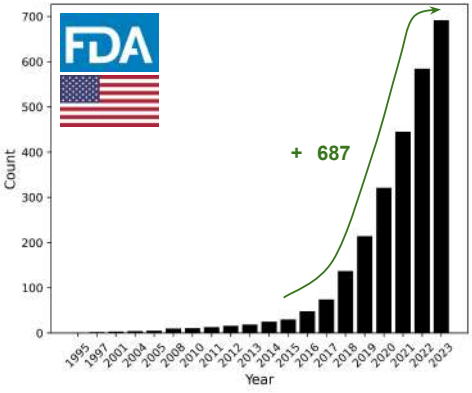
Uncertainty Quantification in Deep Learning-based Medical Image Segmentation

Benjamin Lambert
Supervisors: Michel Dojat, Florence Forbes, Senan Doyle

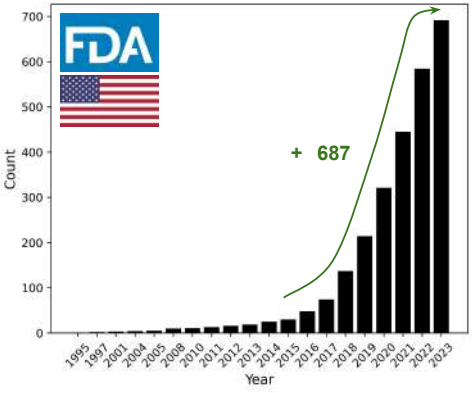
May 16th, 2024, Grenoble

The Raise of AI in Radiology

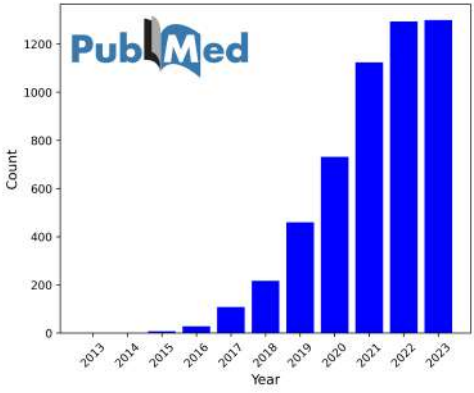


Number of AI-based commercial medical devices approved by the Food and Drug Administration

The Raise of AI in Radiology

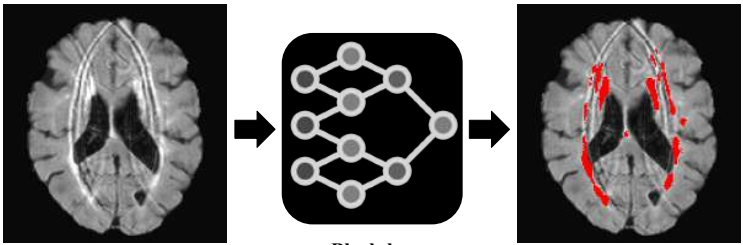


Number of AI-based commercial medical devices approved by the Food and Drug Administration



Number of papers including "Deep Learning" and "Medical Image Analysis"

The black-box effect in Deep Learning

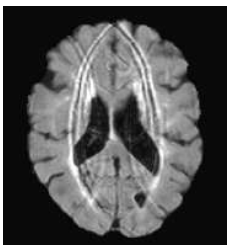


Input MRI

Black-box
Deep neural network

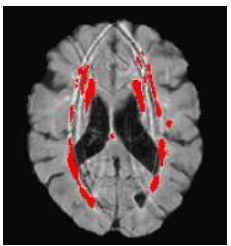
Prediction

The black-box effect in Deep Learning



Input MRI

Black-box
Deep neural network



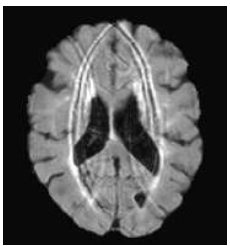
Prediction

End user

Explanation of the result

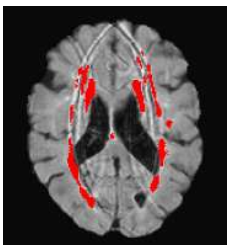
Confidence estimation

The black-box effect in Deep Learning



Input MRI

Black-box
Deep neural network



Prediction

End user

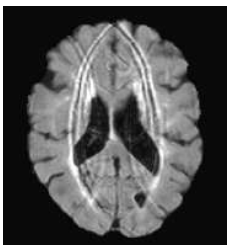
Explainable AI

Explanation of the result

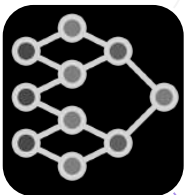
Uncertainty Quantification

Confidence estimation

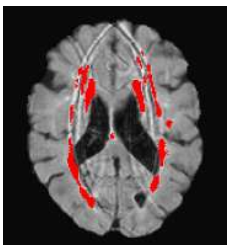
The black-box effect in Deep Learning



Input MRI



Black-box
Deep neural network



Prediction

End user

Explainable AI

Explanation of the result

Prevent silent errors

Uncertainty Quantification

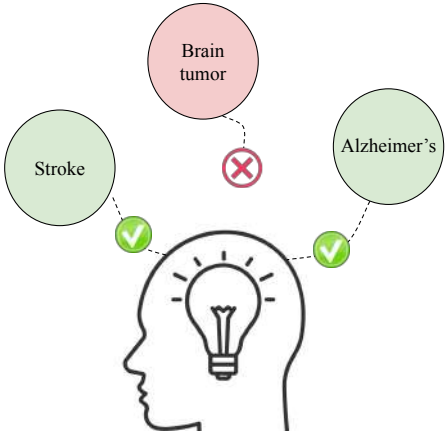
Confidence estimation

What is uncertainty?

Aleatoric Uncertainty

- Inherent randomness in the data.
- Cannot be reduced.

What is uncertainty?



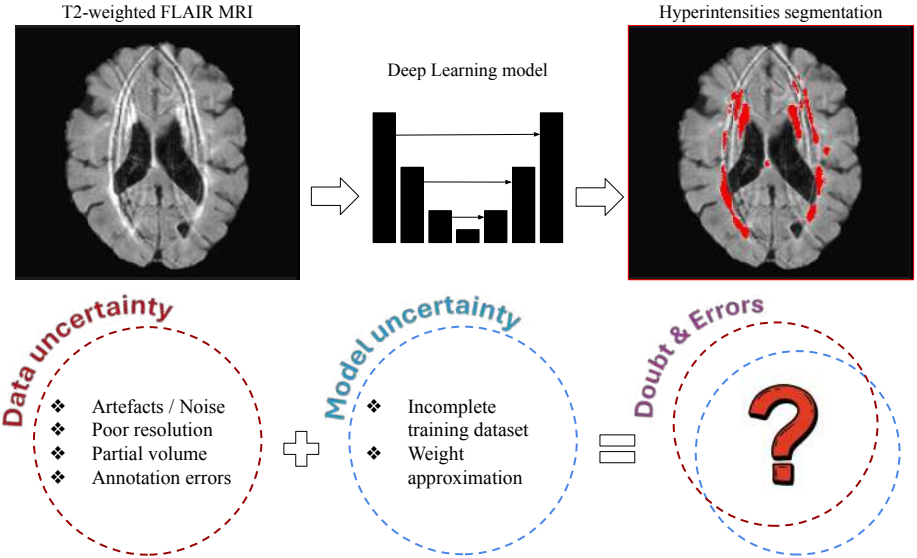
Aleatoric Uncertainty

- Inherent randomness in the data.
- Cannot be reduced.

Epistemic Uncertainty

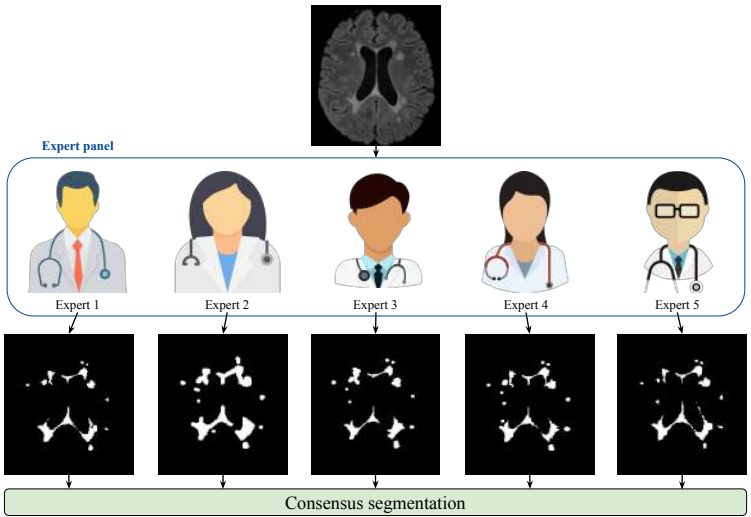
- Lack of knowledge of the model.
- Can be reduced with more data.

