Biomedical Image Segmentation Based on Foundation Models Adapted Without Retraining and With Uncertainty Estimation

Fernando González Salas[a*]

[a]Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Abstract

Two important shortcomings limit the effectiveness of current learning-based solutions for biomedical image segmentation. One major issue is that new segmentation tasks typically demand the training or fine-tuning of new models, a resource-intensive process requiring significant machine learning expertise that is often beyond the reach of medical researchers and clinicians. The second critical limitation is that most existing segmentation methods yield only a single, deterministic segmentation mask, despite the considerable uncertainty often present regarding what constitutes correct segmentation. This uncertainty arises from both inherent data variability (aleatoric) and the model's own knowledge gaps (epistemic). This work specifically addresses the estimation of these uncertainties in the segmentation process. By understanding and quantifying these uncertainties, we can significantly increase the explainability and interpretability of segmentation models, enabling more confident and informed decision-making in vital medical applications. We propose to develop a generalized method to analyze these different uncertainty types without requiring model retraining.

Keywords: *Medical imaging, Segmentation, Artificial intelligence, Machine learning, Deep learning, Monte Carlo Dropout, Uncertainty, Test-Time Augmentation*

1 Introduction

Image segmentation plays a critical role in the medical field, serving as a foundation for numerous applications, including diagnosis, treatment planning, and disease monitoring. Accurate segmentation enables clinicians to identify anatomical structures, delineate pathological regions, and quantify changes over time, offering essential insights that inform clinical decisions. However, the inherent complexity and variability of medical images, combined with the labor-intensive process of annotating them and also the diversity of imaging modalities and target structures, pose significant challenges. These issues are further compounded by privacy regulations and the scarcity of computational capacity, making it difficult to generate the extensive datasets required for training specialized segmentation models.

Segmentation models can be broadly classified into two categories: generalist models and task-specific models. Generalist segmentation models are capable of generalizing across diverse segmentation tasks [1], making them versatile and applicable to a variety of medical imaging problems. However, their performance is often lower than that of task-specific models, which are specialized for a given segmentation task. Task-specific models are typically fine-tuned on specialized datasets and benefit from the expertise of domain knowledge, resulting in higher accuracy. Specialization is often achieved through processes such as transfer learning or fine-tuning, which require extensive computational resources, specialized datasets and significant machine learning expertise.

Despite the potential of task-specific models, the challenges of data availability and computational power remain a major bottleneck. In recent years, **Foundation Models**, which are large pre-trained models capable of adapting to a wide range of tasks with minimal task-specific data, have emerged as a promising alternative. These models offer a strong starting point for various tasks without the need for extensive retraining. By leveraging Foundation Models, it is possible to improve the reliability of segmentation predictions while minimizing the need for specialized datasets and computational resources.

Beyond the challenges of data and computational resources, another critical aspect in biomedical image segmentation is the inherent uncertainty associated with the process. This uncertainty can be broadly categorized into two types: **aleatoric uncertainty** and **epistemic uncertainty**. Aleatoric uncertainty arises from the inherent randomness and variability in the data itself, such as image noise, variations in patient anatomy, or ambiguities in the labeling process. Epistemic uncertainty, on the other hand, reflects the limitations in our model's knowledge or its inability to perfectly represent the underlying data distribution, often due to limited training data or model misspecification.

*Work supervised by Xose Manuel Pardo
Corresponding author: fernando.gonzalez.salas@rai.usc.es
Received: April 21, 2025, **Published:** April 21, 2025

In the domain of image segmentation, uncertainty is typically represented spatially, resulting in a pixel-wise or voxel-wise uncertainty map that aligns with the predicted segmentation mask. These maps provide a visual indication of where the model's prediction is less reliable. The value at each location in the uncertainty map is a scalar score derived from observing variability across multiple predictions for the same input. This variability can be induced through various techniques (e.g., dropout sampling or test-time augmentation).

To compute uncertainty, we use common metrics such as pixel-wise **variance** and **entropy**, defined respectively as:

$$\text{Variance} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}_t - \bar{\mathbf{y}})^2$$

$$\text{Entropy} = -\sum_{c=1}^{C}\bar{p}_c \log_2 \bar{p}_c$$

where $\mathbf{y}_t$ is the predicted probability at a pixel in the $\mathbf{t}$-th prediction, $\bar{\mathbf{y}}$ is the mean probability across $\mathbf{T}$ predictions, and $\bar{\mathbf{p}}_\mathbf{c}$ is the average predicted probability for class $\mathbf{c}$ across those predictions. These metrics provide pixel-wise uncertainty maps that highlight areas of high prediction variability and low confidence.

Given the importance of understanding the reliability of segmentation predictions, this work focuses on the critical aspect of **uncertainty estimation** in biomedical image segmentation, particularly in the context of Foundation Models. We aim to explore and evaluate methods for quantifying both aleatoric and epistemic uncertainty in the predictions of segmentation models. Specifically, the objectives of this work are:

- To develop a method for estimating uncertainty in both generalist (Foundation Models) and task-specific biomedical image segmentation models.
- To enhance segmentation reliability by integrating uncertainty information into the final segmentation output, improving its interpretability and robustness for clinical use.

The remainder of this document is organized as follows. Section 2 reviews related work and presents the main state-of-the-art techniques upon which this study is based. Section 3 introduces the proposed method, including the overall system architecture and the components specifically developed for this work. Section 4 details the experiments carried out and Section 5 discusses the results obtained. Finally, Section 6 outlines the main conclusions and Section 7 gives ideas for possible directions for future research.

2 Related Work

Recent research in uncertainty quantification for image segmentation has taken several complementary approaches to address both aleatoric and epistemic uncertainties. Among the most established strategies are Bayesian methods, which model the posterior distribution of network parameters to account for uncertainty arising from limited data or model misspecification. Valiuddin *et al.* [2] offer a comprehensive review of these techniques in medical image segmentation.

Practical and computationally efficient alternatives include Monte Carlo Dropout and Test-Time Augmentation (TTA). MC Dropout keeps dropout layers active during inference, enabling stochastic predictions with minimal architectural changes [3]. TTA assesses prediction consistency through input transformations [4]. An extension of these ideas is dropout injection, in which dropout layers are added only during inference. Ledda *et al.* [5] show that this enables epistemic uncertainty estimation without retraining, which is particularly useful when annotated data or compute resources are limited.

