Uncertainty Quantification (UQ) in Medical Image Segmentation

Tristan Kirscher
ICUBE IMAGeS, ICANS

Data Science & AI Workshop
May 6, 2025
ICUBE

Information Sciences & Technology
for Healthcare | HealthTech
The interdisciplinary thematic institutes
of the University of Strasbourg & cnrs & Inserm

Why Quantify Uncertainty?

- Organ segmentation for radiotherapy demands **high reliability**.
- UQ enables:
  - **Failure detection** (flag total segmentation failures)
  - **Calibration** (align confidence with true error rates)

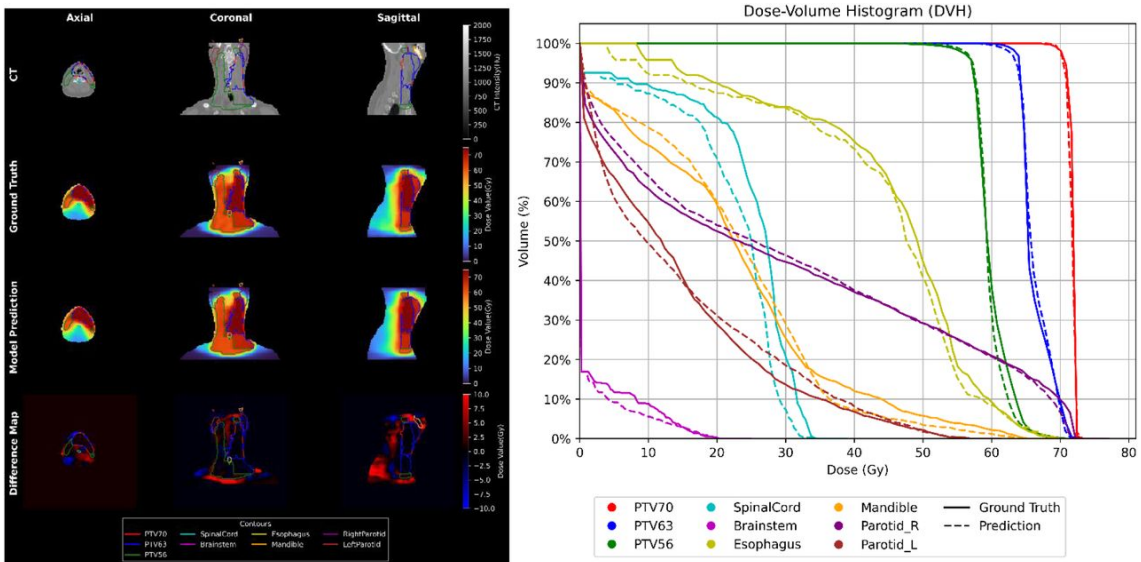


Figure: Dose–Volume Histogram in radiotherapy planning (Chang et al.)

Why is it Hard?

- Manual delineations vary (inter-rater Dice 0.7–0.9).
- **No single "ground truth" mask**; annotation is ambiguous.
- Multiple annotations are required to quantify inter-rater variability.

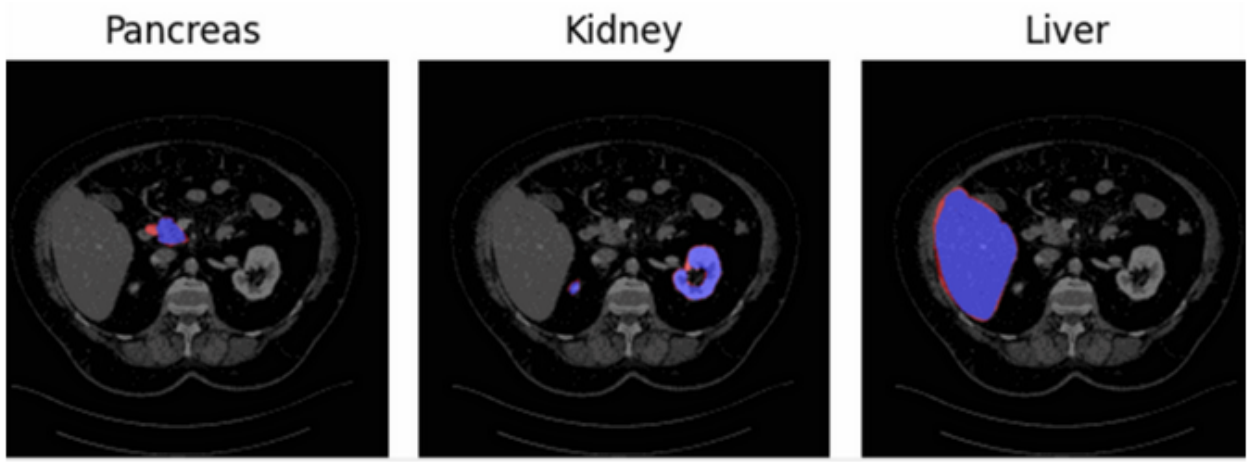


Figure: Inter-rater variability across anatomical structures (CURVAS dataset)

Uncertainty: Aleatoric or Epistemic?