Uncertainty in Medical Image Analysis



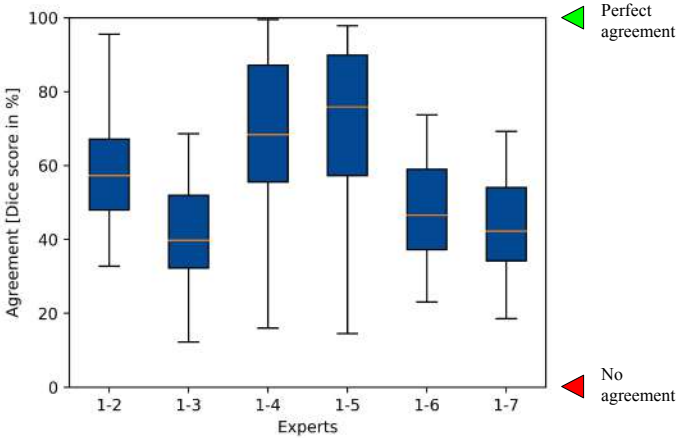
Why measure uncertainty?

Establishing the truth in medical image analysis



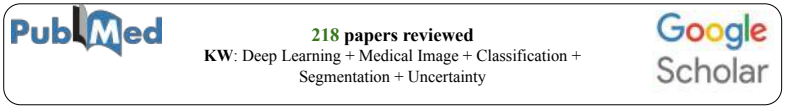
Why measure uncertainty?

Experts agreement on the MSSEG dataset (Commowick et al. 2021).



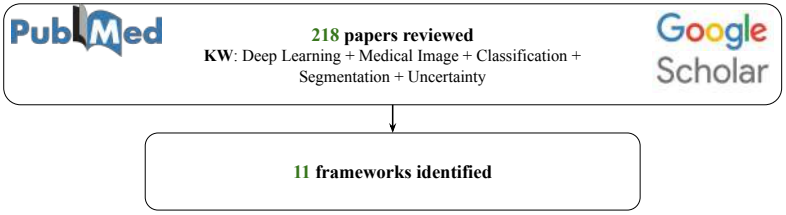
Hard and uncertain tasks even for human experts: high inter-rater variability.

Review of existing work*



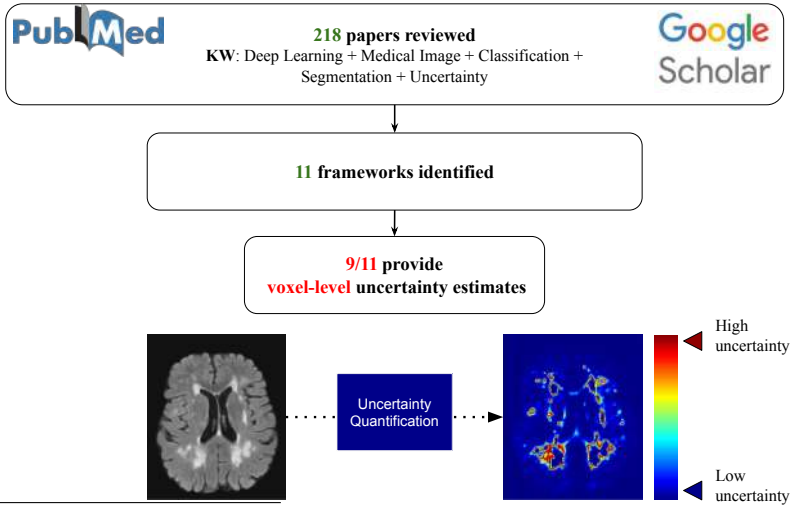
*B. Lambert et al. (2024). "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis". In: *Artificial Intelligence in Medicine*, p. 102830

Review of existing work*



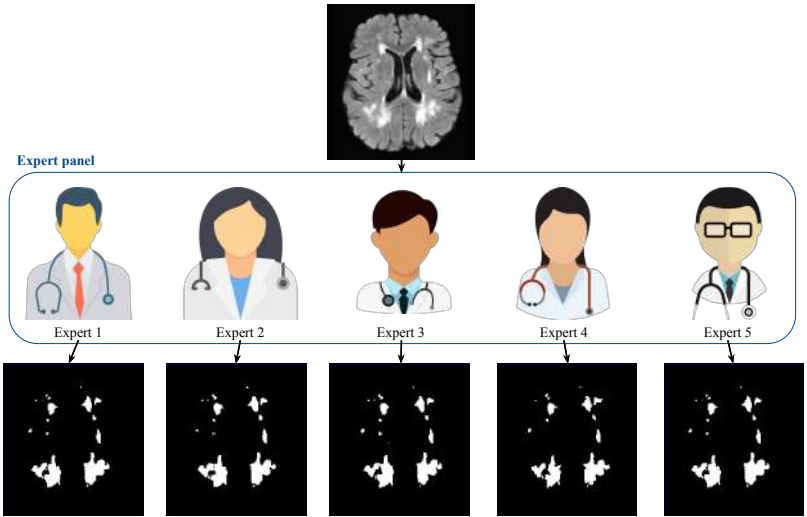
*B. Lambert et al. (2024). "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis". In: *Artificial Intelligence in Medicine*, p. 102830

Review of existing work*

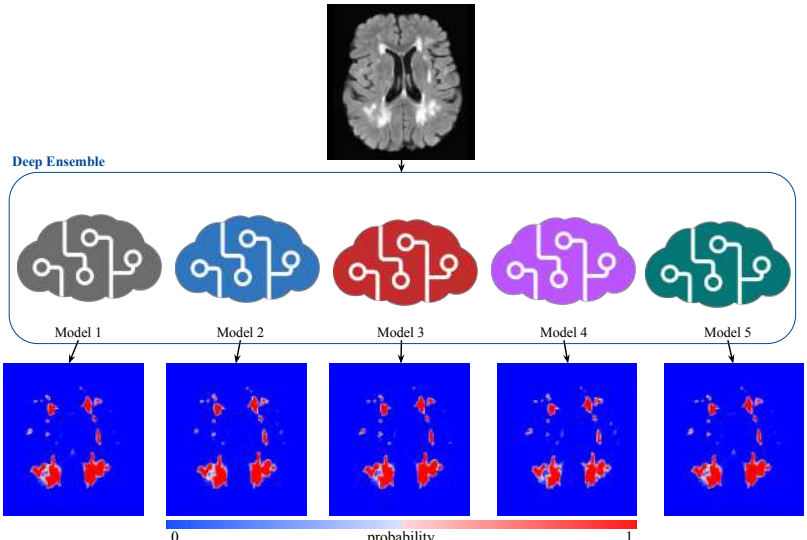


*B. Lambert et al. (2024). "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis". In: *Artificial Intelligence in Medicine*, p. 102830

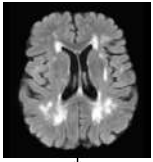
A Popular baseline for voxel uncertainty: Ensembling



A Popular baseline for voxel uncertainty: Ensembling



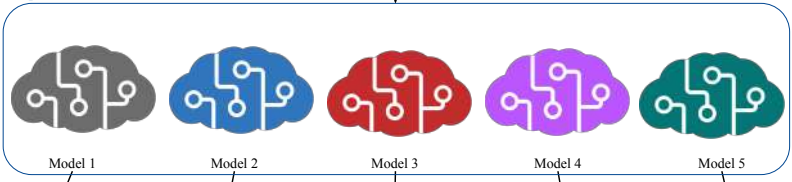
A Popular baseline for voxel uncertainty: Ensembling

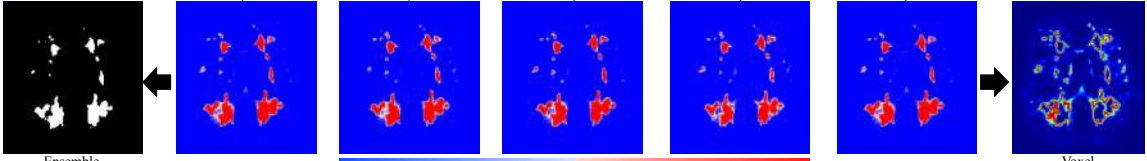


Entropy (C: number of classes, M: number of models)

