Testing the Trust: Verification and Validation of Bayesian Segmentation under Uncertainty

Giuseppina Carannante[1] CARANNANG1@ROWAN.EDU
Nidhal C. Bouaynaya[1] BOUAYNAYA@ROWAN.EDU
Dimah Dera[2] UVW@FOO.AC.UK
Hassan M. Fathallah-Shaykh[3] HFSHAYKH@UABMC.EDU
Ghulam Rasool[4] GHULAM.RASOOL@MOFFITT.ORG

[1] *Department of Electrical and Computer Engineering, Rowan University, NJ, USA*
[2] *Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, NY, USA*
[3] *Department of Neurology, University of Alabama at Birmingham School of Medicine, AL, USA*
[4] *Machine Learning Department, Moffit Cancer Center, FL, USA*

Editors: Under Review for MIDL 2026

Abstract

Deep learning has achieved state-of-the-art performance in medical image segmentation, yet safe clinical deployment requires rigorous verification and validation of model robustness, reliability, and uncertainty behavior. Bayesian segmentation methods are often viewed as more trustworthy because they provide uncertainty estimates that can support human decision-making, flag unreliable predictions, and mitigate risks in downstream clinical workflows. However, most prior studies evaluate these models primarily on clean test data, with limited assessment of robustness to perturbations, and without examining whether the predicted uncertainty meaningfully correlates with segmentation quality.

In this work, we conduct a comprehensive and systematic evaluation of state-of-the-art deterministic and Bayesian segmentation models across multiple datasets, corruption types, and performance metrics. Beyond accuracy-based metrics such as DSC and HD95, we analyze over- and under-segmentation trends, predictive variance, and the relationship between uncertainty and segmentation correctness. Our results show that while all models behave similarly on clean or mildly corrupted data, performance diverges significantly as perturbations increase. Models that learn and propagate uncertainty during training consistently provide both improved robustness and more clinically meaningful uncertainty estimates, making them stronger candidates for safe and reliable deployment in medical imaging applications.

Keywords: Image Segmentation, Trustworthiness, Uncertainty, Validation, Verification.

1. Introduction

Medical image segmentation is central to clinical decision-making, guiding diagnosis, treatment planning, and longitudinal monitoring across modalities such as MRI and CT. For these applications, accuracy alone is insufficient (Galati et al., 2022): clinicians must also trust how models behave when confronted with uncertainty, noise, or out-of-distribution (OOD) inputs. Bayesian segmentation models are often considered better suited for safety-critical settings because they provide uncertainty estimates that can support human–AI decision making. Yet, the majority of segmentation models remain deterministic and are evaluated primarily using accuracy-oriented metrics like the Dice Similarity Coefficient (DSC)

(Menze et al., 2014). While these measures quantify overlap, they fail to capture essential aspects of model reliability, including sensitivity to distributional shifts, resilience to noise, and the interpretability of uncertainty maps.

Uncertainty quantification has recently become a central topic in medical image analysis. While numerous studies have proposed Bayesian or ensemble-based segmentation models to estimate uncertainty (Gawlikowski et al., 2023; Goan and Fookes, 2020), most of these works stop at model development, reporting calibration scores or qualitative examples without considering the broader perspective of Verification and Validation (V&V). In other words, they often ask "How accurate is the model?" rather than "How trustworthy is it under real-world variability?".

Clinical AI deployment demands rigorous V&V: Verification: whether we are building the model right, and Validation: whether we are building the right model for real patients. Bayesian segmentation models, by explicitly modeling uncertainty, offer a principled foundation for such an analysis. Despite this increasing interest in uncertainty quantification methods, few studies have examined Bayesian segmentation models through the lens of V&V, focusing instead on developing new methods rather than systematically testing robustness and reliability.

In this work, we approach Bayesian segmentation through the lens of V&V. We systematically evaluate model robustness, uncertainty calibration, and segmentation consistency under diverse noise and perturbation types. Beyond conventional metrics, we include boundary-sensitive measures such as Hausdorff distance (HD) and over-/under-segmentation ratios, emphasizing that trustworthiness in clinical AI extends beyond Dice similarity. Our findings demonstrate that uncertainty-aware Bayesian approaches provide more interpretable and robust behavior under distributional shifts, contributing to the broader goal of verifiable and trustworthy medical AI systems.

2. Related Work

2.1. Verification and Validation in Deep Learning

Deep learning (DL) models have automated numerous tasks in recent years; however, as their applicability increases, so do concerns regarding their trustworthiness and reliability. These issues are inherently related to the V&V process that models should undergo before deployment. An increased research interest in this area has emerged (Huang et al., 2020).

Some authors have proposed using DL itself to support the verification and validation steps, while others have released Python frameworks to automate parts of this process, from data integrity checking to model reliability assessment (Frounchi et al., 2011; Chorev et al., 2022). Although some authors have pointed out the misleading use of the term *validation* to refer only to hyperparameter tuning (Kim et al., 2020), in general, most studies focus on testing model performance under diverse conditions such as distributional shifts, adversarial perturbations, and OOD data (Christin et al., 2021; Javed et al., 2024; Hong et al., 2024).

Building on these general investigations, a subset of research has concentrated on computer vision tasks, in particular, segmentation, where robustness and generalization are especially critical for downstream decision-making. In this context, several studies have examined how segmentation performance degrades under input corruptions (e.g., noise or adversarial attacks) (Kamann and Rother, 2021), while others address the *verification* as-

pect by analyzing how architectural choices influence robustness and performance (Arnab et al., 2018). More recently, these analyses have extended to large foundation models and their sensitivity to input perturbations (Schiappa et al., 2024).

A similar focus is observed in the medical imaging domain, where segmentation accuracy often underpins diagnostic or treatment decisions. Many researchers have emphasized the need to move beyond maximizing accuracy towards evaluating robustness, generalization, and reliability (Galati et al., 2022). Most studies in medical imaging and segmentation examine robustness and generalization under adversarial attacks (Liu et al., 2021), with some exploring how model architecture impacts adversarial robustness (Paschali et al., 2018; Rodriguez et al., 2022).

Recent reviews summarize strategies for improving robustness and generalizability, highlighting key factors such as appropriate statistical analyses, cross-validation strategies, computational complexity, validation with OOD or adversarial samples, and the role of architecture, data quality, and algorithmic design (Tran et al., 2025; Javed et al., 2024). Collectively, these works underscore that achieving reliable and trustworthy AI systems requires a systematic V&V process encompassing robustness, uncertainty, and generalization.

