formalized through predictive distributions rather than point estimates, and why uncertainty-aware systems are key for trustworthy deployment under ambiguity and domain shift.

Finally, we introduce strategies to generate and visualize uncertainty maps for segmentation outputs. Beyond standard entropy-based representations, we outline additional approaches for producing spatial uncertainty estimates.

2.1. Medical Image Segmentation

Medical image segmentation is a core task in medical image analysis and often a prerequisite for diagnosis support, quantitative measurements, and treatment planning. It consists of partitioning an image into meaningful regions by delineating anatomical structures or pathological findings (e.g., organs, tumors, lesions) in modalities such as CT, MRI, or X-ray [3, 32].

Before deep learning, automated segmentation pipelines were commonly based on classical

image processing techniques, including thresholding, edge detection, and morphological operators, which are typically easy to interpret but tend to rely on hand-crafted rules and expert-tuned parameters. In practice, these approaches often struggle to scale and to remain robust when confronted with the heterogeneity of modern clinical data (different scanners, protocols, resolutions, and patient variability), especially in high-resolution 3D settings[26].

To overcome these limitations, deep learning has become the dominant paradigm for medical image segmentation. In particular, Convolutional Neural Networks (CNNs) can learn hierarchical feature representations directly from data and have shown strong performance across a wide range of segmentation problems. Among CNN-based models, U-Net has emerged as the de facto standard architecture in biomedical segmentation: its encoder–decoder design, coupled with skip connections, supports accurate localization while still capturing high-level semantic context[38].

2.1.1. U-Net Architecture for Biomedical Image Segmentation

One of the most widely used architectures for biomedical image segmentation is U-Net, originally introduced in 2015 by Ronneberger et al. [30]. It was designed to meet typical requirements of medical imaging applications, where annotations are often scarce and segmentation demands fine-grained spatial accuracy. Unlike classification-oriented CNNs, U-Net directly targets dense prediction by assigning a semantic label to each pixel (or voxel in 3D variants)

Architecture overview

U-Net follows a symmetric encoder–decoder design composed of a contracting path (encoder) and an expansive path (decoder). The encoder extracts increasingly abstract features through repeated convolutional blocks and downsampling operations, progressively reducing spatial resolution while increasing the number of channels. The decoder then reconstructs a high-resolution prediction via upsampling, recovering localization capabilities that would otherwise be lost during downsampling.

A defining element of U-Net is the presence of skip connections between corresponding encoder and decoder stages. These connections concatenate high-resolution feature maps from the encoder to the decoder representations at the same scale, enabling the network to combine global context with precise spatial detail. The final prediction is produced by a 1x1 convolution that maps the decoder features to the desired number of segmentation classes.

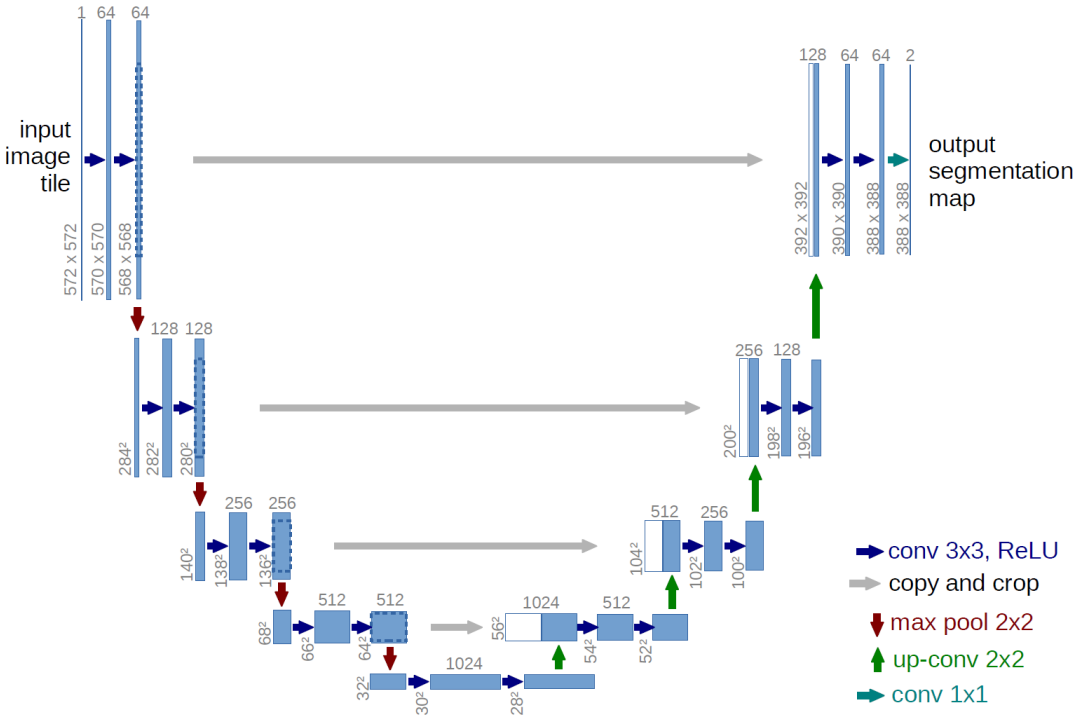

Figure 2.1: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Picture taken from [30].

Training Strategy and Loss Function

To mitigate the limited availability of labeled biomedical data, the original U-Net formulation relied heavily on data augmentation, including elastic deformations to simulate plausible anatomical variability. The network is trained end-to-end using stochastic gradient descent (SGD) with a cross-entropy loss applied pixel-wise. The authors also propose a boundary-aware weighting scheme: a precomputed per-pixel weight map emphasizes difficult regions such as borders between adjacent objects, encouraging better separation in challenging cases.

2.2. nnU-Net: A Self-Adapting Framework for Medical Image Segmentation

While U-Net-based models have become the standard choice for medical image segmentation, achieving strong performance on a new dataset has traditionally required substantial

manual engineering. In practice, researchers often need to tune multiple coupled design decisions, such as preprocessing, network topology, training hyperparameters, inference strategy, and post-processing, leading to many task-specific pipelines and making results harder to reproduce and compare across datasets.

To address this issue, Isensee et al.[11] introduced nnU-Net ("no-new-Net"), a self-adapting framework that deliberately builds on relatively standard U-Net variants, focusing less on architectural novelty and more on a rigorous, automated configuration of the full segmentation pipeline. The key insight is that many reported gains in medical segmentation can be explained by better configuration rather than by complex architectural changes, and that these configuration choices are strongly interdependent.

Self-configuration and model variants

Rather than being a single fixed network, nnU-Net is a framework that automatically instantiates and configures a small set of U-Net-based models depending on dataset properties. In its original formulation, the framework considers three main options: a 2D U-Net, a full-resolution 3D U-Net, and a cascaded 3D approach that performs coarse-to-fine refinement. The 2D variant can be beneficial for strongly anisotropic volumes (e.g., thick-slice acquisitions), while the 3D variant is generally preferred for volumetric data when sufficient context can be captured within feasible patch sizes; for very large volumes, the cascade helps recover global context by first predicting at lower resolution and then refining at full resolution.

A central component of the framework is the dynamic adaptation of network topology to each dataset. This includes automatically selecting patch size, batch size, and the number of pooling operations per axis, trading off model capacity and batch size under GPU memory constraints while ensuring adequate receptive field coverage. In other words, nnU-Net does not assume a "one-size-fits-all" configuration: it adapts to differences in image geometry and spacing across datasets.

End-to-end pipeline: preprocessing, training, inference, post-processing

Beyond model selection, nnU-Net standardizes the end-to-end pipeline in a fully automatic way. In preprocessing, it performs operations such as cropping to the non-zero region, resampling all cases to a target spacing, and applying modality-aware intensity normalization. For example, CT images can be normalized using dataset-level intensity statistics, while other modalities are commonly normalized on a per-case basis. These

steps aim to reduce variability introduced by acquisition protocols and voxel spacing differences, allowing the network to learn spatial semantics more consistently.

Training is also defined as part of the framework "recipe". The original nnU-Net uses a compound objective combining Dice and cross-entropy losses, and relies on strong on-the-fly data augmentation (e.g., geometric transforms and intensity augmentations) to mitigate overfitting on limited medical datasets. It also includes practical sampling strategies to ensure that a sufficient fraction of training patches contains foreground structures, improving stability when classes are small or sparse.

At inference time, nnU-Net applies patch-based (sliding-window) prediction to handle full 3D volumes, and aggregates overlapping patches in a way that reduces border artifacts. It can further employ test-time augmentation and ensembling across multiple models trained in cross-validation, which typically improves robustness. Finally, nnU-Net may apply lightweight dataset-dependent post-processing rules learned from the training labels

From nnU-Net to nnU-Net v2

In this thesis we adopt nnU-Net v2, which preserves the same overall philosophy but provides an updated framework and baselines consistent with modern validation practices. Recent work revisiting nnU-Net emphasizes that, under rigorous benchmarking, well-configured CNN-based U-Net variants within the nnU-Net framework remain highly competitive, and that many claimed improvements from newer architectural families do not necessarily hold when evaluated with strong baselines and fair computational budgets [8, 12].

In summary, nnU-Net (and its v2 evolution) can be seen as a practical "meta-approach" to medical image segmentation: instead of requiring manual dataset-by-dataset engineering, it operationalizes a set of empirically grounded best practices and applies them automatically, enabling reproducible and competitive segmentation pipelines across diverse tasks and imaging modalities.

2.3. Calibration of Predictive Confidence in Deep Learning

In machine learning, especially with deep neural networks, calibration describes how well a model's reported confidence matches the true probability that its prediction is correct [6]. Even when contemporary deep models achieve strong accuracy, their probability outputs are often miscalibrated, meaning that a prediction made with 90% confidence

may not actually be correct about 90% of the time. This gap matters in high-stakes settings such as healthcare, where clinical actions may depend directly on the numerical value of predicted probabilities rather than only on the final class label.

Calibration is also closely related to uncertainty estimation. Calibration does not aim to produce uncertainty, but rather to assess whether the model's confidence (or, equivalently, its implied uncertainty) can be interpreted as a trustworthy probabilistic statement about correctness. In other words, uncertainty methods may indicate how much the model hesitates, due to irreducible noise in the data or limited knowledge captured by the model, while calibration checks whether these confidence levels align with observed reliability in practice[22, 24].

2.3.1. Understanding Calibration

To build intuition for calibration, imagine a classifier that assigns 0.8 confidence to a subset of its predictions and ends up being correct in roughly 80% of those cases. In this ideal situation, the model is perfectly calibrated at the 0.8 confidence level; if, instead, predictions made with 0.8 confidence are correct only 60% of the time, the model is overconfident.

More formally, let a model output a confidence value $p(x)$ for an input x. The model is said to be calibrated if, for every confidence level $p \in [0, 1]$, the conditional probability of a correct prediction matches that confidence:

$$\mathbb{P}\big(Y = \hat{Y} \mid p(x) = p\big) = p. \tag{2.1}$$

The same definition naturally extends to segmentation tasks, where each pixel or voxel is associated with a predicted probability and can be evaluated under the same reliability-based interpretation.

2.3.2. Measuring Calibration

Several quantitative metrics are typically employed to assess how well predictive probabilities are calibrated, each capturing slightly different aspects of probabilistic quality. In this work, we focus on the Expected Calibration Error (ECE), the Brier Score (BS), and the Negative Log-Likelihood (NLL), which together allow us to evaluate both the alignment between confidence and accuracy and the overall usefulness of the predicted probability distributions.

Expected Calibration Error (ECE)

The Expected Calibration Error (ECE) summarizes the discrepancy between predicted confidence and empirical accuracy over a set of predictions[27]. To compute ECE, predictions are partitioned into M disjoint bins according to their confidence scores (for example, $[0, 0.1)$, $[0.1, 0.2)$, ..., $[0.9, 1]$). For each bin, the average predicted confidence is compared with the empirical accuracy.

Formally, let B_m denote the set of indices of predictions whose confidence values fall into bin m, with cardinality $|B_m|$, and let n be the total number of samples. The Expected Calibration Error is defined as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{2.2}$$

where $\mathrm{acc}(B_m)$ is the fraction of correct predictions in bin m, and $\mathrm{conf}(B_m)$ is the mean predicted confidence of the samples in that bin.

A perfectly calibrated model satisfies $\mathrm{acc}(B_m) \approx \mathrm{conf}(B_m)$ for all bins, resulting in an ECE close to zero. Systematic overconfidence or underconfidence leads to larger deviations and thus higher ECE values. In segmentation tasks, the same definition can be applied voxel-wise by treating each voxel prediction as an individual sample, making ECE sensitive to voxel-level miscalibration while still aggregating errors at the bin level.

Brier Score (BS)

The Brier Score provides a complementary perspective by directly measuring the mean squared error between predicted probabilities and the corresponding ground-truth labels [31]. For binary outcomes, and analogously for each voxel in a segmentation map, it is defined as

$$\text{Brier Score} = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_i - y_i)^2, \tag{2.3}$$

where $\hat{p}_i \in [0, 1]$ denotes the predicted probability for the positive class (e.g., the softmax probability assigned to a given label at voxel i), and $y_i \in \{0, 1\}$ is the corresponding ground-truth label. Lower Brier Score values indicate predictions that are, on average, both accurate and well aligned with the true outcomes, whereas larger squared deviations reflect either incorrect classifications or poorly calibrated confidence estimates.

A key difference from ECE is that the Brier Score does not rely on binning and operates at the level of individual predictions. As a result, it mixes two desirable properties into a

single scalar: calibration (probabilities should match empirical frequencies) and resolution (probabilities should move away from 0.5 and concentrate near 0 or 1 when the model is confident and correct). This leads to interesting edge cases: for example, a model that always predicts 0.5 probabilities on a perfectly balanced binary dataset can exhibit a relatively low ECE, because its average confidence matches its average accuracy within each bin, yet still achieve a comparatively high Brier Score due to its lack of discriminative power. In contrast, a model that is both well calibrated and confident where appropriate will tend to obtain low values for both ECE and Brier Score, indicating that it not only aligns probabilities with frequencies but also exploits the full range of the probability scale.

In segmentation, computing the Brier Score voxel-wise makes it particularly suitable for fine-grained analysis. Each miscalibrated voxel is penalized proportionally to the square of its error, making the metric sensitive to overconfident wrong predictions (probability close to 1 for an incorrect class) as well as underconfident correct predictions (probability close to 0.5 or lower for a correct class). This property complements the bin-aggregated view provided by ECE and helps capture local calibration issues that might be smoothed out when aggregating over confidence bins.

Negative Log Likelihood (NLL)

Negative Log-Likelihood (NLL), often referred to as log loss and closely related to cross-entropy, measures how much probability a model assigns to the true outcome. In the multiclass setting, let the predicted class probabilities for sample i be denoted by $\mathbf{p}_i = (\hat{p}_{i,1}, \ldots, \hat{p}_{i,K})$, and let the corresponding ground-truth class label be $y_i \in \{1, \ldots, K\}$. The Negative Log-Likelihood over n samples is then defined as

$$\mathrm{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{p}_{i,y_i}) \,. \tag{2.4}$$

This formulation makes the behavior of NLL intuitive: correct predictions assigned high probability incur a small penalty, whereas incorrect or low-confidence predictions result in a large loss.

Equivalently, using one-hot encoded target vectors $\mathbf{y}_i$, where $y_{i,k} = 1$ if and only if $k = y_i$ and $y_{i,k} = 0$ otherwise, the same quantity can be written as

$$\mathrm{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} y_{i,k} \, \log(\hat{p}_{i,k}) \,. \tag{2.5}$$

This expression highlights that only the log-probability of the true class contributes to the loss, as the one-hot encoding effectively selects the corresponding term in the summation.

NLL is considered calibration-sensitive because it strongly penalizes overconfident mistakes: assigning a near-zero probability to the true class yields a very large contribution due to the logarithm.

2.4. Uncertainty

Uncertainty quantification is increasingly regarded as a key requirement for deploying deep learning systems in clinical environments, where automated predictions may influence diagnosis, treatment planning, and follow-up decisions[2]. In medical image segmentation, a voxel-wise prediction can be accurate on average while still being unreliable in specific regions (e.g., weak boundaries or rare patterns), and uncertainty estimates help identify such cases in a principled way. In the machine learning literature, uncertainty is commonly divided into aleatoric (data) uncertainty and epistemic (knowledge) uncertainty, which capture complementary failure modes of a model.

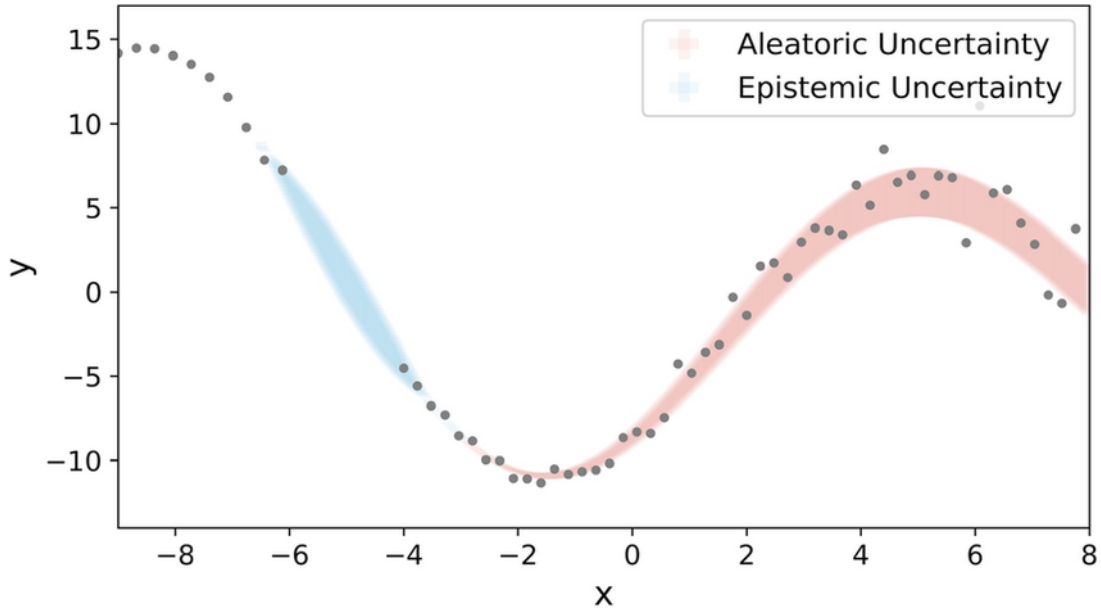

Figure 2.2: An illustration of the difference between aleatoric and epistemic uncertainties. The dots on the plot represent the available data points. Aleatoric uncertainty captures varying degrees of inherent noise in the data, while epistemic uncertainty reflects the ignorance gap due to a lack of data. Picture taken from [37].

2.4.1. Aleatoric Uncertainty

Aleatoric uncertainty, also known as data uncertainty, arises from intrinsic noise and ambiguity in the observed data and therefore cannot be removed simply by collecting more training samples of the same type. In segmentation, it typically concentrates in regions where the image does not provide sufficient evidence to unambiguously assign a class, such as low-contrast interfaces or partially visible structures[21].

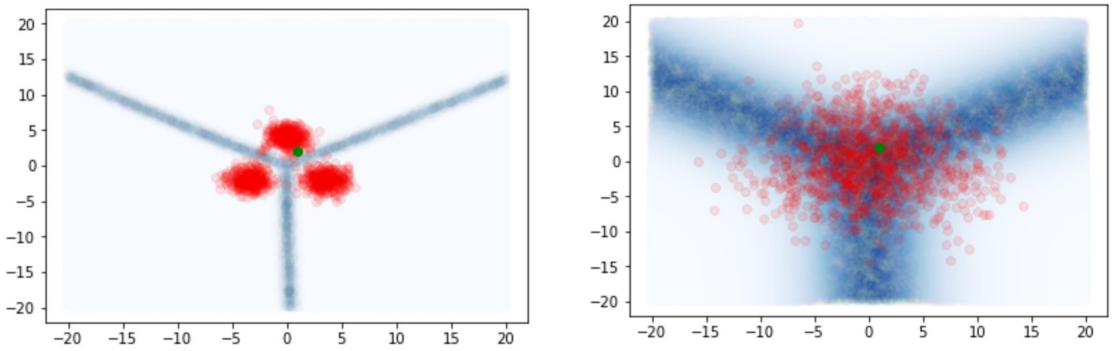

Figure 2.3: In the left example, the classes form compact and well-separated clusters, so a new sample (green point) can be assigned confidently; this corresponds to low aleatoric uncertainty, since the observation is not intrinsically ambiguous. In the right example, the same classes become more diffuse and overlapping because the observations are corrupted by noise; even with a perfect model, the green point remains difficult to attribute to a single class, illustrating high aleatoric uncertainty as ambiguity that is inherent to the data distribution. pictures taken from[28].

A useful distinction, especially in regression settings, is between homoscedastic and heteroscedastic aleatoric uncertainty.
Homoscedastic uncertainty is input-independent: the noise level is assumed approximately constant across the input space (e.g., a uniform measurement noise floor).
Heteroscedastic uncertainty is input-dependent: the noise magnitude varies with the input (e.g., higher uncertainty near ambiguous boundaries and lower uncertainty in clearly defined regions).

Common sources of aleatoric uncertainty in medical imaging include measurement noise and artifacts, ambiguity caused by poor contrast, and variability introduced by the acquisition process. In addition, segmentation ground truth itself may be uncertain due to inter-observer variability, where different experts provide slightly different delineations for the same anatomical boundary, reflecting genuine ambiguity rather than model failure.

2.4.2. Epistemic Uncertainty

Epistemic uncertainty, also known as knowledge uncertainty, reflects limitations of the model caused by insufficient training evidence, limited coverage of the data distribution, or inadequate representation capacity. Unlike aleatoric uncertainty, epistemic uncertainty is, at least in principle, reducible: it can decrease when the model is trained on more diverse and representative data, or when training and modeling choices are improved.

In practice, epistemic uncertainty is often associated with three overlapping situations. First, data sparsity: if certain anatomical configurations or classes are rare, the model may not learn robust decision rules for them and will exhibit higher uncertainty when encountering similar cases. Second, out-of-distribution (OOD) inputs: the model receives inputs that deviate from the training distribution, making predictions unreliable because the model has not learned how to interpret these patterns.

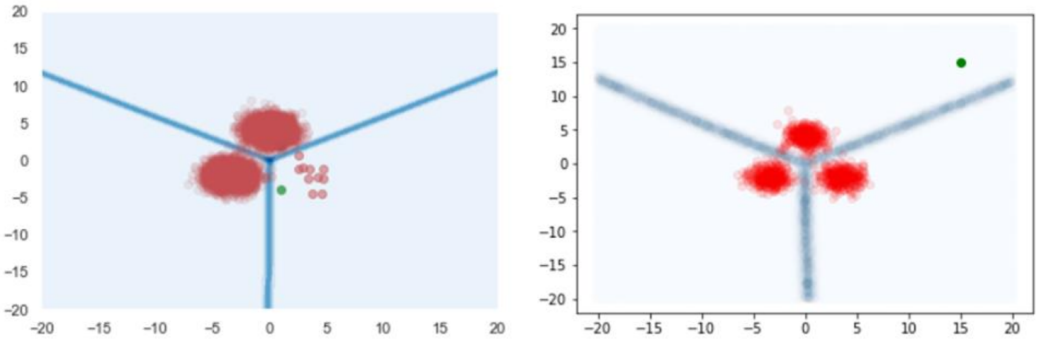

Figure 2.4: Left panel (data sparsity). The green sample lies in a region where the training data provides little or no coverage for the nearby class structure, so the model lacks evidence to learn a reliable decision rule in that area; this corresponds to high epistemic uncertainty due to data sparsity. Right panel (out-of-distribution). The green sample is far from the support of the training distribution (i.e., it does not resemble any of the observed clusters), so it should be treated as out-of-distribution (OOD); this also induces high epistemic uncertainty, since the model is asked to extrapolate beyond what it has learned. Pictures taken from [28].

Third, limited training coverage, which includes the broader phenomenon of domain shift, a change between training and deployment conditions (e.g., different scanners or protocols) that can be interpreted as a form of knowledge uncertainty because the model is exposed to variations it has not learned to handle [16].