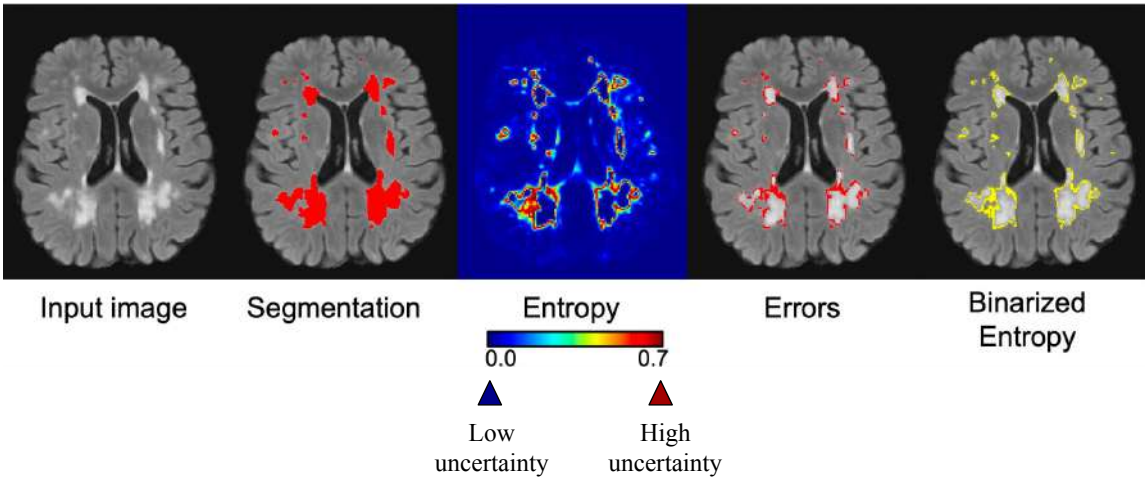
$$H(p) = -\sum_{c=1}^{C} \frac{1}{M} \sum_{m=1}^{M} p_{c,m} \log\left(\frac{1}{M} \sum_{m=1}^{M} p_{c,m}\right)$$

Deep Ensemble





A Popular baseline for voxel uncertainty: Ensembling



Challenges and Objectives

Current issues in Uncertainty Quantification (UQ)

- UQ studies essentially restrict to voxel-level estimates.
- Potential misalignment with the clinical interest, which can be at a higher level (e.g. lesion, volume, or image level).

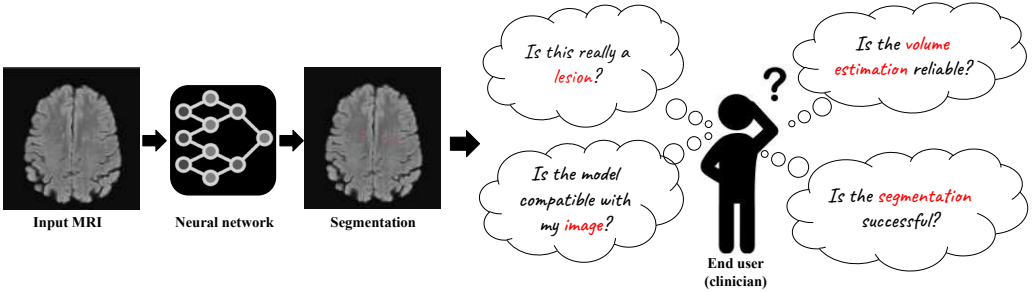
Challenges and Objectives

Current issues in Uncertainty Quantification (UQ)

- UQ studies essentially restrict to voxel-level estimates.
- Potential misalignment with the clinical interest, which can be at a higher level (e.g. lesion, volume, or image level).

Long-term objective

Propose uncertainty estimates aligned with the clinical need.



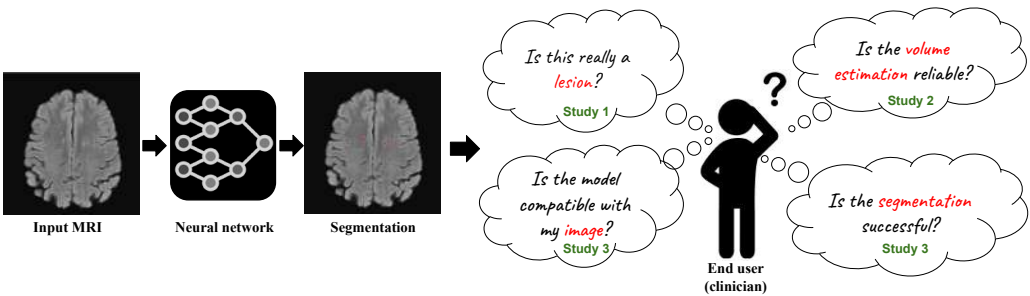
Challenges and Objectives

Current issues in Uncertainty Quantification (UQ)

- UQ studies essentially restrict to voxel-level estimates.
- Potential misalignment with the clinical interest, which can be at a higher level (e.g. lesion, volume, or image level).

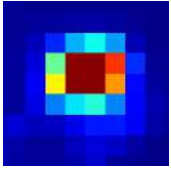
Long-term objective

Propose uncertainty estimates aligned with the clinical need.

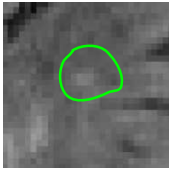


Presentation Outline

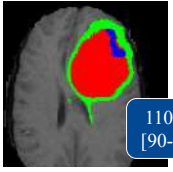
Voxel-level
(Most existing literature)



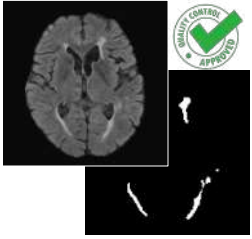
Study 1:
Lesion-level



Study 2:
Predictive Intervals for Volumetry

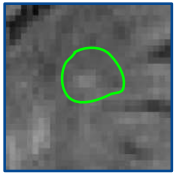


Study 3:
Subject-level

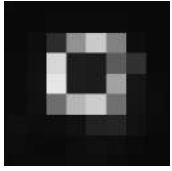


Presentation Outline

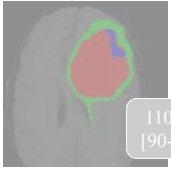
Study 1:
Lesion-level



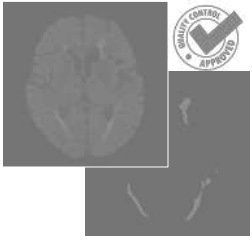
Voxel-level
(Most existing literature)



Study 2:
Predictive Intervals for Volumetry



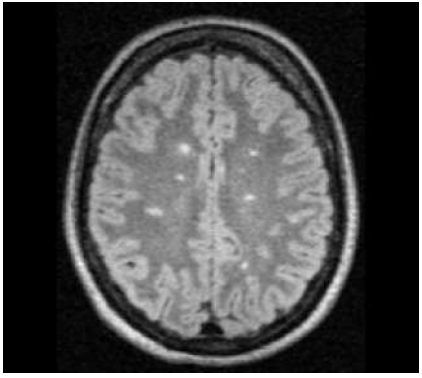
Study 3:
Subject-level



Lesion uncertainty: Motivations

Multiple Sclerosis

- Neurodegenerative disease causing **lesions** in the central nervous system.
- Lesions are disseminated in **space**.

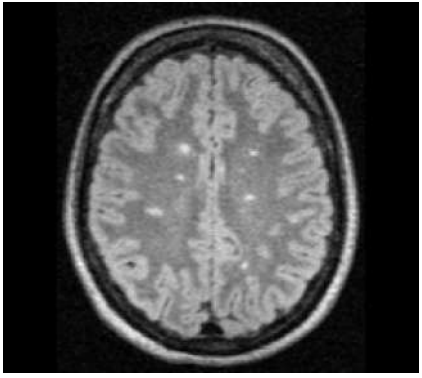


T2-weighted FLAIR MRI

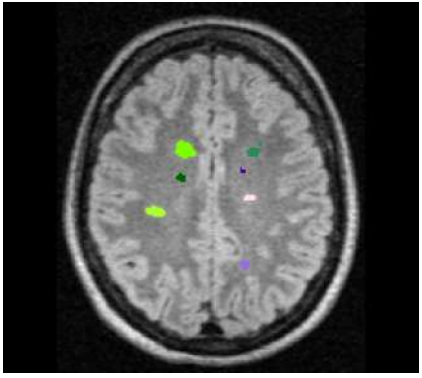
Lesion uncertainty: Motivations

Multiple Sclerosis

- Neurodegenerative disease causing **lesions** in the central nervous system.
- Lesions are disseminated in **space**.