Architectural innovations have also supported uncertainty-aware segmentation. Yang *et al.* [6] propose a multi-decoder U-Net that quantifies annotation variability via decoder divergence. Similarly, Rakic *et al.* [7] present Tyche, a stochastic in-context learning framework for adaptable segmentation, and Boutillon *et al.* [8] highlight the benefits of regularization with shape priors and adversarial constraints.

Ensemble methods provide another path to improved uncertainty estimates. By aggregating predictions from multiple models, they help reduce variance and enhance calibration, as demonstrated by Jungo and Reyes [9]. Pixel-wise metrics such as variance and entropy are commonly employed to quantify predictive disagreement from multiple stochastic predictions. For example, variance captures dispersion in predicted probabilities, while entropy measures the unpredictability across class probabilities. Both are essential for visualizing and quantifying uncertainty across the segmentation mask.

Proper calibration ensures that predicted uncertainty reflects actual confidence. In this context, Lambert *et al.* [10] and Zou *et al.* [11] offer broad overviews of trustworthy AI in medical imaging, analyzing both aleatoric and epistemic uncertainty estimation strategies.

Parallel to the development of uncertainty estimation techniques, foundational segmentation models have transformed the field with highly adaptable, general-purpose frameworks. Notably, MedSAM [12], CellPose [13], UniverSeg [14], and nnU-Net [15]

have demonstrated strong generalization across imaging modalities. Azad *et al.* [16] provide a thorough survey of these models, emphasizing their role in unifying segmentation pipelines.

Recent trends also explore training-free and prompt-based inference strategies, which aim to combine the flexibility of foundation models with uncertainty-aware predictions. Tang *et al.* [17] and Wu *et al.* [18] exemplify this direction, enabling segmentation tasks without the need for task-specific retraining.

Complementary perspectives are provided by studies on segmentation uncertainty in simulation environments [19], topology-aware estimation [20], and domain adaptation [1], which contribute to the robustness and generalizability of segmentation systems. Finally, Conze *et al.* [21] synthesize these developments, outlining how uncertainty-aware techniques and foundational architectures are converging to enable more reliable and adaptive clinical imaging tools.

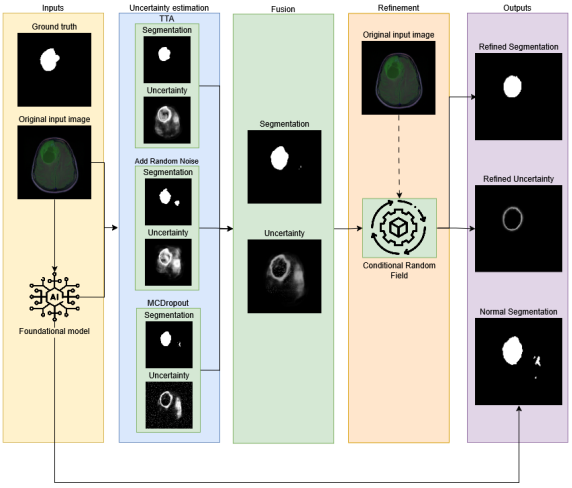

Figure 1: *Illustration of the complete segmentation and uncertainty estimation method. From left to right: input image is processed by a foundational model to obtain initial segmentations and uncertainty maps via stochastic inference (TTA, noise, MC Dropout). These outputs are then fused and refined through a Conditional Random Field (CRF). The green blocks indicate the proposed add-on modules.*

3 Methods

This section describes the proposed methods used to achieve the objectives outlined in the Introduction: adapting pre-trained models for segmentation without fine-tuning and estimating uncertainty in their predictions. We detail the architectural backbones, the stochastic inference techniques used to generate multiple predictions and uncertainty maps, the fusion strategies applied to combine these predictions, and the post-processing steps for refinement. The proposed method is illustrated in Figure 1, depicting the complete method for segmentation and uncertainty estimation. This method integrates stochastic inference techniques, uncertainty-based fusion, and post-processing refinement to enhance prediction robustness and spatial coherence. The green blocks in the diagram highlight the proposed modules added to the base model. The naive inference would use only the yellow block and part of the purple one; as can be seen, the method extends the inference in order to be able to extract the uncertainty estimation.

3.1 Segmentation Backbone and Foundational Models

The method leverages pretrained segmentation models, specifically encoder-decoder architectures such as UNet [22], and advanced foundational models adapted for medical image segmentation, including UniVerSeg [14] and MedSAM [12]. Importantly, the approach does not involve training new models from scratch; instead, it relies on pretrained weights, enabling efficient inference. The UNet model is included for comparative purposes as a representative example of a widely used segmentation architecture, although it is not a foundation model and thus not the primary focus of this study.

The UNet-style backbone utilized follows a symmetric encoder-decoder structure. Its encoder consists of four levels, each doubling the number of feature channels, from 3 channels (input image) up to 512 channels at the bottleneck. The decoder mirrors this structure, progressively reducing channel counts back to a single-channel output, which is typically passed through a sigmoid activation function for binary segmentation. The sigmoid function, defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, maps the output logits to probabilities in the range [0, 1]. While explicitly shown for the UNet here as a common output activation for this architecture in binary tasks, foundational models like UniVerSeg and MedSAM may employ different output mechanisms or post-processing steps tailored to their specific design and pre-training.

In addition, foundational models like UniVerSeg and MedSAM are integrated to capitalize on their capacity for few-shot and zero-shot segmentation, respectively. UniVerSeg uses CrossBlocks to transfer knowledge efficiently, while MedSAM is fine-tuned across diverse imaging modalities, enhancing versatility and clinical applicability.

The full segmentation and uncertainty estimation process is structured into three main stages: **Uncertainty Estimation**, **Fusion**, and **Refinement**, each described below.

3.2 Uncertainty Estimation

The uncertainty estimation stage involves generating multiple segmentation outputs and corresponding uncertainty maps through different stochastic inference techniques. These methods help quantify the model's confidence and reveal areas of ambiguity.

Normal Inference In the normal inference process, a single forward pass is performed through the model. The prediction is obtained directly from the output of the task-specific model $f(x)$, where x is the input image, i.e., $y = f(x)$. No additional hyperparameters are involved. This method does not introduce any variability and therefore does not provide meaningful information regarding model uncertainty, making it less relevant for the objectives of this work.

Monte Carlo (MC) Dropout Monte Carlo (MC) dropout inference enables dropout during inference, allowing the model to sample from its predictive distribution. For each sample i, the model performs a forward pass with dropout enabled with a probability of p: $y_i = f_{dropout}(f(x), p)$. The final prediction is computed as the average over N samples: $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. The key hyperparameters are the dropout probability $p = 0.01$ (configurable) and the number of forward passes N, typically set to 30.