2.2. Bayesian Learning

In Bayesian models, all parameters, i.e., the neural network (NN) weights $\mathcal{W}$, are treated as random variables with a prior distribution $\mathcal{W} \sim p(\mathcal{W})$. Given a training dataset $\mathcal{D} = \{(\mathbf{X}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{n}$, Bayes' theorem allows us to infer the posterior distribution $p(\mathcal{W}|\mathcal{D})$. From this, we can derive the predictive distribution for an unseen input $\mathbf{X}^*$ as: $p(\mathbf{y}^*|\mathbf{X}^*, \mathcal{D}) = \int p(\mathbf{y}^*|\mathbf{X}^*, \mathcal{W})\, p(\mathcal{W}|\mathcal{D})\, d\mathcal{W}$, where $\mathbf{y}^*$ is the associated output. The predictive distribution encapsulates all the information about the model output. Its mean corresponds to the network's prediction, while its predictive variance quantifies the model's uncertainty in that prediction. However, performing exact Bayesian inference in NN is computationally intractable due to the non-linearity of NN and the high dimensionality of the parameter space (Blundell et al., 2015). This has motivated the development of approximation methods for scalable Bayesian learning (Gawlikowski et al., 2023; Goan and Fookes, 2020).

One of the most widely used approximation techniques for Bayesian learning is Variational Inference (VI), which formulates Bayesian inference as an optimization problem (Graves, 2011). In VI, a tractable variational distribution $q_\theta(\mathcal{W})$ is introduced to approximate the true posterior $p(\mathcal{W}|\mathcal{D})$. The optimal parameters θ^* are obtained by minimizing the Kullback–Leibler (KL) divergence between the two distributions, which leads to the Evidence Lower Bound (ELBO) objective: $\mathcal{L}(\theta) = -\mathbb{E}_{q_\theta(\mathcal{W})}\left[\log p(\mathcal{D}|\mathcal{W})\right] + \mathrm{KL}\big(q_\theta(\mathcal{W}) \,\|\, p(\mathcal{W})\big)$. Depending on the choice of prior, the variational family, and the strategy used to approximate the expectations in the ELBO, several practical Bayesian DL formulations have been proposed. For instance, Gal and Ghahramani (2016) demonstrated that applying dropout during training and inference can be interpreted as performing VI. Alternatively, approaches under the umbrella of Variational Density Propagation (VDP) directly propagate both first and second moments through network layers, maintaining explicit representations of mean and covariance (Dera et al., 2021; Carannante et al., 2024). These moment-propagation schemes offer a principled and computationally efficient way to learn predictive uncertainty jointly with the network parameters.

In parallel, non-Bayesian strategies have been explored to capture uncertainty information, such as test-time data augmentation and deep ensembles (Lakshminarayanan et al., 2017; Wang et al., 2019). These approaches, while not grounded in Bayesian theory, have shown competitive empirical performance and are easier to implement in practice. Bayesian and non-Bayesian models have been compared in terms of the quality of their uncertainty estimates, with several studies examining performance under distributional shifts (Ng, 2020) and others relating uncertainty magnitude to prediction correctness (Scalco et al., 2024). Similar ideas have been used to assess prediction quality in the absence of ground truth (Sikha et al., 2025), and to detect distributional shifts or domain changes (Soufi et al., 2025; Ovadia et al., 2019; Carannante et al., 2022).

2.3. Uncertainty Estimation in Segmentation Models

In the context of semantic segmentation, most research has focused on practical approximation techniques. Two of the most widely adopted approaches are Monte Carlo (MC) dropout and model ensembles, which are favored for their simplicity and compatibility with existing architectures (Kendall et al., 2015; Kamnitsas et al., 2017; Ghoshal et al., 2021). In MC-Dropout, the well-known dropout regularization technique is applied not only during training but also at inference time. Multiple forward passes are performed through the network, and uncertainty is quantified using the variance or entropy of the resulting predictions. Similarly, ensemble-based methods train multiple independent (deterministic) networks; during inference, the same input is fed through all models, and the variability across their predictions provides an estimate of the uncertainty.

Recently, we proposed the SUPER-Net approach for segmentation, where the posterior distribution is approximated by its first two moments (mean and covariance) and propagated through the network layers during training (Carannante et al., 2025). At nonlinear layers, the transformation of the mean and covariance is approximated using a first-order Taylor series expansion, enabling the network to jointly learn both predictions and uncertainty in a principled and computationally efficient manner.

While Bayesian models offer a principled framework for uncertainty estimation for segmentation, prior works have typically evaluated them under a limited range of perturbations or datasets. In this work, we frame Bayesian segmentation within the context of *Verification and Validation*, assessing robustness, uncertainty sensitivity, and the correspondence between prediction errors and uncertainty across diverse distributional shifts.

3. Methodology: V&V of Bayesian Segmention Models

Our experiments are structured within a V&V framework for medical image segmentation models. To this end, we evaluate deterministic and Bayesian models on clean test data, study performance under multiple distributional shifts, including several noise types and adversarial attacks, and analyze robustness using complementary metrics.

A key component of our validation is the behavior of uncertainty estimates. We evaluate how predictive uncertainty changes under increasing noise levels and whether uncertainty appropriately reflects prediction errors (i.e., higher uncertainty for incorrect or unstable segmentation). This allows us to assess not only robustness but also the trustworthiness of the uncertainty information provided by Bayesian and approximate-Bayesian models.

Table 1: Model architecture and training details for each dataset.

Dataset	Encoder filters	Decoder filters	Epochs	Batch size
Lungs	16, 32, 64	32, 16	50	10
Hippocampus	32, 64, 128	64, 32	100	20
BraTS	64, 128, 256, 512, 1024	512, 256, 128, 64	100	20

3.1. Datasets

We employ three publicly available medical image segmentation benchmarks: Lung CT, Hippocampus MRI, and Brain Tumor MRI (BraTS) (Ma et al., 2020; Antonelli et al., 2022; Menze et al., 2014). We report results for clinical data in Appendix B. For all datasets, we use an 80/20 split for training, and validation, applied consistently across experiments. Preprocessing includes intensity normalization, removal of empty slices, and resizing or cropping to a fixed spatial resolution.