T2-weighted FLAIR MRI



Manual segmentation of the lesions

Lesion uncertainty: Proposed paradigm

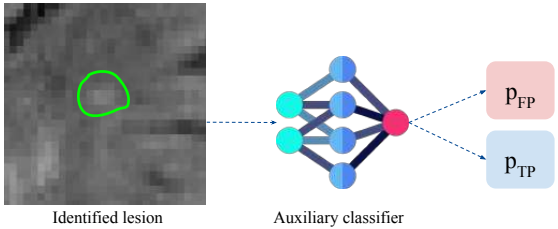
Paradigm

- Lesion-level uncertainty estimates should support the identification of incorrect findings (false positive instances).

Lesion uncertainty: Proposed paradigm

Paradigm

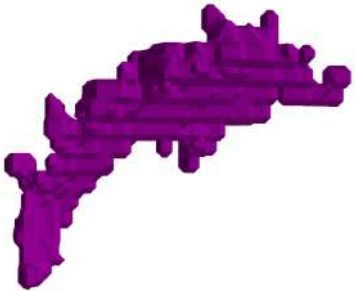
- Lesion-level uncertainty estimates should support the identification of incorrect findings (false positive instances).
- We propose to use p_{FP} the probability that the lesion is a false positive, as lesion uncertainty score.
- In practice, an auxiliary classifier is used to estimate the p_{FP}.



Challenges in classifying lesion instances

Challenges

- Lesions are heterogeneous in shape, size, and position.
- A frugal classifier is desired due to the scarcity of lesion examples.



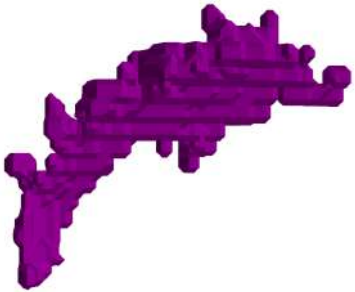
3820 mm^3

71 mm^3

Challenges in classifying lesion instances

Challenges

- Lesions are heterogeneous in shape, size, and position.
- A frugal classifier is desired due to the scarcity of lesion examples.

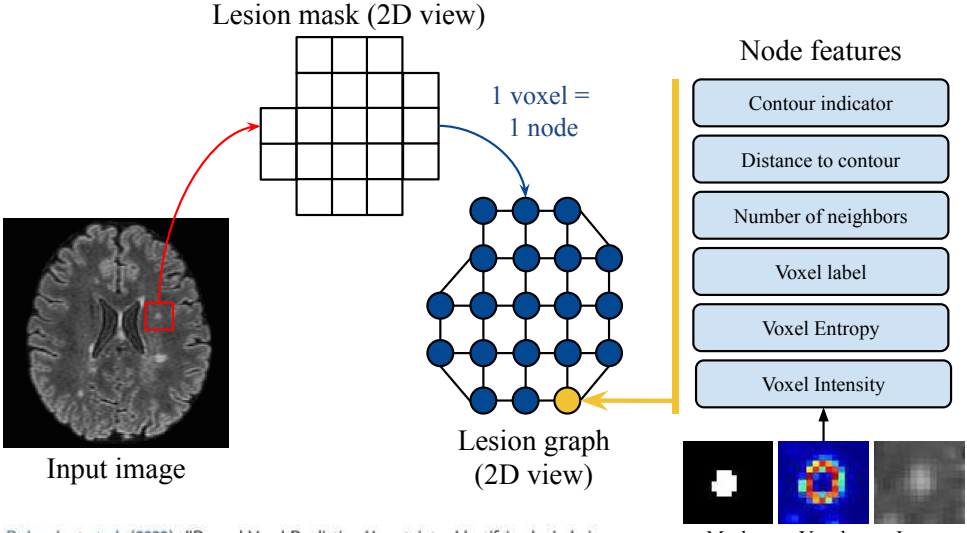


3820 mm^3

71 mm^3

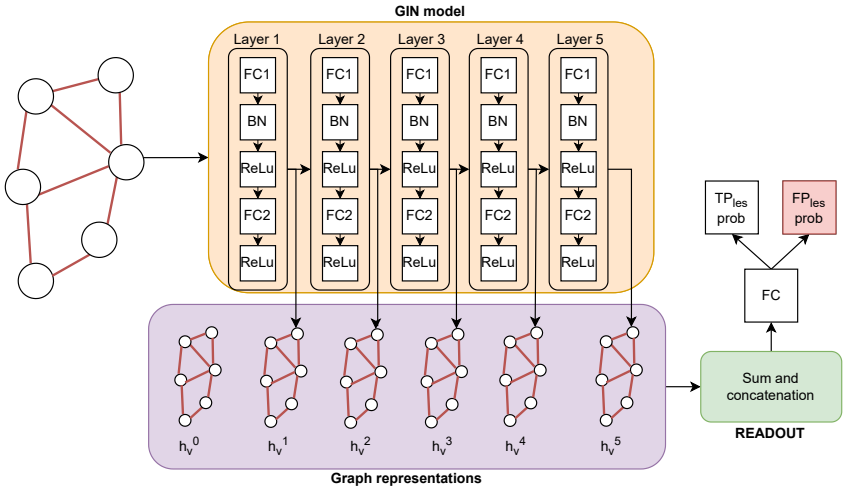
$\Rightarrow$ opt for a graph model approach.

Converting lesions into graphs



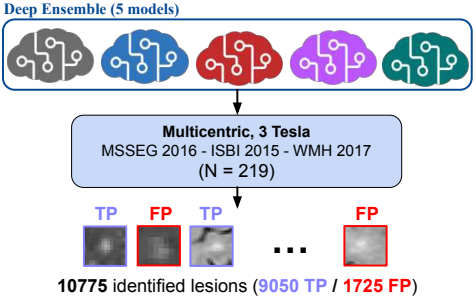
B. Lambert et al. (2022). "Beyond Voxel Prediction Uncertainty: Identifying brain lesions you can trust". In: *iMIMIC 2022, Held in Conjunction with MICCAI 2022*. LNCS 13611, pp. 61–70

Graph classification using a Graph Isomorphism Network (Wang et al. 2022)

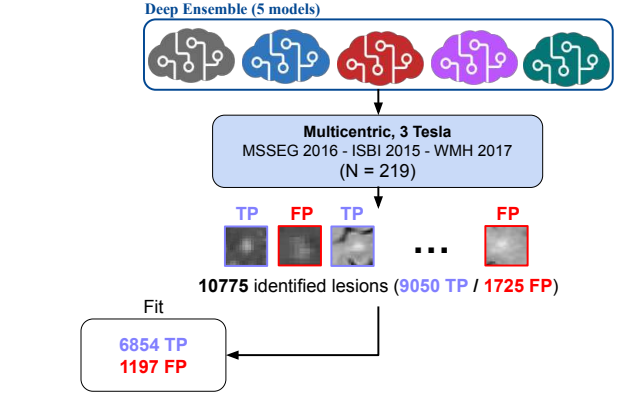


FC: Fully-connected layer. BN: Batch Normalization. ReLU: Rectified Linear Unit. Parameters: **26 700**.

Experimental setting



Experimental setting



Experimental setting

Multicentric, 3 Tesla
MSSEG 2016 - ISBI 2015 - WMH 2017
(N = 219)

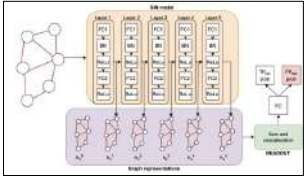
10775 identified lesions (9050 TP / 1725 FP)