Test-Time Augmentation (TTA) In Test-Time Augmentation (TTA), a series of augmentations is applied to the input image, and predictions are obtained for each augmented version. For each transformation T_i, the prediction is given by $y_i = f(T_i(x))$. The final prediction is the average across all M augmentations: $\bar{y} = \frac{1}{M}\sum_{i=1}^{M} y_i$. Typical augmentations include horizontal flips, scaling (e.g., $\{0.5, 1, 2\}$), and multiplicative intensity adjustments (e.g., $\{0.8, 0.9, 1, 1.1, 1.2\}$). The total number of augmentations M is selected accordingly.

Noisy Inference In noisy inference, Gaussian noise is added to the input image to assess robustness. For each sample, the model performs a forward pass with noise added: $y_i = f(x + \epsilon_i)$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$. The final prediction is computed as: $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Key hyperparameters include the noise standard deviation σ_n, adjusted based on the input image scale, and the number of noisy samples N, usually 30.

3.3 Fusion of Predictions

Following stochastic inference, the multiple segmentation probability maps generated (e.g., from MC Dropout, TTA, or Noisy Inference) are aggregated using uncertainty-aware fusion techniques. Each prediction is weighted based on its associated pixel-wise uncertainty, favoring outputs from iterations or augmentations where the model was more confident. This fusion step aims to produce a single, more robust probability map that incorporates information from all stochastic inferences.

The fused prediction probability map, denoted y_{fusion}, is given by: $y_{fusion} = \sum_{i=1}^{K} w_i^{\text{norm}} y_i$, where y_i is an individual probability map prediction from a stochastic inference pass (e.g., a single MC Dropout pass or a prediction from an augmented image), and w_i^{norm} is the normalized weight based on the uncertainty u_i associated with y_i. The number of fusion components K corresponds to the total number of stochastic inference samples or augmentations used. Several weighting strategies are considered:

- **Inverse weighting**: The individual weight is $w_i = \frac{1}{u_i}$. The normalized weight w_i^{norm} is:

$$w_i^{\text{norm}} = \frac{\frac{1}{u_i}}{\sum_j \frac{1}{u_j}}$$

- **Exponential weighting**: The individual weight is $w_i = \exp(-\lambda u_i)$. The normalized weight w_i^{norm} is:

$$w_i^{\text{norm}} = \frac{\exp(-\lambda u_i)}{\sum_j \exp(-\lambda u_j)}$$

- **Power-law weighting**: The individual weight is $w_i = u_i^{-\alpha}$. The normalized weight w_i^{norm} is:

$$w_i^{\text{norm}} = \frac{u_i^{-\alpha}}{\sum_j u_j^{-\alpha}}$$

Uncertainty thresholding methods, such as Otsu's method, percentile thresholds, or mean-std thresholding, are considered as a potential post-processing step after obtaining the fused probability map y_{fusion}. These methods can be applied to the resulting uncertainty map (derived from the variance or entropy of the predictions y_i) to filter or select high-confidence regions, or applied directly to the fused probability map to convert it into a binary segmentation mask. Otsu's method [23] is a common and simple technique for automatic thresholding, suitable for distinguishing foreground from background based on image intensity histograms, and is considered here for its computational efficiency and common use in image processing pipelines to produce a binary mask from a probability output.

3.4 Refinement via Conditional Random Field

To improve spatial consistency in the segmentation maps and refine boundaries, a Conditional Random

Field (CRF) is applied as a post-processing step. The CRF refines the fused probability map y_{fusion} by encouraging label agreement among neighboring pixels based on their appearance and spatial proximity, thereby smoothing the segmentation and better adhering to image edges.

The refined segmentation y_{crf} is computed as: $y_{\text{crf}} = \text{CRF}(y_{\text{fusion}})$. CRF hyperparameters include spatial dimensions defining the neighborhood size for pairwise potentials, channel dimensions for the appearance features (such as pixel intensities of the input image), the number of iterations (typically between 5 and 10), and the standard deviations for the spatial and bilateral kernels used in the CRF formulation.

4 Experiments

We evaluate the segmentation and uncertainty estimation method using the LGG Segmentation Dataset [24], comprising MRI images annotated with brain tumor segmentation masks. Given that the models are used in a pretrained, zero-retraining manner, the dataset is partitioned into validation (15%) and test (15%) subsets for hyperparameter tuning of post-processing steps (like thresholding or CRF) and final evaluation, respectively. Cases without tumor regions were filtered out, resulting in a final dataset of 372 images used for these validation and testing phases. Preprocessing involves converting images to RGB format, normalizing intensities, transforming images and masks into PyTorch tensors, and optionally resizing or cropping for uniform input dimensions. The segmentation backbone and foundational models (UNet, UniVerSeg, MedSAM) described previously in Section 3 are utilized directly without additional training, capitalizing on pretrained weights to ensure efficiency and generalization. The UNet is included for comparison with foundation models, noting its status as a standard architecture rather than a foundation model itself.

4.1 Evaluation Metrics and Comparative Analysis

To rigorously evaluate the segmentation models and their associated uncertainty estimations, we adopt a comprehensive strategy that combines both quantitative metrics and qualitative analyses. This dual perspective enables a deep understanding of model behavior in terms of spatial accuracy, classification reliability, and probabilistic calibration. Furthermore, we assess how the *use* of uncertainty information, particularly through fusion and refinement, impacts the quality of the final segmentation output, addressing the study objective of enhancing segmentation reliability via uncertainty integration.

Segmentation Metrics. These metrics assess how accurately the predicted segmentation masks align with the ground truth annotations. The most prominent among them include:

Intersection over Union (IoU) — also known as the Jaccard Index — quantifies the overlap between the predicted mask $\hat{\mathbf{Y}}$ and the ground truth $\mathbf{Y}$ as:

$$\text{IoU} = \frac{|\mathbf{Y} \cap \hat{\mathbf{Y}}|}{|\mathbf{Y} \cup \hat{\mathbf{Y}}|}$$

Higher IoU values signify better spatial agreement between prediction and ground truth.

Dice Score is a similarity measure that tends to be more sensitive to small structures, which is especially important in medical imaging tasks:

$$\text{Dice} = \frac{2|\mathbf{Y} \cap \hat{\mathbf{Y}}|}{|\mathbf{Y}| + |\hat{\mathbf{Y}}|}$$

It balances precision and recall and is particularly robust in imbalanced datasets where certain classes (e.g., lesions or tumors) are underrepresented.