The Lung dataset consists of 20 chest CT scans with annotations for the left and right lungs, as well as infection regions (Ma et al., 2020). Binary segmentation is performed (lung vs. background). The Hippocampus dataset includes 394 MRI scans from the Medical Segmentation Decathlon (Antonelli et al., 2022), with labels for anterior and posterior hippocampus regions. The BraTS dataset comprises multi-modal MRIs from high-grade glioma patients (Menze et al., 2014). Each case includes five tissue classes; evaluation follows the BraTS convention (whole tumor, core, and enhancing regions).

3.2. Models Specifics

We adopt the U-Net architecture as the backbone for all segmentation experiments (Ronneberger et al., 2015). U-Net is an encoder–decoder convolutional NN with skip connections between symmetric layers, enabling the combination of high-resolution spatial information from the encoder with features from the decoder. Each convolutional block consists of two convolutional layers (kernel size of 3×3) followed by batch normalization and ReLU activation, with max-pooling used for downsampling and upsampling in the decoder. The number of filters in each encoder and decoder stage for each dataset is summarized in Table 1.

As a baseline, we train a deterministic U-Net. We then compare it against three uncertainty-aware variants: (i) MC-Dropout, using 20 MC samples with a dropout probability of $p = 0.5$ applied at the bottleneck layers as in (Kendall et al., 2015); (ii) Ensemble, comprising five independently initialized U-Nets whose predictions are aggregated at inference; and (iii) SUPER-Net, which jointly learns prediction and uncertainty by propagating mean and covariance through all layers as in (Carannante et al., 2025). All models are trained using the specifications listed in Table 1. We employ the Adam optimizer with a learning rate of 0.001, and apply early stopping based on validation performance.

3.3. Distributional Shifts

To evaluate robustness and generalizability, we introduce controlled distributional shifts at test time by corrupting the images with noise and adversarial perturbations. Medical images are known to be affected by acquisition and reconstruction noise, which is commonly modeled using additive white Gaussian noise, though other noise types such as Poisson,

Table 2: Performance Comparison for noise-free Lungs Test Dataset

	Deterministic	**MC-Dropout**	**Ensemble**	**SUPER-Net**
DSC	**.83**	**.83**	**.83**	**.83**
HD95	**3.17**	4.04	5.1	4.94
O_s	.16	**.15**	.17	**.15**
U_s	.04	**.03**	**.03**	**.03**

Table 3: Performance Comparison for the noise-free Hippocampus Test Dataset

	Anterior				**Posterior**			
	Deterministic	**MC-Dropout**	**Ensemble**	**SUPER-Net**	**Deterministic**	**MC-Dropout**	**Ensemble**	**SUPER-Net**
DSC	**.79**	**.79**	**.79**	**.79**	.76	.76	**.77**	.74
HD95	**1.56**	1.68	1.58	1.62	**1.81**	2.11	2.01	2.21
O_s	.15	.13	**.09**	.15	.17	.15	**.11**	.14
U_s	**.14**	.17	.19	.15	**.15**	.19	.21	.23

Table 4: Performance Comparison for noise-free BraTS Test Dataset

	Whole				***Core***				***Enhancing***			
	Det	**MC-Drop**	**Ensemble**	**SUPER-Net**	**Det**	**MC-Drop**	**Ensemble**	**SUPER-Net**	**Det**	**MC-Drop**	**Ensemble**	**SUPER-Net**
DSC	.77	.77	.76	**.83**	.58	.58	.60	**.64**	.57	.57	.63	**.69**
HD95	7.18	6.83	6.38	**3.38**	6.95	6.40	5.53	**4.36**	5.94	5.91	4.09	**3.13**
O_s	.11	.11	**.06**	.10	.29	.24	.14	**.08**	.41	.40	.29	**.25**
U_s	.22	.21	.27	**.14**	**.21**	.26	.34	.34	**.11**	.12	.16	.15

Speckle, and Salt-and-Pepper have also been reported in the literature (Goyal et al., 2018). Accordingly, we generate noisy test sets by adding Gaussian noise at multiple Signal-to-Noise Ratio (SNR) levels, and we additionally include Poisson, Speckle, and Salt-and-Pepper noise for completeness. Noise is applied either to entire image scans or selectively to the structure of interest to simulate localized corruption.

To examine vulnerability to adversarial perturbations, we apply the Fast Gradient Sign Method (FGSM) for untargeted (Liu et al., 2017) and Projected Gradient Descent (PGD) for targeted variants (Madry et al., 2018). These corruptions enable a systematic assessment of performance degradation and uncertainty behavior under a range of perturbations.

3.4. Evaluation Metrics

To assess segmentation performance, we use both region-based and boundary-based metrics. The DSC measures voxel-wise overlap between the prediction and ground truth, capturing overall spatial agreement. Complementarily, the HD quantifies the largest boundary deviation between two segmentations. Because the maximum HD is highly sensitive to isolated outliers, often caused by noise or small contour errors, we report the 95th-percentile Hausdorff Distance (HD95), a robust variant commonly used in medical imaging evaluation. Beyond these standard accuracy metrics, we evaluate over-segmentation (O_s) and under-segmentation (U_s) rates (Mou et al., 2021). These quantify the fraction of incorrectly added or missed voxels, respectively, with both values ranging from 0 to 1, where lower values indicate better performance. Such measures are clinically meaningful, as segmentation errors have distinct consequences depending on their direction. For example, under-segmenting a tumor in radiotherapy planning may lead to undertreatment and increased recurrence risk, whereas over-segmenting may unnecessarily expose healthy tissues to radiation. Like-

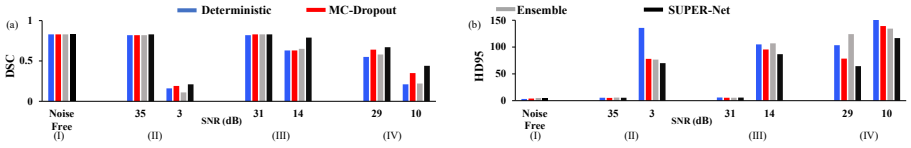

Figure 1: Performance comparison on the Lungs dataset under noise-free conditions (I), Gaussian noise applied to the whole image (II) or lung pixels only (III), and untargeted adversarial attacks (IV). HD95 and DSC are shown on the y-axis of (a) and (b), with SNR values on the x-axis for noisy conditions.