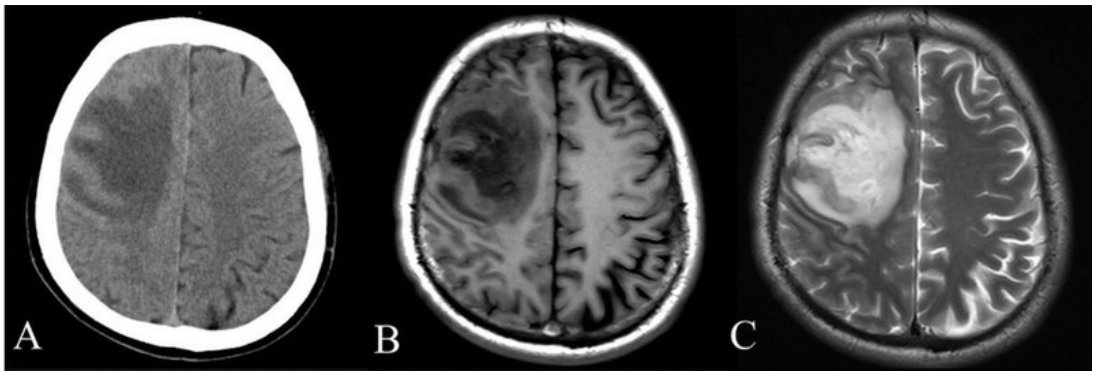

Figure 2.5: A: CT scan of the brain showing hypodensity of the right frontal lobe of the brain. B: T1-weighted MRI scan of the same brain showing hypointensity of the lesion. C: T2-weighted MRI scan of the same brain showing hyperintensity of the lesion. Pictures taken from [10].

From a clinical perspective, epistemic uncertainty is particularly important because it can be interpreted as a warning signal that the model may be operating outside its competence. High epistemic uncertainty can suggest practical actions such as: prioritizing expert review of the corresponding regions/cases; collecting additional annotations for underrepresented patterns; enriching the training set to improve coverage; or investigating whether the data pipeline introduces a shift (e.g., acquisition parameter changes) that should be addressed.

2.5. Methods for Uncertainty Quantification in Deep Learning for Medical Image Segmentation

Reliable deployment of deep learning models for medical image segmentation requires not only accurate predictions, but also an estimate of how trustworthy those predictions are at each spatial location. Uncertainty quantification (UQ) addresses this need by moving beyond point estimates and framing segmentation as a probabilistic prediction problem, where the model aims to characterize a distribution over plausible outputs rather than a single mask. In this perspective, uncertainty is not an auxiliary heuristic but a quantity derived from (an approximation of) the predictive distribution $p(y|x)$, from which voxel-wise uncertainty maps can be computed using measures such as variance or entropy.

A useful way to organize UQ methods is to separate the ideal probabilistic formulation from the practical strategies used in modern segmentation systems. Section 2.6.1 introduces uncertainty as predictive distributions and presents Bayesian Neural Networks

(BNNs) as a theoretical reference, where uncertainty arises naturally by marginalizing predictions over a posterior distribution of model parameters. However, exact Bayesian inference in deep networks is generally intractable, and the posterior predictive integral cannot be evaluated in closed form for realistic architectures, motivating approximate methods.

Section 2.6.2 then focuses on practical approximations that estimate predictive uncertainty through repeated evaluations of a model (or multiple models) and the analysis of the resulting variability. This includes Monte Carlo Dropout, Deep Ensembles, checkpoint-based ensembles, and probabilistic latent-variable segmentation models such as the Probabilistic U-Net, which produce uncertainty estimates by sampling multiple plausible predictions for the same input. Finally, we discuss test-time augmentation (TTA) as an input-driven, non-Bayesian source of variability that can still be leveraged to build uncertainty maps by probing the stability of predictions under plausible transformations of the image acquisition and preprocessing pipeline.

2.5.1. Uncertainty as Predictive Distributions

Uncertainty in segmentation can be framed through the posterior predictive distribution $p(y|x, D)$, which represents the range of plausible masks for an input image given the training data. From this predictive distribution, voxel-wise uncertainty maps are obtained by applying summary measures (e.g., entropy or variance) to its class marginals or to samples drawn from it.

Bayesian Neural Networks as the theoretical ideal

Bayesian Neural Networks (BNNs) provide a coherent probabilistic framework in which uncertainty arises naturally from uncertainty in the model parameters. Instead of learning a single point estimate of the network weights θ, BNNs place a prior distribution $p(\theta)$ over the parameters and update it after observing a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, yielding a posterior distribution:

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}. \tag{2.6}$$

Here, $p(D \mid \theta)$ denotes the likelihood induced by the network and the chosen loss function (e.g., a categorical likelihood in segmentation), while

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta \tag{2.7}$$

is the marginal likelihood (or evidence), which acts as a normalizing constant.

The central object for Bayesian prediction is the posterior predictive distribution. Given a new input image x, predictions are obtained by marginalizing the network output over the posterior uncertainty in the parameters:

$$p(y \mid x, D) = \int p(y \mid x, \theta)\, p(\theta \mid D)\, d\theta. \tag{2.8}$$

This formulation makes explicit why BNNs are often regarded as the theoretical ideal for uncertainty quantification. If the posterior $p(\theta \mid D)$ could be computed exactly and the integral in (2.8) evaluated in closed form, the resulting predictive distribution would jointly account for (i) ambiguity in the input–output mapping, captured by $p(y \mid x, \theta)$, and (ii) uncertainty due to limited knowledge of the model parameters, encoded in $p(\theta \mid D)$ [1].

In medical image segmentation, this perspective is particularly appealing. A single image may admit multiple plausible delineations, for instance due to weak anatomical boundaries or inter-observer variability, while epistemic effects such as domain shift or rare anatomical configurations manifest as increased uncertainty over the model parameters. Through Bayesian marginalization, such parameter uncertainty propagates directly to broader and less confident predictive distributions.

From a practical standpoint, BNNs also clarify the distinction between confidence and uncertainty. The softmax output of a single deterministic network corresponds to a conditional distribution of the form $p(y \mid x, \theta^*)$ for one specific parameter setting θ^*. Bayesian prediction replaces this with the posterior predictive distribution $p(y \mid x, D)$, which can be less sharply peaked when different plausible parameter configurations disagree. This distinction is central in segmentation tasks, where calibrated probability maps are desirable but do not automatically imply epistemic awareness; the Bayesian formulation explicitly encodes parameter uncertainty as an additional and conceptually distinct source of predictive uncertainty.

Intractability of the posterior predictive

Despite its conceptual clarity, the Bayesian Neural Network framework is difficult to realize exactly for modern deep segmentation models. The main limitation is that both the posterior distribution over the parameters, $p(\theta \mid D)$, and the corresponding posterior predictive distribution are generally intractable in high-dimensional neural networks. More specifically, several sources of intractability arise.

First, the posterior over the weights is not available in closed form. For deep networks, the likelihood surface induced by the data is highly non-linear and typically multi-modal, while the dimensionality of the parameter space can easily reach millions of parameters. Under these conditions, an exact Bayesian update cannot be derived analytically, except for highly simplified models.

Second, the marginal likelihood (or evidence) is itself intractable. Computing

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta \tag{2.9}$$

requires integration over the entire parameter space. At the scale of modern deep learning architectures, this integral is computationally prohibitive and cannot be evaluated exactly.

Third, Bayesian prediction requires marginalization over the posterior distribution of the parameters. Even assuming that an explicit expression for $p(\theta \mid D)$ were available, evaluating the posterior predictive distribution

$$p(y \mid x, D) = \int p(y \mid x, \theta)\, p(\theta \mid D)\, d\theta \tag{2.10}$$

still involves a high-dimensional integral. For realistic segmentation networks with dense, voxel-wise outputs, this integral cannot be computed in closed form.

As a consequence, practical Bayesian deep learning methods rely on approximations. The literature commonly distinguishes between several broad families. Variational methods approximate the true posterior with a tractable distribution $q(\theta)$ by optimizing a surrogate objective. Sampling-based approaches, such as Markov Chain Monte Carlo, aim to draw samples from $p(\theta \mid D)$ but typically suffer from poor scalability in deep models. Finally, a class of methods reinterpret standard deep learning components as approximate Bayesian inference mechanisms, such as dropout-based approximations, sacrificing strict theoretical guarantees in favor of computational feasibility.

Across all these approaches, the operational goal remains the same: to approximate the posterior predictive distribution by replacing the intractable integral with a Monte Carlo estimate. In practice, this amounts to computing predictions under multiple realizations of the model parameters or latent variables and aggregating them. This perspective directly motivates the techniques discussed in Section 2.6.2, which can be interpreted as practical strategies for approximating $p(y \mid x, D)$ through repeated predictions and analysis of their disagreement, rather than through exact Bayesian computation.

2.5.2. Practical Approximations of Predictive Uncertainty

In this section, we review a set of practical methods that estimate predictive uncertainty without performing exact Bayesian inference. The common idea is to approximate the posterior predictive $p(y|x, D)$ through an empirical average of multiple predictions obtained for the same input, and to quantify uncertainty by measuring their variability (e.g., via variance- or entropy-based measures).

The methods mainly differ in the source of variability: stochasticity in the model/weights (MC Dropout, Deep Ensembles, Checkpoint Ensembles), sampling in a latent space (Probabilistic U-Net, via samples of z), or controlled perturbations of the input (Test-Time Augmentation), which captures input-driven stability rather than being Bayesian in a strict sense.

Monte Carlo Dropout

Monte Carlo Dropout (MC Dropout) is a post-hoc uncertainty quantification method that leverages dropout to estimate model uncertainty without requiring retraining or architectural modifications. Originally proposed by Gal and Ghahramani [4], MC Dropout keeps dropout active during inference and performs multiple stochastic forward passes, each with a different dropout mask, to approximate model disagreement. Although MC Dropout has theoretical connections to Bayesian inference under specific assumptions, it is primarily employed here as a pragmatic post-hoc technique applicable to any model trained with dropout, irrespective of whether uncertainty estimation was explicitly considered during training.

Method Overview During standard training, dropout randomly deactivates neurons with probability p to prevent co-adaptation and reduce overfitting. At test time, dropout is typically disabled to obtain deterministic predictions from the full network. MC Dropout reverses this practice by keeping dropout active during inference and performing multiple stochastic forward passes, each corresponding to a different randomly sampled dropout mask. Each forward pass can be interpreted as evaluating a different effective subnetwork, thereby inducing model-like variability without maintaining multiple independently trained models.

Formally, given a test input x^*, MC Dropout performs T stochastic forward passes:

$$\left\{ \hat{f}_{w_1}(x^*),\ \hat{f}_{w_2}(x^*),\ \ldots,\ \hat{f}_{w_T}(x^*) \right\}, \tag{2.11}$$

where each $\hat{f}_{w_t}$ denotes the network output under a different dropout realization w_t.

Uncertainty Estimation The ensemble mean prediction is approximated as

$$\mathbb{E}[y^*] \approx \frac{1}{T} \sum_{t=1}^{T} \hat{f}_{w_t}(x^*), \tag{2.12}$$

while the predictive variance is estimated as

$$\mathrm{Var}[y^*] \approx \frac{1}{T} \sum_{t=1}^{T} \hat{f}_{w_t}(x^*)^2 - \left(\mathbb{E}[y^*]\right)^2 . \tag{2.13}$$

In segmentation tasks, voxel-wise variance computed across the T stochastic predictions serves as a proxy for epistemic uncertainty. High variance indicates that different dropout-masked network configurations produce divergent predictions, highlighting regions where the model is uncertain.

Theoretical Grounding and Practical Limitations MC Dropout can be interpreted as an approximate Bayesian neural network under specific theoretical conditions, such as particular choices of dropout placement and weight regularization. In practice, however, these conditions are rarely met, as dropout is typically introduced as a heuristic for regularization rather than as a principled mechanism for posterior inference. Consequently, MC Dropout is best viewed as a cost-effective ensemble approximation rather than a fully Bayesian method. The quality of the resulting uncertainty estimates depends strongly on the chosen dropout rate and the number of stochastic forward passes T, both of which generally require empirical tuning. Moreover, the dropout rate optimized for regularization during training may be suboptimal for uncertainty estimation, and the resulting uncertainty quality is often inferior to that obtained with fully independent ensembles [13, 25].

In summary, MC Dropout provides a practical post-hoc solution for uncertainty estimation when computational constraints prevent the use of fully independent ensembles. However, this convenience comes with a trade-off in uncertainty quality that should be explicitly acknowledged.

Deep Ensembles

One of the most effective and conceptually simple approaches for uncertainty estimation in deep learning is the use of Deep Ensembles. The approach is based on training M

independent models, each initialized with a different random seed, and aggregating their predictions at inference time to estimate both predictive performance and uncertainty.

Theoretical Motivation The motivation for Deep Ensembles stems from the highly non-convex nature of neural network loss landscapes. During training, different random initializations and stochastic optimization trajectories can lead models to converge to distinct local minima that correspond to different functions with comparable training performance. By training multiple independent models, the ensemble implicitly forms a multi-modal approximation to the posterior distribution over functions. Disagreement among ensemble members can therefore be interpreted as a manifestation of epistemic uncertainty, reflecting uncertainty in the model parameters rather than noise intrinsic to the data.

Formally, given M independently trained models with parameters $\{\theta_1, \ldots, \theta_M\}$, the ensemble predictive distribution for an input x is approximated as

$$p(y \mid x, \mathcal{D}) \approx \frac{1}{M} \sum_{m=1}^{M} p(y \mid x, \theta_m). \tag{2.14}$$

Voxel-wise Uncertainty in Segmentation In the context of medical image segmentation, each ensemble member produces a probability vector over K classes at every voxel. Denoting by $p_i^{(m)}$ the predicted probability vector at voxel i from model m, the ensemble mean prediction is given by

$$\bar{p}_i = \frac{1}{M} \sum_{m=1}^{M} p_i^{(m)}. \tag{2.15}$$

A fundamental voxel-wise measure of epistemic uncertainty is the ensemble variance, which quantifies the dispersion of predictions around the ensemble mean:

$$\sigma_i^2 = \frac{1}{M} \sum_{m=1}^{M} \left(p_i^{(m)} - \bar{p}_i \right)^2. \tag{2.16}$$

High variance indicates substantial disagreement among ensemble members regarding the prediction at voxel i, signaling regions where the model is epistemically uncertain. Conversely, low variance suggests strong consensus and higher confidence in the learned representation.

Practical Implementation via Cross-Validation In practice, Deep Ensembles can be efficiently realized through cross-validation. Multiple models are trained on different

folds of the training data, each fold providing a distinct training subset that naturally induces variability in the learned parameters. At inference time, predictions from all trained models are aggregated voxel-wise by averaging their predicted probabilities. The resulting mean probability map can be used both for segmentation, for instance by thresholding at 0.5 in the binary case, and for uncertainty estimation via ensemble-based measures.

This strategy leverages variability arising from both random initialization and differences in training data composition, yielding a practical approximation to posterior sampling. Moreover, cross-validation ensures efficient use of available data, as all samples contribute to training across different folds, while avoiding the need to train entirely independent models beyond those already required for validation.

Practical Advantages and Computational Trade-offs Deep Ensembles are widely considered as one of the most robust and reliable approaches for uncertainty estimation in deep learning. Empirically, they produce well-calibrated uncertainty estimates across a wide range of tasks and often improve baseline predictive performance through model averaging, which mitigates individual model biases and reduces overfitting. The interpretation of uncertainty is also intuitive, as disagreement among independently trained models naturally aligns with the notion of reduced confidence.

These benefits come at a computational cost. Training requires fitting multiple models, increasing overall training time and resource consumption by approximately a factor of M. Inference similarly becomes more expensive, as predictions from all ensemble members must be computed and aggregated. Storage requirements also increase, since multiple complete models must be retained. Nevertheless, empirical evidence suggests that relatively small ensemble sizes, such as $M = 5$, are sufficient to capture most of the uncertainty benefits while remaining computationally tractable.

Checkpoint Ensembles

Checkpoint Ensembles represent a recent and computationally efficient approach that bridges Bayesian uncertainty quantification and practical ensemble-based methods [39]. Instead of training multiple independent models, this technique exploits the inherent stochasticity of stochastic gradient descent (SGD) to collect diverse weight configurations during a single training run.

Theoretical Justification The key insight is that, under stochastic optimization, the training trajectory visits multiple weight configurations that reflect an implicit, algorithm-induced distribution over parameters (driven by random initialization, mini-batch noise,

and the learning-rate schedule). By saving checkpoints along this trajectory, one can form a computationally cheap ensemble whose uncertainty is estimated from the disagreement among these correlated predictors.

Method Description Let w_t denote the network weights at epoch t. After an initial burn-in phase of γT epochs, with $0 < \gamma < 1$ and T denoting the total number of training epochs, the model begins collecting checkpoints $\{w_t\}_{t=\gamma T}^{T}$. The predictive distribution for a new input x^* is then approximated as

$$p(y^* \mid x^*, \mathcal{D}) \approx \frac{1}{n} \sum_{i=1}^{n} p(y^* \mid x^*, w_{t_i}) \,, \tag{2.17}$$

where $\{w_{t_1}, \ldots, w_{t_n}\}$ denotes the subset of n collected checkpoints.

Cyclical Learning Rate and Multi-Modal Posterior To increase the diversity of sampled checkpoints, a cyclical learning rate schedule is typically employed, dividing training into M cycles. At the start of each cycle, the learning rate is reset to a high value, encouraging the optimizer to escape previously visited local minima and explore new regions of the weight space. Within each cycle, the learning rate decays according to a polynomial or cosine schedule. Checkpoints collected after a warm-up phase in each cycle are associated with different local minima, thereby capturing both intra-mode variability (within a single basin) and inter-mode variability (across distinct basins). This procedure enables an implicit approximation of a multi-modal posterior distribution.

Computational Efficiency and Practical Advantages Compared to traditional Deep Ensembles, Checkpoint Ensembles offer substantial computational benefits. First, only a single training run is required, eliminating the need to train multiple independent models. Second, the additional training overhead is negligible, as checkpointing incurs only marginal disk input/output costs. Third, the memory footprint is reduced, since all checkpoints correspond to a single model architecture rather than multiple replicas. Empirically, Checkpoint Ensembles achieve uncertainty estimates that are comparable to those obtained with Deep Ensembles, while requiring significantly fewer computational resources.

Theoretical Limitations and Trade-offs Despite their computational attractiveness, Checkpoint Ensembles rely on weaker theoretical assumptions than independently trained ensembles. The diversity of checkpoints may be limited if the optimizer converges to a narrow region of the weight space, and correlated errors can arise because all checkpoints

originate from a single training trajectory. Moreover, the quality of the posterior approximation depends strongly on the design of the learning rate schedule.

Generative Models and Structured Uncertainty

Advanced approaches such as Probabilistic U-Nets and Stochastic Segmentation Networks aim to model the conditional distribution of segmentations given an input image, generating diverse and spatially coherent predictions that capture both voxel-wise and structural uncertainty [15].

As an example, a Probabilistic U-Net augments a standard U-Net encoder–decoder architecture with a latent variable model. The conditional distribution over segmentations is expressed as

$$p(y \mid x) = \int p(y \mid x, z)\, p(z)\, \mathrm{d}z, \tag{2.18}$$

where z denotes a latent variable that encodes stochasticity in the segmentation process. By sampling from the prior distribution $p(z)$, the model can generate multiple plausible segmentations for the same input image, thereby capturing both aleatoric uncertainty arising from inherent ambiguity in the data and epistemic uncertainty associated with the model.

From a clinical perspective, these methods offer the advantage of producing structured uncertainty estimates at the level of lesions or anatomical objects. Such region-level uncertainty information is often more clinically actionable than purely voxel-wise confidence measures, as it enables probabilistic reasoning over entire anatomical structures rather than isolated voxels.

However, generative segmentation models are computationally demanding, as they require training and inference of stochastic architectures. Moreover, the resulting outputs are inherently more complex to interpret than deterministic segmentations augmented with uncertainty maps, which can limit their practical adoption in routine clinical workflows.

Test-Time Augmentation (TTA)

Test-Time Augmentation (TTA) estimates predictive uncertainty by applying controlled data augmentations at inference time and analyzing the variability of the resulting predictions. Unlike training-time augmentation, which is primarily used as a regularization strategy, TTA directly exploits the sensitivity of the model's predictions to plausible input perturbations in order to quantify uncertainty.

Method Description Given a test input image x and a set of deterministic and reversible augmentations $\mathcal{T} = \{t_1, \ldots, t_N\}$, such as flipping, rotation, or elastic deformation, TTA proceeds by applying each transformation to the input, yielding augmented images $x_n = t_n(x)$ for $n = 1, \ldots, N$. The model then produces a prediction for each augmented input, $y_n = f(x_n)$. To ensure consistency, each prediction is mapped back to the original image space using the inverse transformation, $y_n^{\text{inv}} = t_n^{-1}(y_n)$.

The final prediction is obtained by aggregating the de-augmented outputs, typically through averaging:

$$\bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n^{\text{inv}}, \tag{2.19}$$

while uncertainty is quantified by the variability of the predictions across augmentations:

$$\text{Var}_{\text{TTA}} = \frac{1}{N} \sum_{n=1}^{N} \left(y_n^{\text{inv}} - \bar{y} \right)^2 . \tag{2.20}$$

Uncertainty Semantics The variability induced by test-time augmentations reflects uncertainty that is inherent to the input data. In particular, TTA captures aleatoric uncertainty arising from image noise, low contrast, partial volume effects, or intrinsic ambiguity in anatomical boundaries. Since the model parameters are fixed, TTA does not directly capture epistemic uncertainty, which instead reflects uncertainty in the model itself and is more effectively addressed through ensemble-based approaches.

Computational and Practical Considerations From a practical standpoint, TTA is model-agnostic and can be applied to any trained network without modifying the architecture or retraining the model. It enables the design of clinically meaningful augmentations tailored to specific anatomical structures, provided that care is taken to avoid transformations that violate anatomical plausibility, such as left–right flips in asymmetric organs. TTA is particularly valuable for capturing input-dependent uncertainty that may not be apparent from model disagreement alone, making it complementary to ensemble-based epistemic uncertainty estimates.

The primary trade-off associated with TTA is computational cost, as inference time increases linearly with the number of applied augmentations. Moreover, the choice of augmentations is critical: overly mild transformations may lead to an underestimation of uncertainty, while overly aggressive ones can introduce artifacts and inflate uncertainty in a non-meaningful way. As a result, augmentations must be carefully selected and validated within the clinical context to ensure that the resulting uncertainty estimates are

both reliable and interpretable [17].

2.6. Voxel-wise Uncertainty Measures

Having established the primary uncertainty estimation methods, Now let us describe in detail the most well-known and widely used measures of uncertainty derived from forecasts. These measures are organized into two categories reflecting their computational dependencies and interpretability.

2.6.1. Probability-Based Measures

These measures depend solely on the mean softmax probabilities of the model's output.

Entropy

The Shannon entropy of the predicted probability distribution provides a measure of the inherent uncertainty associated with the ensemble mean prediction [18, 33]. For a voxel i, the entropy is defined as

$$H_i = -\sum_{k=1}^{K} \bar{p}_{i,k} \log(\bar{p}_{i,k} + \epsilon)\,, \tag{2.21}$$

where $\bar{p}_{i,k}$ denotes the probability for class k at voxel i, and $\epsilon = 10^{-8}$ is a small constant introduced to avoid numerical instabilities due to $\log(0)$.

In the case of binary segmentation ($K = 2$), this expression simplifies to

$$H_i = -\bar{p}_i \log(\bar{p}_i + \epsilon) - (1 - \bar{p}_i) \log(1 - \bar{p}_i + \epsilon)\,, \tag{2.22}$$

where $\bar{p}_i$ denotes the probability of the foreground class.

Interpretation Entropy attains its maximum value, $H_{\max} = \log 2 \approx 0.693$ in the binary case, when $\bar{p}_i = 0.5$, corresponding to maximal ambiguity in the mean prediction. Conversely, entropy is minimized ($H_i = 0$) when $\bar{p}_i \in \{0, 1\}$, indicating high confidence in the predicted class.

Confidence

The confidence of the prediction at voxel i is defined as the maximum softmax probability across k classes:

$$C_i = \max_k \bar{p}_{i,k}. \tag{2.23}$$

Interpretation Confidence directly reflects how strongly the prediction favors its most likely class. High confidence values, close to 1, indicate strong confidence in the predicted class, whereas lower values, approaching 0.5 in binary segmentation tasks, indicate ambiguity in the prediction. As such, confidence provides a simple and easily interpretable baseline measure for identifying regions of low predictive certainty.

2.6.2. Ensemble-Based Measures

These measures are computable only with ensemble model. They require access to predictions from individual ensemble members (folds or checkpoints), capturing uncertainty through statistics of disagreement.