Pixel-wise Accuracy computes the ratio of correctly classified pixels over the entire image:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP (True Positives) represents pixels correctly classified as belonging to the positive class, TN (True Negatives) are pixels correctly classified as belonging to the negative class, FP (False Positives) are pixels incorrectly classified as positive, and FN (False Negatives) are pixels incorrectly classified as negative.

While intuitive, accuracy can be misleading when classes are imbalanced, as it may conceal poor performance on minority classes.

Precision and *Recall* offer more nuanced views of model performance for each class c:

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c},$$
$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c},$$

where TP_c denotes true positives for class c, FP_c refers to false positives for class c, and FN_c refers to false negatives for class c.

Precision reflects the proportion of predicted positives for class c that are truly correct, whereas recall indicates the proportion of actual positives of class c that have been successfully identified. These metrics are especially relevant when the model's error type (e.g., over-segmentation or under-segmentation) is of clinical interest.

Uncertainty Metrics. Beyond segmentation accuracy, we assess the reliability of the predicted probability distributions and uncertainty estimations. These metrics are critical in risk-sensitive applications where model confidence informs downstream decisions. While we cannot evaluate "true" uncertainty directly, these metrics allow us to quantify properties of the uncertainty estimates and assess their quality and calibration.

The *Negative Log Likelihood (NLL)* evaluates how well the predicted probabilities match the true labels, penalizing confident but incorrect predictions:

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log p_{i,c},$$

where $y_{i,c}$ is the one-hot encoded ground truth and $p_{i,c}$ the predicted probability for class c at pixel i. Lower values correspond to better probabilistic calibration.

The *Expected Calibration Error (ECE)* quantifies the mismatch between confidence and accuracy by binning predictions and comparing expected and actual accuracies within each bin:

$$\text{ECE} = \sum_{b=1}^{B}\frac{|I_b|}{N}\left|\text{acc}(I_b) - \text{conf}(I_b)\right|,$$

where I_b denotes the indices of predictions in bin b. ECE is useful for detecting systematic overconfidence or underconfidence in model outputs.

The *Brier Score* complements NLL and ECE by measuring the mean squared difference between predicted probabilities and true labels:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - t_i)^2.$$

This metric is bounded and symmetric, offering a smooth estimate of both calibration and confidence sharpness.

Finally, the *Certainty Score* focuses specifically on the model's confidence within clinically relevant regions (e.g., lesion masks). It transforms the uncertainty map into a certainty map and computes the average certainty over foreground pixels:

$$\text{Certainty} = \frac{1}{|T|}\sum_{i\in T}\max\left(0, \min\left(1, 1-u_i\right)\right),$$

where $u_i \in [0,1]$ denotes the uncertainty at pixel i, and T is the set of foreground pixels. If $|T| = 0$, the score is undefined. This measure complements global metrics by offering insight into confidence within high-impact regions.

Comparative Analysis Across Methods

To compare different inference strategies (e.g., deterministic, MC Dropout, TTA, Noisy Inference, uncertainty fusion, and CRF post-processing), we conduct both metric-based evaluations and visual inspections. A key aspect of this comparison is evaluating how the different uncertainty estimation and integration techniques (fusion, thresholding, CRF) affect the *final segmentation performance*, allowing us to determine if utilizing uncertainty information leads to more accurate and reliable segmentation masks, thereby addressing one of the core study objectives.

Quantitative Evaluation. All methods are evaluated using the full suite of segmentation and uncertainty metrics described above. Performance is aggregated across test samples and summarized using descriptive statistics. Additionally, we perform statistical hypothesis testing to assess whether observed differences between methods are significant. Visualization tools such as box plots, violin plots, or confidence intervals are used to illustrate performance variability and robustness.

Qualitative Evaluation. Representative sample images are selected to illustrate the visual characteristics of each method. For each case, we present:

- *Segmentation overlays*: Predicted masks overlaid on input images and ground truth annotations, highlighting areas of agreement and mismatch. This allows visual assessment of segmentation quality and the impact of different methods.
- *Uncertainty maps*: Heatmaps visualizing pixel-wise uncertainty, helping to identify ambiguous or low-confidence regions. This provides insight into where the model is less certain and how uncertainty patterns relate to segmentation errors.

Integrated Discussion and Alignment with Study Objectives. The final stage of the analysis integrates both the quantitative and qualitative findings to provide a coherent interpretation of the model's performance. We examine how uncertainty fusion strategies help mitigate inconsistencies in prediction by aggregating information from multiple inference methods, and we assess the impact of CRF post-processing on enhancing spatial coherence and reducing false positives or boundary errors. This discussion also addresses the specific contexts in which each method performs best, such as cases involving high uncertainty, small-scale structures, or clinically relevant regions. Crucially, the evaluation strategy is directly aligned with the core objectives defined in the introduction: to assess not only segmentation accuracy,

but also the reliability and interpretability of the model's confidence estimates; to understand the effect of uncertainty modeling on robustness and trustworthiness; and to determine the extent to which combining uncertainty-aware methods improves the overall quality and clinical applicability of the predictions. By explicitly linking the evaluation criteria to these guiding goals, we ensure that the analysis remains focused, comprehensive, and directly informative for the broader aims of the study.

5 Results

This section presents the quantitative and qualitative results for the UNet, MedSAM, and UniVerSeg models evaluated using various inference techniques on the LGG Segmentation Dataset. The analysis focuses on how different inference methods, particularly those enabling uncertainty estimation, impact segmentation accuracy, calibration, classification performance, and certainty scores. The results are presented to demonstrate the effectiveness of the proposed framework in estimating uncertainty without retraining and enhancing segmentation reliability by integrating uncertainty information, aligning with the objectives outlined in the Introduction. Both quantitative and qualitative analyses are provided, utilizing a total of 372 images.

5.1 Results for the Trained *U–Net* Model

Quantitative Analysis Metrics from Section 4.1 were applied to the UNet model. Figure 2 presents the distributions for six inference variants: *Normal*, *MC Dropout*, *TTA*, *Noisy*, *Fusion*, and *CRF*.