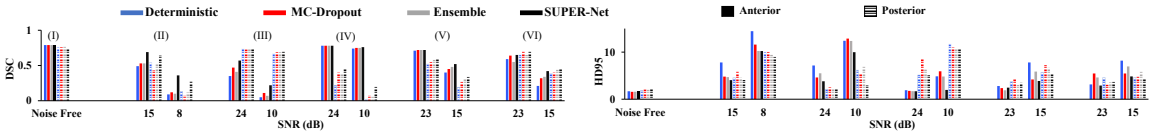

Figure 2: Performance comparison on the Hippocampus dataset for noise-free data (I), Gaussian noise applied to the whole image (II), anterior pixels (III), or posterior pixels (IV), and targeted adversarial attacks (V–VI). HD95 and DSC are shown on the y-axis, with SNR on the x-axis for noisy cases. Full-color bars correspond to the anterior structure, and dashed bars to the posterior structure.

wise, systematic O_s or U_s in longitudinal tumor monitoring can lead to misinterpretation of growth or stability.

4. Results and Discussion

4.1. Performance Comparison

To establish a baseline, we first evaluate all models, Deterministic, MC-Dropout, Ensemble, and SUPER-Net, on noise-free test data. Tables 2, 3, and 4 report DSC, HD95, O_s,

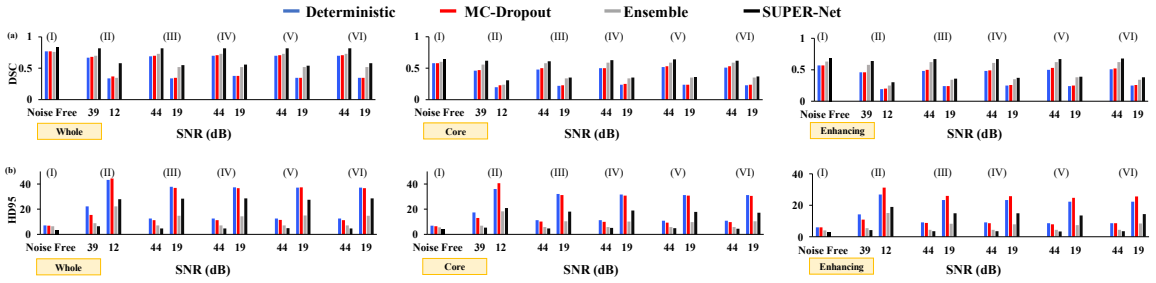

Figure 3: Performance comparison on BraTS for (I) noise-free, (II) untargeted, and (III–VI) targeted attacks. Subplots show DSC and HD95 vs. SNR for whole tumor, core, and enhancing regions.

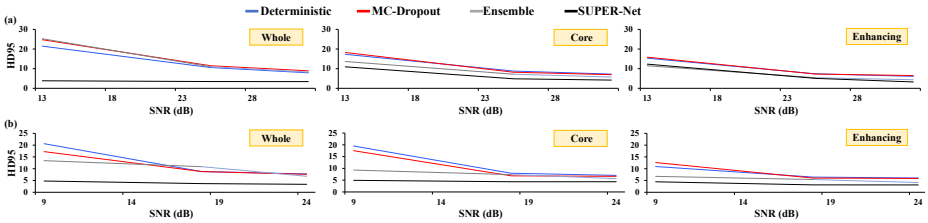

Figure 4: Performance comparison under Gaussian noise applied to (a) tumor pixels only and (b) entire scans in the BraTS test data. Subplots show HD95 across SNR levels for whole tumor, core, and enhancing regions.

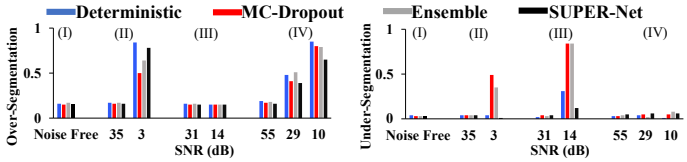

Figure 5: Performance comparison on the Lungs dataset for noise-free (I), Gaussian noise applied to the whole image (II) or lung pixels only (III), and untargeted adversarial attacks (IV). Subplots show O_s and U_s across SNR levels.

and U_s for the Lung, Hippocampus, and BraTS datasets. No single method consistently achieves the best performance across all metrics or anatomical structures, underscoring the importance of multi-metric evaluation.

Figures 1, 2, and 3 summarize segmentation performance under several perturbation scenarios, including multiple levels of Gaussian noise (applied either to entire images or only to target structures) and both targeted and untargeted adversarial attacks. Across datasets, all models maintain comparable accuracy under mild perturbations, but performance deteriorates substantially as noise or adversarial strength increases, particularly for HD95, which is highly sensitive to boundary distortions. In general, DSC and HD95 provide consistent signals of degradation, although the relative ordering of models may differ at high corruption levels. For BraTS, the robustness trends are further illustrated in Fig. 4, which

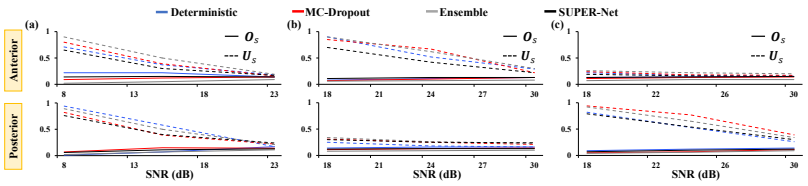

Figure 6: Performance comparison on the Hippocampus dataset with Gaussian noise applied to the whole image (I), anterior pixels (II), or posterior pixels (III). O_s and U_s are plotted versus SNR for the anterior and posterior structures.

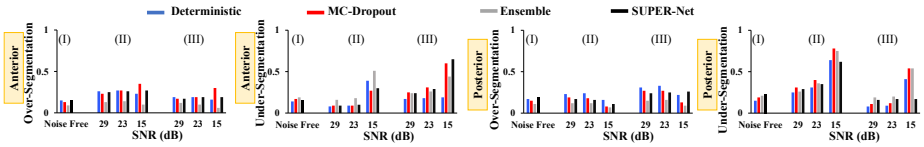

Figure 7: Performance comparison on the Hippocampus dataset for (I) noise-free, and (II–III) adversarial attacks. Panels (II) and (III) show targeted attacks. Subplots report O_s and U_s versus SNR for the anterior and posterior structures.

