Quantification of Uncertainty with Adversarial Models in Medical Image Segmentation

Hana Jebril[⋆1,2][0000−0002−6593−5662], Thomas Pinetz[⋆1,2][0000−0002−6100−2136]✉, Günter Klambauer[3,4][0000−0003−2861−5552], and Hrvoje Bogunović[1,2][0000−0002−9168−0894]

[1] Institute of Artificial Intelligence, Center for Medical Data Science, Medical University of Vienna, Austria
[2] Comprehensive Center for AI in Medicine, Medical University of Vienna, Austria
[3] ELLIS Unit Linz, LIT AI Lab and Institute for Machine Learning, Johannes Kepler University Linz, Austria
[4] Clinical Research Center for Medical AI, Johannes Kepler University Linz, Austria
hana.jebril@meduniwien.ac.at thomas.pinetz@meduniwien.ac.at
klambauer@ml.jku.at hrvoje.bogunovic@meduniwien.ac.at

Abstract. Reliable pixel-level uncertainty quantification holds the potential to transform clinical workflows by enabling high-fidelity longitudinal monitoring and distinguishing true pathological changes from artifacts. Ideally, these models provide the stability required for critical treatment planning and surgical intervention. However, standard deep learning models often suffer from miscalibration, yielding overconfident predictions that mask underlying vulnerabilities at subtle pathological boundaries. To address this, we propose QUAM-SM, a post-hoc framework using targeted adversarial search to identify "adversarially fragile" pixels. By actively seeking perturbations that expose predictive instability, our method highlights regions where decisions are most vulnerable to being flipped. Importantly, the framework disentangles epistemic uncertainty from aleatoric uncertainty. Experiments on two public datasets with multiple expert annotations demonstrate that QUAM-SM outperforms both standard and recent uncertainty estimation approaches in terms of reliability and boundary sensitivity. Code is available at https://github.com/HanaJebril/quam_sm

Keywords: Aleatoric Uncertainty · Adversarial Search · Segmentation.

1 Introduction

The deployment of deep learning models in clinical workflows requires not only accurate segmentation but also reliable pixel-level uncertainty quantification [5]. In applications such as longitudinal lesion monitoring, small segmentation errors may lead to incorrect clinical decisions, for example, when distinguishing true progression from pseudo-progression [3,17].

⋆ Equal Contribution

Despite their strong performance, standard segmentation networks often produce overconfident predictions that mask underlying failure modes, particularly at object boundaries and in subtle pathological regions. Uncertainty in medical image segmentation arises from multiple sources, including limited training data and model uncertainty (epistemic uncertainty) [10], as well as annotation variability and inherent data ambiguity (aleatoric uncertainty) [7,12]. A growing body of work has explored the relationship between inter-rater variability and these uncertainty components [15], emphasizing the clinical importance of disentangling their effects. Prior work has highlighted the clinical relevance of these components and their relationship to inter-rater variability [15]. Existing approaches to improve model reliability, such as Deep Ensembles [9], Bayesian approximations (e.g., Monte Carlo Dropout) [2], and recent self-supervised evidential methods like SURE [11], have shown promise. However, most methods rely on stochastic sampling within limited regions of the posterior, restricting their ability to capture the full spectrum of potential prediction failures.

To address this, we propose QUAM-SM, which leverages the principle that uncertain pixels are inherently easier for an adversarial agent to perturb into a false class. By conducting a post-hoc adversarial search, our method identifies "adversarially fragile" pixels where the model's decision is easily flipped, thereby providing a more sensitive map of potential errors compared to traditional density-based uncertainty estimation. We extend the concept of Quantification of Uncertainty with Adversarial Models [16] to the segmentation domain, achieving improved coverage of the posterior by highlighting pixels that are most susceptible to being "surely false". For classification tasks, targeted attacks are defined over the discrete class labels. However, in semantic segmentation, each pixel represents an independent classification instance, leading to a much larger space of targeted attack possibilities. This pixel-wise formulation allows the adversarial search to drive predictions toward diverse alternative segmentation hypotheses. Our key contributions are as follows:

1. We propose a novel post-hoc uncertainty quantification framework for medical image segmentation based on adversarial models that identifies fragile pixel-wise predictions.
2. We distinguish between aleatoric and epistemic uncertainty, and show improved consistency with inter-rater variability on multiple public datasets.
3. We evaluate the proposed method on multiple-observer segmentation tasks, highlighting its effectiveness in capturing uncertain pixels.