Variance and Standard Deviation

The ensemble variance at voxel i quantifies the degree of disagreement among ensemble members [20]. Given an ensemble of M models, each producing a probability distribution over C classes for a voxel x_i, the predicted probability for class k from the m-th ensemble member is denoted by $p_{i,k,m}$. Although class probabilities are constrained to sum to one and are therefore statistically dependent, variance can be computed marginally for each class and aggregated across classes to obtain a scalar measure of dispersion that captures the overall disagreement among ensemble members. In this work, variance is estimated using the moment-based formulation:

$$\sigma_i^2 = \frac{1}{C} \sum_{k=1}^{C} \left(\frac{1}{M} \sum_{m=1}^{M} p_{i,k,m}^2 - \left(\frac{1}{M} \sum_{m=1}^{M} p_{i,k,m} \right)^2 \right), \tag{2.24}$$

The corresponding standard deviation, $\sigma_i = \sqrt{\sigma_i^2}$, conveys the same information on the original probability scale and is often more interpretable in practice [5, 7].

Interpretation High variance values indicate strong disagreement among ensemble members regarding the prediction at voxel i, serving as a direct indicator of epistemic uncertainty. Conversely, zero variance implies perfect consensus across the ensemble,

reflecting high confidence in the learned model representation.

3 | Dataset and Models

The dataset used in this thesis is a collection of full-body computed tomography (CT) scans acquired at the IRCCS Humanitas Research Hospital in Milan, Italy. All scans were acquired in a clinical setting for diagnostic purposes and have been subsequently de-identified to remove any personally identifiable information, thereby complying with privacy regulations and ethical standards for medical data research.

All CT volumes included in this study were manually segmented by expert radiologists and radiation oncologists in order to localize the lymphatic system within each scan. The specialists delineated lymphatic structures throughout the entire volume, generating binary masks that identify lymphatic tissue against the surrounding anatomical background. These expert annotations constitute the ground truth used throughout this thesis for model training, validation, and quantitative evaluation.

The clinical significance of accurate lymphatic system segmentation lies in its relevance to oncology and radiotherapy planning. The lymphatic system plays a central role in cancer staging, treatment planning, and assessment of metastatic spread. However, lymph nodes are inherently challenging to visualize on standard CT protocols, they are small, diffuse, and often exhibit poor contrast relative to surrounding tissue. This anatomical reality is directly reflected in the dataset characteristics and motivates the focus of this work on uncertainty estimation, as clinicians require not only predictions but also confidence maps that highlight potentially unreliable regions.

3.1. Visual Overview of CT Scans and Ground Truth Annotations

Figure 3.1 shows a representative CT volume from the dataset, visualized along the three principal anatomical planes: coronal, sagittal, and axial. The first row displays the original CT slices, while the second row overlays the corresponding ground-truth segmentation masks in red, highlighting the annotated lymphatic structures.

This example illustrates two key characteristics of the dataset. First, each CT scan spans from the head to the legs, leading to a large proportion of background tissue without lymphatic annotations. This full-body acquisition protocol induces a pronounced class imbalance, which has a substantial impact on both model training and performance evaluation. Second, the lymphatic structures appear as thin, tortuous, and poorly contrasted regions within the surrounding mediastinal and abdominal tissues. Such low contrast is typical of standard venous-phase CT imaging and represents a major source of aleatoric uncertainty, as well as inter-observer variability in manual annotations.

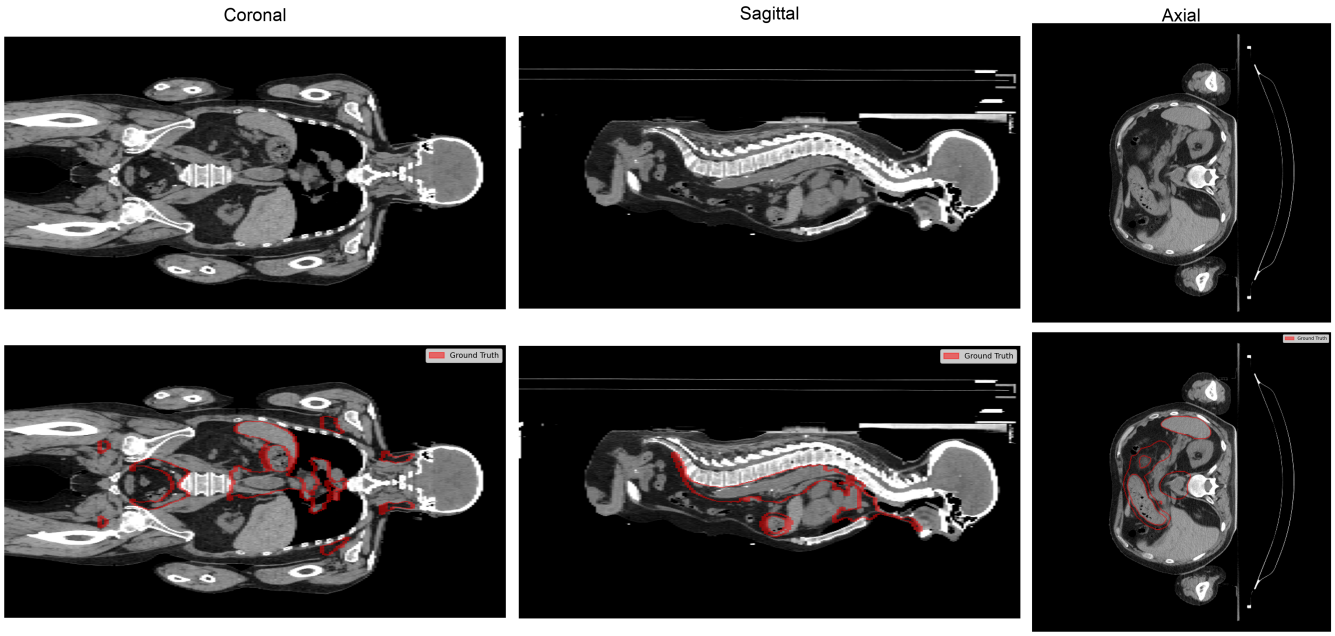

Figure 3.1: Example of a CT with ground truth annotation on 3 different planes.

3.2. Volume Shape and Spatial Heterogeneity

All volumes are stored in the NIfTI (Neuroimaging Informatics Technology Initiative) format, which is a standard representation in medical imaging and neuroimaging research. The in-plane spatial resolution is consistent across the dataset, with axial and sagittal dimensions fixed at 512×512 voxels. In contrast, the number of slices along the cranio–caudal (Z) axis varies substantially between patients.

Figure 3.2 reports the distribution of volume shapes within the dataset by showing the frequency of each distinct depth, expressed as the number of slices along the Z-axis. The depth ranges approximately from 192 to 398 voxels, with several values occurring more frequently than others, notably around 219, 232, and 243 slices. This variability can be attributed to differences in patient body size, acquisition parameters such as field-of-view

settings and table speed, and the clinical protocols adopted at the time of scanning.

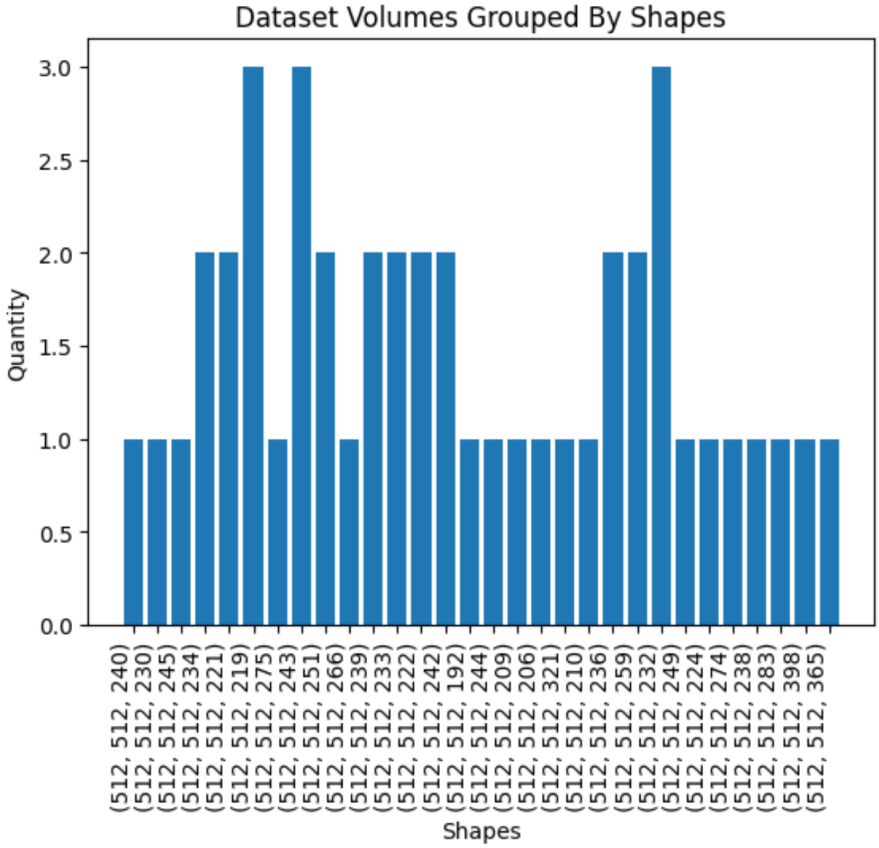

Figure 3.2: Histogram showing the distribution of dataset volumes grouped by shape (width, height, depth).

Such heterogeneity in volume depth introduces several practical challenges during preprocessing. Deep learning models typically require fixed-size inputs, which implies that all volumes must be resampled, padded, or cropped to satisfy the architectural constraints of the segmentation network. These operations may introduce interpolation artifacts or lead to partial loss of anatomical information. In addition, cases with a smaller number of slices may correspond to scans covering a limited anatomical extent, thereby reducing contextual information and potentially increasing the difficulty of the segmentation task. Finally, variability in depth directly affects the total number of voxels per volume, influencing memory usage and computational cost during training, inference, and evaluation.

The nnU-Net framework adopted in this thesis mitigates these issues by dynamically configuring its preprocessing pipeline. In particular, volumes are resampled to a common target spacing and cropped to remove non-informative background regions, allowing the network to handle variable-depth inputs in a robust and computationally efficient manner.

3.3. Voxel Spacing and Anisotropy

Figure 3.3 depicts the distribution of volumes grouped according to their voxel spacing configuration, expressed in millimeters. The in-plane voxel size along the X and Y axes varies approximately between 0.976 mm and 1.5625 mm, whereas the inter-slice spacing along the Z-axis assumes only three discrete values: 3.0 mm, 5.0 mm, or 7.5 mm.

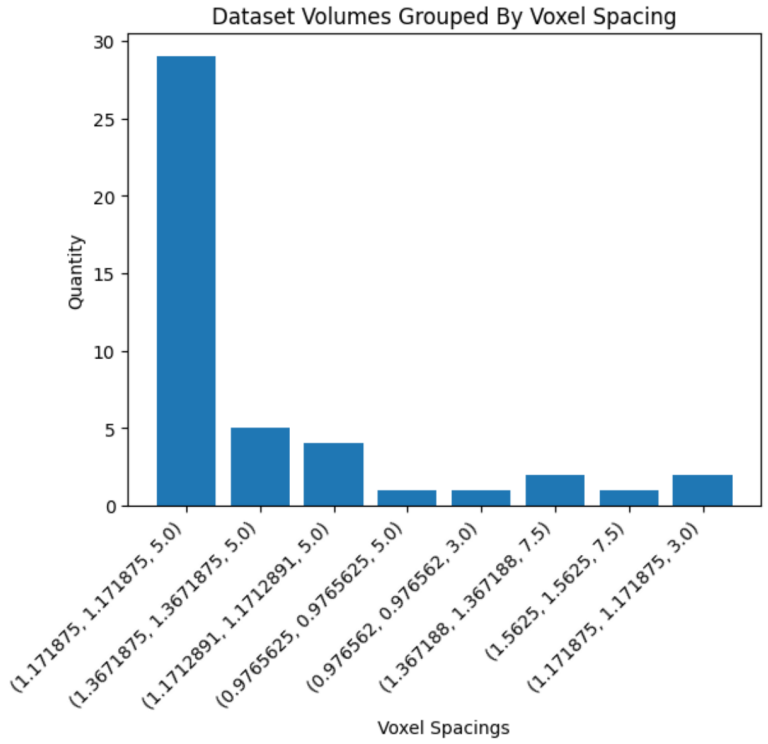

Figure 3.3: Histogram showing the number of dataset volumes grouped by voxel spacing configuration.

The most prevalent voxel spacing in the dataset is $(1.171875, 1.171875, 5.0)$ mm, which occurs in 29 cases. Other spacing configurations are less frequent but nonetheless present, contributing to the overall spatial heterogeneity of the dataset.

This heterogeneity has several important implications. First, the data exhibit marked anisotropic sampling: while the in-plane resolution is on the order of 1.17 mm, the slice thickness ranges from 5 mm to 7.5 mm depending on the case. Consequently, the resolution along the Z-axis is approximately four to six times coarser than that in the X–Y plane. Such anisotropy can adversely affect the performance of 3D convolutional neural networks, which generally assume near-isotropic voxel dimensions and may struggle to extract meaningful volumetric features when information is sparsely sampled along one axis.

Second, fine anatomical structures are particularly at risk. Lymphatic vessels and small lymph nodes often appear as thin, elongated structures, and a coarse inter-slice spacing, especially at 7.5 mm, may fail to capture their full spatial extent. This can lead to partial or incomplete segmentations, even when the in-plane resolution is adequate.

For these reasons, preprocessing is essential. The nnU-Net framework addresses this issue by automatically resampling all volumes to a common target spacing, chosen as the median spacing across the dataset. This normalization step enforces a consistent input geometry and reduces variability during model training, at the cost of potentially introducing minor interpolation artifacts in scans acquired at substantially different native resolutions.

3.4. Foreground Coverage and Class Imbalance

One of the most defining characteristics of this dataset is the severe class imbalance between lymphatic tissue (foreground) and background. Since each CT scan covers the entire body, the annotated lymphatic structures occupy only a very small fraction of the total volume.

To quantify this imbalance, the coverage of the foreground is defined as

$$\text{coverage}(i) = \frac{V_{\text{fg}}(i)}{V_{\text{tot}}(i)}, \tag{3.1}$$

where $V_{\text{fg}}(i)$ denotes the number of voxels labeled as lymphatic tissue, and $V_{\text{tot}}(i)$ represents the total number of voxels in the volume, computed as the product of its spatial dimensions. Figure 3.4 shows the histogram of foreground coverage percentages across all cases, with values rounded to one decimal place, corresponding to bins of 0.1%.

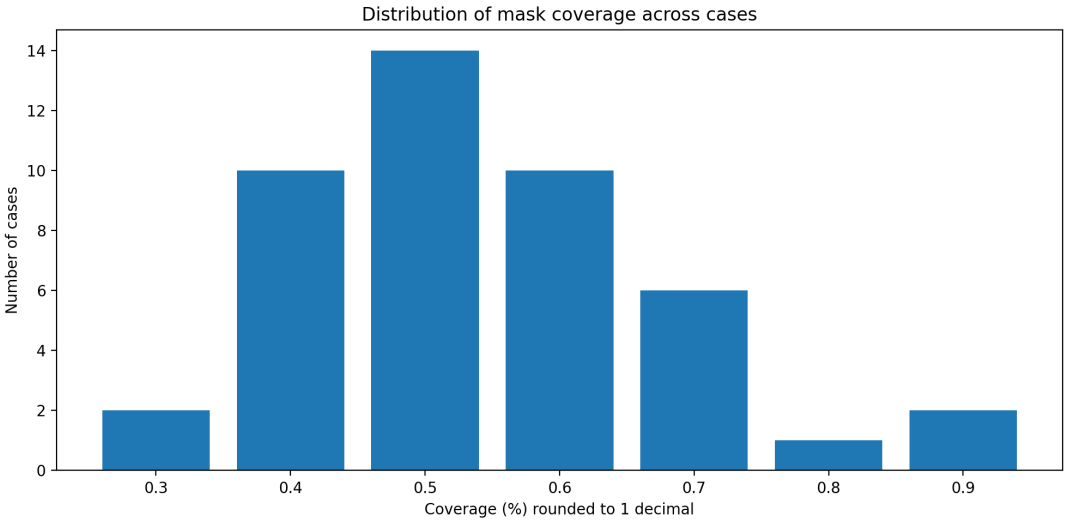

Figure 3.4: Histogram showing the coverage of the foreground grouped by percentage bin.

The resulting distribution highlights several relevant patterns. First, coverage values are uniformly low across the dataset, with all cases exhibiting less than 1.0% foreground voxels. The observed range spans approximately from 0.3% to 0.9%, indicating that lymphatic tissue accounts for fewer than one voxel per hundred. Second, the distribution is approximately unimodal, with a pronounced concentration around 0.5% (about 14 cases) and secondary peaks at 0.4% and 0.6%. This suggests that the overall extent of the annotated lymphatic system is relatively consistent across patients, despite variations in body size and scan depth. Finally, a small number of cases lie in the tails of the distribution, with particularly low coverage (around 0.3%) or higher coverage (approximately 0.8–0.9%), potentially reflecting differences in pathology, annotation practices, or scan coverage.

This extreme class imbalance has important implications for both model training and evaluation. During training, standard loss functions based on cross-entropy or Dice formulations can become biased toward the dominant background class, allowing a model to achieve high apparent performance by predicting background almost everywhere. To mitigate this effect, segmentation pipelines commonly rely on class weighting, focal losses, or sampling strategies that emphasize rare or difficult foreground examples. From an evaluation perspective, voxel-wise accuracy is largely uninformative in this setting, as it is dominated by correct background predictions. Instead, foreground-oriented metrics such as the Dice Similarity Coefficient, Hausdorff distance, and precision–recall curves are required for a meaningful assessment of segmentation quality.

Class imbalance also affects calibration and uncertainty analysis. When calibration met-

rics such as the Expected Calibration Error (ECE) are computed over the entire volume, the overwhelming prevalence of background voxels can lead to artificially optimistic confidence estimates. For this reason, ECE and related measures are typically evaluated within a spatial region of interest surrounding the ground-truth segmentation. Finally, the scarcity of foreground voxels limits the amount of information available for learning the shape and boundaries of lymphatic structures, naturally leading to elevated epistemic uncertainty within lymphatic regions and increased aleatoric uncertainty at their boundaries, where soft-tissue contrast in CT is inherently low.

3.5. Models Training and Implementation

The methodological choices adopted in this work are grounded in the experimental framework established by Mondo [23], who conducted a systematic investigation of uncertainty estimation techniques for medical image segmentation on the same dataset. His study highlighted that combining multiple strategies, namely ensembling, calibration, and test-time augmentation, can lead to improved segmentation accuracy and more reliable confidence estimates. Rather than attempting a full replication of Mondo's pipeline, the present work builds upon his findings by adopting a subset of these strategies that are natively supported and robustly implemented within the nnU-Net v2 framework.

To systematically assess how calibrations and uncertainty related strategies interact with different modeling paradigms, the experiments in this work are conducted on three configurations: a baseline single model, a checkpoint ensemble, and a deep ensemble. This comparative setup enables the analysis of how the proposed methods behave across varying levels of model diversity and whether consistent improvements in performance and reliability can be observed.

The nnU-Net architecture was deliberately selected for its capacity to autonomously adapt its configuration, including architecture, training scheme, and preprocessing pipeline, to the specific characteristics of the dataset at hand. This automated adaptation is particularly advantageous when working with limited annotated data, as it is in our case, where only 45 annotated volumes were available.

To enable robust comparison across multiple independent configurations, each model is implemented 4 times using 4 different partitions of the 45 volumes (in this thesis referred to as Datasets 005, 006, 007, and 008), created maintaining a mutual exclusivity constraint: no two test sets across different partitions contain the same cases.

3.5.1. Deep Ensemble Training Strategy

The Deep Ensemble training protocol implemented uses a 5-fold cross-validation strategy within the nnU-Net framework. From the total of 45 annotated volumes, 5 were reserved as a held-out test set for final evaluation, leaving 40 volumes for the training and validation pipeline. These 40 volumes were then systematically divided into five equally-sized folds, with each fold containing 8 volumes designated for validation and the remaining 32 for training. Crucially, each validation set is mutually exclusive, no two validation folds share overlapping cases, ensuring that the five resulting models are trained on complementary data splits and with different random initializations. At inference time the 5 predictions of all models are combined to obtain the final prediction.

Each ensemble was trained using the 3d_fullres configuration, which nnU-Net automatically selected based on the dataset fingerprint, with all preprocessing steps (resampling, normalization, cropping) and architectural decisions (depth, filter numbers) managed automatically by the framework.

3.5.2. Checkpoint Ensemble Training Strategy

Instead of training multiple independent networks on different data splits, the checkpoint ensemble leverages multiple snapshots of a single model taken at different stages of training, thus approximating posterior sampling at a substantially reduced computational cost compared to a full deep ensemble.

As in the deep ensemble setup, 5 volumes out of the 45 annotated scans were reserved as a strictly held-out test set, while the remaining 40 volumes were used for training and validation (respectively 32 and 8 samples). For the checkpoint ensemble, a single nnU-Net fold was trained on these 40 cases. The network was trained using the 3d_fullres configuration automatically selected by nnU-Net based on the dataset fingerprint, ensuring full consistency in preprocessing, architectural design, and training scheme with the other experimental configurations.

The training procedure adopts a cyclical learning rate schedule designed to encourage exploration of multiple regions of the weight space across different cycles. The total training budget is $T = 1200$ epochs and it is divided into $M = 3$ cycles of equal length $T_c = 400$, and within each cycle the learning rate is reset to a high restart value and then decays polynomially over epochs. The learning rate at epoch t within a cycle is computed

via a polynomial decay:

$$\alpha(t) = \begin{cases} \alpha_r, & \text{if } t_c = 0, \\ \alpha_0 \left[1 - \dfrac{\min(t_c, T_c)}{T_c}\right]^{\varepsilon}, & \text{otherwise,} \end{cases}$$

where $\alpha_r = 0.1$ is the initial restart learning rate, $\alpha_0 = 1 \times 10^{-2}$ is the base learning rate, $\gamma = 0.8$, $\varepsilon = 0.9$, and $T_c = T/M$ is the number of epochs per cycle. This schedule promotes traversal of distinct weight regions in each cycle.

During training, candidate 2 checkpoints are collected in the final phase of each cycle, when the learning rate has decayed and the model has converged toward a local basin in the loss landscape. For the experiments in this thesis those six checkpoints are used for inference, capturing both intra-cycle and inter-cycle variability.

3.5.3. Test-Time Augmentation in the Present Work

While test-time augmentation is often discussed in the literature as a mechanism for estimating aleatoric uncertainty through input perturbations, its role in this framework is primarily to improve inference robustness and segmentation performance rather than to explicitly quantify uncertainty. In line with this perspective, the present work adopts the native test-time augmentation mechanism provided by nnU-Net v2. This mechanism applies spatial reflections of the input CT volume along the anatomically valid axes. Specifically, the input volume is augmented using the following seven flip combinations:

- Flip along X
- Flip along Y
- Flip along Z
- Flip along XY
- Flip along XZ
- Flip along YZ
- Flip along XYZ

Inference is performed independently on the original volume and on each of the seven mirrored versions. The predicted probability maps obtained from the augmented inputs are subsequently reoriented back to the original spatial configuration. All predictions are then aggregated through voxel-wise summation and averaged to produce the final

probability map. This mirroring-based TTA does not introduce geometric distortions such as rotations or elastic deformations; instead, it exploits anatomical symmetry to improve robustness while preserving spatial consistency. The approach is computationally efficient, fully integrated within the nnU-Net inference pipeline, and requires no custom implementation. Crucially, all uncertainty measures analyzed in subsequent sections are derived from ensemble variability and calibrated probability outputs, not from variability across augmented inputs.

4 | Temperature Scaling for Confidence Calibration

In the context of clinical segmentation, calibration plays a central role beyond classification accuracy. In our setting, voxel-wise probabilities are not only used to generate binary segmentations, but also to derive uncertainty maps and support downstream error analysis. Poor calibration directly affects the reliability of these probabilistic outputs: overconfident predictions may conceal clinically relevant errors, while underconfident predictions may inflate uncertainty and reduce interpretability.

Temperature scaling represents a remarkably simple yet effective post-processing technique for addressing this problem. Unlike many calibration methods that require retraining the model or access to additional labeled data, temperature scaling operates as a parameter-free post-hoc correction applied to the logit scores. The method requires only a validation set, minimal computational overhead (approximately one millisecond per prediction), and can be implemented in two lines of code. Despite its simplicity, temperature scaling has demonstrated state-of-the-art calibration performance across diverse datasets and network architectures[35]. In this work we restrict calibration to an anatomically meaningful region of interest (ROI) to avoid background dominance and ensure clinical relevance.

4.1. Theoretical Background and Motivation

Temperature scaling modifies the logit scores z produced by a neural network before applying the softmax function. For a single temperature parameter $T > 0$, the calibrated probability estimate becomes:

$$p_{\text{calib}}(y|x) = \text{softmax}\left(\frac{z}{T}\right)$$

The temperature parameter acts as a simple smoothing factor: values $T > 1$ suppress con-

fidence (flattening the softmax distribution), while values $T < 1$ sharpen it (concentrating probability mass). A temperature of $T = 1$ corresponds to the original uncalibrated predictions.