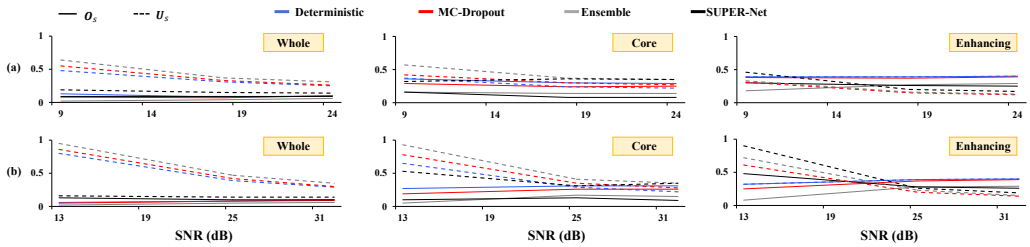

Figure 8: Performance comparison on BraTS under Gaussian noise applied to (a) the entire image and (b) tumor pixels only. The y-axis shows O_s and U_s versus SNR for whole tumor, core, and enhancing regions.

displays HD95 for Gaussian noise applied either globally or to tumor regions only. SUPER-Net exhibits the smallest increase in HD95 as noise intensity grows, indicating improved boundary stability relative to the other methods.

We additionally examine O_s and U_s behavior under both Gaussian noise and adversarial attacks (Figs. 5, 6, 8, 7, 9). No approach dominates across all conditions, highlighting the intrinsic tradeoff between O_s and U_s. Some models systematically under-segment, while others shift between under- and over-segmentation depending on dataset and corruption level. However, we observe that low O_s values at high noise levels may correspond to severe under-segmentation, emphasizing the need to interpret these metrics jointly with DSC and visual inspection. SUPER-Net generally achieves a favorable balance across metrics and often attains the lowest combined segmentation error, especially at low SNRs.

In addition, we provide a $k-$fold cross-validation analysis for SUPER-Net in Appendix A. Results are consistent across folds, indicating robustness is not due to fold-specific effects.

4.2. Noise Analysis

We extend the noise analysis to additional corruption types commonly observed in medical imaging, Speckle, Salt & Pepper, and Poisson, to further assess model behavior under realistic acquisition conditions. Performance under these noise types is reported in Table 5 and Figures 10, 11, and 12. Across all datasets and noise types, the models behave similarly under clean and low-noise conditions. However, as corruption severity increases, performance degrades sharply, especially for deterministic and approximate Bayesian methods.

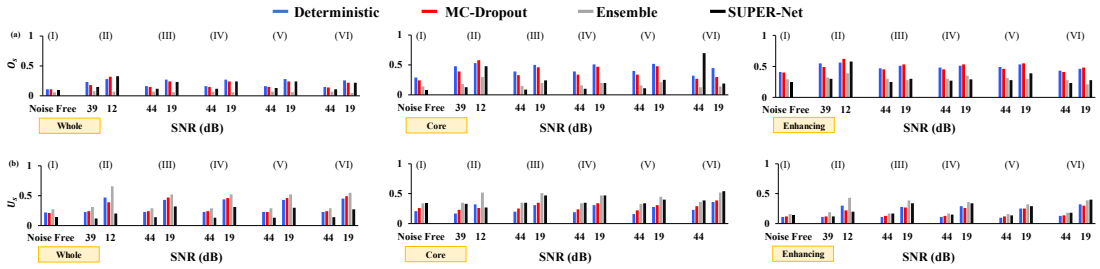

Figure 9: Performance of the four networks on BraTS for (I) noise-free, (II) untargeted, and (III–VI) targeted attacks. Subplots show O_s and U_s vs. SNR for whole tumor, core, and enhancing regions.

Table 5: DSC for Hippocampus data corrupted with Speckle and Poisson noise.

	Anterior				Posterior			
	Deterministic	**MC-Dropout**	**Ensemble**	**SUPER-Net**	**Deterministic**	**MC-Dropout**	**Ensemble**	**SUPER-Net**
Speckle noise added to entire image								
SNR ≈ 20 dB	.77	.77	**.78**	**.78**	.73	.74	**.75**	.73
SNR ≈ 14 dB	.58	.65	.65	**.68**	.53	.54	.60	**.63**
SNR ≈ 10 dB	.18	.20	.23	**.48**	.09	.10	.12	**.37**
Poisson noise added to entire image								
SNR ≈ 20 dB	.74	.75	.75	**.77**	**.73**	.71	.72	.71
SNR ≈ 11 dB	.39	.38	.39	**.56**	.39	.38	.36	**.52**
SNR ≈ 8 dB	.22	.20	.22	**.40**	.23	.21	.21	**.36**

SUPER-Net consistently exhibits greater robustness at low SNRs, maintaining higher DSC and more stable degradation trends, suggesting that propagating uncertainty helps increase robustness under heavy corruption.

4.3. Uncertainty Analysis

We first examine whether uncertainty responds meaningfully to corruption severity. Figure 13 shows DSC and predictive variance for SUPER-Net under Gaussian noise and targeted adversarial attacks. As expected, uncertainty increases as DSC decreases, indicating that the model expresses lower confidence when its predictions deteriorate. We then compare uncertainty behavior across all approaches. In Figure 14, we report the average predictive variance for correctly and incorrectly classified pixels under Gaussian, Poisson, and adversarial perturbations. All methods assign higher uncertainty to incorrect predictions, but only SUPER-Net exhibits a monotonically increasing uncertainty trend with increasing corruption levels, a desirable property for reliability under distributional shifts.

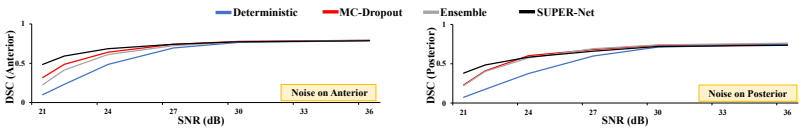

Figure 10: Performance comparison under various levels of Speckle noise applied to the anterior and posterior hippocampus. Subplots display DSC across SNR levels.