2 Method

The Quantification of Uncertainty with Adversarial Models in Medical Image Segmentation (QUAM-SM) framework (Fig. 1) first generates a reference segmentation y_{ref} using a fixed pre-trained model, assuming access to the training data or training distribution to maintain consistency with the learned data manifold. We then perform a post-hoc adversarial search (Sec. 2.1) to identify "adversarially fragile" pixel regions where the model's decision can be easily

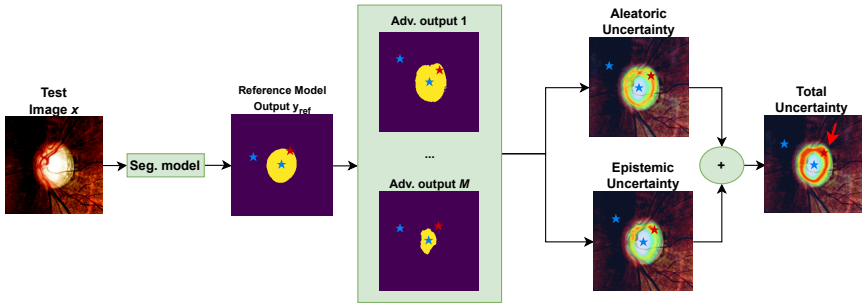

Fig. 1: QUAM-SM Overview: Following initial segmentation (y_{ref}), QUAM-SM utilizes post-hoc adversarial search across M adversarial model outputs to identify predictive fragility. This fragility is decomposed into aleatoric and epistemic components and combined into a Total Uncertainty map, effectively distinguishing boundary instability (red star) from robust certain regions (blue stars).

flipped to an incorrect class. Finally, uncertainty is quantified by measuring this predictive instability (Sec. 2.2), highlighting pixels most likely to be erroneous and supporting safer clinical interpretation.

2.1 Adversarial Model Search Algorithm

The adversarial search aims to generate predictions that remain consistent with the training data while diverging from the reference model on test samples, enabling broader posterior coverage and improved uncertainty characterization. To achieve this, we optimize two complementary loss terms (Alg. 1). The first is a penalty loss that enforces consistency with the training distribution to ensure posterior plausibility, which requires access to training data. The second is an adversarial loss that drives the prediction toward a targeted label y_{mask}, encouraging exploration of alternative segmentation hypotheses.

Targeted Attacks: To alter the adversarial model predictions, adversarial attacks are applied. In a targeted attack, the optimization explicitly drives the model toward a predefined output class by minimizing the loss with respect to the desired target [20,4]. In this work, we investigate multiple targeted attack strategies to broaden posterior exploration. Formally, let y_{ref} be the reference prediction and $\mathcal{D}_{\text{train}}$ the set of training masks. We define three targeted adversarial labels y_{mask}:

1. **Global extreme masks (GEM)**: representing pixel-wise constant segmentation maps where all pixels belong to either the background or the foreground, which is a straight-forward extension to QUAM approach [16]: $y_{\text{mask}} \in \{\mathbf{0}_{H\times W},\ \mathbf{1}_{H\times W}\}$

2. **Training-based masks (TBM)**: encouraging the adversarial model to explore realistic anatomical configurations; $y_{\text{mask}} \sim \mathcal{D}_{\text{train}}$ (ω selected masks).
3. **Morphology-augmented masks (MAM)**: $\mathcal{M}(\cdot)$ denotes a morphological operation such as dilation or erosion, probing local structural variability near boundaries which are eliminated; $y_{\text{mask}} = \mathcal{M}(y_{\text{ref}})$

Algorithm 1: Adversarial Model Search Algorithm

Input: Test image x, training set D, reference model w, reference loss $\mathcal{L}_{ref}$,targeted label y_{mask}, penalty scheduler η, divergence $D(\cdot,\cdot)$, parameters c_0, M, γ, Compute $y_{ref} = p(y|x,w)$