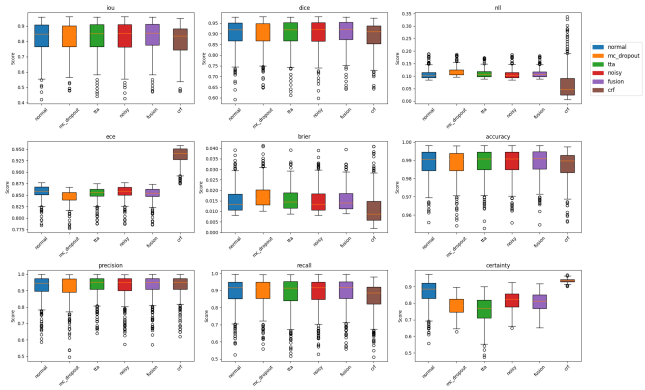

Figure 2: *Box-plot comparison of segmentation and uncertainty metrics for the UNet model on the LGG dataset.*

Segmentation metrics. Median IoU and Dice are virtually identical for Normal, MC Dropout, TTA, Noisy, and Fusion, indicating that these stochastic methods and their fusion do not significantly degrade segmentation overlap compared to the deterministic approach. CRF shows a slightly lower median and a wider IQR, driven by a handful of difficult cases, suggesting potential issues with spatial coherence in some instances.

Calibration metrics. NLL is highest for Normal inference, indicating poor calibration. MC Dropout, TTA, Noisy, and Fusion all significantly reduce NLL, ECE, and Brier Score, demonstrating improved probabilistic calibration through uncertainty estimation. CRF offers a moderate improvement over the baseline Normal inference.

Pixel-wise classification metrics. Accuracy remains near 1.0 for every setting. Precision and Recall display broader spreads, yet stochastic methods do not degrade performance, highlighting their ability to provide uncertainty without sacrificing classification ability. CRF occasionally lowers Recall for fine peripheral structures.

Certainty score. Normal inference yields the highest Certainty inside the tumour mask. MC Dropout, TTA, and Noisy reduce it, reflecting their ability to capture uncertainty. Fusion lies in between, showing a balance between the deterministic and stochastic methods. CRF produces the lowest, reflecting its sharper boundary adjustments and potentially overconfident predictions in some areas.

Summary. Stochastic inference techniques (MC Dropout, TTA, Noisy) and their Fusion markedly enhance calibration without harming overlap metrics for the UNet model. This demonstrates their effectiveness in estimating uncertainty at inference time. CRF improves spatial coherence but offers limited calibration gains and, rarely, a small IoU/Dice drop, suggesting it may not be as effective for uncertainty-aware refinement in all cases.

Qualitative Analysis Qualitative analysis for the U-Net is shown in Figure 3. Normal inference captures tumour extent accurately and with well-defined borders, closely matching the GT. Fusion cleans minor artefacts and precisely highlights boundary ambiguity, demonstrating excellent reliability through uncertainty integration. CRF sharpens contours to an almost perfect circle, but introduces a minor shape deviation from the GT, while its uncertainty map shows a thin, accurate halo along the boundary. Overall, stochastic inference plus Fusion yields an excellent trade-off between visual fidelity and explicit uncertainty representation for the U-Net, aligning with the objective of enhancing reliability with uncertainty.

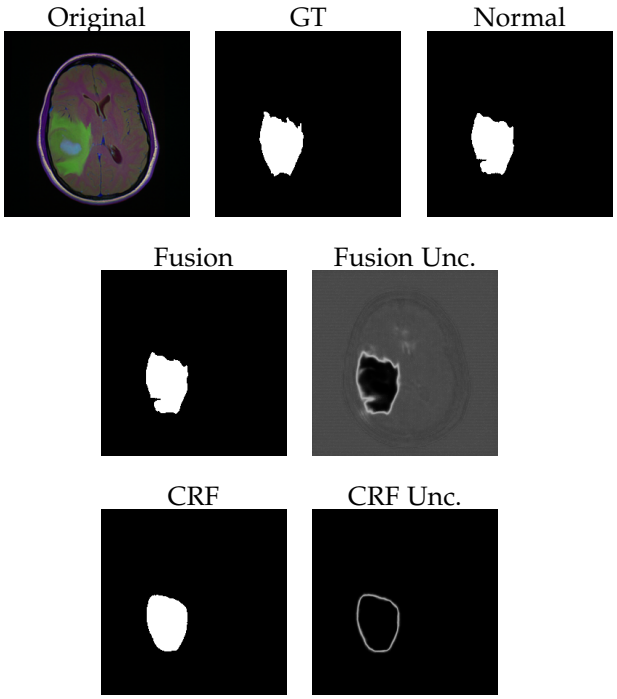

Figure 3: *Qualitative comparison for the U-Net.*

5.2 MedSAM Results

Quantitative Analysis Figure 4 summarises the distributions for IoU, Dice, NLL, ECE, Brier, Accuracy, Precision, Recall, and Certainty for six inference modes: *Normal*, *MC Dropout*, *TTA*, *Noisy*, *Fusion*, and *CRF*, applied to the MedSAM model.

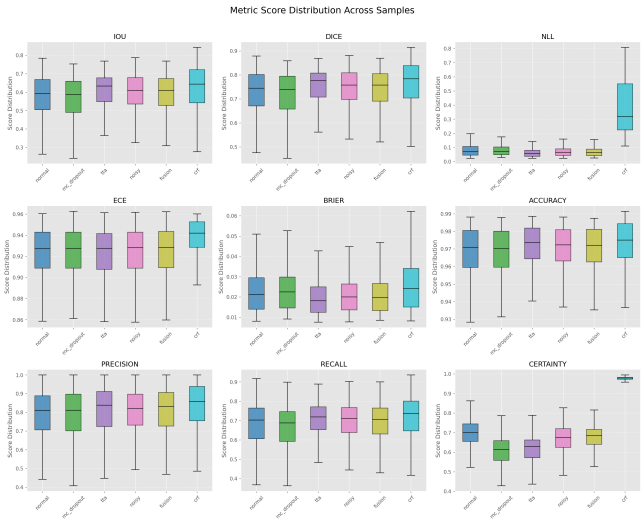

Figure 4: *Metric distributions for MedSAM on the LGG dataset.*

Segmentation. Median IoU/Dice are moderate (∼0.70–0.80) for MedSAM and almost unchanged across Normal, MC Dropout, TTA, Noisy, and Fusion. This indicates that, similar to UNet, stochastic inference and fusion maintain segmentation accuracy. CRF shows a slight dip and a broader IQR, with outliers driving the tail, suggesting it can negatively impact segmentation for some cases.

Calibration. Normal inference yields the highest NLL for MedSAM. MC Dropout, TTA, Noisy, and Fusion markedly lower NLL, ECE, and Brier, demonstrating substantial improvements in calibration through uncertainty estimation techniques. CRF inflates NLL and Brier, indicating mis-calibration relative to all other settings, suggesting it is not suitable for improving calibration with MedSAM.