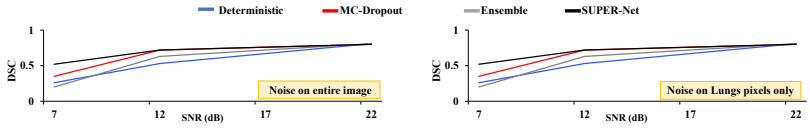

Figure 11: Performance comparison under various levels of Salt & Pepper noise applied to the entire image or the lung pixels only. DSC vs. SNR is plotted.

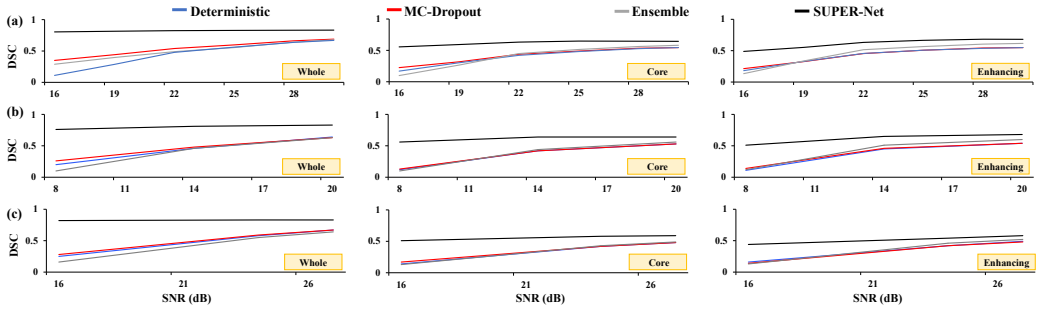

Figure 12: Performance comparison under Speckle noise added to (a) tumor pixels and (b) all pixels, and under (c) Salt & Pepper noise on the BraTS test data. Subplots show DSC across SNR levels for whole tumor, core, and enhancing regions.

Finally, we assess whether uncertainty meaningfully identifies erroneous pixels. In Figure 15, we report the change in DSC when "uncertain" pixels are removed. Ideally, models should be uncertain only for incorrectly segmented pixels, leading to improved DSC after removal ($\Delta\text{DSC} > 0$). However, several approaches show a DSC decrease, suggesting that they also flag correctly classified pixels as uncertain. SUPER-Net shows the most consistent positive or stable ΔDSC, indicating more informative uncertainty estimates.

5. Conclusion

Accurate and reliable segmentation is essential for many clinical tasks, including diagnosis, treatment planning, and long-term disease monitoring. Yet in practice, medical images are affected by noise, artifacts, scanner variability, and unexpected data shifts, all of which can cause segmentation models to fail silently. Reliable medical image segmentation therefore requires more than high accuracy on clean test data, it demands models that remain trust-

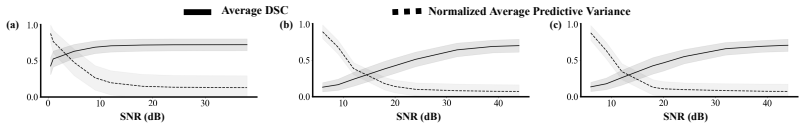

Figure 13: SUPER-Net DSC and predictive variance vs. SNR for the Hippocampus data. Variance is normalized between 0 and 1. Results are shown for (a) Gaussian noise, and targeted adversarial attacks (b - c).

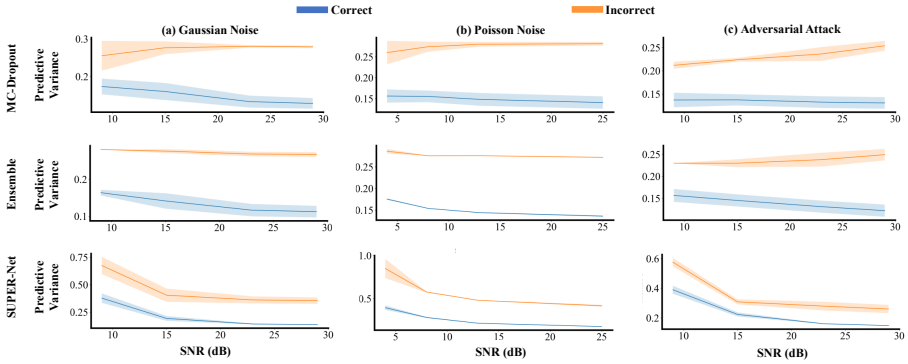

Figure 14: Average predictive variance versus SNR for: 1) correctly labeled pixels (blue) and 2) misclassified ones (orange) for (a) Gaussian noise, (b) Poisson noise and (c) adversarial attacks applied to the Hippocampus data.

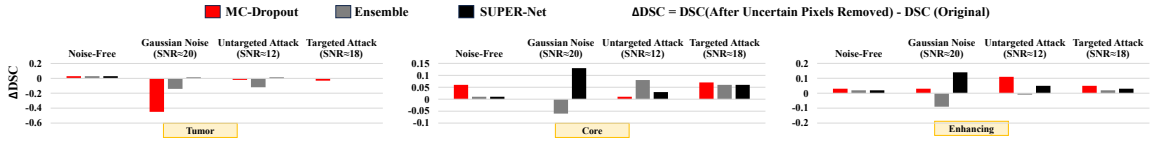

Figure 15: Change in DSC (ΔDSC) after the removal of *uncertain* pixels.

worthy under real-world variability and provide uncertainty estimates that meaningfully reflect prediction reliability for clinical use.

Through a comprehensive V&V framework, we compared deterministic, approximate Bayesian, and uncertainty-propagating segmentation models across multiple datasets and clinically relevant perturbations. Our analysis shows that approaches producing uncertainty as part of the forward pass, such as SUPER-Net, offer more stable performance under noise and adversarial conditions and generate uncertainty values that consistently flag incorrect predictions. Such behavior is crucial in clinical decision-making, where overconfident errors may directly contribute to misdiagnosis or suboptimal treatment.

Our findings underscore that model selection for clinical deployment must integrate robustness analysis, boundary-sensitive metrics, and principled uncertainty evaluation, rather than relying solely on accuracy. By framing segmentation assessment within a V&V perspective, this work emphasizes that uncertainty-aware modeling is central to building safe, interpretable, and clinically actionable AI systems. Evaluating models under distributional shifts and scrutinizing their uncertainty behavior should become standard practice in the development of trustworthy clinical AI. Ultimately, this work moves toward AI tools that better support clinicians by providing not only accurate predictions but also clear indications of when those predictions can be trusted.