Output: Optimized adversarial model $\tilde{w}$

Initialize $\tilde{w} \leftarrow w$, $c \leftarrow c_0$;

for $m = 1$ **to** M **do**

- Compute penalty loss: $\mathcal{L}_{pen} = \frac{1}{K}\sum_{i=1}^{K}\left[\ell(p(y_i|x_i,\tilde{w}), y_i) - (\mathcal{L}_{ref} + \gamma)\right]$;
- Compute $y_{\text{mask}} = \mathcal{M}(y_{\text{ref}})$;
- Compute adversarial loss: $\mathcal{L}_{adv} = -D(y_{mask}, p(y|x,\tilde{w}))$;
- Total loss: $\mathcal{L} = \mathcal{L}_{adv} + c\,\mathcal{L}_{pen}$;
- Update parameters: $\tilde{w} \leftarrow \eta(\tilde{w})$, $c \leftarrow \eta(c)$;
- **if** $\mathcal{L}_{pen} \leq 0$ *and* $\mathcal{L}_{adv}$ *improved* **then**
  - store $\tilde{w}$ and break ;

return $\tilde{w}$;

2.2 Predictive Instability Uncertainty Quantification

The output of the adversarial search, a set of plausible model weights $\{\mathbf{w}_t\}_{t=1}^{N}$, forms a sparse pseudo-posterior from which we estimate the predictive distribution $P(Y \mid \mathbf{x}, \mathcal{D})$.

Mixture Importance Sampling (MIS) Due to the adversarial bias in sample generation, we employ Mixture Importance Sampling (MIS) [6] to derive a statistically consistent estimate of the expectation $\mathbb{E}_{\mathbf{w}}[\cdot]$ over the true posterior $P(\mathbf{w} \mid \mathcal{D})$. The expectation is approximated by a weighted sum over the sampled models $\mathbf{w}_t$:

$$\mathbb{E}_{\mathbf{w}}[\mathcal{F}(\mathbf{w})] \approx \sum_{t=1}^{N} W_t \cdot \mathcal{F}(\mathbf{w}_t) \quad \text{where} \quad W_t = \frac{\exp(-\mathcal{L}_{\text{pen}}(\mathbf{w}_t)/T)}{\sum_{j=1}^{N}\exp(-\mathcal{L}_{\text{pen}}(\mathbf{w}_j)/T)} \tag{1}$$

The weight W_t is the MIS importance weight derived from the model's Penalty Loss ($\mathcal{L}_{\text{pen}}$), following the original QUAM work [16]. In particular, the term $\exp(-\mathcal{L}_{\text{pen}}(\mathbf{w}_t)/T)$ ensures that models $\mathbf{w}_t$ which achieve low penalized loss on

the training data contribute proportionally more to the final predictive distribution. T is the temperature hyperparameter controlling the sharpness of the weight distribution. We define the Bayesian Model Average (BMA) weighted prediction $P_{\mathcal{D}} = \sum_t W_t P(Y \mid \mathbf{x}, \mathbf{w}_t)$. The resulting Weighted BMA prediction yields more robust and accurate estimates for both Aleatoric and Epistemic uncertainty.

Total Uncertainty Decomposition We decompose the total predictive uncertainty $H[P_{\mathcal{D}}]$ into its aleatoric and epistemic components on a per-pixel basis using the established Expected Entropy (EE) framework. This standard decomposition leverages the properties of Mutual Information and conditional entropy, where the Epistemic Uncertainty is identified with the Mutual Information between the predicted outcome Y and the model parameters $\mathbf{w}$:

$$H[P_{\mathcal{D}}] = \underbrace{\mathbb{E}_{\mathbf{w}}\left[H[P(Y \mid \mathbf{x}, \mathbf{w})]\right]}_{\text{Aleatoric Uncertainty}} + \underbrace{\mathbb{I}[Y; \mathbf{w} \mid \mathbf{x}, \mathcal{D}]}_{\text{Epistemic Uncertainty}} \tag{2}$$