The theoretical justification for temperature scaling is grounded in the principle of minimizing negative log-likelihood (NLL), a proper scoring rule. Proper scoring rules have the fundamental property that they are minimized when predicted probabilities equal true frequencies; they incentivize honest probability estimation and cannot be exploited by strategic manipulation. NLL, defined as:

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|x_i)$$

directly measures the plausibility of observed labels under predicted probability distributions. By selecting the temperature that minimizes NLL on a validation set, we obtain a principled estimate of the parameter that best aligns predicted confidence with true accuracy, without arbitrarily choosing between competing calibration metrics.

A critical observation is that temperature scaling affects only the probabilistic outputs, not the hard predictions. The argmax operation which, determines the class assignment, is invariant to scaling by a positive constant, so $\arg\max(z/T) = \arg\max(z)$ for any $T > 0$. Therefore, temperature scaling cannot improve classification accuracy; it can only redistribute confidence across correct and incorrect predictions to better reflect the underlying uncertainty.

A second constraint is that temperature scaling uses a single global parameter across all samples and predictions. This architectural simplicity makes the method efficient and interpretable, but it also means the parameter must be estimated from a sufficiently large and representative calibration set. Naive selection on test data leads to overfitting; selection on training data leads to optimistic estimates. Thus, robust estimation of T requires a properly held-out calibration dataset.

4.2. Calibration Set

The challenge in selecting T is obtaining a validation set that is both (i) large enough to estimate a single global parameter reliably, and (ii) truly independent of both training and test data. In the case of the Base Model and the Checkpoint Ensemble, unfortunately there is not much we can do, due to the scarcity of data, the fact that we cannot reuse the training data and that it is in our best interest to keep the test cases for future analysis,

for each implementation, we are forced to use only its 8 validation cases. In the context of deep ensemble models based on cross-validation, an elegant solution exists: the out-of-fold (OOF) predictions.

In a K-fold cross-validation ensemble, each training example is held out from exactly one fold during training. The predictions of that fold on the held-out examples are, by construction, independent of those specific examples. By collecting predictions from each fold on its held-out validation set, we obtain a full-coverage set of predictions on all training examples without any leakage. Importantly, this OOF dataset is disjoint from the final test set but large enough (in this study 40 training cases with hundreds of thousands of voxels per case) to provide stable parameter estimates.

Concretely, the OOF calibration procedure proceeds as follows:

1. For each training case i, identify the fold k_i where case i was held out during validation.
2. Extract the logit scores z_i produced by fold k_i when predicting on case i.
3. Collect all OOF logits and corresponding ground-truth labels into a single calibration dataset.
4. Estimate the global temperature T^* by optimizing an objective function (NLL) on this aggregated OOF dataset.

Thanks to this method, the OOF dataset comprises 40 different cases, selecting predictions made by individual folds that have not seen those cases during training, thereby avoiding data leakage while providing a realistic view of ensemble behavior on unseen data.

4.3. Region-of-Interest-Aware Calibration

In volumetric medical imaging, the background class typically dominates the volume by orders of magnitude. A segmentation model trained on these imbalanced voxel-wise labels often produces very high confidence for background voxels, which, while correct, are informatively poor for assessing calibration. When computing metrics such as expected calibration error (ECE) over the entire volume, background pixels contribute disproportionately, masking calibration problems in clinically relevant structures.

To address this, we restrict calibration to a region of interest (ROI) defined as all voxels within a clinically meaningful distance from the segmentation boundary. Specifically, the ROI is constructed as the set of voxels within 15 mm of the border separating positive and negative classes (as determined by the ground-truth annotation or the prediction). This distance threshold of 15 mm is chosen because it represents the clinically relevant margin around anatomical structures: predictions near the boundary are where diagnostic

confidence matters most, both for clinical decision-making and for detecting systematic calibration failures.

The ROI is defined in physical space (millimeters) using the voxel spacing encoded in the image header, ensuring spatial consistency across cases with different resolutions and acquisition protocols. Voxels are included in the ROI if they are in the prediction or ground truth mask or if their Euclidean distance to the nearest boundary point is ≤ 15 mm.

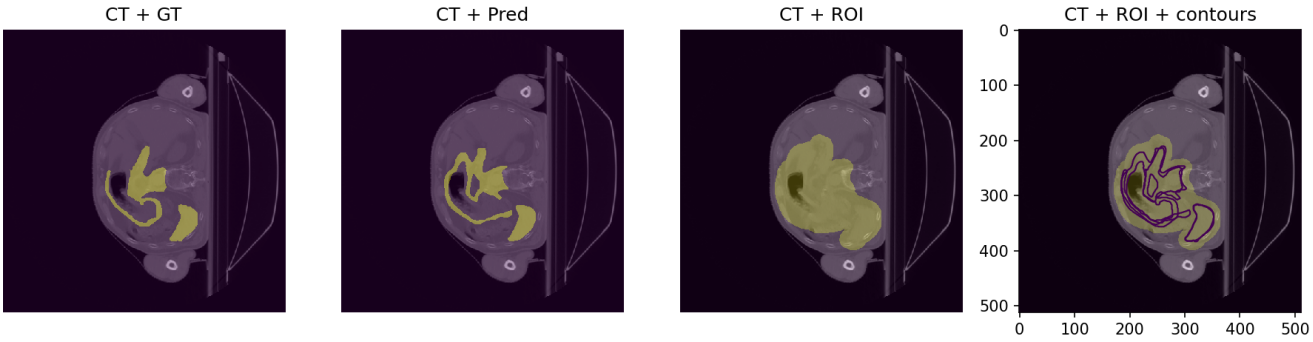

Figure 4.1: Pictures showing GT, prediction and ROI masks.

By performing all calibration steps on ROI-restricted voxels, we obtain a more informative and clinically grounded estimate of T^* that reflects the model's confidence in the regions where clinicians actually rely on the network's uncertainty quantification.

4.4. Temperature Optimization and Evaluation Metrics

Given the calibration dataset restricted to the region of interest (ROI), we estimate the optimal temperature T^* using a two-stage grid search procedure.

Coarse search. We evaluate the NLL across 25 logarithmically spaced temperature values in the range $[1, 15.0]$. This logarithmic spacing ensures uniform exploration across orders of magnitude, appropriate for an exponentiated scaling parameter.

Fine refinement. From the coarse search results, we identify the three temperature values that yield the lowest NLL. We then define a refined log-spaced grid between the

minimum and maximum of these three values and re-evaluate 25 points in this narrower interval to locate the optimum with higher precision.

The final estimate T^* is selected as the temperature minimizing NLL. This choice emphasizes theoretical soundness by optimizing a proper scoring rule rather than relying on ad-hoc calibration metrics. The NLL directly reflects the log-likelihood of the model's probabilistic predictions on held-out data.

Calibration metrics. Although the temperature T^* is selected by minimizing ROI NLL, we also track ECE and the Brier score as complementary calibration indicators. This serves as a consistency check: improvements in NLL should be accompanied by improvements ECE and in overall probabilistic accuracy, BS. This is to verify that the optimization is not limited to improving a single parameter, but produces consistent calibration gains across multiple criteria.

4.5. Results: Temperature Estimation and Test Generalization

In all three configurations, *base model*, *deep ensemble*, and *checkpoint ensemble*, the temperature sweep produced a well-defined minimum of NLL as a function of T, with a broad interval of values yielding consistent improvements across metrics (NLL, ECE, Brier). The optimal temperatures clustered around $T^* \approx 2.2$–2.8, with mean values of 2.62 ± 0.03 for the base model, 2.83 ± 0.05 for the deep ensemble, and 2.20 ± 0.04 for the checkpoint ensemble. This tight range suggests that all three predictors share a similar over-confidence pattern that can be effectively corrected by a single global scaling factor.

The temperature sweep for each configuration exhibited a well-defined minimum in negative log-likelihood, with highly consistent behavior across all models (Figure 4.2). All of them demonstrated similar curves: NLL, ECE, and Brier score all decrease as temperature increases from $T = 1$, reach a minimum near T^*, then increase again for higher temperatures. Expected calibration error showed the most substantial absolute improvements around the optimum, while Brier score remained relatively flat but still decreased. The consistency of sweep profiles across independent models validates that the calibration deficit is a systematic property of the nnUNet training procedure over our dataset, not an artifact of individual model quirks.

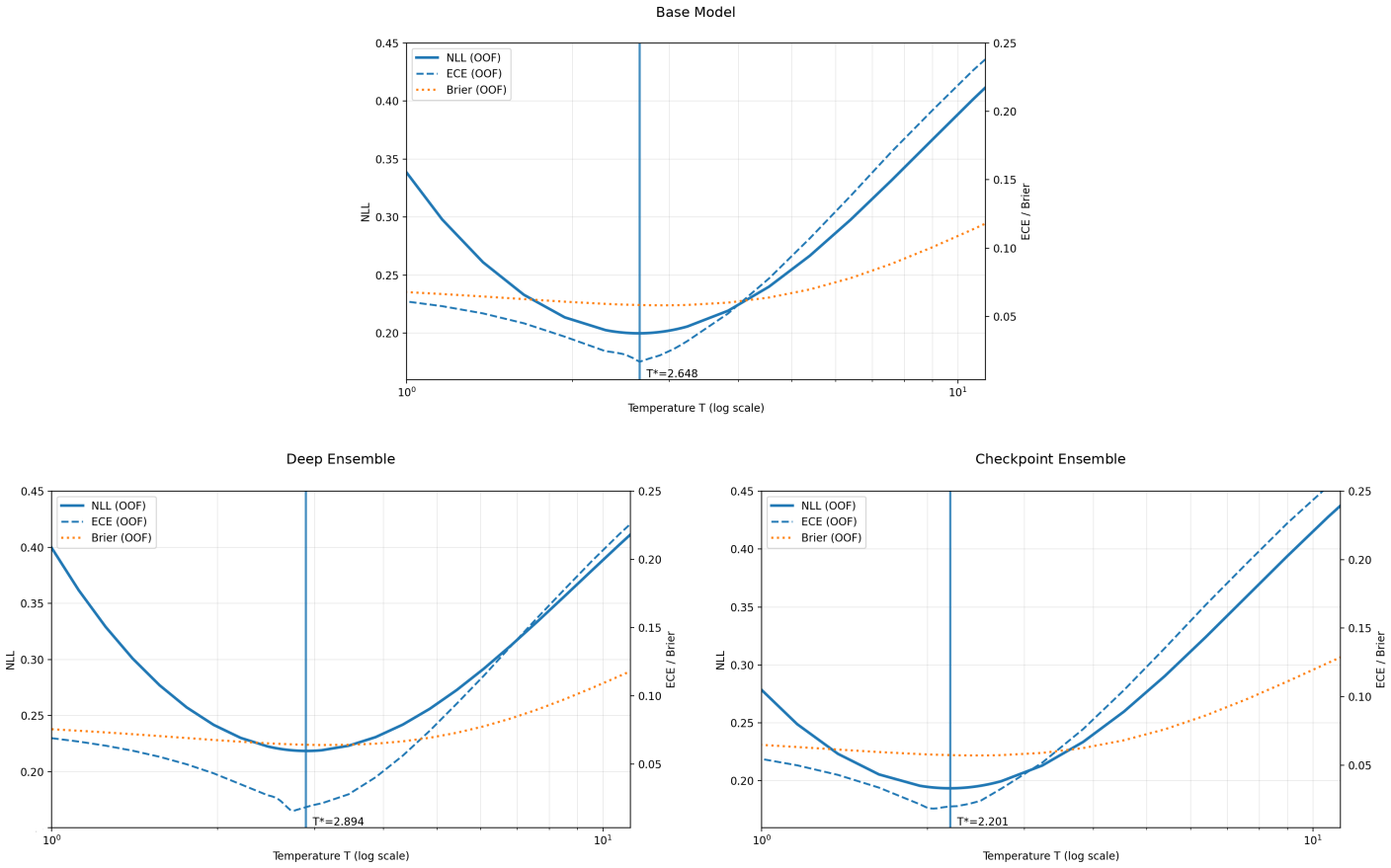

Figure 4.2: Temperature sweep across all configuration over the partition D005. Each panel shows NLL (left axis, solid blue), ECE (left axis, dashed blue), and Brier score (right axis, dotted orange) as functions of temperature T on logarithmic scale. Vertical lines mark the temperature T^* minimizing NLL for each ensemble.

Test Set Evaluation On the test set, temperature scaling substantially improved calibration for all methods. Tables 4.1, 4.2 and 4.3 report the mean calibration metrics before and after applying T^*.

Method	$\mathbf{T}^*$	**NLL**(T=1)	**NLL**(T^*)
Base Model	2.6197 ± 0.0308	0.3423 ± 0.1012	0.2012 ± 0.0367
Deep Ensemble	2.8269 ± 0.0479	0.3172 ± 0.0902	0.1934 ± 0.0322
Checkpoint Ensemble	2.2011 ± 0.0381	0.2814 ± 0.0786	0.1959 ± 0.0356

Table 4.1: Mean Negative Log-Likelihood (NLL) before and after temperature scaling for the three model configurations.

Method	T^*	ECE(T=1)	ECE(T^*)
Base Model	2.6197 ± 0.0308	0.0612 ± 0.0147	0.0270 ± 0.0088
Deep Ensemble	2.8269 ± 0.0479	0.0572 ± 0.0146	0.0281 ± 0.0086
Checkpoint Ensemble	2.2011 ± 0.0381	0.0544 ± 0.0142	0.0262 ± 0.0089

Table 4.2: Mean Expected Calibration Error (ECE) before and after temperature scaling for the three model configurations.

Method	T^*	BS(T=1)	BS(T^*)
Base Model	2.6197 ± 0.0308	0.0680 ± 0.0139	0.0586 ± 0.0114
Deep Ensemble	2.8269 ± 0.0479	0.0650 ± 0.0137	0.0562 ± 0.0107
Checkpoint Ensemble	2.2011 ± 0.0381	0.0652 ± 0.0131	0.0578 ± 0.0110

Table 4.3: Mean Brier Score (BS) before and after temperature scaling for the three model configurations.

For the base model, NLL decreased from 0.3423 ± 0.1012 to 0.2012 ± 0.0367, ECE from 0.0612 ± 0.0147 to 0.0270 ± 0.0088, and Brier score from 0.0680 ± 0.0139 to 0.0586 ± 0.0114. The deep ensemble and the checkpoint ensemble showed very similar absolute values, with post-calibration NLL in the range 0.19–0.20 and ECE around 0.026–0.028.

The grouped bar plot in Figure 4.3 highlights that, for each model class, all three metrics consistently improve after temperature scaling.

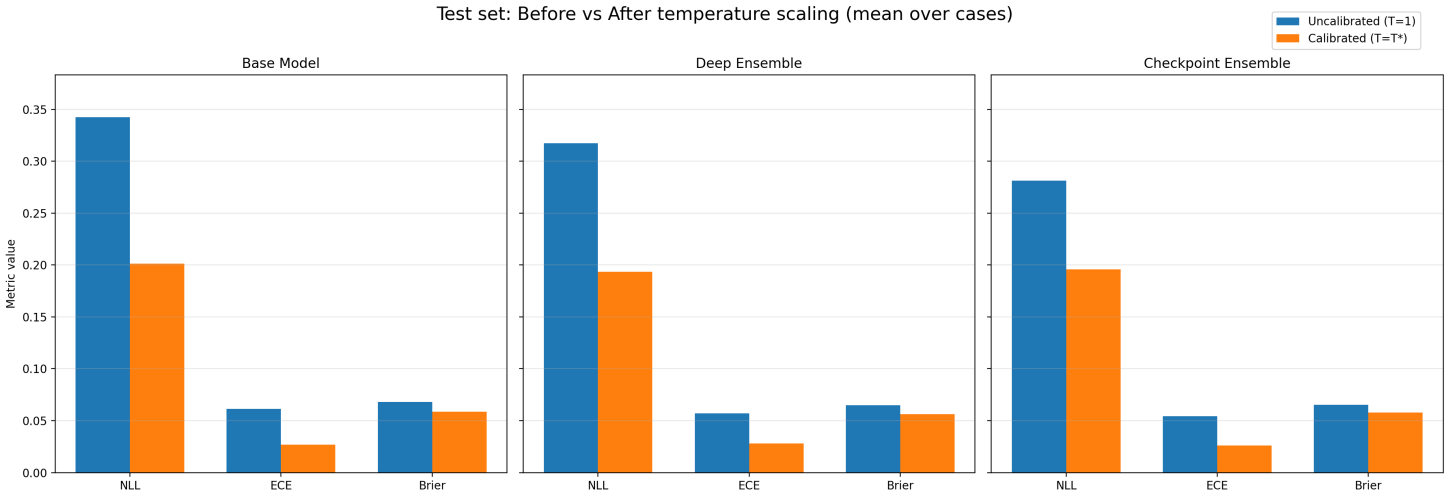

Figure 4.3: Temperature comparison across all 3 methods. Each panel shows mean NLL, ECE and Brier score variations from T to T^*.

To better compare the three predictors, Table 4.4 summarizes the relative percentage change of each metric between $T = 1$ and T^*.

Method	ΔNLL [%]	ΔECE [%]	ΔBS [%]
Base Model	−41.2	−55.9	−13.8
Deep Ensemble	−39.0	−50.9	−13.5
Checkpoint Ensemble	−30.4	−51.8	−11.3

Table 4.4: Relative percentage change of NLL, ECE and Brier Score between the uncalibrated model (T=1) and the calibrated model (T^*) for the three configurations. Negative values indicate an improvement after temperature scaling.

Applying T^* achieved substantial improvements across all metrics and methods. The base model achieves the strongest relative gain, with NLL, ECE, and Brier reduced by 41.2%, 55.9%, and 13.8%, respectively. The ensembles show improvements of similar magnitude, every test case showed improvements and no degradation occurred.

Overall, temperature scaling proves to be a remarkably simple yet powerful post-hoc calibration technique that generalizes seamlessly across the base model, deep ensemble, and checkpoint ensemble configurations. Notably, all three methods exhibit consistent improvements across all calibration metrics (NLL, ECE, Brier score), with no observed trade-offs. Post-calibration, the ensembles achieve marginally better performance than the calibrated base model (NLL 0.193–0.196 vs. 0.201; ECE 0.026–0.028 vs. 0.027; BS 0.0562–0.0578 vs. 0.0586), but these gains, are unlikely to justify their substantially higher computational cost (5× training and inference for the deep ensemble, 6× inference for the checkpoint ensemble) in most practical scenarios. This demonstrates that a single global temperature parameter T^* can effectively correct systematic miscalibration across fundamentally different model implementations, making temperature scaling an attractive baseline for uncertainty-aware medical image segmentation. In subsequent chapters, we adopt the calibrated versions of all three predictors for spatial uncertainty analysis.

5 | Uncertainty Estimation and Visualization

In this chapter, we move from global calibration metrics to spatially resolved uncertainty estimates. The goal is to provide clinicians not only with a predicted segmentation on CT scans, but also with voxel-wise maps that highlight where the model is more likely to be wrong. Such maps are intended as decision-support tools: they should draw attention to regions that warrant careful review and potential manual correction, rather than treating all voxels as equally trustworthy.

Building on the uncertainty measures introduced in Section 2.6, we derive several voxel-wise uncertainty maps from models' predictions, all normalized in the range [0, 1]. We first consider "base" maps that directly reflect local predictive ambiguity, such as entropy or ensemble disagreement. We then introduce distance-aware uncertainty representations that explicitly de-emphasize the trivial uncertainty occurring at the segmentation boundary and prioritize atypical, high-uncertainty patterns deeper inside and outside the predicted structures.

5.1. Voxel-Wise Uncertainty Map Computation

Voxel-wise uncertainty maps are derived from the predictive probabilities produced by the segmentation model. Given an input CT volume, the model outputs a voxel-wise probability map for the foreground class. Depending on the configuration, this predictive information may come from a single deterministic predictor (Base Model) or from multiple predictors (Deep Ensemble or Checkpoint Ensemble), whose outputs approximate a predictive distribution when aggregated. Some of the uncertainty measures that we are going to see are disagreement-based measures that rely on variability across multiple predictors and are therefore defined only when ensemble-type methods are available.

5.1.1. Base uncertainty measures

The first group of uncertainty maps consists of base measures computed directly from the models' predictions, without incorporating explicit spatial context. These maps quantify how ambiguous or unstable the model's predictions are at each voxel, either due to uncertainty in the ensemble mean prediction or due to disagreement among ensemble members.

In addition to the uncertainty measures introduced in the Background chapter, in this work we introduce new ensemble-based measures to study uncertainty from different points of view.

Mutual Information

"Mutual Information (MI) measures the information shared between the ensemble member index and the predicted output." [19]. In the context of ensemble-based segmentation, it is defined as the difference between the entropy of the ensemble mean and the average entropy of individual predictions, thereby quantifying the epistemic uncertainty attributable to inter-model variability. Formally, the mutual information at voxel i is:

$$\mathrm{MI}_i = H(\bar{p}_i) - \frac{1}{M}\sum_{m=1}^{M} H\left(p_i^{(m)}\right), \tag{5.1}$$

where $H(p) = -\sum_k p_k \log(p_k)$ denotes the Shannon entropy, $\bar{p}_i$ is the ensemble mean probability vector at voxel i, and $p_i^{(m)}$ is the prediction of the m-th ensemble member.

The two terms in this expression have distinct interpretations. The entropy of the ensemble mean, $H(\bar{p}_i)$, measures the ambiguity of the consensus prediction, reflecting how uncertain the aggregated output is. In contrast, the average entropy of the individual ensemble members, $\frac{1}{M}\sum_m H(p_i^{(m)})$, captures the typical uncertainty present within each single prediction.

If all ensemble members produce identical predictions, their individual entropies match the entropy of the ensemble mean, resulting in $\mathrm{MI}_i = 0$, which indicates perfect consensus and no epistemic uncertainty. Conversely, MI becomes large when ensemble members are individually confident, exhibiting low entropy, but disagree strongly with one another. In such cases, the ensemble mean may appear ambiguous, yet this ambiguity is entirely due to model disagreement rather than noise in the data.

Expected Posterior KL Divergence

Expected Pairwise Kullback–Leibler divergence (EPKL) quantifies the average divergence between the predictions of individual ensemble members and the ensemble mean. It provides an information-theoretic measure of how much, on average, each model's prediction differs from the consensus distribution. Formally, the EPKL at voxel i is defined as

$$\mathrm{EPKL}_i = \frac{1}{M} \sum_{m=1}^{M} \mathrm{KL}\left(p_i^{(m)} \,\|\, \bar{p}_i \right), \tag{5.2}$$

where $\mathrm{KL}(p\|q) = \sum_k p_k \log(p_k/q_k)$ denotes the Kullback–Leibler divergence, $p_i^{(m)}$ is the probability distribution predicted by the m-th ensemble member at voxel i, and $\bar{p}_i$ is the ensemble mean probability distribution.

The KL divergence measures how strongly an individual member's prediction deviates from the ensemble mean, with larger deviations incurring higher penalties. EPKL averages this divergence across all ensemble members and is therefore zero when all models produce identical predictions. As an epistemic uncertainty measure, EPKL captures disagreement arising from model variability rather than from ambiguity intrinsic to the data.

Unlike mutual information, which measures the potential information gain from knowing which model generated a prediction, EPKL directly quantifies the magnitude of disagreement between each ensemble member and the consensus.
From a clinical perspective, EPKL is particularly useful for identifying voxels where individual models substantially deviate from the consensus. Such regions highlight cases in which reliance on the ensemble mean alone may be misleading, flagging areas where the segmentation is unreliable due to inconsistent model behavior and therefore warrants closer inspection.

Maximum Fold Deviation

Maximum Fold Deviation (MFD) captures the most extreme disagreement between any ensemble member and the ensemble mean prediction at a given voxel. It is defined as

$$\mathrm{MFD}_i = \max_m \left| p_i^{(m)} - \bar{p}_i \right|, \tag{5.3}$$

where $p_i^{(m)}$ denotes the prediction of the m-th ensemble member at voxel i and $\bar{p}_i$ is the ensemble mean.

MFD measures the worst-case epistemic deviation within the ensemble. While variance

summarizes the average spread of predictions and can therefore be attenuated by agreement among most models, MFD explicitly highlights the presence of a single strongly dissenting prediction. A high MFD indicates that at least one ensemble member substantially disagrees with the consensus, even if the remaining models agree closely.

From a clinical perspective, MFD is particularly relevant for robustness assessment. A single confident dissenting prediction may signal an ambiguous region, a systematic model bias, or a potential error case in which the ensemble mean masks an important alternative interpretation. Consequently, voxels with high MFD warrant closer inspection, as the segmentation outcome may depend sensitively on model initialization or training variability.