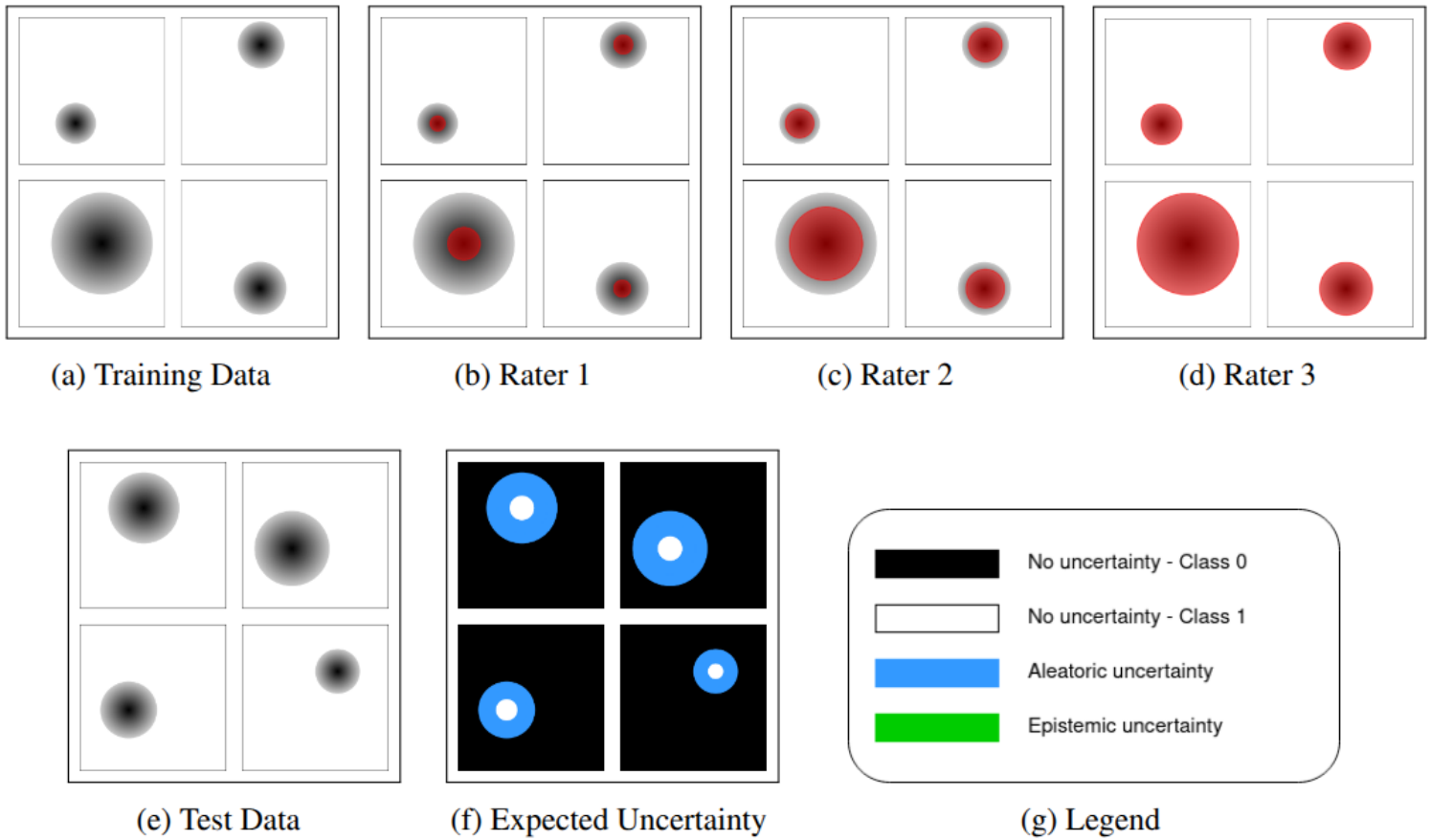
Notation: x input image, y mask, D dataset, θ model parameters.

$$\text{PU} = H[p(y \mid x, D)] = \underbrace{\mathbb{E}_{\theta \sim p(\theta|D)}[H(p(y \mid x, \theta))]}_{\text{Aleatoric}} + \underbrace{I(y, \theta \mid x, D)}_{\text{Epistemic}}$$

- **Aleatoric:** irreducible noise from imaging and annotation.
- **Epistemic:** reducible model uncertainty due to data/model limits.