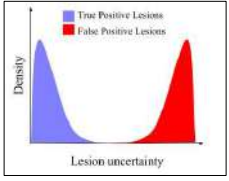
Fit

6854 TP
1197 FP

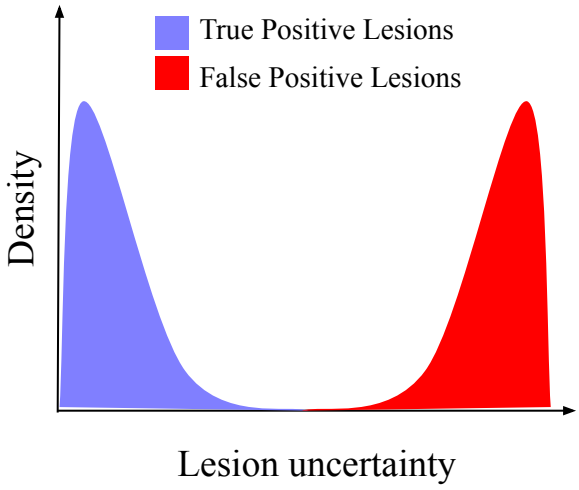
Test

2196 TP
528 FP

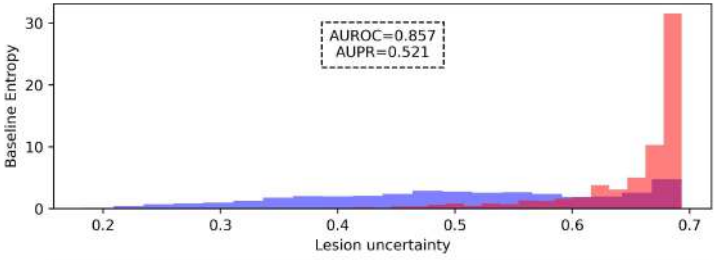




Ideal lesion uncertainty estimates

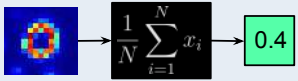


Results - Lesion uncertainty



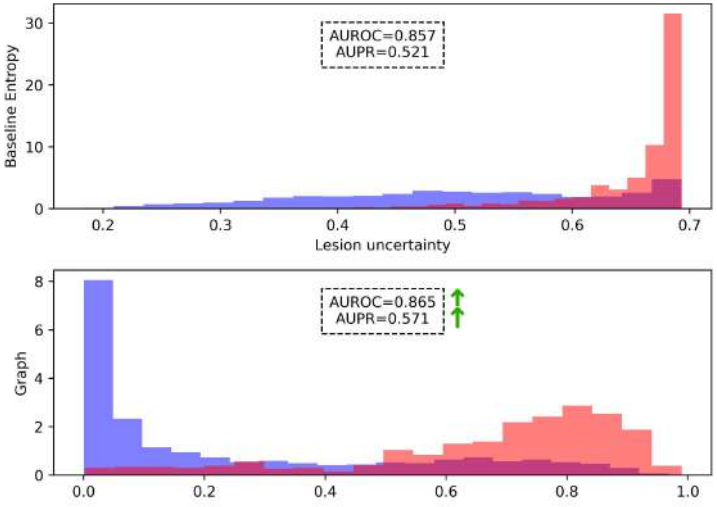
Baseline

Average of voxel-level uncertainty scores.



Densities of uncertainty scores for True and False Positive lesions.

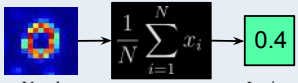
Results - Lesion uncertainty



Densities of uncertainty scores for True and False Positive lesions.

Baseline

Average of voxel-level uncertainty scores.

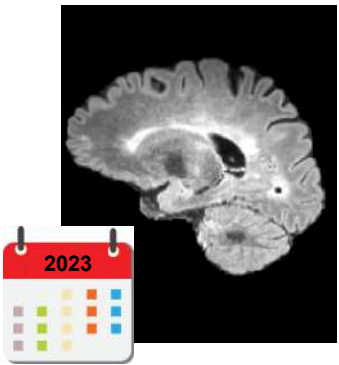


Analysis

- Slight gain over the baseline.
- The proposed score is easily interpretable.

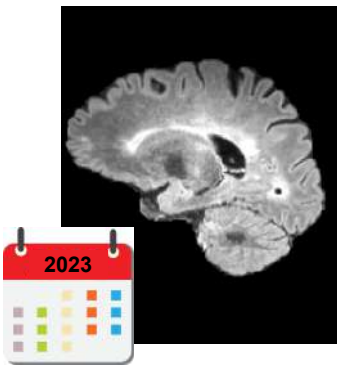
Towards longitudinal lesion uncertainty

Cross-sectional Analysis

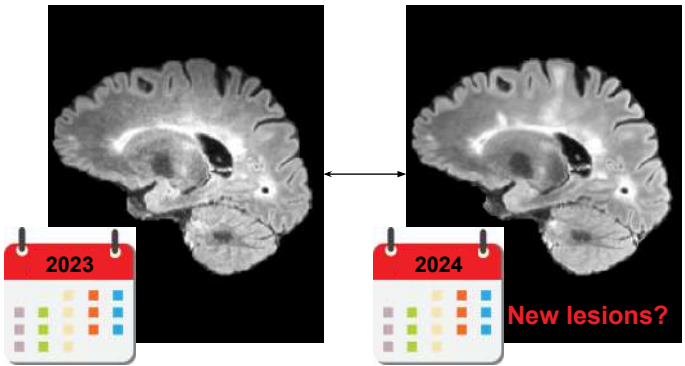


Towards longitudinal lesion uncertainty

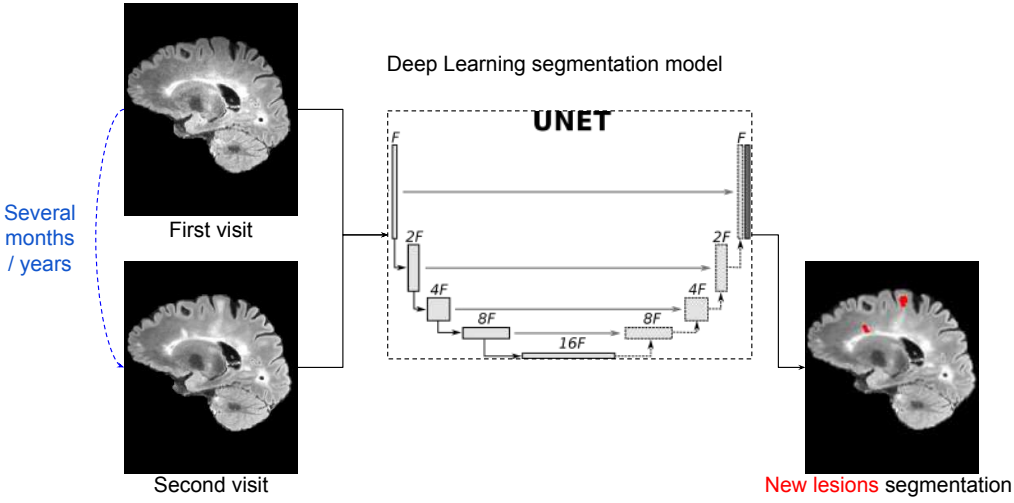
Cross-sectional Analysis



Longitudinal Analysis



Towards longitudinal lesion uncertainty



Experimental setting - Longitudinal lesion uncertainty

Longitudinal, Multicentric, 3 Tesla
MSSEG-2
(N = 100, 2 time points)

TP FP TP ... FP

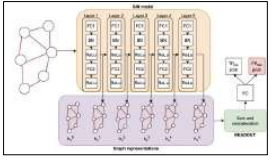
437 identified lesions (289 TP / 148 FP)
(versus **10775** for the 1st experiment, **25x less**)

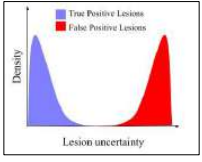
Fit

103 TP
28 FP

Test

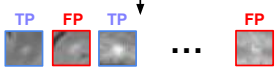
186 TP
120 FP





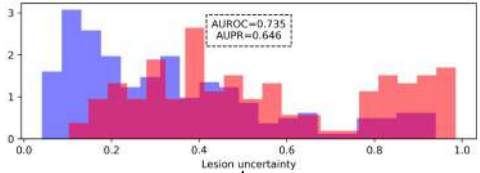
Experimental setting - Longitudinal lesion uncertainty

Longitudinal, Multicentric, 3 Tesla
MSSEG-2
(N = 100, 2 time points)



437 identified lesions (289 TP / 148 FP)
(versus **10775** for the 1st experiment, **25x less**)

Poor discrimination of True and False Positive lesions

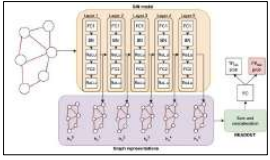


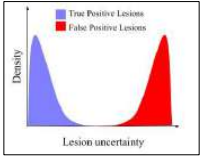
Fit

103 TP
28 FP

Test

186 TP
120 FP





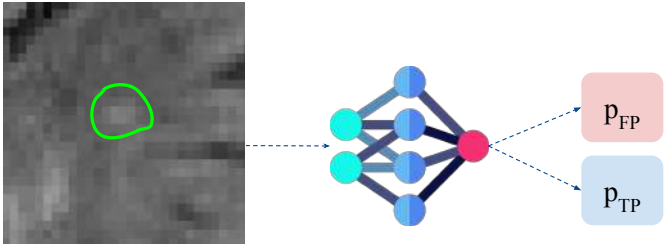
Experimental setting - Longitudinal lesion uncertainty

Identified limit:

- Need enough examples of True and False Positive lesions to train the GIN model.
  - **Cross-sectional** Multiple Sclerosis: $\approx$ 10000 instances for train / test.
  - **Longitudinal** Multiple Sclerosis: $\approx$ 450 lesion instances for train / test $\not\Rightarrow$ supervised training

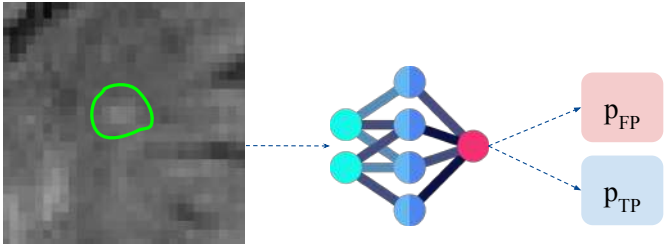
Study summary: lesion uncertainty module

- We propose to use an **auxiliary classifier** to estimate uncertainty at the lesion level.