Boundary Disagreement

Boundary Disagreement (BD) identifies voxels for which the ensemble exhibits a fundamental split at the decision boundary, namely cases where some ensemble members predict the presence of a structure while others predict its absence. It is defined as

$$\mathrm{BD}_i = \left(\max_m p_i^{(m)} - \min_m p_i^{(m)}\right) \cdot 1\left[\max_m p_i^{(m)} > 0.5 \ \wedge \ \min_m p_i^{(m)} < 0.5\right], \tag{5.4}$$

where $p_i^{(m)}$ denotes the prediction of the m-th ensemble member at voxel i, and $\mathbb{1}[\cdot]$ is an indicator function.

The indicator activates only when ensemble predictions cross the decision threshold, meaning that at least one model predicts inclusion and at least one predicts exclusion. In such cases, BD quantifies the magnitude of disagreement as the span between the most confident positive and most confident negative prediction. If all ensemble members lie on the same side of the threshold, BD is zero, regardless of any internal variability in probability values.

BD captures decision-level epistemic uncertainty, focusing specifically on uncertainty about binary class membership rather than numerical dispersion alone. From a clinical perspective, voxels with high BD correspond to unstable segmentation boundaries where the inclusion or exclusion of tissue is contested across models. These regions are particularly error-prone and therefore represent natural candidates for expert review, manual correction, or automated quality assurance flags.

Anti-Confidence

Anti-Confidence metrics are designed to quantify asymmetric decision-level risk in ensemble-based segmentation, focusing on situations in which ensemble members disagree across

the decision boundary. Unlike symmetric disagreement measures, these metrics explicitly distinguish between the risk of false negative and false positive decisions, thereby providing uncertainty estimates that are directly aligned with clinically meaningful error modes. Let $\hat{Y}$ denote the predicted binary segmentation and Y the ground truth. Voxel-wise error masks are defined using the standard confusion components as follows:

- **False positives (FP):** voxels for which $\hat{Y} = 1$ and $Y = 0$,
- **False negatives (FN):** voxels for which $\hat{Y} = 0$ and $Y = 1$.

Anti-Confidence FN (False Negative Risk) Anti-Confidence FN targets voxels for which the ensemble mean prediction indicates absence of the structure, while at least one ensemble member assigns a high probability to its presence. Formally, it is defined as

$$\text{AC-FN}_i = \max_m p_i^{(m)} \cdot 1\left[\bar{p}_i \leq 0.5\right], \tag{5.5}$$

where $\bar{p}_i$ is the ensemble mean probability at voxel i.

This measure is activated only when the ensemble mean supports a negative decision. Its magnitude reflects the strongest dissenting opinion in favor of inclusion, and therefore quantifies the risk of a false negative. A high AC-FN value indicates that, despite a negative consensus, at least one model is confident that the structure is present, suggesting that exclusion may be unreliable.

Anti-Confidence FP (False Positive Risk) Conversely, Anti-Confidence FP identifies voxels for which the ensemble mean prediction is positive, but one or more ensemble members confidently predict absence of the structure. It is defined as

$$\text{AC-FP}_i = \left(1 - \min_m p_i^{(m)}\right) \cdot 1[\bar{p}_i > 0.5]\,. \tag{5.6}$$

Here, the metric is activated only when the ensemble mean favors inclusion. Its value captures the strongest evidence against inclusion among the ensemble members, reframed as confidence in absence. High AC-FP values therefore indicate an elevated risk of false positive segmentation.

Interpretation and Relation to Other Measures Both Anti-Confidence measures capture asymmetric epistemic uncertainty by conditioning on the ensemble's mean decision and quantifying how strongly individual models disagree in a direction that threatens that decision. While measures such as Boundary Disagreement identify whether ensemble

predictions cross the decision threshold and quantify the overall spread, Anti-Confidence explicitly encodes which type of error is more likely, missing a structure or incorrectly including it. This directional information makes Anti-Confidence particularly informative in clinical settings.

From a practical perspective, high AC-FN values in regions predicted as negative suggest the need to verify whether relevant anatomy may have been missed, whereas high AC-FP values in regions predicted as positive indicate potential over-segmentation. By distinguishing between these two error modes, Anti-Confidence metrics provide more actionable uncertainty information than symmetric disagreement measures, supporting targeted review and quality assurance in clinical workflows.

Anti-Confidence Total To gain a broader view, it is possible to combine anti-confidence FN and FP into a single map so that areas at risk of false positives and false negatives can be seen at a glance.

Sample Mean Uncertainty

Sample Mean Uncertainty (SMU) is a composite voxel-wise measure designed to capture two complementary sources of unreliability in ensemble predictions: disagreement among ensemble members and marginality of the ensemble mean prediction. While ensemble variance reflects how strongly individual models disagree, serving as a proxy for epistemic uncertainty, it does not account for situations in which models agree on a prediction that is itself weak or close to the decision boundary. SMU explicitly addresses this limitation by penalizing low-confidence consensus predictions.

Formally, for a voxel i, SMU is defined as

$$\mathrm{SMU}_i = \sigma_i^2 + \left(1 - \max_k \bar{p}_{i,k}\right), \tag{5.7}$$

where σ_i^2 denotes the ensemble variance at voxel i, and $\bar{p}_{i,k}$ is the ensemble mean probability for class k. The second term acts as a confidence penalty: it approaches zero when the ensemble mean prediction is highly confident, and increases as the maximum class probability decreases, reaching its highest values near the decision boundary (e.g., around 0.5 in binary segmentation).

In relation to other uncertainty measures, entropy captures only the ambiguity of the ensemble mean prediction, while variance captures only the dispersion of individual predictions around that mean. SMU combines both effects and therefore becomes large either

when ensemble members disagree substantially or when they agree on an intrinsically ambiguous prediction. As a result, SMU tends to emphasize regions such as segmentation boundaries and other difficult areas where predictions are either unstable across models or only marginally confident even in consensus.

From a practical standpoint, SMU is computationally inexpensive once the ensemble variance and mean probabilities are available, as it requires only a maximum operation over classes and a simple summation.

When visualizing voxel-wise uncertainty maps derived from ensemble predictions, it is important to note that different uncertainty measures often highlight similar spatial regions. In the examples reported in this chapter, most uncertainty maps emphasize areas close to the predicted segmentation boundaries and anatomically ambiguous regions. This behavior is expected and reflects the fact that all measures are computed from the same underlying ensemble predictions and are therefore sensitive to shared sources of uncertainty, such as low contrast, partial volume effects, and class imbalance.

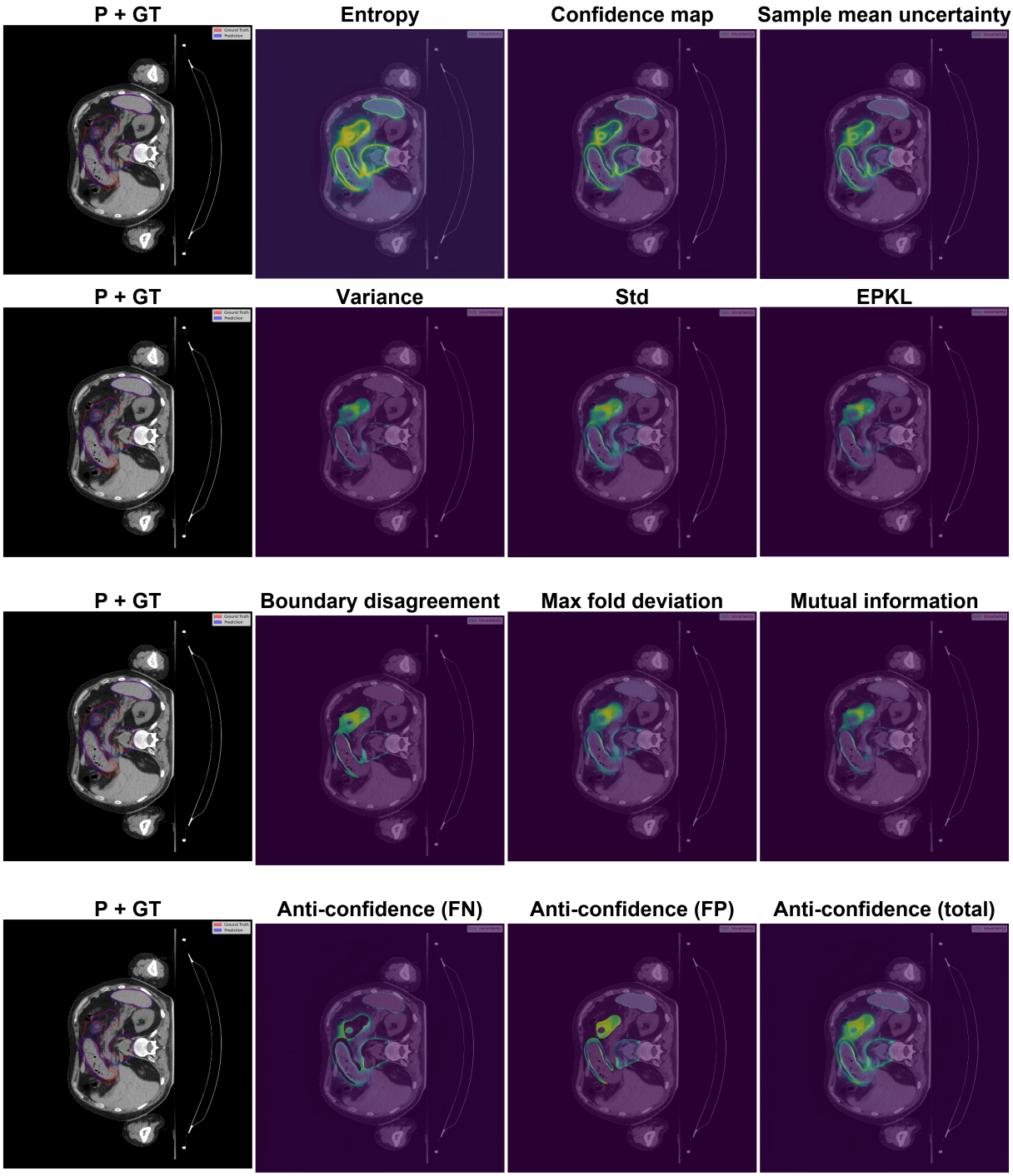

Figure 5.1: Comparison example between prediction–ground truth overlay and different uncertainty maps for case AUTOMI_00006 (axial slice 110) computed from a deep ensemble model.

However, visual similarity does not imply conceptual equivalence. Each uncertainty map encodes a distinct notion of model unreliability. Probability-based measures, such as

entropy and confidence, primarily capture ambiguity in the ensemble mean prediction, whereas ensemble-based measures, including variance, mutual information, and related disagreement metrics, reflect epistemic uncertainty arising from disagreement among ensemble members. Decision-oriented measures, such as boundary disagreement and anti-confidence, explicitly focus on instability around the classification threshold and differentiate between false-positive and false-negative risk.

The fact that most uncertainty maps emphasize areas close to the predicted segmentation boundaries is correct and clinically meaningful, however they represent a form of trivial uncertainty: high uncertainty near boundaries is a common and largely unavoidable phenomenon in medical image segmentation, even two experienced doctors may note two different borders on the same CT scan. As a result, raw uncertainty maps tend to be dominated by boundary effects, potentially obscuring more subtle but clinically relevant uncertainty patterns occurring away from the segmentation border. High uncertainty deep inside a predicted structure or well outside it may indicate anomalous model behavior, systematic failure modes, or regions that warrant closer inspection. To address this limitation, we introduce distance-aware uncertainty representations that explicitly account for the voxel's distance from the predicted segmentation boundary. The key idea is to interpret uncertainty not in absolute terms, but relative to what is expected at a given spatial location.

5.1.2. Distance-Aware Expected Uncertainty

These novel measures incorporate geometric context by conditioning uncertainty estimates on the voxel's distance to the predicted segmentation boundary. In this thesis, we decided to experiment with this approach by focusing on entropy, as it is a well-known, widely used type of uncertainty in many fields and can be computed with the base model too, but any type of voxel-wise uncertainty measure can be used.

Distance-to-Boundary Concept

Given a predicted binary segmentation mask M, the signed distance to the boundary is defined voxel-wise as

$$d_i = \begin{cases} +\operatorname{dist}(i, \partial M), & \text{if } i \in M \quad \text{(inside)}, \\ -\operatorname{dist}(i, \partial M), & \text{if } i \notin M \quad \text{(outside)}, \end{cases} \tag{5.8}$$

where ∂M denotes the set of voxels lying on the predicted structure boundary, typically identified via morphological erosion, and $\operatorname{dist}(\cdot, \cdot)$ represents the Euclidean distance com-

puted in physical space (e.g., millimeters) to account for anisotropic voxel spacing. By definition, voxels on the boundary itself satisfy $d_i = 0$.

Baseline Computation A baseline is constructed from training and validation data. For each signed distance bin (1 mm resolution), the expected value and percentiles (10th, 50th, 90th, 99th) are computed separately for voxels inside and outside the predicted mask, capturing the asymmetry in uncertainty patterns across the boundary.

More formally, for a distance bin $[d, d+1)$ (measured in millimeters), the baseline expected entropy is computed as

$$\mathbb{E}[H \mid d] = \frac{1}{|V_d|} \sum_{i \in V_d} H_i, \tag{5.9}$$

where V_d denotes the set of voxels whose distance to the boundary falls within the specified bin.

Clinical Motivation In medical image segmentation, predictive uncertainty often varies systematically with distance from the structure boundary. Voxels close to the boundary are inherently more ambiguous due to partial volume effects, noise, and gradual anatomical transitions, whereas interior voxels are generally expected to be more stable. Consequently, uncertainty occurring deep inside a structure or far outside it may indicate anomalous or clinically relevant failure modes. Distance-aware representations explicitly encode this spatial context, enabling a more meaningful interpretation of uncertainty patterns in relation to anatomical boundaries.

Distance-Expected Uncertainty (DEU)

Distance-Expected Uncertainty (DEU) quantifies how the observed uncertainty at a given voxel deviates from what is typically expected at the same distance from the predicted segmentation boundary. It is defined as

$$\mathrm{DEU}_i = H_i - \mathbb{E}[H \mid d_i]\,, \tag{5.10}$$

where H_i is the entropy at voxel i, d_i is its signed distance to the boundary, and $\mathbb{E}[H \mid d_i]$ denotes the expected entropy conditioned on that distance saved in the baseline.

Positive DEU values indicate uncertainty that is higher than expected for the given spatial context, while negative values indicate lower than expected uncertainty. As a result, DEU is particularly effective for detecting anomalous uncertainty patterns that deviate from typical segmentation behavior.

Exceedance Based Contextual Uncertainty (EBCU)

A more robust, percentile-based extension of Distance-Expected Uncertainty is provided by the Exceedance Based Contextual Uncertainty (EBCU), which measures how strongly the observed entropy at a voxel exceeds what is considered typical for its distance from the segmentation boundary. Rather than relying on the mean expected entropy, EBCU compares the observed value to high-percentile thresholds derived from baseline data, making it less sensitive to global calibration effects.

Formally, EBCU at voxel i is defined as

$$\mathrm{EBCU}_i = \max\left(0, \frac{H_i - q_{90}(d_i)}{q_{99}(d_i) - q_{90}(d_i) + \epsilon}\right), \tag{5.11}$$

where H_i is the entropy at voxel i, d_i is the signed distance to the boundary, and $q_{90}(d)$ and $q_{99}(d)$ denote the 90th and 99th percentiles of entropy conditioned on distance d, estimated from baseline data. The small constant ϵ prevents numerical instability when the percentile range is narrow. The choice of the 90th percentile $q_{90}(d)$ as the lower exceedance threshold is motivated by the construction of the baseline distributions. The distributions are estimated from predictions on training and validation volumes, *i.e.*, on in-distribution cases that the model has already seen or closely resembles. This results in relatively low entropy levels and, consequently, in an optimistic baseline for the expected uncertainty at each distance d. Selecting an high percentile such as $q_{90}(d)$ ensures that EBCU is not triggered by typical uncertainty values encoded in this optimistic baseline, but only by outlier voxels whose entropy lies in the upper tail of the distance-conditioned distribution.

By construction, EBCU yields values in the range $[0, 1]$. Values close to zero indicate that the observed entropy lies at or below the 90th percentile for the given distance and is therefore consistent with normal uncertainty patterns. Values approaching one indicate entropy levels at or beyond the 99th percentile, corresponding to highly atypical and potentially problematic uncertainty. Intermediate values linearly interpolate the degree of exceedance, providing a normalized and interpretable anomaly score.

From a practical and clinical perspective, EBCU offers several advantages. Because it relies on empirical percentiles rather than absolute entropy values, it is robust to systematic overconfidence or underconfidence of the model. Its normalized scale facilitates visualization and threshold-based filtering, while the explicit conditioning on distance naturally focuses attention on spatial regions that are most critical for segmentation quality, particularly near anatomical boundaries. In practice, high EBCU values in boundary-adjacent

regions often correlate with genuine segmentation errors, making this measure especially useful for quality control and prioritization of expert review.

Distance-aware uncertainty maps

Building on these distance-aware formulations, we derive corresponding voxel-wise uncertainty maps by evaluating the baseline profiles and exceedance scores for each voxel in a predicted CT volume. These maps retain the original spatial resolution of the segmentation and can be visualized as overlays on the CT images, providing a spatially contextualized view of model uncertainty that highlights deviations from the expected distance–uncertainty relationship.

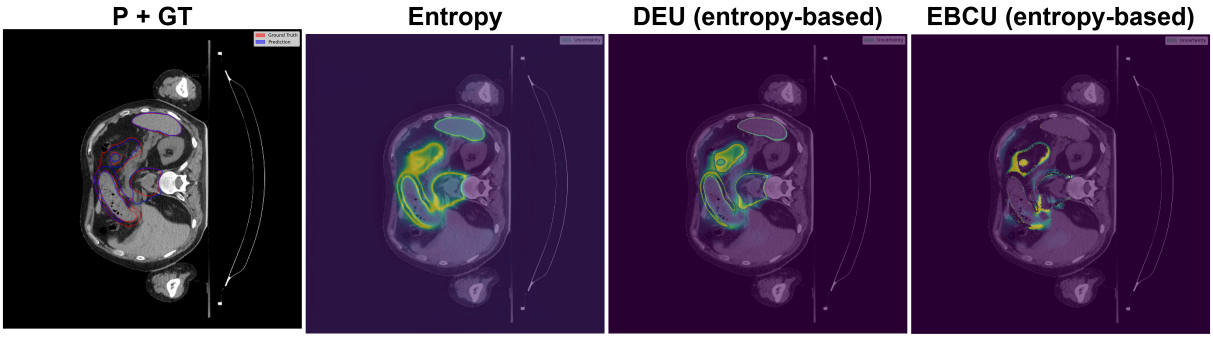

Figure 5.2: Comparison between prediction–ground truth overlay and different distance-aware uncertainty maps for case AUTOMI_00006 (axial slice 110) computed from a deep ensemble model.

By construction, these distance-aware representations suppress trivial boundary-related uncertainty and emphasize atypical uncertainty patterns that are more likely to correspond to clinically relevant errors. Figure 5.2 illustrates how DEU redistributes uncertainty relative to raw entropy, while EBCU isolates sparse, high-confidence regions of abnormal uncertainty that warrant closer clinical inspection.

While the distance-aware uncertainty maps introduced in this section provide a more informative and structured representation of model uncertainty, they remain maps covering the entire volume, even when insignificant in a clinical setting, and they make the display very messy. To enable practical use, uncertainty values must be transformed into local or sparse indicators that clearly identify regions requiring closer inspection. This requires the definition of an appropriate uncertainty threshold, whose selection cannot be arbitrary and must reflect the intended goal of localizing segmentation errors. The following section therefore focuses on the calibration of uncertainty thresholds, with the

objective of maximizing the spatial overlap between high-uncertainty regions and actual segmentation errors.

5.2. Threshold Calibration

For uncertainty to be clinically actionable, the continuous maps must be converted into indicators that identify regions requiring closer inspection. This conversion requires the definition of an uncertainty threshold, above which voxels are flagged as uncertain. Since the scale and distribution of uncertainty values differ substantially across measures, a dedicated calibration procedure is performed independently for each uncertainty map.
For a given uncertainty measure, the threshold is selected by optimizing its ability to localize actual segmentation errors. To this end, uncertainty values are evaluated within a predefined region of interest (ROI), restricting the calibration process to spatial locations that are relevant for error analysis and avoiding dilution effects caused by large background areas. The ROI is calculated in the same way as the one described in section 4.3 and shown on picture 4.1.

Segmentation errors are defined voxel-wise as the union of false positives and false negatives with respect to the ground truth segmentation. For a candidate threshold τ, an uncertainty map is binarized by marking all voxels with uncertainty values greater than or equal to τ as uncertain. The quality of this binary uncertainty mask is then assessed by measuring its spatial overlap with the error mask. Specifically, an overlap-based objective function is used to quantify how well regions of high uncertainty coincide with actual segmentation errors.

The optimal threshold for a given uncertainty map is defined as the value that maximizes this uncertainty–error overlap. This procedure favors thresholds that simultaneously capture a large fraction of true errors while minimizing the inclusion of correctly segmented regions, thereby aligning the calibration objective with the intended clinical use of uncertainty maps as error localization tools.

Formally, let $U_i(\mathbf{x})$ denote the normalized uncertainty value at voxel $\mathbf{x}$ for case i, and let $E_i(\mathbf{x})$ be the binary error mask identifying false positives and false negatives with respect to the ground truth. Given a candidate threshold τ, the uncertainty map is binarized as

$$\hat{U}_i^{\tau}(\mathbf{x}) = 1[U_i(\mathbf{x}) \geq \tau]\,,$$

where $1[\cdot]$ denotes the indicator function. The spatial agreement between the binarized

uncertainty map and the error mask is quantified using an overlap-based score defined as

$$\mathrm{UEO}_i(\tau) = \frac{2\,|\hat{U}_i^\tau \cap E_i|}{|\hat{U}_i^\tau| + |E_i|},$$

which corresponds to a Dice coefficient measuring the overlap between high-uncertainty regions and actual segmentation errors. The threshold calibration objective is defined as the maximization of the mean overlap score across cases,

$$\tau^* = \arg\max_\tau \; \frac{1}{|\mathcal{I}_\tau|} \sum_{i \in \mathcal{I}_\tau} \mathrm{UEO}_i(\tau),$$

where $\mathcal{I}_\tau$ denotes the subset of cases for which the overlap score is well-defined. This formulation ensures that the selected threshold balances sensitivity to true errors with specificity to correctly segmented regions in a rigorous and data-driven manner.

To ensure robustness and prevent overfitting to individual cases, threshold calibration is performed in a leave-one-out fashion across the dataset. For each held-out case, the optimal threshold is estimated using the remaining cases only.