ValUES: *A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation* [Kahl et al.(2024)]

Uncertainty Types: Aleatoric Example

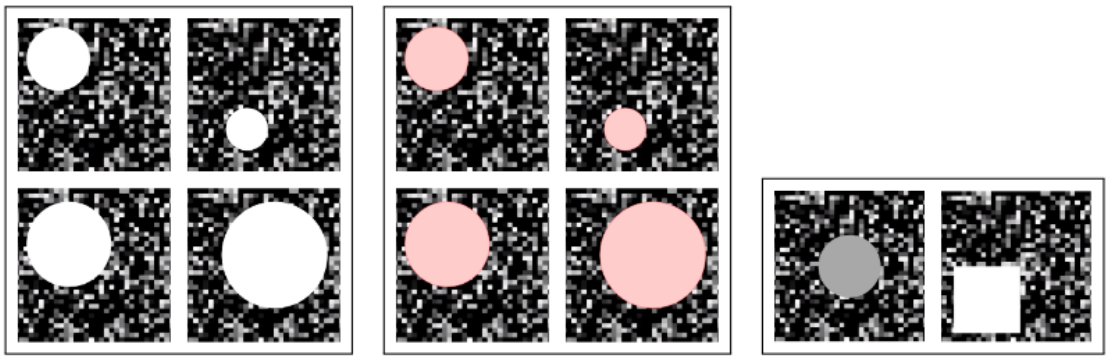


(a) Training Data (b) Rater 1 (c) Rater 2 (d) Rater 3

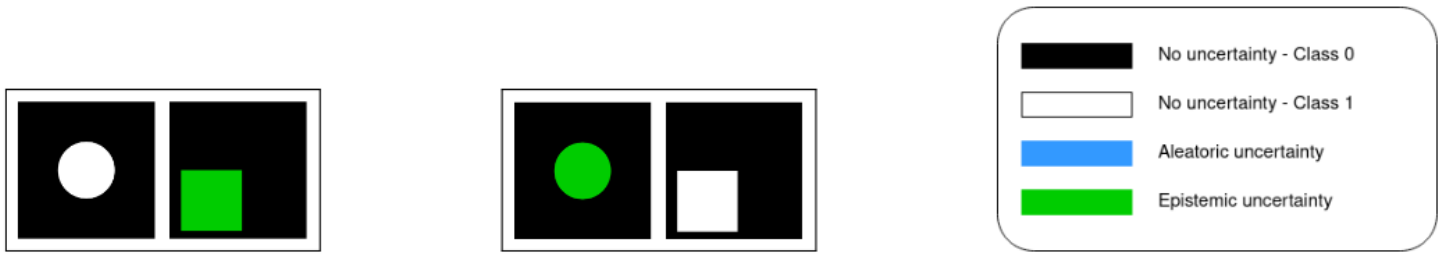
(e) Test Data (f) Expected Uncertainty (g) Legend

Toy example of annotation noise (ValUES paper)

Uncertainty Types: Epistemic Example



(a) Training Data (b) Ground Truth (c) Test Data



(d) Expected Uncertainty - Network Learned Shape (e) Expected Uncertainty - Network Learned Intensity (f) Legend

Toy example of model uncertainty under domain shift (adapted from ValUES paper)