Let N be the number of sampled model weights, $\hat{P}_t \in \mathbb{R}^C$ the softmax probability vector from the t-th model, and W_t its corresponding MIS weight ($\sum_{t=1}^{N} W_t = 1$). C is the total number of segmented classes. The uncertainty components are estimated using the weighted expectations:

- **Total Uncertainty ($\mathbf{H_{Total}}$)**: The entropy of the Weighted Bayesian Model Averaging (BMA) prediction.

$$\mathbf{H}_{\text{Total}} = -\sum_{c=1}^{C} \left(\sum_{t=1}^{N} W_t \cdot \hat{P}_{t,c} \right) \log \left(\sum_{t=1}^{N} W_t \cdot \hat{P}_{t,c} \right) \tag{3}$$

- **Aleatoric Uncertainty ($\mathbf{H_{Aleatoric}}$)**: The expected entropy of the individual predictions, estimated by the weighted average of individual entropies.

$$\mathbf{H}_{\text{Aleatoric}} = \sum_{t=1}^{N} W_t \cdot \left(-\sum_{c=1}^{C} \hat{P}_{t,c} \log(\hat{P}_{t,c}) \right) \tag{4}$$

- **Epistemic Uncertainty ($\mathbf{H_{Epistemic}}$)**: Calculated as the weighted measure of disagreement (KL divergence).

$$\mathbf{H}_{\text{Epistemic}} = \sum_{t=1}^{N} W_t \cdot \left[\sum_{c=1}^{C} \hat{P}_{t,c} \log \left(\frac{\hat{P}_{t,c}}{\sum_{k=1}^{N} W_k \cdot \hat{P}_{k,c}} \right) \right] \tag{5}$$

3 Experiments and Results

3.1 Datasets and Implementation Details

To evaluate our proposed QUAM-SM, we use two public datasets with multiple annotations available. (1) The REFUGE dataset ($\mathcal{D}_R$) [14] consists of 400 color

fundus photography (CFP) images for training, 400 for validation, and 400 for testing. Each image is accompanied by seven independent segmentation annotations of the optic disc and optic cup, provided by multiple expert observers. (2) The QUBIQ2021 prostate tumour dataset ($\mathcal{D}_Q$) [1] consists of 45 magnetic resonance imaging (MRI) volumes for training and 7 volumes for testing, each has one selected 2D slice and annotated by six experts. All images were cropped to a spatial resolution of 512×512 before training and evaluation.

Comparison Study We compare the proposed QUAM-SM against several widely used uncertainty-aware segmentation baselines, namely Monte Carlo (MC) Dropout [2], Deep Ensembles (DE) [9], Probabilistic U-Net (PUnet) [8], and the Uncertainty-Supervised Interpretable and Robust Evidential Segmentation (SURE) framework [11]. For MC, DE, and PUnet, aleatoric and epistemic uncertainty maps are computed using the formulations in Eqs. (3–5), while omitting the proposed weighting term.

For the evidential-based SURE model, uncertainty estimation follows the formulation introduced in [18]. Specifically, let $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^{K}$ denote the Dirichlet concentration parameters over K classes and $S = \sum_{i=1}^{K} \alpha_i$. Aleatoric and epistemic uncertainties are computed as:

$$\mathbf{u}_{\text{aleatoric}} = \sum_{i=1}^{K} \frac{\alpha_i(S-\alpha_i)}{S(S+1)}, \quad \mathbf{u}_{\text{epistemic}} = \sum_{i=1}^{K} \frac{\alpha_i(S-\alpha_i)}{S^2(S+1)}. \tag{6}$$

Implementation details: We employed the Attention U-Net [13] as a backbone for our algorithm and other baselines, and Adam optimizer with a learning rate of 10^{-4}. The batch size was set to 4 for both datasets. To evaluate the quality of uncertainty estimation, we measure the agreement between the predicted uncertainty maps and the reference entropy map derived from multiple annotators. We report the Pearson Correlation Coefficient (PCC) and the coefficient of determination (R^2) to assess correlation and goodness of fit. In addition, the Soft Dice (SDice) is used to evaluate the segmentation.