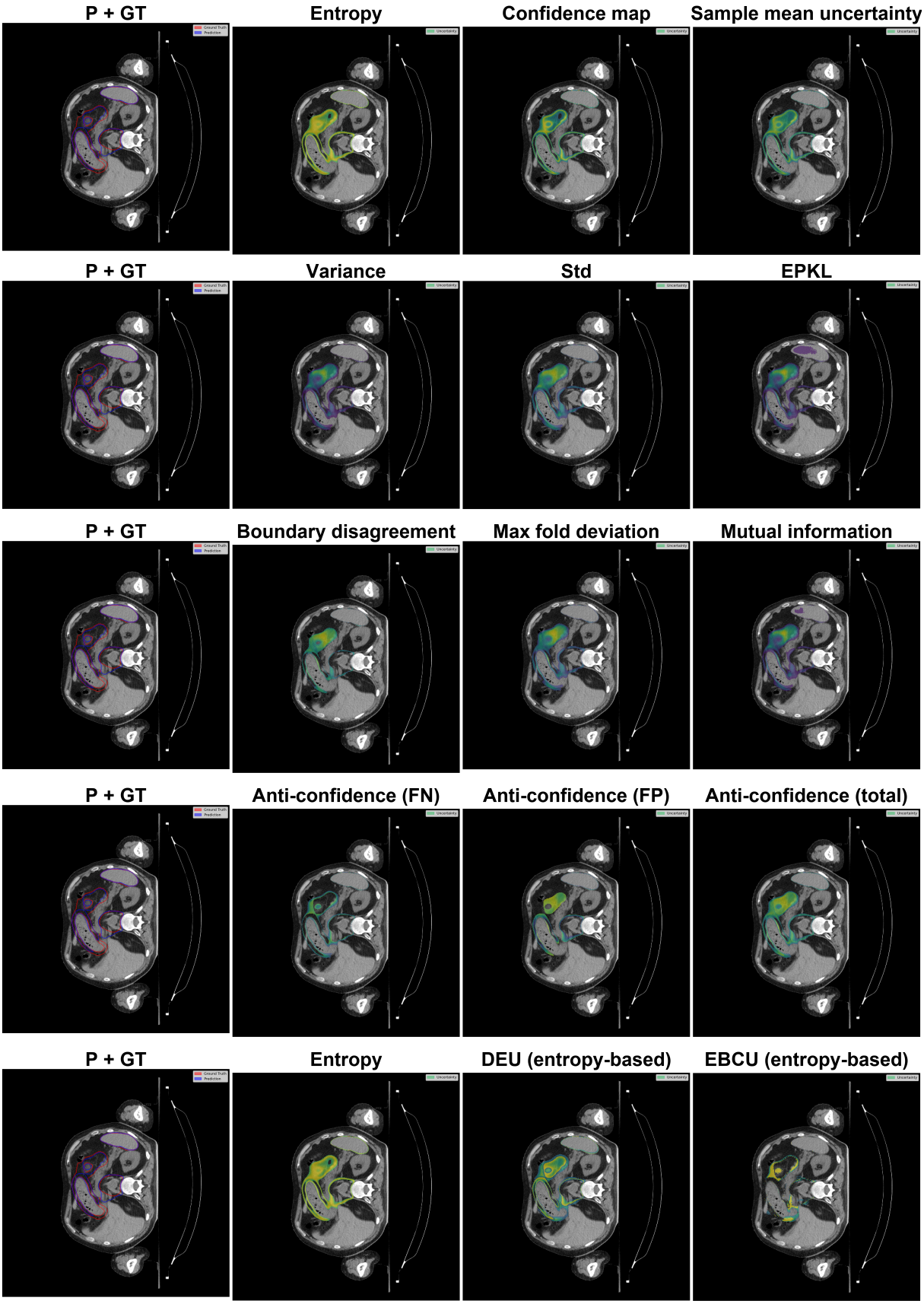

Figure 5.3: Comparison between prediction–ground truth overlay and different uncertainty maps for case AUTOMI_00006 (axial slice 110) computed from a deep ensemble model, with the computed threshold.

After threshold calibration, most base uncertainty maps exhibit similar spatial patterns. Measures such as entropy, variance-based uncertainties, epkl, mutual information, and ensemble disagreement-based uncertainties tend to highlight comparable regions, primarily located around segmentation boundaries and in anatomically ambiguous areas. Once thresholded, their differences in scale and formulation have a limited impact on the qualitative localization of high-uncertainty regions.
In contrast, distance-aware uncertainty maps show a qualitatively different behavior. By explicitly accounting for the expected relationship between uncertainty and distance from the segmentation boundary, Distance-Expected Uncertainty (DEU) reduces the dominance of boundary-related uncertainty and redistributes uncertainty toward regions that deviate from typical spatial patterns. This effect is further amplified in Exceedance-Based Contextual Uncertainty (EBCU), which produces sparse uncertainty masks that suppress most boundary voxels and selectively highlight regions of high uncertainty occurring away from the segmentation border.
Notably, EBCU emphasizes uncertainty in areas where segmentation errors are less expected based purely on spatial proximity to the boundary, making these regions more salient for clinical review. Compared to base uncertainty measures, EBCU provides a clearer separation between trivial boundary uncertainty and atypical, potentially clinically relevant uncertainty.

5.3. Evaluation

After threshold calibration, each uncertainty map U is binarized into an uncertainty mask that flags voxels deemed unreliable. The goal of the evaluation is to measure how well this map acts as an error detector: ideally, most voxels flagged as uncertain should correspond to actual segmentation errors, while correct voxels should remain unflagged. Once again, in order to prevent the large amount of negative background from negatively influencing the measurements and to better focus on the clinically interesting areas, a ROI was taken into consideration in this phase as well, defined as the union of prediction and ground truth masks, extended with a 15 mm buffer in physical space. In addition, we also report an optional "no-border" ROI that excludes a narrow band (2 mm) around the prediction boundary to explicitly remove the trivially uncertain border region, enabling a more stringent assessment of distance-aware measures.

We evaluate a thresholded uncertainty mask $U(t)$ as an error detection tool, i.e., as a binary indicator of which voxels should be considered at risk and prioritized for review. Let FP and FN denote the voxel-wise false positive and false negative masks of the

segmentation, and let $E = FP \cup FN$ be the overall error mask. Furthermore, let $C = \neg E$ be the mask of correct voxels.

Given the overlap between the uncertainty mask and the error masks, we define the following metrics.

recall_FP (coverage of false positive errors):

$$\text{recall}_{FP} = \frac{|FP \cap U|}{|FP|}.$$

This metric answers: what fraction of the segmentation false positives is flagged as uncertain?

recall_FN (coverage of false negative errors):

$$\text{recall}_{FN} = \frac{|FN \cap U|}{|FN|}.$$

This metric answers: what fraction of the segmentation false negatives is flagged as uncertain?

recall_total (overall error recall):

$$\text{recall}_{\text{total}} = \frac{|E \cap U|}{|E|}.$$

This metric measures the fraction of all segmentation errors captured by the uncertainty mask, independent of whether they are false positives or false negatives.

precision (reliability of the uncertainty mask):

$$\text{precision} = \frac{|E \cap U|}{|U|}.$$

Precision answers: among the voxels flagged as uncertain, which fraction are truly segmentation errors?

false_alarm_rate (fraction of correct voxels incorrectly flagged):

$$\text{false\_alarm\_rate} = \frac{|C \cap U|}{|C|}.$$

This metric quantifies the review cost induced by the uncertainty map: high values indicate that many correct voxels are unnecessarily highlighted as unreliable.

To summarize the trade-off between precision and overall error coverage, we report the **F1-score**, computed using precision and $\text{recall}_{\text{total}}$:

$$F_1 = \frac{2\,(\text{precision} * \text{recall}_{\text{total}})}{\text{precision} + \text{recall}_{\text{total}}}.$$

In this work, the F1-score corresponds to the main uncertainty–error overlap objective used to calibrate the threshold: it increases only when the uncertainty mask captures a large fraction of true errors (high $\text{recall}_{\text{total}}$) while avoiding excessive over-flagging (high precision).

5.4. Results

This section reports the quantitative evaluation of the thresholded uncertainty maps across all three model configurations: the Base model, the Deep Ensemble, and the Checkpoint Ensemble. Results are organized into three evaluation sections: the standard ROI (15 mm), the no-border ROI (15 mm excluding a 2 mm boundary band), and a cross-ROI analysis comparing performance across the two settings. For each setting, we assess the ability of each uncertainty formulation to localize segmentation errors, and we examine how the choice of model configuration influences error detection performance. The Base model is limited to probability-based and distance-aware measures (entropy, confidence map, EBCU, DEU), since ensemble-based measures such as variance, mutual information, or anti-confidence require multiple predictions.

5.4.1. Standard ROI (15 mm)

Tables 5.1, 5.2, 5.3, 5.4, 5.5 and 5.6 report the performance of the thresholded uncertainty maps evaluated within the standard ROI for the three model configurations. Figure 5.4 shows the top-10 rankings by metric.

Map (Base)	recall_FP	recall_FN	recall_total
entropy	0.704 ± 0.106	0.533 ± 0.098	0.601 ± 0.094
confidence_map	0.707 ± 0.107	0.537 ± 0.101	0.605 ± 0.096
DEU_entropy	0.792 ± 0.081	0.498 ± 0.103	0.621 ± 0.086
ebcu_entropy	0.107 ± 0.049	0.374 ± 0.105	0.263 ± 0.078

Table 5.1: Recall (mean ± std) for base model, ROI 15mm. The top 3 values are reported in red for each column.

Map (Base)	precision	f1_score	false_alarm_rate
entropy	0.358 ± 0.039	0.442 ± 0.022	0.091 ± 0.020
confidence_map	0.357 ± 0.039	0.442 ± 0.022	0.092 ± 0.020
DEU_entropy	0.288 ± 0.043	0.388 ± 0.032	0.131 ± 0.033
ebcu_entropy	0.201 ± 0.043	0.220 ± 0.039	0.094 ± 0.045

Table 5.2: Precision, F1 Score, False Alarm Rate (mean ± std) for base model, ROI 15mm. The top 3 values are reported in red for each column.

Map (Deep Ensemble)	recall_FP	recall_FN	recall_total
entropy	0.721 ± 0.112	0.557 ± 0.103	0.618 ± 0.098
confidence_map	0.720 ± 0.110	0.557 ± 0.099	0.618 ± 0.096
sample_mean_uncertainty	0.714 ± 0.100	0.572 ± 0.092	0.626 ± 0.088
variance	0.684 ± 0.096	0.615 ± 0.096	0.640 ± 0.090
std	0.682 ± 0.092	0.615 ± 0.092	0.639 ± 0.086
epkl	0.552 ± 0.095	0.526 ± 0.109	0.535 ± 0.100
boundary_disagreement	0.487 ± 0.127	0.398 ± 0.120	0.432 ± 0.119
max_fold_deviation	0.678 ± 0.099	0.614 ± 0.097	0.637 ± 0.092
mutual_information	0.574 ± 0.085	0.544 ± 0.100	0.554 ± 0.090
anti_confidence_FN	0.000 ± 0.000	0.659 ± 0.088	0.392 ± 0.092
anti_confidence_FP	0.814 ± 0.094	0.000 ± 0.000	0.326 ± 0.066
anti_confidence_total	0.711 ± 0.098	0.591 ± 0.088	0.636 ± 0.083
DEU_entropy	0.805 ± 0.089	0.476 ± 0.137	0.608 ± 0.100
ebcu_entropy	0.089 ± 0.051	0.387 ± 0.092	0.268 ± 0.070

Table 5.3: Recall (mean ± std) for deep model, ROI 15mm. The top 3 values are reported in red for each column.

Map (Deep Ensemble)	precision	f1_score	false_alarm_rate
entropy	0.349 ± 0.054	0.437 ± 0.036	0.097 ± 0.035
confidence_map	0.349 ± 0.055	0.437 ± 0.035	0.097 ± 0.036
sample_mean_uncertainty	0.353 ± 0.057	0.443 ± 0.040	0.097 ± 0.037
variance	0.328 ± 0.057	0.426 ± 0.045	0.112 ± 0.045
std	0.328 ± 0.057	0.426 ± 0.045	0.111 ± 0.043
epkl	0.327 ± 0.069	0.394 ± 0.044	0.097 ± 0.044
boundary_disagreement	0.392 ± 0.074	0.393 ± 0.042	0.062 ± 0.043
max_fold_deviation	0.323 ± 0.058	0.421 ± 0.045	0.113 ± 0.040
mutual_information	0.327 ± 0.065	0.402 ± 0.046	0.099 ± 0.042
anti_confidence_FN	0.330 ± 0.056	0.354 ± 0.063	0.064 ± 0.012
anti_confidence_FP	0.311 ± 0.112	0.310 ± 0.081	0.067 ± 0.038
anti_confidence_total	0.346 ± 0.058	0.441 ± 0.044	0.101 ± 0.037
DEU_entropy	0.284 ± 0.062	0.378 ± 0.047	0.131 ± 0.048
ebcu_entropy	0.219 ± 0.054	0.233 ± 0.046	0.086 ± 0.047

Table 5.4: Precision, F1 Score, False Alarm Rate (mean ± std) for deep model, ROI 15mm. The top 3 values are reported in red for each column.

Map (Checkpoint Ensemble)	recall_FP	recall_FN	recall_total
entropy	0.710 ± 0.108	0.559 ± 0.091	0.619 ± 0.089
confidence_map	0.707 ± 0.108	0.557 ± 0.091	0.617 ± 0.089
sample_mean_uncertainty	0.705 ± 0.110	0.566 ± 0.094	0.622 ± 0.092
variance	0.680 ± 0.103	0.606 ± 0.103	0.634 ± 0.096
std	0.679 ± 0.095	0.604 ± 0.096	0.633 ± 0.088
epkl	0.597 ± 0.087	0.563 ± 0.101	0.575 ± 0.089
boundary_disagreement	0.373 ± 0.064	0.305 ± 0.062	0.331 ± 0.054
max_fold_deviation	0.680 ± 0.094	0.608 ± 0.098	0.636 ± 0.090
mutual_information	0.607 ± 0.099	0.571 ± 0.108	0.584 ± 0.098
anti_confidence_FN	0.000 ± 0.000	0.665 ± 0.092	0.379 ± 0.092
anti_confidence_FP	0.808 ± 0.100	0.000 ± 0.000	0.344 ± 0.065
anti_confidence_total	0.702 ± 0.102	0.586 ± 0.090	0.632 ± 0.086
DEU_entropy	0.781 ± 0.105	0.452 ± 0.183	0.595 ± 0.133
ebcu_entropy	0.091 ± 0.044	0.326 ± 0.109	0.229 ± 0.080

Table 5.5: Recall (mean ± std) for checkpoint model, ROI 15mm. The top 3 values are reported in red for each column.

Map (Checkpoint Ensemble)	precision	f1_score	false_alarm_rate
entropy	0.354 ± 0.044	0.444 ± 0.025	0.096 ± 0.029
confidence_map	0.355 ± 0.044	0.444 ± 0.025	0.096 ± 0.029
sample_mean_uncertainty	0.353 ± 0.043	0.443 ± 0.024	0.098 ± 0.030
variance	0.329 ± 0.039	0.427 ± 0.025	0.111 ± 0.033
std	0.329 ± 0.037	0.428 ± 0.025	0.110 ± 0.030
epkl	0.320 ± 0.038	0.405 ± 0.026	0.104 ± 0.027
boundary_disagreement	0.413 ± 0.034	0.363 ± 0.028	0.040 ± 0.012
max_fold_deviation	0.324 ± 0.040	0.424 ± 0.025	0.113 ± 0.032
mutual_information	0.321 ± 0.038	0.408 ± 0.027	0.106 ± 0.030
anti_confidence_FN	0.303 ± 0.054	0.333 ± 0.061	0.072 ± 0.015
anti_confidence_FP	0.327 ± 0.106	0.328 ± 0.076	0.065 ± 0.028
anti_confidence_total	0.344 ± 0.040	0.439 ± 0.024	0.103 ± 0.029
DEU_entropy	0.289 ± 0.065	0.375 ± 0.041	0.132 ± 0.060
ebcu_entropy	0.216 ± 0.053	0.210 ± 0.043	0.077 ± 0.049

Table 5.6: Precision, F1 Score, False Alarm Rate (mean ± std) for checkpoint model, ROI 15mm. The top 3 values are reported in red for each column.

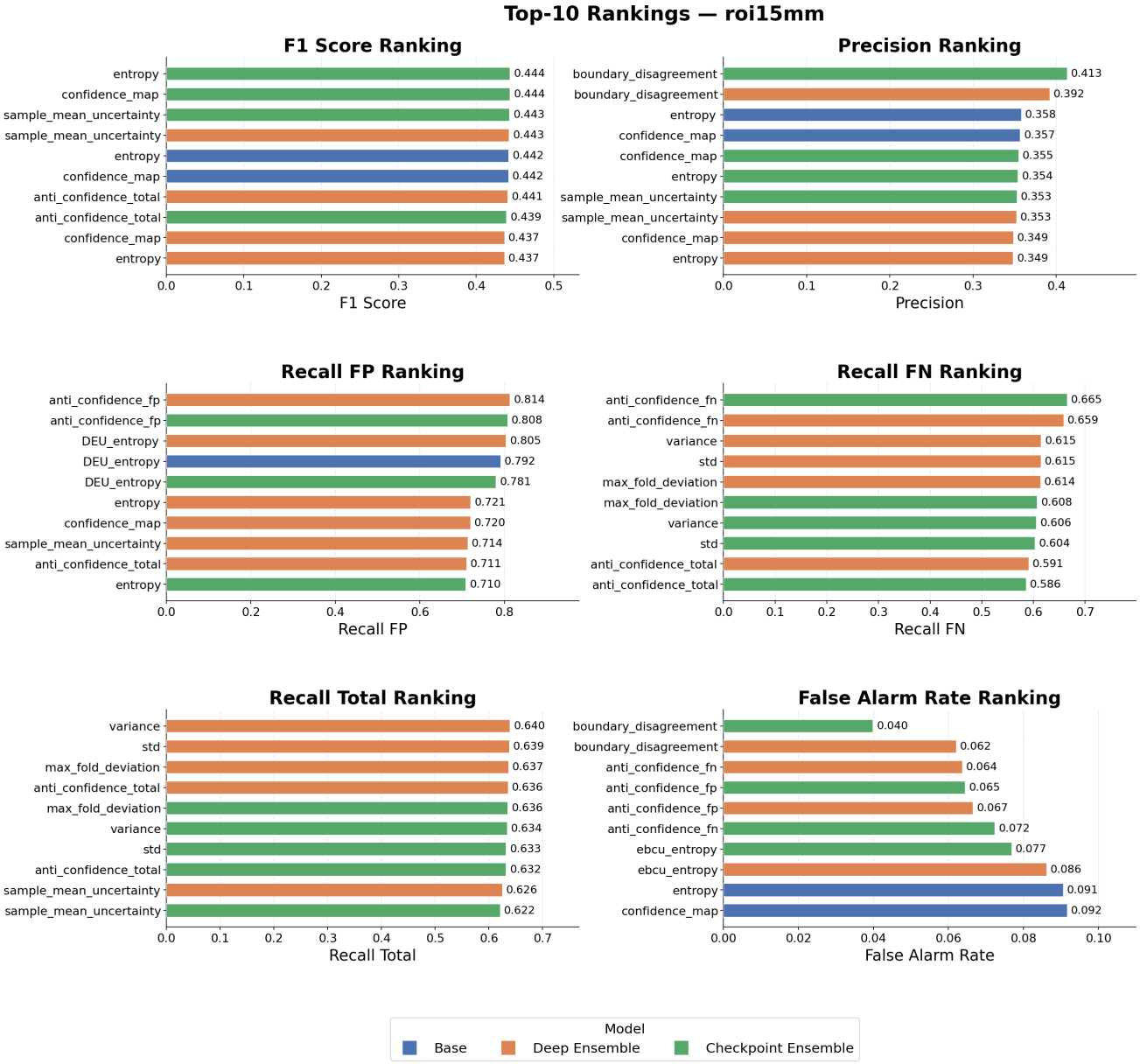

Figure 5.4: Rankings of the different uncertainty maps grouped by metrics for ROI 15mm.

Overall F1-score The top-10 F1 scores fall within a narrow range (0.437–0.444), with no Base model configuration clearly dominating. However, a consistent pattern emerges: entropy and confidence map appear in the top-10 for all three model configurations, including the Base model, which is limited to probability-based measures. This suggests that prediction-level uncertainty, which captures ambiguity in the mean softmax output, provides the most balanced trade-off between error coverage and precision, regardless of the underlying model. Ensemble-specific measures such as anti-confidence total and sample mean uncertainty achieve comparable F1 scores (0.439–0.443) but do not surpass

probability-based measures, indicating that the additional disagreement information does not translate into a better precision–recall balance within this evaluation setting.

Precision and False Alarm Rate Boundary Disagreement is by far the most precise measure, achieving 0.413 (Checkpoint) and 0.392 (Deep Ensemble) substantially higher than any other uncertainty type. It also achieves the lowest false alarm rates (0.040 and 0.062, respectively). However, this selectivity comes at a considerable cost: boundary disagreement achieves a recall total of only 0.331 (Checkpoint) and 0.432 (Deep), making it the most conservative measure in the evaluation. In practice, it flags very few voxels as uncertain, and those it flags are highly likely to correspond to actual errors, but the majority of segmentation errors go undetected. This profile makes boundary disagreement suitable for high-confidence flagging, but insufficient as a standalone error localization tool.

Error Recall The recall rankings reveal the clearest advantage of ensemble models. The Base model is virtually absent from the top-10 in all recall categories, except for DEU entropy in recall FP, where it ranks 4th (0.792) thanks to the distance-aware formulation rather than ensemble diversity. Recall FP. Anti-confidence FP dominates for both ensembles (0.814 Deep, 0.808 Checkpoint), followed by DEU entropy across all three models. These two measures capture complementary mechanisms but yet they both are very effective, DEU in this case compensates for the lack of multiple predictions for the base model. Recall FN. Anti-confidence FN leads both ensemble configurations (0.665 Checkpoint, 0.659 Deep), followed by variance, std, and max fold deviation, all ensemble-based disagreement measures. The Base model's best recall FN is only 0.537 (confidence map), a gap of roughly 12 percentage points compared to the ensembles' best. This difference is particularly relevant from a clinical perspective: false negatives represent missed pathological structures, which are arguably the most dangerous errors and the hardest to identify without explicit uncertainty guidance. Recall Total. The top-10 is entirely populated by ensemble configurations, led by Deep Ensemble variance (0.640) and std (0.639), closely followed by Checkpoint Ensemble equivalents. Anti-confidence total maintains competitive recall (0.636 Deep, 0.632 Checkpoint) by combining FP and FN detection into a single map. The Base model's best recall total (DEU entropy, 0.621) falls approximately 2 percentage points below the ensemble top.

Model Comparison While F1 scores and precision are largely comparable across the three model configurations, the recall analysis shows that ensemble models provide a clear and consistent advantage in error detection, particularly for false negatives. The gap be-

tween Deep Ensemble and Checkpoint Ensemble is small and inconsistent in direction, suggesting that the cheaper Checkpoint Ensemble provides competitive uncertainty estimation. The Base model, despite being limited to probability-based measures, achieves strong F1 scores through high precision, but its lower recall, especially on false negatives, indicates that it misses a larger fraction of clinically relevant errors. Overall, the additional computational cost of training an ensemble appears justified by the improved ability to localize segmentation errors.

5.4.2. No-border ROI (15 mm, excluding 2 mm boundary band)

By excluding voxels within 2 mm of the predicted segmentation boundary, this evaluation setting removes the trivial boundary-related uncertainty that typically dominates all uncertainty maps, and focuses on errors occurring deeper inside or outside the predicted structures. Tables 5.7, 5.8, 5.9, 5.10, 5.11 and 5.12 report the full numerical results. Figure 5.5 shows the top-10 rankings by metric.

Map (Base Model)	recall_FP	recall_FN	recall_total
entropy	0.441 ± 0.132	0.321 ± 0.094	0.360 ± 0.100
confidence_map	0.447 ± 0.134	0.326 ± 0.097	0.365 ± 0.103
DEU_entropy	0.598 ± 0.113	0.532 ± 0.123	0.553 ± 0.110
ebcu_entropy	0.174 ± 0.060	0.455 ± 0.129	0.358 ± 0.101

Table 5.7: Recall (mean ± std) for base model, ROI 15mm (no border 2mm). The top 3 values are reported in red for each column.

Map (Base Model)	precision	f1_score	false_alarm_rate
entropy	0.333 ± 0.050	0.332 ± 0.039	0.039 ± 0.017
confidence_map	0.331 ± 0.050	0.332 ± 0.040	0.040 ± 0.018
DEU_entropy	0.231 ± 0.044	0.317 ± 0.033	0.099 ± 0.034
ebcu_entropy	0.177 ± 0.041	0.227 ± 0.040	0.093 ± 0.047

Table 5.8: Precision, F1 Score, False Alarm Rate (mean ± std) for base model, ROI 15mm (no border 2mm). The top 3 values are reported in red for each column.

Map (Deep Ensemble)	recall_FP	recall_FN	recall_total
entropy	0.475 ± 0.123	0.352 ± 0.103	0.388 ± 0.102
confidence_map	0.470 ± 0.125	0.350 ± 0.100	0.385 ± 0.101
sample_mean_uncertainty	0.473 ± 0.111	0.380 ± 0.101	0.408 ± 0.096
variance	0.525 ± 0.106	0.489 ± 0.112	0.499 ± 0.102
std	0.526 ± 0.103	0.490 ± 0.108	0.500 ± 0.099
epkl	0.430 ± 0.097	0.432 ± 0.121	0.431 ± 0.106
boundary_disagreement	0.278 ± 0.121	0.242 ± 0.116	0.254 ± 0.112
max_fold_deviation	0.527 ± 0.108	0.493 ± 0.113	0.502 ± 0.103
mutual_information	0.446 ± 0.092	0.446 ± 0.114	0.445 ± 0.100
anti_confidence_FN	0.000 ± 0.000	0.503 ± 0.103	0.337 ± 0.100
anti_confidence_FP	0.651 ± 0.117	0.000 ± 0.000	0.212 ± 0.061
anti_confidence_total	0.503 ± 0.104	0.422 ± 0.099	0.447 ± 0.090
DEU_entropy	0.617 ± 0.125	0.535 ± 0.151	0.561 ± 0.134
ebcu_entropy	0.143 ± 0.064	0.490 ± 0.131	0.377 ± 0.106

Table 5.9: Recall (mean ± std) for deep model, ROI 15mm (no border 2mm). The top 3 values are reported in red for each column.