Approaches to Uncertainty Quantification

Aggregation of plausible segmentations

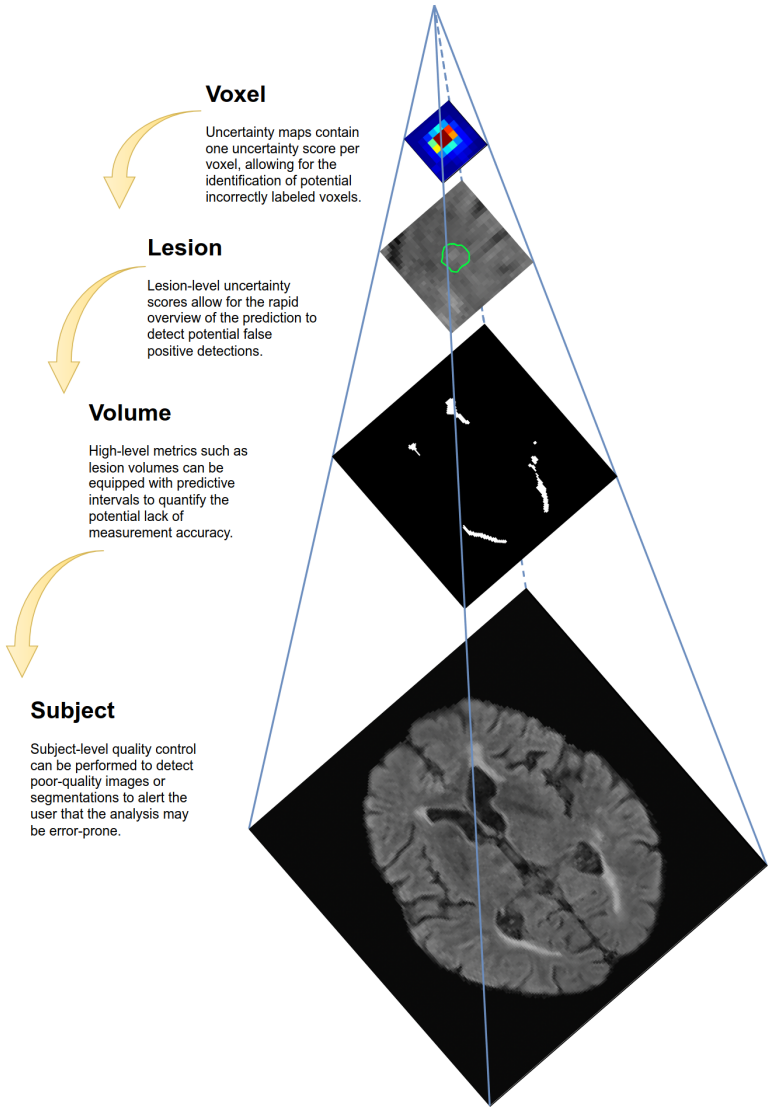
- **Bayesian Neural Networks:** [Blundell et al.(2015)]
  Approximate $p(\theta \mid D)$ (e.g. variational inference)
  $\rightarrow$ sample M plausible segmentation masks by drawing weights $\{\theta_m\}$
- **MC-Dropout:** [Gal & Ghahramani(2016)]
  At test time, perform M stochastic forward passes with dropout
  $\rightarrow$ sample M plausible masks
- **Deep Ensembles:** [Lakshminarayanan et al.(2017)]
  Train M independent networks with different random seeds
  $\rightarrow$ sample M masks (one from each model)

Calibration set

- **Conformal Prediction:** [Angelopoulos & Bates(2020)]
  Uses a held-out calibration set to construct prediction sets with guaranteed coverage, without requiring sampling or aggregation.