3.2 Experimental Results

Table 1 summarizes the quantitative results on the multi-annotator datasets REFUGE ($\mathcal{D}_R$) and QUBIQ2021 prostate tumour ($\mathcal{D}_Q$). We compare different uncertainty estimation methods by computing the correlation between each predicted uncertainty map and the multi-observer entropy map. Our method (QUAM-SM) achieves the highest PCC and R^2 values across epistemic, aleatoric, and total uncertainty, indicating superior alignment with the reference uncertainty. In terms of segmentation accuracy, QUAM-SM method also consistently outperforms all competing approaches for all uncertainty types, as reflected by the SDice scores. Notably, the improvement for both datasets is most pronounced for aleatoric uncertainty, which is expected since the multiple annotation task

primarily reflects aleatoric uncertainty arising from inter-observer variability. Figure 2 presents representative qualitative results from each dataset for both aleatoric and total uncertainty. The visualizations demonstrate that QUAM-SM closely matches the ground-truth entropy maps, accurately capturing regions of high inter-annotator uncertainty.

Table 1: Comparison of uncertainty methods $\mathcal{M}$ (DE [9], MC [2], PUnet [8], SURE [11], TTA [19]) on REFUGE ($\mathcal{D}_R$) and QUBIQ2021 ($\mathcal{D}_Q$).

$\mathcal{D}$	$\mathcal{M}$	Epistemic			Aleatoric			Total		
		R^2	PCC	SDice	R^2	PCC	SDice	R^2	PCC	SDice
$\mathcal{D}_R$	[9]	$.18 \pm .07$	$.41 \pm .10$	$.23 \pm .07$	$.39 \pm .10$	$.62 \pm .09$	$.55 \pm .08$	$.34 \pm .10$	$.57 \pm .11$	$.43 \pm .08$
	[2]	$.28 \pm .11$	$.52 \pm .12$	$.33 \pm .09$	$.39 \pm .10$	$.62 \pm .10$	$.51 \pm .09$	$.36 \pm .11$	$.59 \pm .11$	$.42 \pm .09$
	[8]	$.10 \pm .11$	$.25 \pm .20$	$.30 \pm .15$	$.13 \pm .11$	$.31 \pm .17$	$.30 \pm .15$	$.13 \pm .11$	$.31 \pm .17$	$.29 \pm .14$
	[11]	$.07 \pm .11$	$.20 \pm .18$	$.25 \pm .1$	$.08 \pm .09$	$.18 \pm .06$	$.23 \pm .11$	$.08 \pm .10$	$.19 \pm .16$	$.24 \pm .09$
	[19]	$.18 \pm .07$	$.41 \pm .11$	$.19 \pm .06$	$.40 \pm .11$	$.62 \pm .10$	$.52 \pm .08$	$.36 \pm .11$	$.59 \pm .11$	$.42 \pm .08$
	Ours	$\mathbf{.54 \pm .11}$	$\mathbf{.73 \pm .09}$	$\mathbf{.73 \pm .08}$	$\mathbf{.64 \pm .11}$	$\mathbf{.80 \pm .08}$	$\mathbf{.80 \pm .07}$	$\mathbf{.63 \pm .11}$	$\mathbf{.79 \pm .07}$	$\mathbf{.78 \pm .08}$
$\mathcal{D}_Q$	[9]	$.04 \pm .02$	$.17 \pm .08$	$.20 \pm .07$	$.21 \pm .09$	$.44 \pm .12$	$.41 \pm .11$	$.17 \pm .08$	$.39 \pm .12$	$.33 \pm .10$
	[2]	$.10 \pm .09$	$.29 \pm .14$	$.32 \pm .11$	$.12 \pm .07$	$.33 \pm .10$	$.28 \pm .09$	$.12 \pm .09$	$.34 \pm .11$	$.33 \pm .11$
	[8]	$.08 \pm .07$	$.27 \pm .12$	$.15 \pm .07$	$.14 \pm .08$	$.36 \pm .10$	$.38 \pm .09$	$.14 \pm .08$	$.36 \pm .10$	$.38 \pm .09$
	[11]	$.13 \pm .15$	$.31 \pm .20$	$.35 \pm .17$	$.13 \pm .14$	$.31 \pm .20$	$.34 \pm .18$	$.13 \pm .14$	$.31 \pm .20$	$.35 \pm .18$
	[19]	$.15 \pm .11$	$.37 \pm .13$	$.38 \pm .10$	$.37 \pm .12$	$.60 \pm .11$	$.59 \pm .09$	$.33 \pm .10$	$.57 \pm .10$	$.54 \pm .09$
	Ours	$\mathbf{.20 \pm .09}$	$\mathbf{.44 \pm .13}$	$\mathbf{.39 \pm .13}$	$\mathbf{.40 \pm .16}$	$\mathbf{.62 \pm .14}$	$\mathbf{.62 \pm .13}$	$\mathbf{.36 \pm .13}$	$\mathbf{.58 \pm .13}$	$\mathbf{.57 \pm .13}$