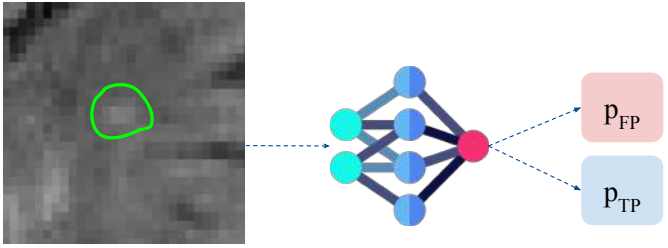
Study summary: lesion uncertainty module

- We propose to use an **auxiliary classifier** to estimate uncertainty at the lesion level.
- Lesions are first converted into **graphs** and then processed by a Graph neural network.

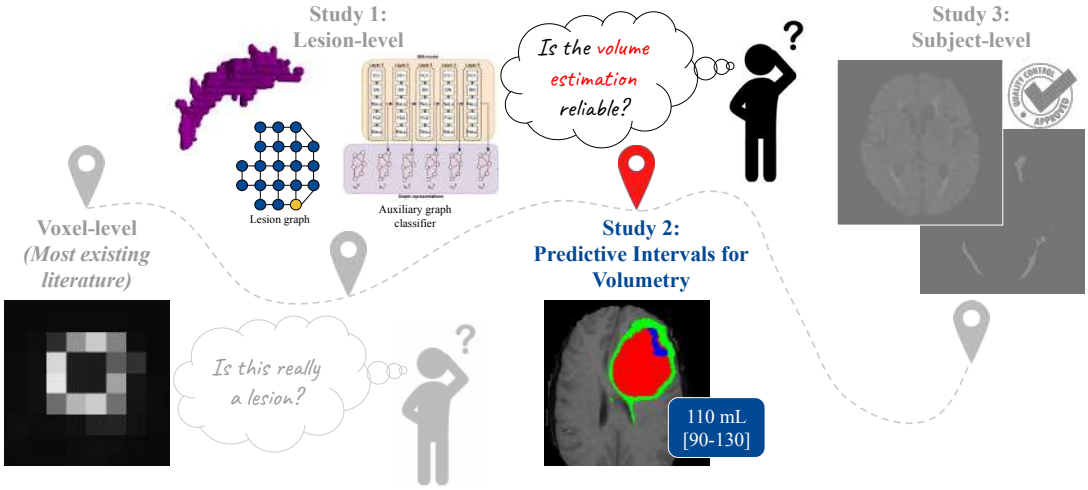


Study summary: lesion uncertainty module

- We propose to use an **auxiliary classifier** to estimate uncertainty at the lesion level.
- Lesions are first converted into **graphs** and then processed by a Graph neural network.
- The model is trained in a standard binary classification setting to distinguish between **True and False Positive** lesions.



Outline


Study 1:
Lesion-level
Lesion graph
Auxiliary graph classifier
Is the volume estimation reliable?
Study 3:
Subject-level
QUALITY CONTROL
APPROVED
Voxel-level
(Most existing literature)
Is this really a lesion?
Study 2:
Predictive Intervals for Volumetry
110 mL
[90-130]

Pixyl Analysis Reports

PIXYL — **Pixyl.Neuro.MS**
Longitudinal report

Patient Information

Name: **Jane Doe** | Sex: **F** | Born in: **1989** | ID: **2622**
Visit Date: **Oct 5, 2023**, Prior Visit Date: **Oct 6, 2022**

Quality Control

	Observations
Pass	-

THIS AUTOMATED REPORT DOES NOT REPLACE MEDICAL EXPERTISE.
PLEASE REFER TO THE RADIOLOGY REPORT.

Disease Activity

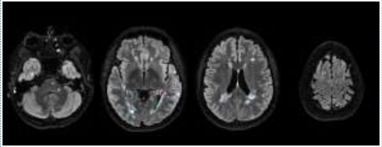
T2 FLAIR lesions			
New	7	Enlarging	2

Lesion Load

	Volume(ml)	Change(ml)	Lesion count *
Periventricular	9.87	0.61	≥ 1
Juxtacortical	2.27	0.47	≥ 1
Infratentorial	0.25	-0.09	≥ 1
Deep WM	0.95	0.1	≥ 1
Whole Brain	**13.34**	**1.09**	**≥ 9 ****

* The lesion count is based on the 2017 revision of the McDonald criteria.
** The Barkhof MRI criteria for MS diagnosis includes at least 9 lesions on T2-weighted images.

FLAIR Segmentation



New | Enlarging | Stable

Report automatically generated on Feb 12, 2024. Not approved for clinical use. Please visit the instructions for use https://pixyl.ai/7ab7f6ec1be9ca3e29108a0f587375d3 for more information.

1/1 pixyl.ai
Pixyl.Neuro.MS v2.1.0 Pixyl SAS 5 av du Grand Sablon, 38700 La Tronche, France - contact@pixyl.ai UDI: (01)03770028144013(8012)v2.1.0

PIXYL — **Pixyl.Neuro.BV**
Longitudinal report

Patient Information

Name: **John Smith** | Sex: **M** | Born in: **1945** | ID: **2620**
Visit Date: **Jan 1, 2020**, Prior Visit Date: **Jan 1, 2015**

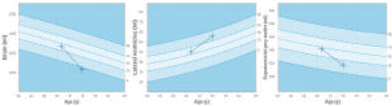
Quality Control

	Observations
Pass	-

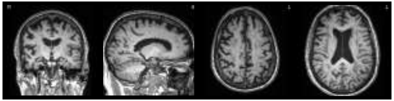
THIS AUTOMATED REPORT DOES NOT REPLACE MEDICAL EXPERTISE.
PLEASE REFER TO THE RADIOLOGY REPORT.

Brain T1 volumetry and comparison with normative population values

	Prior visit	Current visit		
	Volume(ml)	Volume(ml)	Change(%)	Normal range(ml)
Brain	1167.37	**1106.92**	-5.18%	1113.29 - 1202.98
Supratentorial grey matter	541.59	517.07	-4.53%	504.91 - 572.4
Supratentorial white Matter	495.43	461.44	-6.86%	433.03 - 509.91
Cerebellum GM+WM	130.35	128.41	-1.49%	128.34 - 167.72
Left lateral ventricle	19.93	27.89	39.94%	13.02 - 32.26
Right lateral ventricle	19.96	27.87	39.63%	12.39 - 29.96



The normative distribution is calculated over 2700+ normal subjects. Volumes are normalized by the volume of the intracranial cavity when compared with the normative population of the same age. The curves displayed correspond to the 5th, 25th, 50th, 75th and 95th percentile for healthy subjects.



Report automatically generated on Feb 12, 2024. Not approved for clinical use. Please visit the instructions for use https://pixyl.ai/7ab7f6ec1be9ca3e29108a0f587375d3 for more information.

1/2 pixyl.ai
Pixyl.Neuro.BV v2.1.0 Pixyl SAS 5 av du Grand Sablon, 38700 La Tronche, France - contact@pixyl.ai UDI: (01)03770028144013(8012)v2.1.0

Pixyl Analysis Reports

PIXYL — **Pixyl.Neuro.MS** Longitudinal report

Patient Information
Name: **Jane Doe** | Sex: **F** | Born in: **1989** | ID: **2622**
Visit Date: **Oct 5, 2023**, Prior Visit Date: **Oct 6, 2022**

Quality Control

	Observations
Pass	-

THIS AUTOMATED REPORT DOES NOT REPLACE MEDICAL EXPERTISE.
PLEASE REFER TO THE RADIOLOGY REPORT.