Map (Deep Ensemble)	precision	f1_score	false_alarm_rate
entropy	0.322 ± 0.077	0.334 ± 0.051	0.047 ± 0.035
confidence_map	0.322 ± 0.078	0.332 ± 0.047	0.047 ± 0.035
sample_mean_uncertainty	0.325 ± 0.084	0.345 ± 0.057	0.049 ± 0.036
variance	0.278 ± 0.073	0.344 ± 0.059	0.073 ± 0.044
std	0.277 ± 0.073	0.344 ± 0.059	0.073 ± 0.042
epkl	0.272 ± 0.089	0.316 ± 0.061	0.067 ± 0.042
boundary_disagreement	0.363 ± 0.111	0.269 ± 0.070	0.032 ± 0.036
max_fold_deviation	0.271 ± 0.073	0.339 ± 0.059	0.074 ± 0.040
mutual_information	0.273 ± 0.085	0.323 ± 0.062	0.068 ± 0.041
anti_confidence_FN	0.299 ± 0.067	0.310 ± 0.069	0.039 ± 0.011
anti_confidence_FP	0.254 ± 0.124	0.219 ± 0.076	0.041 ± 0.039
anti_confidence_total	0.305 ± 0.078	0.350 ± 0.061	0.057 ± 0.037
DEU_entropy	0.233 ± 0.062	0.315 ± 0.048	0.101 ± 0.050
ebcu_entropy	0.196 ± 0.049	0.248 ± 0.049	0.086 ± 0.051

Table 5.10: Precision, F1 Score, False Alarm Rate (mean ± std) for deep model, ROI 15mm (no border 2mm). The top 3 values are reported in red for each column.

Map (Checkpoint Ensemble)	recall_FP	recall_FN	recall_total
entropy	0.468 ± 0.132	0.362 ± 0.097	0.396 ± 0.098
confidence_map	0.465 ± 0.132	0.360 ± 0.097	0.393 ± 0.098
sample_mean_uncertainty	0.467 ± 0.134	0.376 ± 0.103	0.405 ± 0.103
variance	0.491 ± 0.114	0.459 ± 0.125	0.468 ± 0.113
std	0.488 ± 0.108	0.456 ± 0.116	0.466 ± 0.104
epkl	0.430 ± 0.091	0.440 ± 0.120	0.436 ± 0.103
boundary_disagreement	0.155 ± 0.051	0.149 ± 0.068	0.150 ± 0.058
max_fold_deviation	0.495 ± 0.114	0.464 ± 0.120	0.473 ± 0.110
mutual_information	0.439 ± 0.109	0.447 ± 0.130	0.443 ± 0.116
anti_confidence_FN	0.000 ± 0.000	0.513 ± 0.113	0.330 ± 0.105
anti_confidence_FP	0.646 ± 0.133	0.000 ± 0.000	0.226 ± 0.060
anti_confidence_total	0.482 ± 0.120	0.411 ± 0.102	0.433 ± 0.097
DEU_entropy	0.587 ± 0.151	0.506 ± 0.191	0.537 ± 0.167
ebcu_entropy	0.147 ± 0.055	0.407 ± 0.151	0.320 ± 0.123

Table 5.11: Recall (mean ± std) for checkpoint model, ROI 15mm (no border 2mm). The top 3 values are reported in red for each column.

Map (Checkpoint Ensemble)	precision	f1_score	false_alarm_rate
entropy	0.316 ± 0.056	0.338 ± 0.030	0.048 ± 0.028
confidence_map	0.317 ± 0.056	0.337 ± 0.031	0.047 ± 0.028
sample_mean_uncertainty	0.312 ± 0.054	0.338 ± 0.030	0.050 ± 0.029
variance	0.271 ± 0.046	0.333 ± 0.033	0.068 ± 0.032
std	0.272 ± 0.046	0.333 ± 0.032	0.068 ± 0.029
epkl	0.255 ± 0.048	0.312 ± 0.034	0.068 ± 0.026
boundary_disagreement	0.363 ± 0.059	0.203 ± 0.045	0.015 ± 0.012
max_fold_deviation	0.266 ± 0.048	0.330 ± 0.031	0.071 ± 0.032
mutual_information	0.257 ± 0.047	0.315 ± 0.035	0.069 ± 0.029
anti_confidence_FN	0.266 ± 0.059	0.288 ± 0.066	0.047 ± 0.014
anti_confidence_FP	0.264 ± 0.119	0.232 ± 0.072	0.040 ± 0.028
anti_confidence_total	0.293 ± 0.049	0.339 ± 0.028	0.057 ± 0.029
DEU_entropy	0.232 ± 0.059	0.304 ± 0.039	0.103 ± 0.060
ebcu_entropy	0.193 ± 0.051	0.226 ± 0.043	0.077 ± 0.053

Table 5.12: Precision, F1 Score, False Alarm Rate (mean ± std) for checkpoint model, ROI 15mm (no border 2mm). The top 3 values are reported in red for each column.

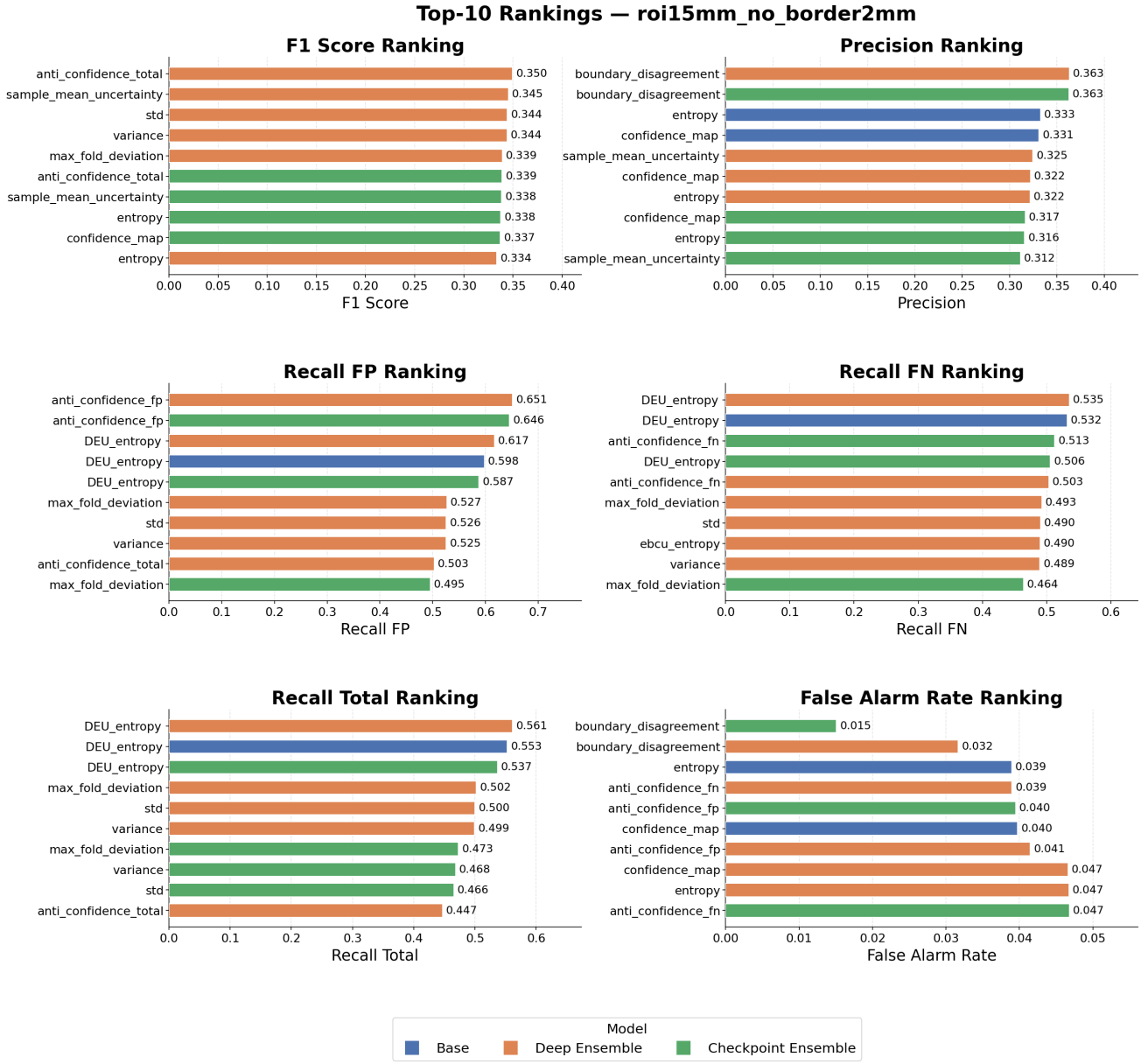

Figure 5.5: Rankings of the different uncertainty maps grouped by metrics for ROI 15mm no border 2mm.

F1 Score This top-10 is populated exclusively by ensemble models, with the Deep Ensemble occupying all top-5 positions: anti-confidence total (0.350), sample mean uncertainty (0.345), std (0.344), variance (0.344), and max fold deviation (0.339). The Checkpoint Ensemble follows with anti-confidence total (0.339) and sample mean uncertainty (0.338). The Base model's best F1, confidence map (0.332) and entropy (0.332), falls just outside the top-10. This indicates that ensemble disagreement becomes more informative when boundary effects are removed and errors are harder to detect from the

mean prediction alone.

Precision and False Alarm Rate Boundary disagreement is the most precise measure (0.363 Deep, 0.363 Checkpoint), closely followed by the Base model's entropy (0.333) and confidence map (0.331). However, boundary disagreement's recall is at the bottom, 0.254 (Deep) and 0.150 (Checkpoint), making it extremely conservative in this setting. Its false alarm rate is also the lowest (0.015 Checkpoint, 0.032 Deep), confirming that the few voxels it flags are almost always actual errors, but the vast majority of non-boundary errors go undetected. This suggests that boundary disagreement, by construction, loses most of its utility when boundary voxels are excluded from the evaluation.

Error Recall Recall FP: anti-confidence FP leads for both ensembles (0.651 Deep, 0.646 Checkpoint), confirming its effectiveness in detecting over-segmentation errors also away from boundaries. DEU entropy ranks 3rd–5th across all three model configurations (0.617 Deep, 0.598 Base, 0.587 Checkpoint), emerging as the strongest non-ensemble-specific measure for false positive detection in this setting. Recall FN: this ranking highlights the key advantage of the distance-aware formulation: DEU entropy achieves the two highest scores (0.535 Deep, 0.532 Base), outperforming even anti-confidence FN (0.513 Checkpoint, 0.503 Deep). The Base model's DEU entropy (0.532) surpasses most ensemble-based measures, confirming that conditioning on boundary distance is particularly effective for identifying missed structures in non-boundary regions. Ensemble disagreement measures (max fold deviation, std, variance) occupy the remaining top-10 positions, maintaining competitive recall FN between 0.464 and 0.493. Recall Total: DEU entropy dominates the top-3 across all model configurations: Deep Ensemble (0.561), Base (0.553), and Checkpoint (0.537). The gap to the next best measure, Deep Ensemble max fold deviation (0.502), is notable ($\sim$5 percentage points), confirming that DEU's distance-aware design provides a decisive advantage when boundary voxels are excluded. The remaining top-10 positions are filled entirely by ensemble disagreement measures (std, variance, max fold deviation), with the Base model appearing only through DEU.

Model Comparison The Deep Ensemble consistently ranks above the Checkpoint Ensemble across all metrics, and the Base model is excluded from the F1 top-10 entirely. The exception is DEU entropy, which enables even the Base model to achieve recall levels competitive with ensemble disagreement measures, confirming that spatial contextualization can partially compensate for the lack of ensemble diversity. Nevertheless, for overall error detection in non-boundary regions, ensemble models, and the Deep Ensemble in particular, provide a measurable and consistent advantage.

5.4.3. Cross-ROI Analysis

To characterize how each uncertainty formulation responds to the removal of boundary-adjacent voxels, we compute the percentage change in every metric when moving from the standard ROI (15 mm) to the no-border ROI (15 mm, excluding 2 mm). Tables 5.13, 5.14, 5.15, 5.16, 5.17 and 5.18 report these percentage changes for all three model configurations; Figure 5.6 provides a scatter-plot comparison of F1-score and recall total across the two settings.

Map (Base Model)	recall_FP	recall_FN	recall_total
entropy	-37.335	-39.835	-40.138
confidence_map	-36.819	-39.361	-39.654
DEU_entropy	-24.453	6.727	-10.950
ebcu_entropy	61.753	21.523	36.101

Table 5.13: Percentage change (%) for 15 mm ROI without 2 mm border compared to 15 mm ROI values (recall metrics). Only improved values are reported in red.

Map (Base Model)	precision	F1	false alarm rate
entropy	-7.134	-25.006	-57.015
confidence_map	-7.268	-24.783	-56.627
DEU_entropy	-19.802	-18.179	-24.283
ebcu_entropy	-11.994	3.580	-0.186

Table 5.14: Percentage change (%) for 15 mm ROI without 2 mm border compared to 15 mm ROI values (precision/F1/FAR). Only improved values are reported in red.

Map (Deep Ensemble)	recall_FP	recall_FN	recall_total
entropy	-34.136	-36.746	-37.165
confidence_map	-34.679	-37.138	-37.603
sample_mean_uncertainty	-33.754	-33.540	-34.696
variance	-23.131	-20.425	-21.909
std	-22.891	-20.220	-21.689
epkl	-22.195	-17.774	-19.412
boundary_disagreement	-42.911	-39.196	-41.053
max_fold_deviation	-22.260	-19.766	-21.142
mutual_information	-22.320	-17.989	-19.637
anti_confidence_FN	0.000	-23.589	-14.057
anti_confidence_FP	-19.988	0.000	-35.010
anti_confidence_total	-29.295	-28.561	-29.758
DEU_entropy	-23.357	12.373	-7.714
ebcu_entropy	61.167	26.674	40.588

Table 5.15: Percentage change (%) for 15 mm ROI without 2 mm border compared to 15 mm ROI values (recall metrics). Only improved values are reported in red.

Map (Deep Ensemble)	precision	F1	false alarm rate
entropy	-7.758	-23.643	-51.745
confidence_map	-7.701	-23.983	-51.910
sample_mean_uncertainty	-7.894	-22.009	-49.351
variance	-15.299	-19.191	-34.953
std	-15.359	-19.131	-34.749
epkl	-16.849	-19.807	-30.342
boundary_disagreement	-7.506	-31.418	-49.023
max_fold_deviation	-16.098	-19.437	-34.031
mutual_information	-16.499	-19.517	-31.147
anti_confidence_FN	-9.348	-12.432	-38.861
anti_confidence_FP	-18.190	-29.444	-37.738
anti_confidence_total	-11.795	-20.762	-43.417
DEU_entropy	-17.796	-16.567	-23.056
ebcu_entropy	-10.908	6.266	0.246

Table 5.16: Percentage change (%) for 15 mm ROI without 2 mm border compared to 15 mm ROI values (precision/F1/FAR). Only improved values are reported in red.

Map (Checkpoint Ensemble)	recall_FP	recall_FN	recall_total
entropy	-34.066	-35.210	-36.052
confidence_map	-34.319	-35.376	-36.257
sample_mean_uncertainty	-33.710	-33.589	-34.827
variance	-27.801	-24.195	-26.154
std	-28.011	-24.429	-26.387
epkl	-27.906	-21.834	-24.288
boundary_disagreement	-58.449	-51.317	-54.590
max_fold_deviation	-27.202	-23.689	-25.587
mutual_information	-27.576	-21.713	-24.099
anti_confidence_FN	0.000	-22.963	-12.955
anti_confidence_FP	-20.124	0.000	-34.330
anti_confidence_total	-31.431	-29.845	-31.451
DEU_entropy	-24.779	11.794	-9.719
ebcu_entropy	62.117	25.016	39.812

Table 5.17: Percentage change (%) for 15 mm ROI without 2 mm border compared to 15 mm ROI values (recall metrics). Only improved values are reported in red.

Map (Checkpoint Ensemble)	precision	F1	false alarm rate
entropy	-10.895	-23.953	-50.362
confidence_map	-10.845	-24.075	-50.557
sample_mean_uncertainty	-11.754	-23.651	-48.884
variance	-17.463	-22.049	-38.303
std	-17.459	-22.014	-38.509
epkl	-20.112	-22.937	-34.496
boundary_disagreement	-12.262	-44.034	-62.209
max_fold_deviation	-18.098	-22.202	-37.166
mutual_information	-19.790	-22.780	-34.730
anti_confidence_FN	-12.189	-13.534	-35.432
anti_confidence_FP	-19.302	-29.266	-38.778
anti_confidence_total	-14.649	-22.888	-44.512
DEU_entropy	-19.928	-18.954	-21.906
ebcu_entropy	-10.585	7.296	-0.083

Table 5.18: Percentage change (%) for 15 mm ROI without 2 mm border compared to 15 mm ROI values (precision/F1/FAR). Only improved values are reported in red.

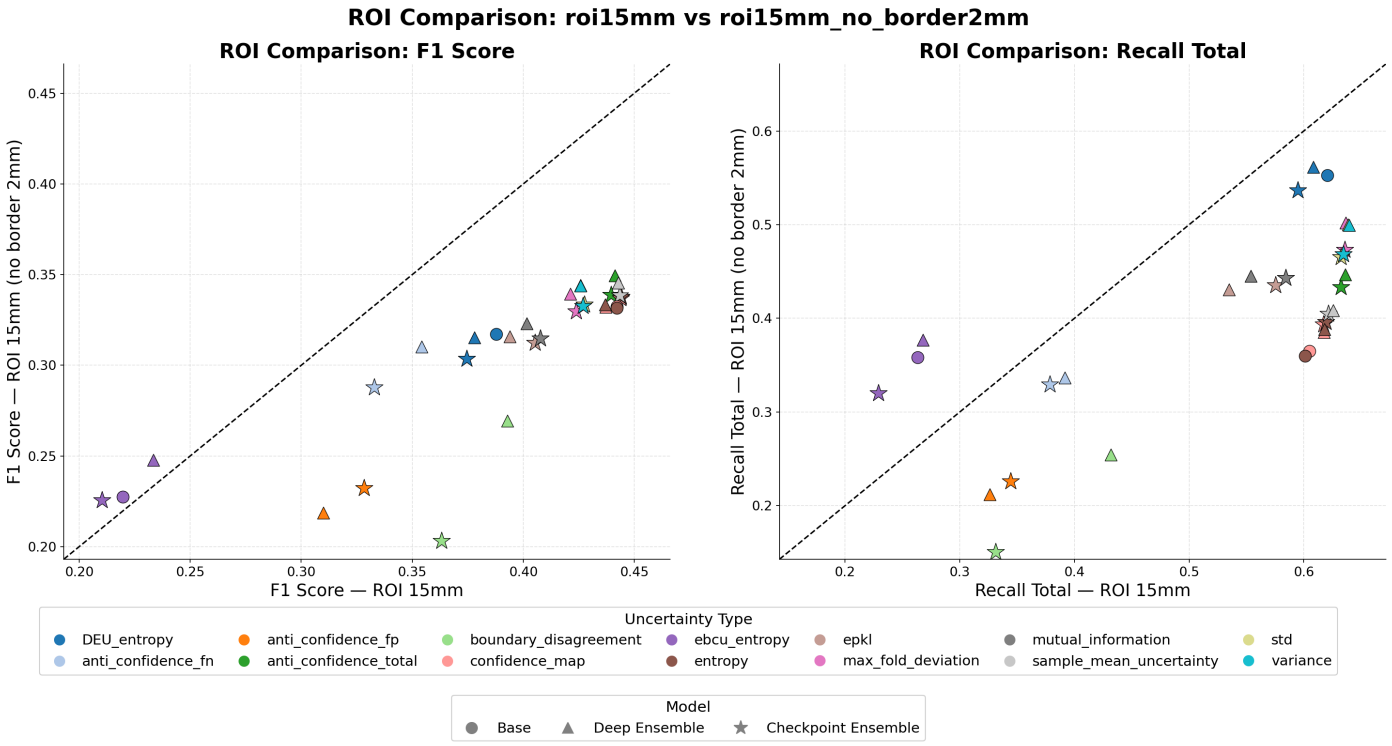

Figure 5.6: Cross-ROI comparison of F1-score (left) and total error recall (right) for the different uncertainty maps and different models. Each point represents one uncertainty formulation in one model setting, plotted by its performance in the standard ROI (x-axis) versus the no-border ROI (y-axis). The dashed diagonal indicates equal performance in both evaluation settings. Points below the diagonal correspond to maps whose performance decreases when boundary-adjacent voxels are excluded, while points above the diagonal indicate improved performance in the no-border ROI.

The dominant trend is a widespread decline in recall: excluding the 2 mm boundary band removes the spatial region where most uncertainty mass concentrates, reducing the fraction of errors that the thresholded maps can capture. However, the magnitude of this decline varies substantially across uncertainty categories, revealing important structural differences in what each measure actually detects. Probability-based measures (entropy, confidence map) suffer on average the largest recall total drops, ranging from -36% to -40% across all models. This confirms that these measures are predominantly driven by boundary ambiguity: once that region is excluded, their ability to localize errors degrades sharply. Ensemble disagreement measures (variance, std, max fold deviation) degrade more moderately. This smaller drop indicates that ensemble disagreement captures a broader spatial distribution of errors, extending further from the boundary than

probability-based uncertainty. Notably, the Deep Ensemble degrades less than the Checkpoint Ensemble, suggesting that the greater diversity of independently trained models produces uncertainty signals that are less boundary-concentrated.

Boundary disagreement exhibits the most severe degradation of all measures. Its recall_total drops by -41.1% (Deep) and -54.6% (Checkpoint), confirming that it is, by construction, almost entirely dependent on boundary-adjacent voxels. In the no-border setting, this measure retains very few flagged voxels, and its already low recall becomes even more limited. The Checkpoint Ensemble is particularly affected, with recall total falling to just 0.150.

The distance-aware formulations exhibit a qualitatively different behavior, confirming the design logic behind their construction. DEU entropy is the most resilient conventional measure. Its recall_total drops by only -7.7% (Deep), -9.7% (Checkpoint), and -10.9% (Base), far less than other measures. More importantly, DEU entropy improves recall FN when moving to the no-border ROI: +12.4% (Deep), +11.8% (Checkpoint), and +6.7% (Base). This asymmetry between FN improvement and FP decline reveals that many false negatives previously missed were located near the boundary. Conversely, the reduction in recall FP (-23% to -25%) indicates that a substantial fraction of false positives flagged under the standard ROI were boundary-driven artifacts. EBCU entropy, even if it gets very low scores, is the only measure to improve across all recall metrics and all models when moving to the no-border ROI. Recall total increases by +40.6% (Deep), +39.8% (Checkpoint), and +36.1% (Base); recall FP surges by over +61% across all configurations, and recall FN increases by +21% to +27%. EBCU also achieves a positive F1 change (+6.3% Deep, +7.3% Checkpoint, +3.6% Base), making it the only measure whose overall error localization quality improves when boundary voxels are excluded. This behavior is fully consistent with its percentile-based construction: EBCU suppresses uncertainty that is expected at a given boundary distance and activates only for outlier voxels. In the standard ROI, boundary-adjacent voxels dilute its effectiveness; once they are removed, the remaining flagged voxels are more precisely aligned with actual errors, but still fails to achieve top 10 scores .

Almost all measures benefit from reduced false alarm rates in the no-border setting, confirming that boundary regions generate a high density of false alarms across all formulations. Probability-based measures see the largest FAR reductions (-51% to -57%), while ensemble disagreement measures improve by -30% to -39%. DEU entropy, despite its strong recall resilience, achieves a more modest FAR improvement (-23% to -24%), reflecting the fact that its higher overall recall comes at the cost of some residual false alarms in non-boundary regions.

A consistent hierarchy emerges across the ROI transition. The base model degrades more than ensemble models for probability-based measures, confirming that ensemble diversity provides a buffer against boundary dependence. Between the two ensembles, the Deep Ensemble degrades less than the Checkpoint Ensemble across most measures and metrics, particularly for variance-based and disagreement-based formulations. This suggests that independently trained models produce more spatially diverse uncertainty patterns compared to checkpoints from a single training trajectory. However, distance-aware measures partially equalize the models: DEU and EBCU show similar percentage changes across all three configurations, indicating that spatial contextualization operates independently of ensemble diversity and can benefit even base-model pipelines.

In summary, the cross ROI analysis shows that the performance of most uncertainty maps is strongly influenced by boundary related uncertainty. When boundary adjacent regions are excluded, their ability to localize errors decreases substantially, revealing a limited sensitivity to non trivial errors away from the segmentation edge. Distance aware formulations exhibit a different behavior: `EBCU_entropy` in particular benefits from boundary exclusion, while `DEU_entropy` partially focuses uncertainty toward false negatives occurring farther from the boundary.