3.3 Ablation Study

We evaluate the effectiveness of three targeted adversarial attacks introduced in Sec. 2.1. As shown in Table 2, top part, the targeted attack based on morphological operations consistently outperforms the other attack strategies across epistemic, aleatoric, and total uncertainty measures. We also investigate the impact of the structuring element (kernel) size and the number of iterations in the morphology-based targeted attack. Table 2, second part, shows that kernel sizes of 3 and 5 yield the best performance across uncertainty maps, with erosion and dilation applied for three iterations. Furthermore, increasing the number of iterations M in the QUAM-SM algorithm (Sec. 2.1) consistently improves performance as shown in the last part of Table 2. The maximum number of iterations corresponds to the length of the training loader, i.e., $M = (\text{len}(\text{train-loader})/x) \times D$, where $x \in \{1, \ldots, |\text{train-loader}|\}$ and D denotes the number of adversarial models. For the REFUGE ($\mathcal{D}_R$) dataset, we set $|\text{train-loader}| = 100$ and $D = 2$.

4 Conclusions

In this paper, we presented QUAM-SM, a post-hoc uncertainty quantification framework for medical image segmentation based on targeted adversarial search.

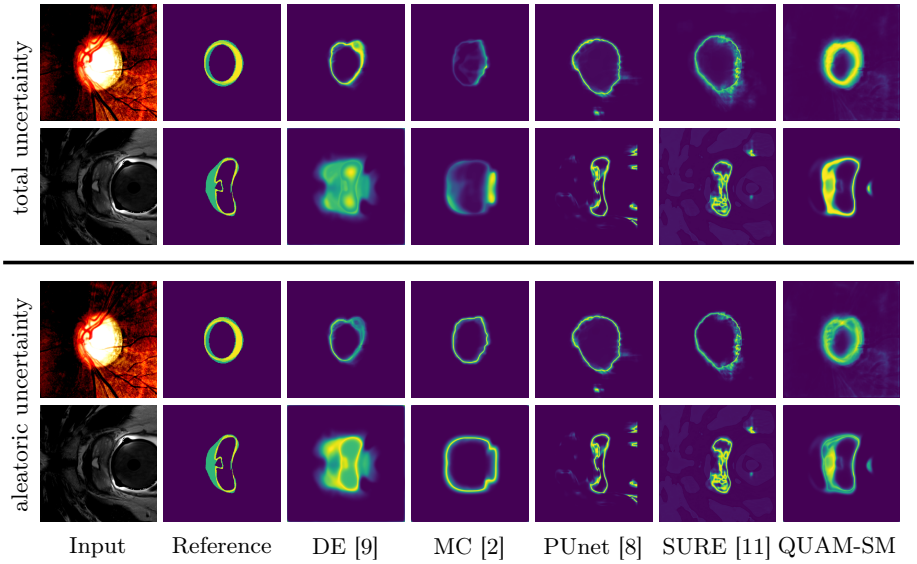

Fig. 2: Qualitative comparison between our method and state-of-the-art for total and aleatoric uncertainty on the same cases.

Table 2: Ablation study for the selection of *kernel size* k and M on REFUGE ($\mathcal{D}_R$). We set $\omega = 5$ for **TBM**.