Acknowledgments

This work was supported by the National Science Foundation awards NSF 1903466, NSF 2008690, and NSF 2234468.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 888–897, 2018.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

Giuseppina Carannante, Dimah Dera, Orune Aminul, Nidhal C Bouaynaya, and Ghulam Rasool. Self-assessment and robust anomaly detection with bayesian deep learning. In *2022 25th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2022.

Giuseppina Carannante, Nidhal Bouaynaya, Ghulam Rasool, and Lyudmila Mihaylova. Basis-net: From point estimate to predictive distribution in neural networks-a bayesian sequential importance sampling framework. *Transactions on Machine Learning Research*, 2024.

Giuseppina Carannante, Nidhal C Bouaynaya, Dimah Dera, Hassan M Fathallah-Shaykh, and Ghulam Rasool. Super-net: Trustworthy image segmentation via uncertainty propagation in encoder-decoder networks. *Pattern Recognition*, page 112503, 2025.

Shir Chorev, Philip Tannor, Dan Ben Israel, Noam Bressler, Itay Gabbay, Nir Hutnik, Jonatan Liberman, Matan Perlmutter, Yurii Romanyshyn, and Lior Rokach. Deepchecks: A library for testing and validating machine learning models and data. *Journal of Machine Learning Research*, 23(285):1–6, 2022.

Sylvain Christin, Éric Hervet, and Nicolas Lecomte. Going further with model verification and deep learning. *Methods in Ecology and Evolution*, 12(1):130–134, 2021.

Dimah Dera, Nidhal Carla Bouaynaya, Ghulam Rasool, Roman Shterenberg, and Hassan M Fathallah-Shaykh. Premium-cnn: Propagating uncertainty towards robust convolutional neural networks. *IEEE Transactions on Signal Processing*, 69:4669–4684, 2021.

H. M. Fathallah-Shaykh, A. DeAtkine, E. Coffee, E. Khayat, A. K. Bag, X. Han, P. P. Warren, M. Bredel, J. Fiveash, J. Markert, N. Bouaynaya, and L. B. Nabors. Diagnosing growth in low-grade gliomas with and without longitudinal volume measurements: A retrospective observational study. *PLoS Med*, 16(5):e1002810, 05 2019.

Kambiz Frounchi, Lionel C Briand, Leo Grady, Yvan Labiche, and Rajesh Subramanyan. Automating image segmentation verification and validation by learning test oracles. *Information and Software Technology*, 53(12):1337–1348, 2011.

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

Francesco Galati, Sébastien Ourselin, and Maria A Zuluaga. From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review. *Applied Sciences*, 12(8):3936, 2022.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56 (Suppl 1):1513–1589, 2023.

Biraja Ghoshal, Allan Tucker, Bal Sanghera, and Wai Lup Wong. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence*, 37(2):701–734, 2021.

Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. In *Case Studies in Applied Bayesian Data Science*, pages 45–87. Springer, 2020.

Bhawna Goyal, Sunil Agrawal, and BS Sohi. Noise issues prevailing in various types of medical images. *Biomedical & Pharmacology Journal*, 11(3):1227, 2018.

Alex Graves. Practical Variational Inference for Neural Networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011.

Zesheng Hong, Yubiao Yue, Yubin Chen, Lele Cong, Huanjie Lin, Yuanmei Luo, Mini Han Wang, Weidong Wang, Jialong Xu, Xiaoqi Yang, et al. Out-of-distribution detection in medical image analysis: A survey. *arXiv preprint arXiv:2404.18279*, 2024.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.

Haseeb Javed, Shaker El-Sappagh, and Tamer Abuhmed. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust ai applications. *Artificial Intelligence Review*, 58(1):12, 2024.

Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *International journal of computer vision*, 129(2):462–483, 2021.

K Kamnitsas, W Bai, E Ferrante, S McDonagh, M Sinclair, N Pawlowski, M Rajchl, M Lee, B Kainz, D Rueckert, et al. Ensembles of multiple models and architectures for robust

brain tumour segmentation. In *International MICCAI brainlesion workshop*, pages 450–462. Springer, 2017.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Dong Wook Kim, Hye Young Jang, Yousun Ko, Jung Hee Son, Pyeong Hwa Kim, Seon-Ok Kim, Joon Seo Lim, and Seong Ho Park. Inconsistency in the use of the term "validation" in studies reporting the performance of deep learning algorithms in providing diagnosis from medical imaging. *Plos one*, 15(9):e0238908, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Zheng Liu, Jinnian Zhang, Varun Jog, Po-Ling Loh, and Alan B McMillan. Robustifying deep networks for medical image segmentation. *Journal of digital imaging*, 34(5):1279–1293, 2021.

Jun Ma, Cheng Ge, Yixin Wang, Xingle An, Jiantao Gao, Ziqi Yu, Minqing Zhang, Xin Liu, Xueyuan Deng, Shucheng Cao, Hao Wei, Sen Mei, Xiaoyu Yang, Ziwei Nie, Chen Li, Lu Tian, Yuntao Zhu, Qiongjie Zhu, Guoqiang Dong, and Jian He. Covid-19 ct lung and infection segmentation dataset, April 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Lei Mou, Yitian Zhao, Huazhu Fu, Yonghuai Liu, Jun Cheng, Yalin Zheng, Pan Su, Jianlong Yang, Li Chen, Alejandro F Frangi, et al. Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging. *Medical image analysis*, 67:101874, 2021.

Matthew Ng. *Estimating uncertainty in neural networks for cardiac mri segmentation*. University of Toronto (Canada), 2020.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 493–501. Springer, 2018.

David Rodriguez, Tapsya Nayak, Yidong Chen, Ram Krishnan, and Yufei Huang. On the role of deep learning model complexity in adversarial robustness for medical images. *BMC Medical Informatics and Decision Making*, 22(Suppl 2):160, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas 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.

Elisa Scalco, Silvia Pozzi, Giovanna Rizzo, and Ettore Lanzarone. Uncertainty quantification in multi-class segmentation: Comparison between bayesian and non-bayesian approaches in a clinical perspective. *Medical Physics*, 51(9):6090–6102, 2024.