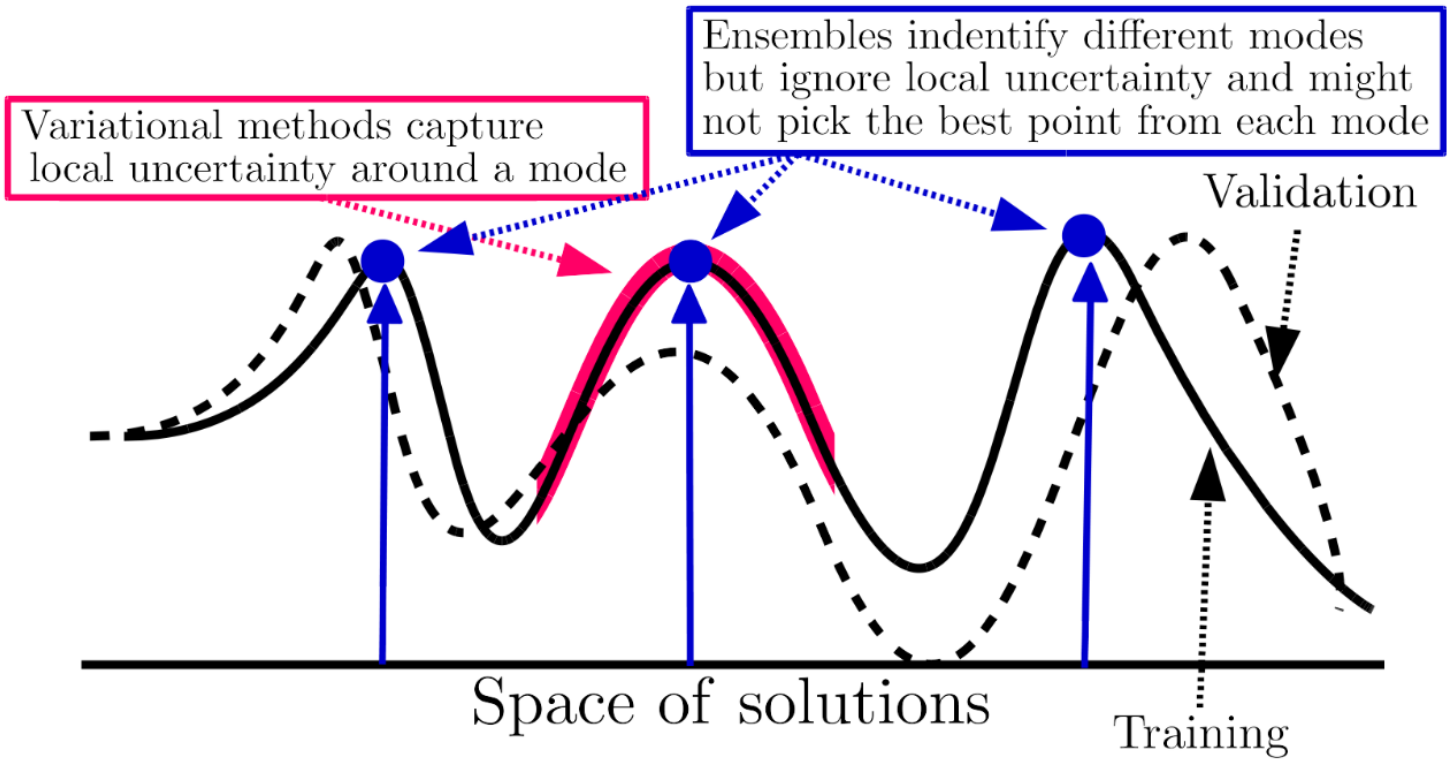
Focus on **Deep Ensembles** for state-of-the-art performance.

How to Aggregate Uncertainty Across Levels?



Uncertainty scales: voxel $\rightarrow$ lesion $\rightarrow$ volume $\rightarrow$ subject [Lambert(2024)]

Why Deep Ensembles Excel?



Different seeds converge to distinct modes [Fort et al.(2019)]

Key Takeaways and Research Directions

- **Aleatoric vs Epistemic**: data noise (irreducible) vs model uncertainty (reducible)
- **UQ methods**: Point estimate $\rightarrow$ distribution of segmentations $\rightarrow$ quantified uncertainty (Bayesian, Dropout, Deep Ensembles...)
- **Deep Ensembles**: multi-mode coverage $\rightarrow$ more diversity for reliable UQ (Lakshminarayanan '17; Fort '19)

Next steps:

1. How to quantify and maximize functional diversity in ensembles?
2. How to efficiently aggregate 1) predictions 2) voxel uncertainty ?
3. Integrate UQ into radiotherapy pipeline and measure DVH impact

Ready for questions!

References I

K.-C. Kahl *et al.*

ValUES: A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation.
ICLR 2024.
https://arxiv.org/abs/2401.08501

B. Lakshminarayanan, A. Pritzel & C. Blundell.

Simple and scalable predictive uncertainty estimation using deep ensembles.
NeurIPS 2017.
https://arxiv.org/abs/1612.01474

S. Fort, H. Hu & B. Lakshminarayanan.

Deep Ensembles: A Loss Landscape Perspective.
arXiv:1912.02757, 2019.
https://arxiv.org/abs/1912.02757

Y. Gal & Z. Ghahramani.

Dropout as a Bayesian approximation: Representing model uncertainty in deep learning.
ICML 2016.
https://arxiv.org/abs/1506.02142

C. Blundell *et al.*

Weight uncertainty in neural networks.
ICML 2015.
https://arxiv.org/abs/1505.05424

A. Angelopoulos & S. Bates.

A gentle introduction to conformal prediction and distribution-free uncertainty quantification.
arXiv:2002.04182, 2020.
https://arxiv.org/abs/2002.04182

References II

B. Lambert.

Quantification et caractérisation de l'incertitude de segmentation d'images médicales par des réseaux profonds.
PhD thesis, Université Grenoble Alpes, 2024.
https://theses.hal.science/tel-04673383