$\mathbf{y_{mask}}$	Epistemic			Aleatoric			Total		
	R^2	PCC	SDice	R^2	PCC	SDice	R^2	PCC	SDice
GEM	.13 ± .09	.33 ± .14	.29 ± .12	.33 ± .15	.56 ± .14	.56 ± .12	.33 ± .15	.56 ± .14	.56 ± .12
TBM	.12 ± .10	.32 ± .14	.27 ± .11	.34 ± .16	.56 ± .14	.57 ± .13	.32 ± .16	.55 ± .14	.56 ± .13
MAM	**.38 ± .13**	**.61 ± .11**	**.62 ± .10**	**.51 ± .13**	**.71 ± .09**	**.71 ± .08**	**.50 ± .13**	**.70 ± .09**	**.69 ± .10**
k	Epistemic			Aleatoric			Total		
1	.33 ± .14	.56 ± .13	.53 ± .11	.52 ± .13	.71 ± .10	.70 ± .08	.49 ± .14	.69 ± .11	.70 ± .11
3	**.39 ± .14**	**.61 ± .12**	.61 ± .10	**.53 ± .13**	**.72 ± .09**	**.72 ± .08**	**.52 ± .14**	**.71 ± .10**	**.71 ± .11**
5	.38 ± .13	.60 ± .11	**.62 ± .10**	.51 ± .13	.71 ± .09	.71 ± .08	.50 ± .13	.70 ± .09	.69 ± .10
7	.35 ± .12	.59 ± .10	.60 ± .09	.47 ± .12	.68 ± .09	.69 ± .09	.47 ± .11	.68 ± .08	.66 ± .10
9	.33 ± .11	.56 ± .09	.58 ± .09	.44 ± .13	.65 ± .10	.66 ± .09	.44 ± .10	.66 ± .08	.63 ± .10
M	Epistemic			Aleatoric			Total		
2	.33 ± .13	.56 ± .12	.56 ± .11	.47 ± .14	.67 ± .11	.66 ± .09	.46 ± .13	.67 ± .10	.66 ± .10
4	.39 ± .13	.61 ± .11	.61 ± .10	.53 ± .13	.72 ± .09	.71 ± .08	.51 ± .13	.71 ± .09	.70 ± .10
10	.45 ± .13	.67 ± .09	.66 ± .08	.60 ± .11	.77 ± .07	.77 ± .07	.58 ± .11	.76 ± .07	.76 ± .08
20	.48 ± .13	.69 ± .10	.69 ± .09	.60 ± .11	.77 ± .07	.78 ± .07	.58 ± .11	.76 ± .07	.75 ± .08
40	.53 ± .12	.72 ± .09	.72 ± .08	**.64 ± .10**	**.80 ± .07**	**.80 ± .07**	**.63 ± .11**	**.79 ± .07**	**.78 ± .07**
100	.53 ± .12	.72 ± .08	.72 ± .08	.63 ± .10	.79 ± .07	.79 ± .07	.62 ± .10	.78 ± .07	.77 ± .07
200	**.54 ± .12**	**.73 ± .08**	**.73 ± .08**	.63 ± .11	.79 ± .07	.79 ± .07	.61 ± .10	.78 ± .07	.77 ± .07

The method identifies adversarially fragile pixels to reveal prediction instability and mitigate the clinical risks of overconfident model decisions. QUAM-SM produces interpretable epistemic and aleatoric uncertainty maps for more reliable model assessment. While the approach requires access to the training distribution, experiments on realistic clinical benchmarks show that the estimated aleatoric uncertainty correlates more strongly with inter-rater variability than total uncertainty, highlighting its relevance for capturing annotation ambiguity and supporting safer clinical decision-making.

Acknowledgments This work was supported in part by the Austrian Science Fund (FWF) grant 10.55776/FG9. TP is funded by the European Union (I-SCREEN, grant no. 101130093), EIC-2023-PATHFINDEROPEN-01. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Disclosure of Interests The authors have no competing interests to declare.