Madeline Chantry Schiappa, Shehreen Azad, Sachidanand Vs, Yunhao Ge, Ondrej Miksik, Yogesh S Rawat, and Vibhav Vineet. Robustness analysis on foundational segmentation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1786–1796. IEEE, 2024.

OK Sikha, Meritxell Riera-Marin, Adrian Galdran, Javier Garcia Lopez, Julia Rodriguez-Comas, Gemma Piella, and Miguel A Gonzalez Ballester. Uncertainty-aware segmentation quality prediction via deep learning bayesian modeling: Comprehensive evaluation and interpretation on skin cancer and liver segmentation. *Computerized Medical Imaging and Graphics*, 123:102547, 2025.

Mazen Soufi, Yoshito Otake, Makoto Iwasa, Keisuke Uemura, Tomoki Hakotani, Masahiro Hashimoto, Yoshitake Yamada, Minoru Yamada, Yoichi Yokoyama, Masahiro Jinzaki, et al. Validation of musculoskeletal segmentation model with uncertainty estimation for bone and muscle assessment in hip-to-knee clinical ct images. *Scientific reports*, 15(1): 125, 2025.

Anh T Tran, Tal Zeevi, and Seyedmehdi Payabvash. Strategies to improve the robustness and generalizability of deep learning segmentation and classification in neuroimaging. *BioMedInformatics*, 5(2):20, 2025.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Table 6: Cross-validation results for noise-free BraTS test data

	Whole	Core	Enhancing
DSC (mean ± std)	0.82 ± 0.02	0.66 ± 0.02	0.69 ± 0.02
HD95 (mean ± std)	3.42 ± 0.36	3.82 ± 0.35	2.93 ± 0.34

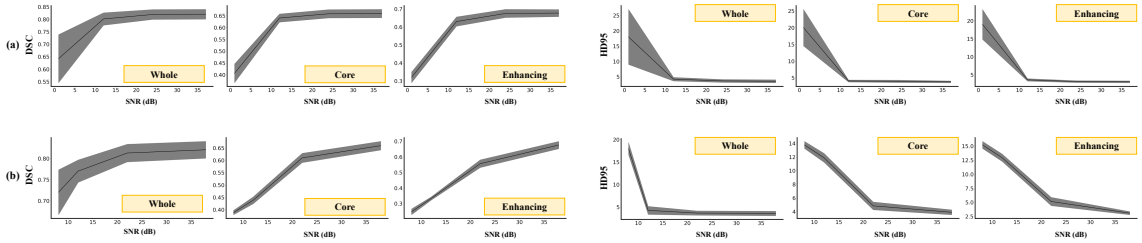

Figure 16: DSC and HD95 for SUPER-Net for Gaussian noise applied to (a) the entire image and (b) tumor pixels only of the BraTS dataset. Curves show the mean across $k = 5$ cross-validation models, with shaded regions indicating standard deviation, providing insight into the variability of the model performance.

Appendix A. Cross-Validation

To comprehensively evaluate the performance of SUPER-Net for medical image segmentation, we conduct a cross-validation study. In our study, we opt for the $k-$fold cross-validation method with $k = 5$. Cross-validation mitigates potential biases in training and test set selection by randomly sampling different splits of the data for each iteration.

For both the Lung and Hippocampus data, the model presented above falls within the range obtained with the cross-validation results. For the Lung data, the mean DSC for the k-fold models is 0.83 with a standard deviation of 0.01. For the Hippocampus data, the mean and standard deviation DSC are 0.78 and 0.02 for the anterior structure, and 0.74 and 0.01 for the posterior structure.

We report the cross-validation results for the BraTS data in Table 6, which are similar to the one-split results reported in Section 4.1. Additionally, we test the reliability of the results when the k-fold models are tested under noisy conditions. In Figure 16 we show the DSC and HD95 vs. SNR for the k-fold models. The line represents the average performance, while the shaded area refers to the standard deviation.

We observe that there is not much variation among the models under low-noise conditions, while there is higher variation under high-noise conditions. Yet, it is interesting to observe that all the SUPER U-Net models from the k-fold cross-validation perform better than other approaches.

Appendix B. Clinical Dataset

We include a clinical MRI dataset acquired at the University of Alabama at Birmingham (UAB), consisting of 627 FLAIR volumes from patients with grade II glioma (Fathallah-Shaykh et al., 2019). Each volume includes manual tumor annotations provided by an expert

Table 7: Model architecture and training details for the clinical dataset.

Encoder filters	Decoder filters	Epochs	Batch size	Optimizer	Learning Rate
16, 32, 64, 128, 256	128, 64, 32, 16	100	10	Adam	0.001

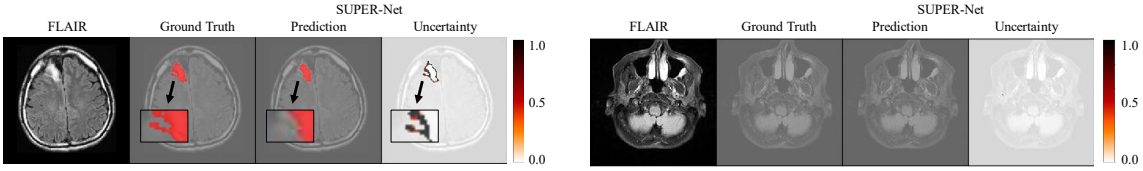

Figure 17: Two Sample scans from the clinical data. Both images show (left to right) the FLAIR input, the ground-truth segmentation, the SUPER U-Net prediction and the uncertainty map overlaid on the input scan. We zoom on regions incorrectly classified by the network and the corresponding regions in the uncertainty maps.

neuroradiologist. As with the other datasets in this study, we apply standard preprocessing (intensity normalization and removal of empty slices) and split the data into an 80/20 train–validation partition.

For this dataset, we use a U-Net backbone (details in Table 7) within the SUPER-Net framework. The model achieves a DSC of 86% on the validation set. Notably, it is also able to handle empty slices, i.e., scans without visible tumor, even though such cases were not included in the training set. Figure 17 illustrates two representative examples. In the first case, we highlight regions that are incorrectly segmented and compare them with the corresponding areas in the uncertainty map (computed from the predictive variance). As expected, the misclassified pixels exhibit high uncertainty. In the second example, we show an unseen empty scan. The model correctly predicts the absence of tumor, and, importantly, the associated uncertainty remains very low, indicating high confidence in its predictions.