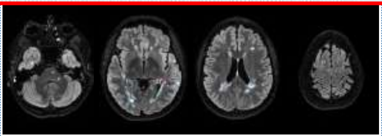
Disease Activity

T2 FLAIR lesions			
New	7	Enlarging	2

Lesion Load

	Volume(ml)	Change(ml)	Lesion count *
Periventricular	9.87	0.61	≥ 1
Juxtacortical	2.27	0.47	≥ 1
Infratentorial	0.25	-0.09	≥ 1
Deep WM	0.95	0.1	≥ 1
Whole Brain	**13.34**	**1.09**	**≥ 9 ****

* The lesion count is based on the 2017 revision of the McDonald criteria.
** The Barkhof MRI criteria for MS diagnosis includes at least 9 lesions on T2-weighted images.



New | Enlarging | Stable

Report automatically generated on Feb 12, 2024. Not approved for clinical use. Please visit the instructions for use https://pixyl.ai/7ab7f6ec1be9ca3e29108a0f587375d3 for more information.

1/1 pixyl.ai

Pixyl.Neuro.MS v2.1.0 | Pixyl SAS 5 av du Grand Sablon, 38700 La Tronche, France - contact@pixyl.ai | UDI: (01)03770028144013(8012)v2.1.0

PIXYL — **Pixyl.Neuro.BV** Longitudinal report

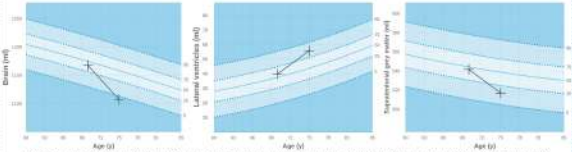
Patient Information
Name: **John Smith** | Sex: **M** | Born in: **1945** | ID: **2620**
Visit Date: **Jan 1, 2020**, Prior Visit Date: **Jan 1, 2015**

Quality Control

Observations

Brain T1 volumetry and comparison with normative population values

	Prior visit	Current visit		
	Volume(ml)	Volume(ml)	Change(%)	Normal range(ml)
Brain	1167.37	**1106.92**	-5.18%	1113.29 - 1202.98
Supratentorial grey matter	541.59	517.07	-4.53%	504.91 - 572.4
Supratentorial white Matter	495.43	461.44	-6.86%	433.03 - 509.91
Cerebellum GM+WM	130.35	128.41	-1.49%	128.34 - 167.72
Left lateral ventricle	19.93	27.89	39.94%	13.02 - 32.26
Right lateral ventricle	19.96	27.87	39.63%	12.39 - 29.96



The normative distribution is calculated over 2700+ normal subjects. Volumes are normalized by the volume of the intracranial cavity when compared with the normative population of the same age. The curves displayed correspond to the 5th, 25th, 50th, 75th and 95th percentile for healthy subjects.

Report automatically generated on Feb 12, 2024. Not approved for clinical use. Please visit the instructions for use https://pixyl.ai/7ab7f6ec1be9ca3e29108a0f587375d3 for more information.

1/2 pixyl.ai

Pixyl.Neuro.BV v2.1.0 | Pixyl SAS 5 av du Grand Sablon, 38700 La Tronche, France - contact@pixyl.ai | UDI: (01)03770028144013(8012)v2.1.0

Predictive Intervals in segmentation-based volumetry

Definition

- $X \in \mathbb{R}^{N-1}$ are estimates of the true volumes $Y \in \mathbb{R}^{N-1}$, obtained from the segmentation.
- A predictive interval $\Gamma_\alpha(X)$ is a range of values intended to encompass Y with a specified degree of confidence $1-\alpha$ (e.g. 90%, 95%), so that $P(Y \in \Gamma_\alpha(X)) \geq 1-\alpha$

Predictive Intervals in segmentation-based volumetry

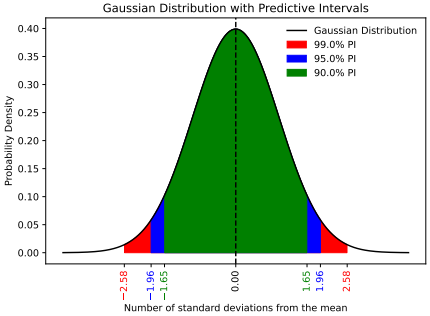
Definition

- $X \in \mathbb{R}^{N-1}$ are estimates of the true volumes $Y \in \mathbb{R}^{N-1}$, obtained from the segmentation.
- A predictive interval $\Gamma_\alpha(X)$ is a range of values intended to encompass Y with a specified degree of confidence $1-\alpha$ (e.g. 90%, 95%), so that $P(Y \in \Gamma_\alpha(X)) \geq 1-\alpha$

Sampling-based approaches

- Sample a set of estimated volumes $X_1, ..., X_K$ for the given image.
- Estimate the mean $\mu(X)$ and standard deviation $\sigma(X)$.
- Assuming $Y|X \sim \mathcal{N}(\mu(X), \sigma(X))$:

$$\Gamma_\alpha(X) = [\mu(X) - z\sigma(X), \mu(X) + z\sigma(X)] \quad (1)$$



Predictive Intervals in segmentation-based volumetry

Definition

- $X \in \mathbb{R}^{N-1}$ are estimates of the true volumes $Y \in \mathbb{R}^{N-1}$, obtained from the segmentation.
- A predictive interval $\Gamma_\alpha(X)$ is a range of values intended to encompass Y with a specified degree of confidence $1-\alpha$ (e.g. 90%, 95%), so that $P(Y \in \Gamma_\alpha(X)) \geq 1-\alpha$

Sampling-based approaches

- Sample a set of estimated volumes $X_1, ..., X_K$ for the given image.
- Estimate the mean $\mu(X)$ and standard deviation $\sigma(X)$.
- Assuming $Y|X \sim \mathcal{N}(\mu(X), \sigma(X))$:

$$\Gamma_\alpha(X) = [\mu(X) - z\sigma(X), \mu(X) + z\sigma(X)] \quad (1)$$

Limitations

- Inference time, due to the sampling procedure.
- The normality assumption, which may not always hold.
- Lack of flexibility, as intervals are symmetrical by design.

Predictive Intervals in segmentation-based volumetry

Definition

- $X \in \mathbb{R}^{N-1}$ are estimates of the true volumes $Y \in \mathbb{R}^{N-1}$, obtained from the segmentation.
- A predictive interval $\Gamma_\alpha(X)$ is a range of values intended to encompass Y with a specified degree of confidence $1-\alpha$ (e.g. 90%, 95%), so that $P(Y \in \Gamma_\alpha(X)) \geq 1-\alpha$

Sampling-based approaches

- Sample a set of estimated volumes $X_1, ..., X_K$ for the given image.
- Estimate the mean $\mu(X)$ and standard deviation $\sigma(X)$.
- Assuming $Y|X \sim \mathcal{N}(\mu(X), \sigma(X))$:

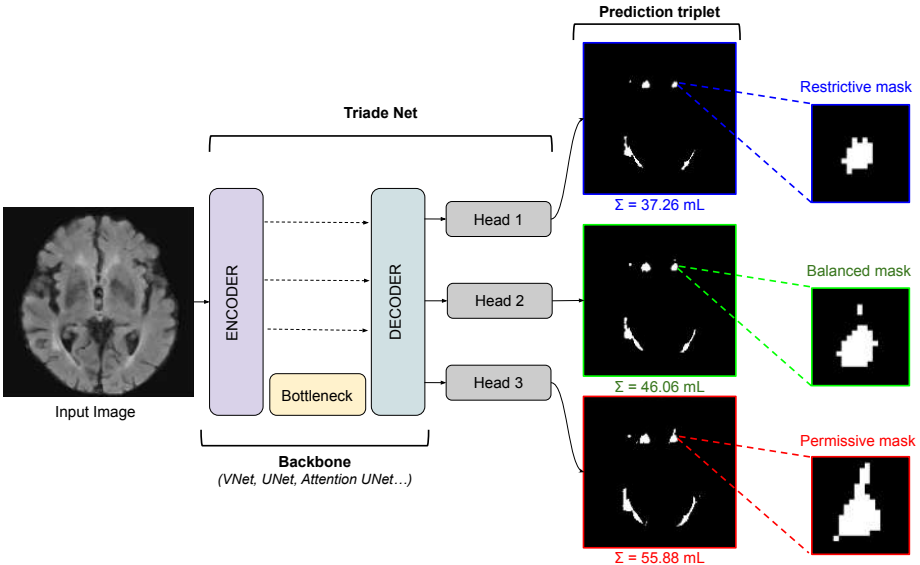
$$\Gamma_\alpha(X) = [\mu(X) - z\sigma(X), \mu(X) + z\sigma(X)] \quad (1)$$