References

1. Becker, A.S., Chaitanya, K., Schawkat, K., Muehlematter, U.J., Hötker, A.M., Konukoglu, E., Donati, O.F.: Variability of manual segmentation of the prostate in axial T2-weighted MRI: A multi-reader study. European Journal of Radiology **121**, 108716 (2019)
2. Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P., Kooi, E., van der Lugt, A., de Bruijne, M.: Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In: Sudre, C.H., Fehri, H., Arbel, T., Baumgartner, C.F., Dalca, A., Tanno, R., Van Leemput, K., Wells, W.M., Sotiras, A., Papiez, B., Ferrante, E., Parisot, S. (eds.) Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. pp. 32–41. Springer International Publishing, Cham (2020)
3. Diaz-Hurtado, M., Martínez-Heras, E., Solana, E., Casas-Roma, J., Llufriu, S., Kanber, B., Prados, F.: Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review. Neuroradiology **64**(11), 2103–2117 (2022)
4. Gao, L., Cheng, Y., Zhang, Q., Xu, X., Song, J.: Feature space targeted attacks by statistic alignment. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI). pp. 671–677 (8 2021), `https://doi.org/10.24963/ijcai.2021/93`
5. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**(Suppl 1), 1513–1589 (2023)
6. Hesterberg, T.: Weighted average importance sampling and defensive mixture distributions. Technometrics **37**(2), 185–194 (1995)

7. Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., Reyes, M.: On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 682–690. Springer (2018)
8. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic U-Net for segmentation of ambiguous images. Advances in neural information processing systems **31** (2018)
9. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
10. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. Artif. Intell. Medicine **150**, 102830 (2024)
11. Li, Y., Sui, A., Wu, F., Zhuang, X.: Uncertainty-supervised interpretable and robust evidential segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 649–658. Springer Nature Switzerland, Cham (2026)
12. Liu, X., Xing, F., Marin, T., El Fakhri, G., Woo, J.: Variational inference for quantifying inter-observer variability in segmentation of anatomical structures. In: Medical Imaging 2022: Image Processing. vol. 12032, pp. 438–443. SPIE (2022)
13. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning where to look for the pancreas. In: Medical Imaging with Deep Learning (2018), https://openreview.net/forum?id=Skft7cijM
14. Orlando, J.I., Fu, H., Barbosa Breda, J., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., Lee, J., Li, X., Liu, P., Lu, S., Murugesan, B., Naranjo, V., Phaye, S.S.R., Shankaranarayana, S.M., Sikka, A., Son, J., van den Hengel, A., Wang, S., Wu, J., Wu, Z., Xu, G., Xu, Y., Yin, P., Li, F., Zhang, X., Xu, Y., Bogunović, H.: REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Analysis **59**, 101570 (Jan 2020), http://dx.doi.org/10.1016/j.media.2019.101570
15. Roshanzamir, P., Rivaz, H., Ahn, J., Mirza, H., Naghdi, N., Anstruther, M., Battié, M.C., Fortin, M., Xiao, Y.: How inter-rater variability relates to aleatoric and epistemic uncertainty: a case study with deep learning-based paraspinal muscle segmentation. In: International workshop on uncertainty for safe utilization of machine learning in medical imaging (UNSURE). pp. 74–83. Springer (2023)
16. Schweighofer, K., Aichberger, L., Ielanskyi, M., Klambauer, G., Hochreiter, S.: Quantification of uncertainty with adversarial models. Advances in Neural Information Processing Systems **36**, 19446–19484 (2023)
17. Shi, Y., Zu, C., Yang, P., Tan, S., Ren, H., Wu, X., Zhou, J., Wang, Y.: Uncertainty-weighted and relation-driven consistency training for semi-supervised head-and-neck tumor segmentation. Knowledge-Based Systems **272**, 110598 (2023)
18. Tan, H.S., Wang, K., Mcbeth, R.: Uncertainty-error correlations in evidential deep learning models for biomedical segmentation. In: International Conference on Technologies and Applications of Artificial Intelligence. pp. 91–105. Springer (2024)
19. Wang, G., Li, W., Ourselin, S., Vercauteren, T.: Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In: Crimi,

A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 61–72. Springer International Publishing, Cham (2019)
20. Yu, P., Song, K., Lu, J.: Generating adversarial examples with conditional generative adversarial net. In: 24th International Conference on Pattern Recognition (ICPR). pp. 676–681 (2018). https://doi.org/10.1109/ICPR.2018.8545152