Classification. Accuracy is consistently high; Precision and Recall vary more but do not deteriorate under stochastic inference. CRF introduces a few low-Recall outliers.

Certainty. Inside the tumour mask, Certainty ranks: Normal > Fusion > (MC Dropout, TTA, Noisy) > CRF. The low Certainty under CRF echoes its degraded calibration metrics, reinforcing that CRF does not effectively represent confidence for MedSAM.

Summary. MedSAM attains only moderate overlap scores compared to UNet, yet MC Dropout, TTA, Noisy, and Fusion substantially improve calibration without harming segmentation accuracy. This highlights the benefit of uncertainty estimation methods for foundational models like MedSAM. CRF delivers little benefit and can even be detrimental to both Certainty and calibration.

Qualitative Analysis Qualitative analysis for MedSAM is shown in Figure 5. Normal inference captures tumour extent but with coarse borders and false positives. Fusion cleans minor artefacts and highlights boundary ambiguity, demonstrating improved reliability through uncertainty integration. CRF sharpens contours but introduces low-confidence halos and sporadic omissions, mirroring its poor Certainty and calibration scores. Overall, stochastic inference plus Fusion yields the best trade-off between visual fidelity and explicit uncertainty representation for MedSAM, aligning with the objective of enhancing reliability with uncertainty.

5.3 UniVerSeg Results

Quantitative Analysis Figure 6 reports the distributions for IoU, Dice, NLL, ECE, Brier, Accuracy, Precision, Recall, and Certainty for six inference modes: *Normal*, *MC Dropout*, *TTA*, *Noisy*, *Fusion*, and *CRF*, applied to the UniVerSeg model.

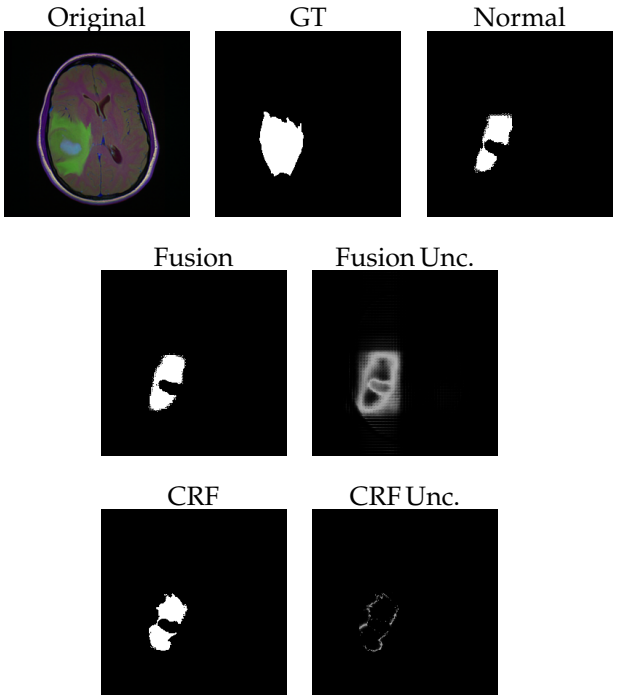

Figure 5: *Qualitative comparison for MedSAM.*

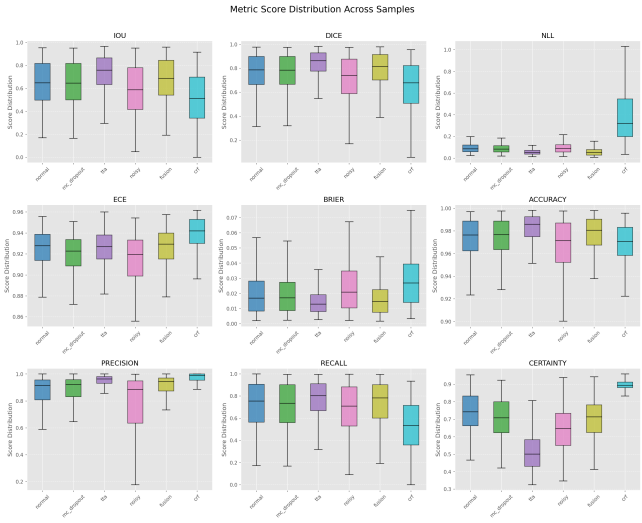

Figure 6: *Metric distributions for UniVerSeg on the LGG dataset.*

Segmentation. UniVerSeg attains the highest median IoU/Dice of the three backbones (∼0.80–0.90), demonstrating its strong raw segmentation performance. Normal, MC Dropout, TTA, Noisy, and Fusion show near-identical medians and IQRs, indicating these methods maintain UniVerSeg's high accuracy. CRF trails slightly and displays a longer lower tail.

Calibration. Normal inference produces the worst NLL for UniVerSeg. MC Dropout, TTA, Noisy, and Fusion all lower NLL, ECE, and Brier, demonstrating improved calibration through uncertainty estimation. CRF again degrades NLL and Brier, mirroring its behaviour with MedSAM and suggesting it's not suitable for calibrating UniVerSeg's outputs.

Classification. Accuracy approaches 1.0 for every mode. Precision and Recall are high but more dispersed; Noisy shows a mild dip, while Fusion recovers the spread. CRF's distribution is comparable, albeit with a few low-end outliers.

Certainty. Certainty ranks: Normal > Fusion > (MC Dropout, TTA, Noisy) > CRF. The pronounced drop for CRF aligns with its poor calibration metrics, similar to MedSAM.

Summary. UniVerSeg delivers the best raw segmentation scores among the evaluated models. MC Dropout, TTA, Noisy, and Fusion further enhance calibration without sacrificing accuracy, demonstrating that uncertainty estimation methods are effective even for high-performing foundational models. Fusion offers the best overall trade-off between accuracy and calibration. CRF yields limited benefit and undermines calibration.