In practice, an uncertainty map that maximizes quantitative scores may not necessarily be the most useful for an experienced clinician. Distance-aware maps such as DEU and EBCU may be preferred in a clinical setting because, even if they do not flag every erroneous voxel, they highlight the regions that warrant closer inspection without overloading the uncertainty map, unlike standard, non distance-aware maps that emphasize trivial boundary errors and increase visual noise. For this reason, a qualitative evaluation involving expert clinicians, where different uncertainty maps are directly compared in terms of interpretability and usefulness during manual review, would represent a valuable complement to the quantitative analysis presented in this work.

6 | Interactive Viewer

While the previous sections focused on quantitative evaluation of uncertainty maps, numerical metrics alone are insufficient to assess their practical usefulness in a clinical context. Measures such as recall of error voxels or aggregate performance scores do not fully capture how uncertainty is spatially distributed, how visually interpretable it is, or whether it highlights clinically meaningful discrepancies.
To address this limitation, an interactive visualization tool was developed to enable systematic qualitative inspection of segmentation results and associated uncertainty maps. The viewer allows direct comparison between prediction, ground truth, and uncertainty estimates at voxel level, thereby implementing the qualitative evaluation proposed at the end of the previous chapter.

6.1. System Overview

The viewer is designed as an interactive application, implemented in Streamlit [14], for inspection of 3D medical volumes. For each clinical case, it displays:

- The original CT scan in grayscale.
- The ground truth segmentation.
- The model prediction.
- A selected uncertainty map, automatically thresholded using the optimal value estimated for the specific (case, uncertainty type) pair.
- Optional multi-organ segmentations generated with TotalSegmentator to provide additional anatomical context.

The interface supports navigation across axial, coronal, and sagittal planes, as well as slice-wise scrolling within each plane. All visual components are rendered as layered overlays, enabling direct spatial comparison.

When an optimal threshold is available, it is applied automatically. Manual adjustment

of the threshold is disabled, ensuring consistency with the calibration procedure described in the previous chapter and preventing arbitrary user-dependent interpretations.

6.2. Visualization and Interaction Design

The visualization is organized into multiple controllable layers, allowing structured and selective inspection of each component.

Anatomical Background and Navigation. The CT scan constitutes the reference layer. Users can select the viewing plane and scroll through slice indices, enabling localized analysis of segmentation behavior and uncertainty distribution in three-dimensional space.

Segmentation Overlays. Ground truth and prediction are displayed as overlaid masks. This facilitates identification of false positives, false negatives, and boundary discrepancies.

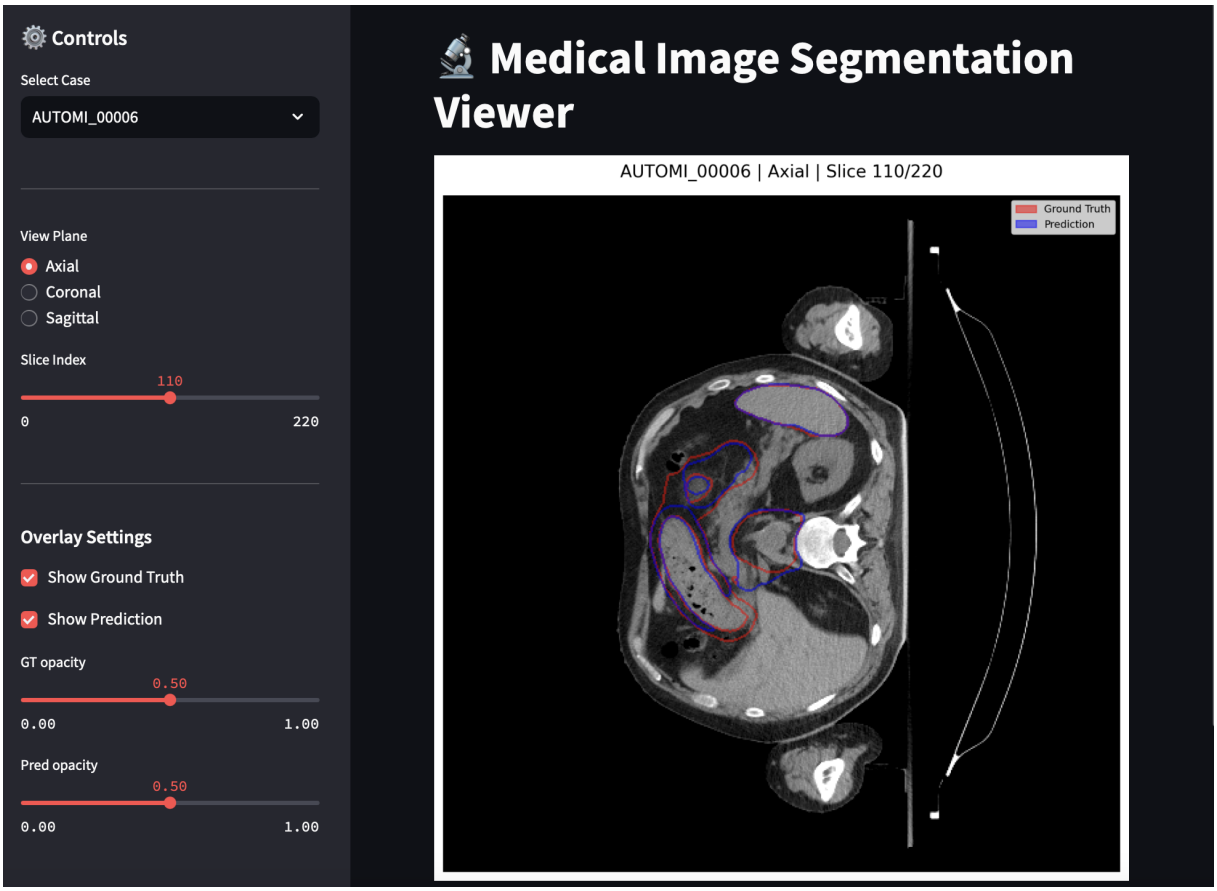

Figure 6.1: Viewer with prediction and GT

Uncertainty Maps. Different uncertainty estimators can be selected for visualization. The chosen map is overlaid on the CT image and displayed after application of its optimal threshold. The resulting highlighted regions emphasize voxels classified as uncertain under calibrated conditions.

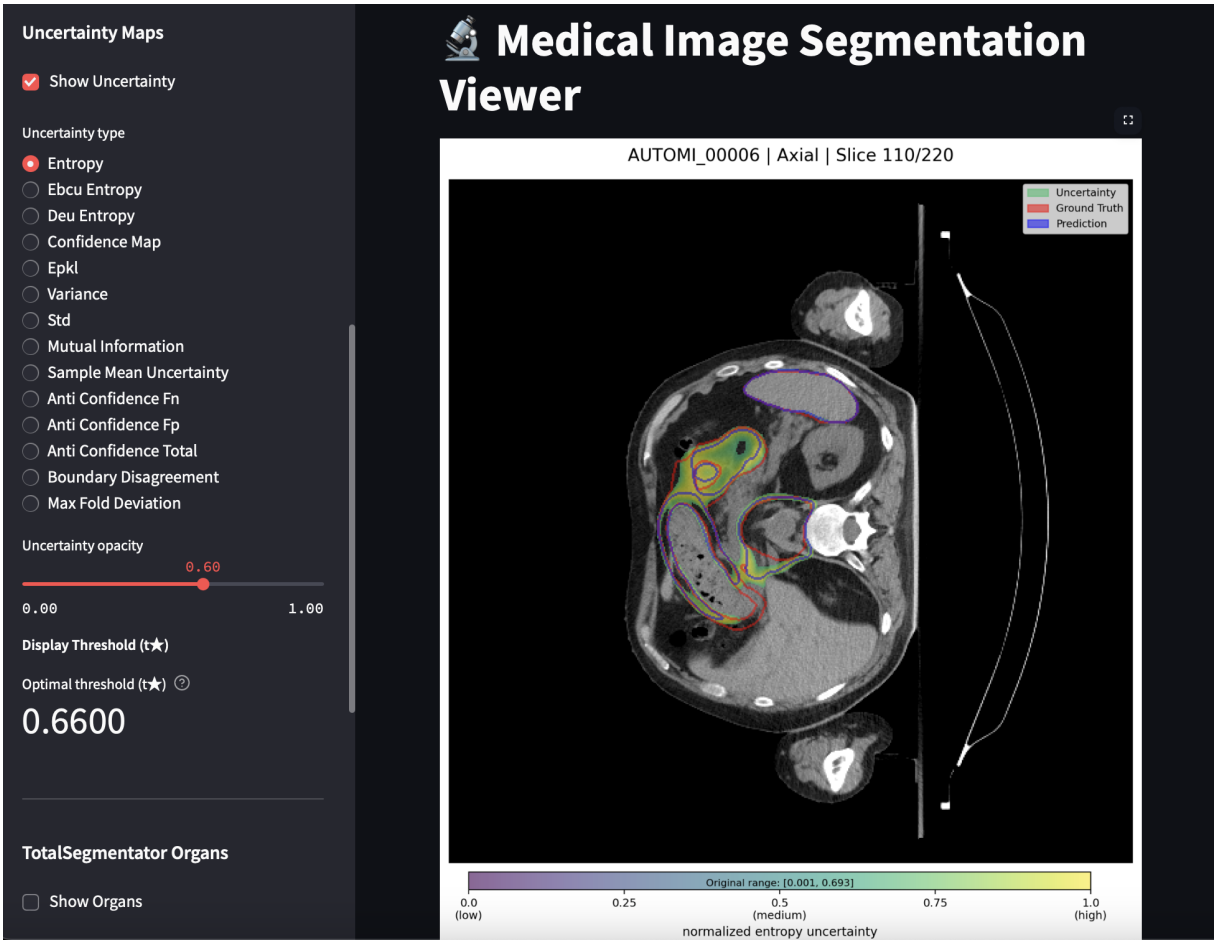

Figure 6.2: Viewer with uncertainty map

Anatomical Context via TotalSegmentator. Segmentations of additional organs, computed by TotalSegmentator [36], can be selectively activated. Each organ can be displayed independently, allowing the user to assess whether uncertainty or segmentation errors correlate with specific anatomical structures or complex boundaries.

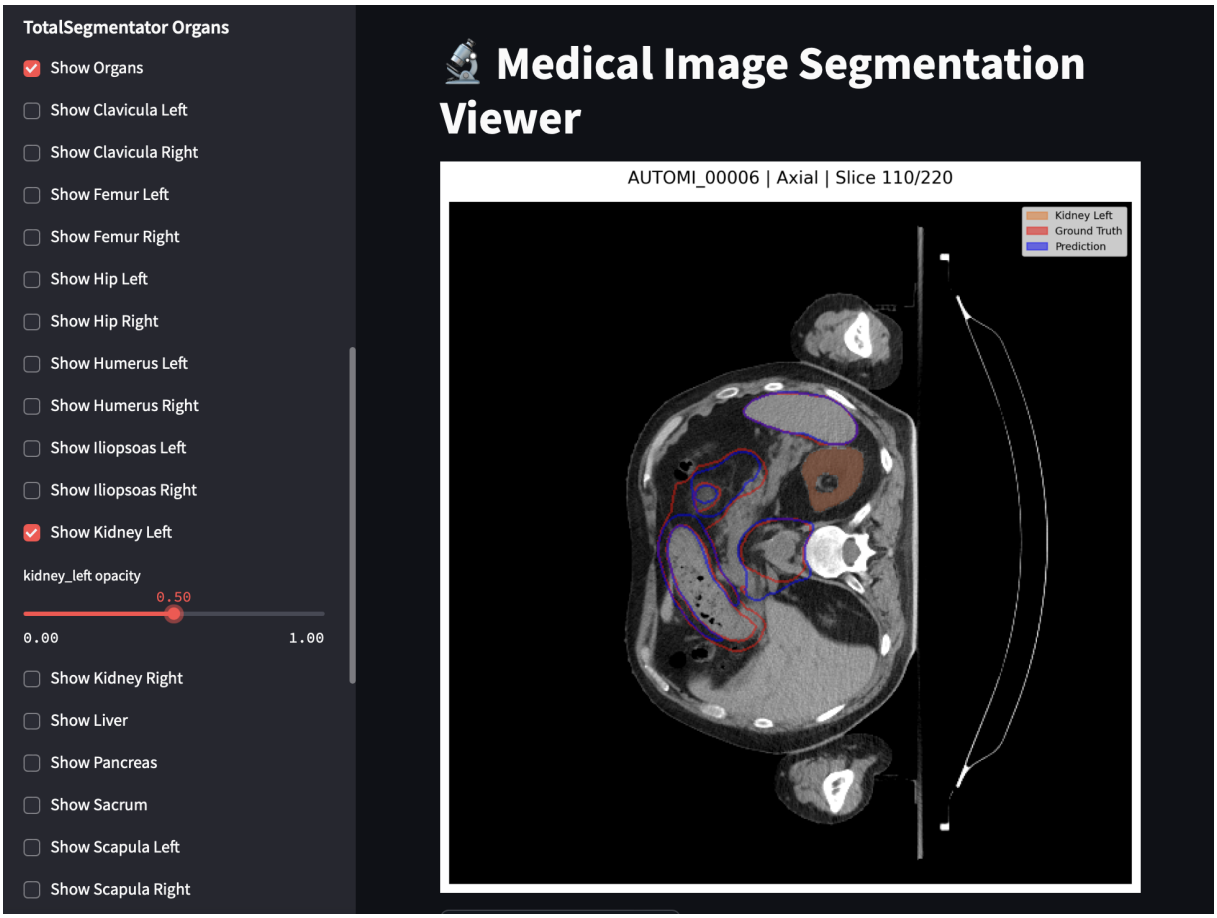

Figure 6.3: Viewer with organ segmentation

Interaction Design. Overall, the interaction design supports controlled exploration. The user can:

- Select the clinical case.
- Switch between planes.
- Scroll through slices.
- Toggle prediction, ground truth, uncertainty maps, and organ segmentations.
- Adjust opacity of each layer.
- Switch between available uncertainty maps.

This structure ensures flexibility in visual inspection while preserving methodological coherence with the quantitative calibration framework.

6.3. Role in the Evaluation Pipeline

The viewer serves as a bridge between numerical performance analysis and clinical interpretability. It enables verification of whether high-performing uncertainty maps effectively localize clinically meaningful segmentation errors and whether distance-aware approaches reduce visual clutter compared to standard maps. Beyond illustrative purposes, the tool provides the foundation for future expert-based qualitative evaluation. By enabling controlled side-by-side comparison of uncertainty strategies, it supports structured assessment of interpretability, usability, and practical relevance in realistic clinical review scenarios. In this way, the viewer complements the quantitative analysis by embedding uncertainty estimation within a clinically oriented visual inspection framework.

7 | Conclusions and future developments

7.1. Summary of Contributions

This thesis investigated uncertainty estimation and confidence calibration for deep learning-based medical image segmentation of lymphatic structures in full-body CT scans, with the goal of making model predictions more transparent and clinically actionable.

A region-of-interest-aware temperature scaling procedure was developed and shown to consistently improve calibration metrics (NLL, ECE, Brier Score) across all model configurations, base, deep ensemble, and checkpoint ensemble, while preserving segmentation accuracy (chapter 4).

A comprehensive set of voxel-wise uncertainty maps was computed within the nnU-Net v2 framework, including both base measures (entropy, ensemble variance, mutual information, boundary disagreement, anti-confidence, SMU) and two novel distance-aware measures, DEU and EBCU, which explicitly account for the expected relationship between uncertainty and distance from the segmentation boundary (chapter 5).

A threshold calibration procedure based on leave-one-out optimisation was introduced to convert continuous uncertainty maps into local maps, bridging the gap between abstract uncertainty values and actionable clinical feedback.

Evaluation under both standard and no-border ROI settings demonstrated that distance-aware measures, particularly DEU-entropy, maintain or improve error detection performance when the trivially uncertain boundary band is excluded, whereas base measures tend to degrade in this more stringent setting.

Finally, an interactive Streamlit-based viewer was developed to support real-time exploration of segmentation results, uncertainty heatmaps, and ground truth overlays on a slice-by-slice basis (chapter 6).

7.2. Future Developments

The present work was conducted on a private dataset of 45 full-body CT scans targeting a single anatomical structure. A natural next step is validation on public and multi-organ benchmarks, which would strengthen the evidence for the generalisability of both the distance-aware uncertainty measures and the ROI-aware calibration strategy across different institutions, scanners, and segmentation tasks.

Similarly, all experiments relied exclusively on nnU-Net v2 as the segmentation backbone. Extending the proposed pipeline to alternative architectures, such as Transformer-based models, would clarify whether the observed benefits of DEU, EBCU, and threshold calibration are architecture-dependent or hold more broadly.

The distance-aware measures DEU and EBCU were instantiated only on top of entropy. Applying the same framework to other base measures such as ensemble variance or mutual information could reveal complementary strengths and further improve error localization performance.

EBCU relies on percentile-based expected uncertainty curves derived from the training set, which may not remain stable under significant distribution changes. Evaluating the robustness of the full pipeline under controlled domain shift scenarios, for example, data acquired with different scanners or protocols, would clarify the practical boundaries of the current approach and motivate adaptive recalibration strategies.

Finally, while the interactive viewer was designed to support clinical workflows, no formal evaluation with medical experts was performed. A structured clinical user study, measuring review time, inter-observer agreement, and error interception rates with and without uncertainty overlays, would provide direct evidence of the practical value of the proposed tools and guide their refinement for real-world deployment.

Bibliography

[1] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/blundell15.html`.

[2] M. Chua, D. Kim, J. Choi, N. G. Lee, V. Deshpande, J. Schwab, M. H. Lev, R. G. Gonzalez, M. S. Gee, and S. Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6):711–718, 2023.

[3] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen, and D. Visvikis. Current and emerging trends in medical image segmentation with deep learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6):545–569, 2023. doi: 10.1109/TRPMS.2023.3265863.

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

[5] B. Ghoshal, A. Tucker, B. Sanghera, and W. Lup Wong. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence*, 37(2):701–734, 2021.

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

[7] C. J. Holder and M. Shafique. Efficient uncertainty estimation in semantic segmentation via distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3087–3094, October 2021.

[8] L. Huang, A. Miron, K. Hone, and Y. Li. Segmenting medical images: from unet

to res-unet and nnunet. In *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 483–489. IEEE, 2024.

[9] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[10] ImageCare Centers. What is the difference between a ct scan and an mri? `https://imagecarecenters.com/blog/what-is-the-difference-between-a-ct-scan-and-an-mri/`. Accessed: 28 January 2026.

[11] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[12] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.

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

[14] M. Khorasani, M. Abdou, and J. H. Fernández. Web application development with streamlit. *Software Development*, 498:507, 2022.

[15] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.

[16] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein, and E. Burnaev. Domain shift in computer vision models for mri data analysis: an overview. In *Thirteenth International Conference on Machine Vision*, volume 11605, pages 126–133. SPIE, 2021.

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

[18] A. J. Larrazabal, C. Martínez, J. Dolz, and E. Ferrante. Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation.

In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 273–283, Cham, 2023. Springer Nature Switzerland.

[19] P. E. Latham and Y. Roudi. Mutual information. *Scholarpedia*, 4(1):1658, 2009.

[20] H. J. Lee, S. T. Kim, H. Lee, N. Navab, and Y. M. Ro. Efficient ensemble model generation for uncertainty estimation with bayesian approximation in segmentation. *arXiv preprint arXiv:2005.10754*, 2020.

[21] T. Löhr, M. Ingrisch, and E. Hüllermeier. Towards aleatoric and epistemic uncertainty in medical image classification. In *International Conference on Artificial Intelligence in Medicine*, pages 145–155. Springer, 2024.

[22] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic. Revisiting the calibration of modern neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/8420d359404024567b5aefda1231af24-Paper.pdf`.

[23] L. Mondo. Uncertainty estimation in medical image segmentation with deep learning. Master's thesis, Politecnico di Milano, 2025. URL `https://hdl.handle.net/10589/240159`.

[24] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

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

[26] D. L. Pham, C. Xu, and J. L. Prince. Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337, 2000.

[27] N. Posocco and A. Bonnefoy. Estimating expected calibration errors. In *International conference on artificial neural networks*, pages 139–150. Springer, 2021.

[28] V. Raina. Introduction into uncertainty and domain shift. `https://github.com/agaldran/uqinmia-miccai-2023/blob/main/2024/part1_introduction/`

part1_general_introduction_into_uncertainty_and_domain_shift.pdf, 2023. Technical report, accessed January 28, 2026.

[29] K. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest. A review of medical image segmentation algorithms. *EAI Endorsed Transactions on Pervasive Health & Technology*, 7(27), 2021.

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

[31] K. Rufibach. Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, 63(8):938–939, 2010.

[32] N. Salpea, P. Tzouveli, and D. Kollias. Medical image segmentation: A review of modern architectures. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022.

[33] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf.

[34] F. Shi, W. Hu, J. Wu, M. Han, J. Wang, W. Zhang, Q. Zhou, J. Zhou, Y. Wei, Y. Shao, et al. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. *Nature communications*, 13(1):6566, 2022.

[35] C. Wang. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*, 2023.

[36] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5): e230024, 2023.

[37] Wikipedia contributors. Uncertainty quantification. https://en.wikipedia.org/wiki/Uncertainty_quantification, 2026. Accessed: 28 January 2026.

[38] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie. From cnn to transformer: A review of medical image segmentation models. *Journal of Imaging Informatics in Medicine*, 37(4):1529–1547, 2024.

[39] Y. Zhao, C. Yang, A. Schweidtmann, and Q. Tao. Efficient bayesian uncertainty estimation for nnu-net. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 535–544. Springer, 2022.

List of Figures

List of Tables

Acknowledgements

Con questa tesi concludo una fantastica avventura iniziata cinque anni e mezzo fa qui al Politecnico di Milano. Un periodo di formazione lungo e difficile che ha di certo richiesto sacrifici ma che alla fine mi ha dato grandissime soddisfazioni. Ma di certo non ci sarei mai riuscito da solo, e per questo il primo ringraziamento non può che andare alla mia preziosa famiglia che per tutto questo tempo mi ha sostenuto nei miei studi, permettendomi di fare esperienze uniche ed indimenticabili. Per tutto questo mi sento incredibilmente fortunato ed estremamente grato. Grazie mamma, grazie papà, grazie Ale e grazie Tommy.

Ma lontano da casa è facile sentirsi un po' tristi e soli... eppure a me, non è mai successo. E di questo devo ringraziare tutti voi di Gabibbo, un gruppo straordinario su cui ho sempre potuto contare, fatto dalle migliori persone che io potessi mai conoscere in questa università. Studiare insieme a voi ed essere così affiatati mi ha spinto a dare il meglio di me.

Vorrei poi fare una menzione d'onore ai City Boooys e Fra Rivitti, amici con i quali fin da subito ho creato un legame speciale, profondo, un legame che è andato ben oltre le quattro mura universitarie, oltre Milano, oltre l'Italia e incredibilmente anche oltre l'Europa. Un legame che spero vada ancora una volta oltre il tempo e le distanze che ci separeranno.

Vorrei continuare poi ringraziando chi in soli 6 mesi ha reso il mio Erasmus un' esperienza che mi ha cambiato la vita. Uppsala è una città straordinaria ma sarebbe stata niente senza gli incredibili amici che ho trovato. Ho condiviso il mio corridoio con i migliori coinquilini che potessi mai chiedere e le mie avventure quotidiane con un folle gruppo di italiani, unici nel loro genere, con cui ho conquistato la Svezia.

Voglio continuare ringraziando i miei amici di Roma, amici di una vita. Nonostante la distanza la nostra amicizia non si è mai affievolita e ogni volta che c'è la possibilità abbiamo la voglia e il piacere di rincontrarci. Grazie a voi so che ovunque io vada Roma rimane un porto sicuro con persone sempre pronte ad accogliermi a braccia aperte.

Un ringraziamento va anche ad Andrea. Dopo vari traslochi con te ho condiviso molte cene, risate e podcast nel primo luogo, a Milano, che per me è stato degno di chiamarsi

casa.

E infine, ma di certo non in ordine di importanza, voglio ringraziare una persona molto speciale: la mia incredibile ragazza Costanza. Con il tuo inesauribile buon umore, la tua dolcezza contagiosa e lo splendido sorriso sei per me un punto di riferimento costante. Mi hai sostenuto durante tutta la stesura di questa tesi dandomi la forza per impegnarmi e la serenità per superare i momenti più difficili. Mi sento infinitamente fortunato ad averti al mio fianco, ti amo.

Grazie a tutti!