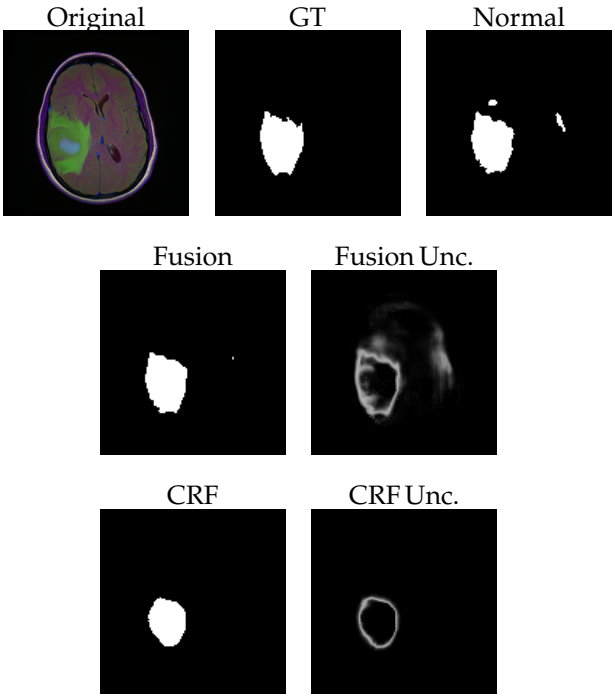

Figure 7: *Qualitative comparison for UniVerSeg.*

Qualitative Analysis Qualitative analysis for UniVerSeg is shown in Figure 7. Normal inference captures tumour extent but with coarse borders and false positives. Fusion cleans minor artefacts and-cribes improved reliability through uncertainty integration. CRF sharpens contours but introduces

low-confidence halos and sporadic omissions, mirroring its poor Certainty and calibration scores. Overall, stochastic inference plus Fusion yields the best trade-off between visual fidelity and explicit uncertainty representation for UniVerSeg, aligning with the objective of enhancing reliability with uncertainty

5.4 Comparative Analysis and Discussion

The results across the three models (UNet, MedSAM, UniVerSeg) and six inference techniques demonstrate consistent trends related to uncertainty estimation and its impact on segmentation reliability.

Stochastic inference methods (MC Dropout, TTA, Noisy) consistently improved the probabilistic calibration of the models, as evidenced by lower NLL, ECE, and Brier scores compared to normal deterministic inference. This highlights their effectiveness in quantifying uncertainty at inference time without requiring model retraining, addressing a key objective of this work. These methods also generally maintained or had minimal impact on segmentation accuracy metrics (IoU, Dice, Accuracy, Precision, Recall) across all models.

The Fusion approach, which integrates predictions based on uncertainty, generally provided a good balance between maintaining segmentation accuracy and improving calibration. This method effectively leverages the uncertainty estimates from multiple stochastic inferences to produce a more robust final segmentation, thereby enhancing reliability. Qualitatively, Fusion tended to produce smoother masks and uncertainty maps that highlighted ambiguous boundary regions, providing valuable insights for interpretation.

The application of CRF as a post-processing step showed mixed results. While it improved spatial coherence and sharpened boundaries in some cases, particularly for UNet, it often degraded calibration metrics and resulted in lower Certainty scores for the foundational models (MedSAM and UniVerSeg). This suggests that the interaction between CRF and the outputs of these advanced models requires further investigation or alternative calibration-aware post-processing techniques may be necessary.

UniVerSeg demonstrated the best overall raw segmentation performance, achieving the highest median IoU and Dice scores. This underscores the potential of foundational models for generalization without task-specific training. Both foundational models, MedSAM and UniVerSeg, benefited significantly from the uncertainty inference techniques in terms of calibration.

In summary, the study successfully demonstrated a method for estimating uncertainty in biomedical image segmentation using methods adaptable without retraining, fulfilling a key objective. The improved calibration offered by MC Dropout, TTA, and Noisy inference, and the balanced performance of the Fusion method, provide a more trustworthy basis for interpreting segmentation results and support more informed decision-making in clinical applications. Future work should explore the integration of uncertainty more deeply and investigate alternative post-processing methods compatible with calibration goals.

6 Conclusions

This work investigated the feasibility and effectiveness of adapting both a standard UNet architecture and advanced foundational models for biomedical image segmentation without requiring model retraining, specifically focusing on incorporating uncertainty estimation. Our findings demonstrate that uncertainty quantification techniques such as MC Dropout, TTA, and perturbing inputs with noise can be successfully applied during inference to provide valuable insights into model confidence, even when the base model was not originally trained with these methods in mind.

The quantitative results consistently showed that while these uncertainty-aware inference methods largely maintained segmentation accuracy metrics at levels comparable to or better than standard deterministic inference across all evaluated models, they significantly improved the probabilistic calibration of the outputs. This was evidenced by substantially lower NLL and improved ECE and Brier Scores, particularly for the UNet and MedSAM models. These methods successfully captured and expressed uncertainty, resulting in lower 'Certainty' scores within the segmented regions compared to standard inference, indicating areas where the model is less confident.

The Fusion approach, which combined predictions from multiple stochastic inferences, generally provided a good balance between segmentation accuracy and improved calibration. It takes advantage of uncertainty quantification to potentially produce more robust final segmentations.

Notably, the application of CRF refinement as a post-processing step showed mixed results. While CRFs are often used to improve spatial coherence in segmentation, our quantitative analysis indicated that applying CRF after obtaining the mean prediction could negatively impact calibration metrics and result in very low Certainty scores for the foundational models on this dataset. This suggests that the interplay between the output characteristics of these advanced models and traditional post-processing methods like CRF needs further investigation, or that alter-

native calibration-aware post-processing techniques are necessary.

UniVerSeg demonstrated strong segmentation performance on the LGG dataset, competitive with or exceeding the task-specific trained UNet, highlighting the potential of foundational models for generalization without task-specific training. MedSAM also showed promising capabilities, and both foundational models benefited significantly from the uncertainty inference techniques in terms of calibration.

Overall, this study successfully developed a framework for estimating uncertainty in biomedical image segmentation using methods adaptable without retraining, addressing a critical need for accessible uncertainty quantification in clinical settings where computational resources and labeled data are often limited. The improved calibration offered by MC Dropout, TTA, and Noisy inference provides a more trustworthy basis for interpreting segmentation results and supports more informed decision-making.

7 Future Work

Building upon the insights and limitations of this study, several promising directions for future research can be identified. One key area involves the exploration of alternative uncertainty quantification methods that can be applied at inference time without the need for retraining. This includes techniques such as ensemble approaches with pre-trained models or novel perturbation strategies, with a focus on comparing their effectiveness and computational demands.

Another important direction is a deeper investigation into the interaction between Conditional Random Fields (CRFs) and calibration performance, particularly for foundational models. Understanding why CRF refinement may degrade calibration could involve analyzing output probability characteristics, testing different CRF parameters or formulations, and exploring alternative post-processing strategies more compatible with calibration goals.

Evaluating the clinical utility of the generated uncertainty maps also holds significant value. Collaborations with clinicians could help determine whether such maps assist in detecting unreliable segmentations, reducing review times, increasing diagnostic confidence, or improving patient outcomes.

To assess the broader applicability of the proposed framework, it is essential to test it across a variety of biomedical segmentation tasks, anatomical regions, and imaging modalities such as CT and ultrasound. This would provide a comprehensive view of its generalizability and robustness.

Research should also delve into more advanced strategies for fusing predictions and integrating uncertainty estimates from multiple methods. Approaches like confidence-weighted averaging or learned fusion mechanisms may lead to more accurate and better-calibrated outputs.

Improving computational efficiency is another priority, particularly for inference methods that require multiple forward passes, such as Monte Carlo Dropout, Test-Time Augmentation (TTA), or noisy inference. Optimizing these techniques is crucial for enabling their use in time-sensitive clinical environments.

Additionally, there is a need to develop uncertainty quantification approaches tailored specifically to the architectures and training schemes of foundational models, which may offer unique opportunities for refinement.

Finally, a detailed analysis correlating regions of high uncertainty with specific segmentation error types—such as boundary inaccuracies, omissions, or false positives—could provide deeper insights into the meaning and utility of uncertainty maps. Collectively, these research directions aim to advance the development of more reliable, interpretable, and clinically effective segmentation models that operate without extensive retraining.

Acknowledge

To see the code of this work access the following link: https://github.com/Fyrthuz/Codigo_TFM.

References

[1] Marco Toldo et al. “Unsupervised Domain Adaptation in Semantic Segmentation: A Review”. In: *Technologies* 8.2 (2020). ISSN: 2227-7080. DOI: 10.3390/technologies8020035. URL: https://www.mdpi.com/2227-7080/8/2/35.

[2] A. Valiuin and Collaborators. “Bayesian Uncertainty Quantification in Medical Image Segmentation”. In: *arXiv preprint arXiv:2411.16370* (2024). Preprint.

[3] J. Doe and A. Smith. “Monte Carlo Dropout and Test-Time Augmentation for Robust Medical Image Segmentation”. In: *arXiv preprint arXiv:2310.06873* (2023). Preprint.

[4] N. Moshkov, B. Mathe, A. Kertesz-Farkas, et al. "Test-Time Augmentation for Deep Learning-Based Cell Segmentation on Microscopy Images". In: *Scientific Reports* 10 (2020). https://doi.org/10.1038/s41598-020-61808-3, p. 5068. DOI: 10.1038/s41598-020-61808-3.

[5] Emanuele Ledda, Giorgio Fumera, and Fabio Roli. "Dropout Injection at Test Time for Post Hoc Uncertainty Quantification in Neural Networks". In: *Information Sciences* 645 (2023). https://doi.org/10.1016/j.ins.2023.119356. DOI: 10.1016/j.ins.2023.119356.

[6] H. Yang et al. "Multi-Decoder U-Net for Uncertainty Quantification in Medical Image Segmentation". In: *Proceedings of MICCAI*. 2023, pp. 450–459.

[7] Marianne Rakic et al. "Tyche: Stochastic In-Context Learning for Medical Image Segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 11159–11173.

[8] Arnaud Boutillon et al. "Multi-Structure Bone Segmentation in Pediatric MR Images with Combined Regularization from Shape Priors and Adversarial Network". In: *Artificial Intelligence in Medicine* 132 (2022). https://doi.org/10.1016/j.artmed.2022.102364. DOI: 10.1016/j.artmed.2022.102364.

[9] J. Jungo and R. Reyes. "Confidence Calibration in Fully Convolutional Networks for Medical Image Segmentation". In: *Medical Image Analysis* 65 (2020), pp. 101–112.

[10] Benjamin Lambert et al. "Trustworthy Clinical AI Solutions: A Unified Review of Uncertainty Quantification in Deep Learning Models for Medical Image Analysis". In: *Artificial Intelligence in Medicine* 150 (2024). https://doi.org/10.1016/j.artmed.2024.102830. DOI: 10.1016/j.artmed.2024.102830.

[11] Ke Zou et al. "A Review of Uncertainty Estimation and Its Application in Medical Imaging". In: *Meta-Radiology* 1.1 (2023). https://doi.org/10.1016/j.metrad.2023.100003. DOI: 10.1016/j.metrad.2023.100003.

[12] Jun Ma et al. "Segment anything in medical images". In: *Nature Communications* 15.1 (Jan. 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-024-44824-z. URL: http://dx.doi.org/10.1038/s41467-024-44824-z.

[13] Carsen Stringer and Marius Pachitariu. "Cellpose3: one-click image restoration for improved cellular segmentation". In: *Nature Methods* 22.3 (2025). https://doi.org/10.1038/s41592-025-02595-5, pp. 592–599. DOI: 10.1038/s41592-025-02595-5.

[14] Victor Ion Butoi et al. *UniverSeg: Universal Medical Image Segmentation*. 2023. arXiv: 2304.06131 [cs.CV]. URL: https://arxiv.org/abs/2304.06131.

[15] Fabian Isensee et al. *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation*. 2018. arXiv: 1809.10486 [cs.CV]. URL: https://arxiv.org/abs/1809.10486.

[16] Bobby Azad et al. *Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision*. 2023. arXiv: 2310.18689 [cs.CV]. URL: https://arxiv.org/abs/2310.18689.

[17] L. Tang, P. T. Jiang, H. Xiao, et al. "Towards Training-Free Open-World Segmentation via Image Prompt Foundation Models". In: *International Journal of Computer Vision* (2024). https://doi.org/10.1007/s11263-024-02185-6. DOI: 10.1007/s11263-024-02185-6.

[18] Junde Wu et al. *One-Prompt to Segment All Medical Image*. 2024. arXiv: 2305.10300 [cs.CV]. URL: https://arxiv.org/abs/2305.10300.

[19] P. Rosenberg and Collaborators. "Impact of Segmentation Uncertainty on Image-Based Simulations". In: *Nature Communications* 12.1 (2021), pp. 1–10.

[20] X. Author. "Topology-Aware Uncertainty Estimation in Image Segmentation". In: *Proceedings of NeurIPS*. 2023.

[21] Pierre-Henri Conze et al. "Current and Emerging Trends in Medical Image Segmentation with Deep Learning". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 7.6 (2023). hal-04075794v2, pp. 545–569. DOI: 10.1109/TRPMS.2023.3265863.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[23] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.

[24] Baoxian Zhou. *Brain MRI Segmentation*. 2024. DOI: 10.21227/pv7k-b062. URL: https://dx.doi.org/10.21227/pv7k-b062.