Direct approaches

- Directly estimate the quantiles $\hat{t}_{\alpha/2}(X)$ and $\hat{t}_{1-\alpha/2}(X)$.
- The PI is computed as:

$$\Gamma_\alpha(X) = [\hat{t}_{\alpha/2}(X), \hat{t}_{1-\alpha/2}(X)] \quad (2)$$

Fast Predictive Intervals using a multi-head segmentation model: TriadNet



Training TriadNet through Tversky loss parametrization

The Soft Dice loss (Milletari et al. 2016)

$$\mathcal{L} \approx 1 - \frac{TP}{TP + 0.5FP + 0.5FN} \tag{3}$$

TP: True Positive, FP: False Positive, FN: False Negative

Training TriadNet through Tversky loss parametrization

The Soft Dice loss (Milletari et al. 2016)

$$\mathcal{L} \approx 1 - \frac{TP}{TP + 0.5FP + 0.5FN} \tag{3}$$

TP: True Positive, FP: False Positive, FN: False Negative

The Tversky loss (Salehi et al. 2017)

$$\mathcal{T}_{\gamma,\beta} \approx 1 - \frac{TP}{TP + \gamma FP + \beta FN} \tag{4}$$

Training TriadNet through Tversky loss parametrization

The Soft Dice loss (Milletari et al. 2016)

$$\mathcal{L} \approx 1 - \frac{TP}{TP + 0.5FP + 0.5FN} \tag{3}$$

TP: True Positive, FP: False Positive, FN: False Negative

The Tversky loss (Salehi et al. 2017)

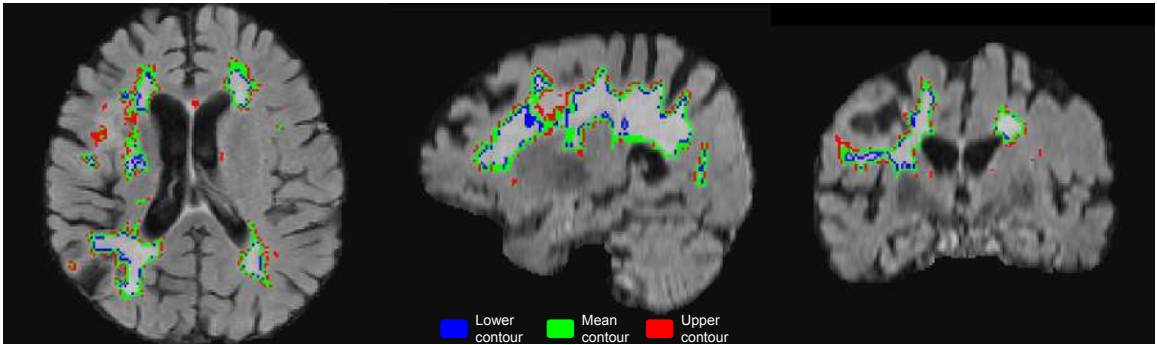
$$\mathcal{T}_{\gamma,\beta} \approx 1 - \frac{TP}{TP + \gamma FP + \beta FN} \tag{4}$$

Definition

Writing p_{lower}, p_{mean} and p_{upper} the 3 output heads of the TriadNet:

$$\text{TriadLoss} = \mathcal{T}_{0.8,0.2}(p_{lower,y}) + \mathcal{T}_{0.5,0.5}(p_{mean,y}) + \mathcal{T}_{0.2,0.8}(p_{upper,y}) \tag{5}$$

Visualization of the bounds predicted by TriadNet



Conformal calibration of Predictive Intervals

Algorithm Pseudocode for Predictive Intervals calibration (Angelopoulos et al. 2023)

Input: Trained TriadNet model

Input: Calibration dataset with N pairs of image and ground truth segmentation

Input: User-defined coverage level $(1-\alpha)\%$

Output: Corrective factor $\hat{q}$ to calibrated PIs at the $(1-\alpha)\%$ level

1: **for** $j = 1$ to N **do**
2: Estimate the volume X_j, lower bound volume $L(X_j)$, upper bound volume $U(X_j)$ using TriadNet.
3: Compute score function: $s_j(X, Y) = \max\{L(X_j) - Y, Y - U(X_j)\}$
4: **end for**
5: Compute corrective factor $\hat{q} = \text{Quantile}(s_1, s_2, ..., s_N; \frac{\lceil (N+1)(1-\alpha) \rceil}{N})$

Conformal calibration of Predictive Intervals

Algorithm Pseudocode for Predictive Intervals calibration (Angelopoulos et al. 2023)

Input: Trained TriadNet model

Input: Calibration dataset with N pairs of image and ground truth segmentation

Input: User-defined coverage level $(1-\alpha)\%$

Output: Corrective factor $\hat{q}$ to calibrated PIs at the $(1-\alpha)\%$ level

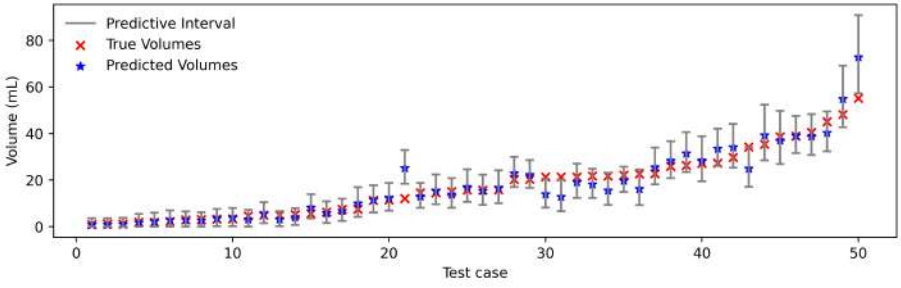
1: **for** $j = 1$ to N **do**
2: Estimate the volume X_j, lower bound volume $L(X_j)$, upper bound volume $U(X_j)$ using TriadNet.
3: Compute score function: $s_j(X, Y) = \max\{L(X_j) - Y, Y - U(X_j)\}$
4: **end for**
5: Compute corrective factor $\hat{q} = \text{Quantile}(s_1, s_2, ..., s_N; \frac{\lceil (N+1)(1-\alpha) \rceil}{N})$

Calibrated PIs at test time

$$\Gamma_\alpha(X_{test}) = [L(X_{test}) - \hat{q}, U(X_{test}) + \hat{q}]$$

Application 1: lesion load estimation in Multiple Sclerosis patients

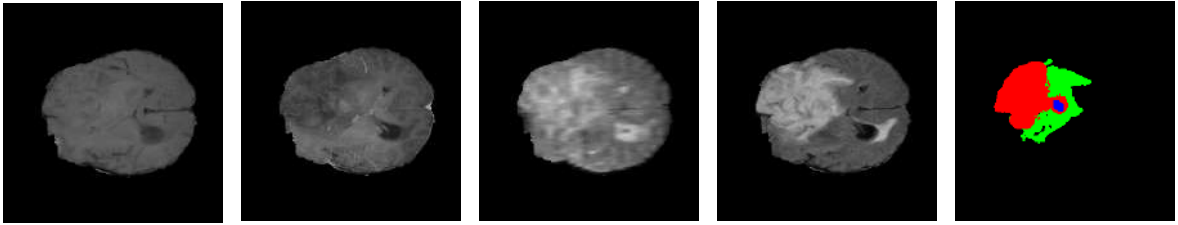
- 120 subjects for training, 40 for calibration and 50 for in-distribution testing. (Multicentric - 3 Tesla: MSSEG 2016 / WMH 2017 / ISBI 2015)
- Intervals calibrated for a target coverage of 90%.
- Metrics (bootstrapping, $M = 15000$):
  - Mean Average Error: 3.08 ± 0.46 mL
  - Empirical Coverage: $92.06 \pm 5.34\%$



Application 2: tumor volume estimation in glioblastoma patients

Pathology Description

- Prevalent form of brain tumor, associated with poor prognosis (Grech et al. 2020).
- Estimation of the tumor volume is crucial for treatment planning.
- Quantification of the tumor can be performed through a 3-classes segmentation: necrosis, edema, gadolinium-enhancing tumor.



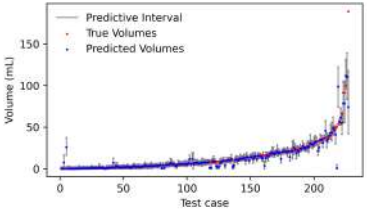
T1 T1 injected T2 FLAIR Ground truth

Application 2: tumor volume estimation in glioblastoma patients

- 679 subjects for training, 227 for calibration, and 227 for testing (BraTS 2023 dataset)
- Intervals calibrated for a target coverage of 90%.

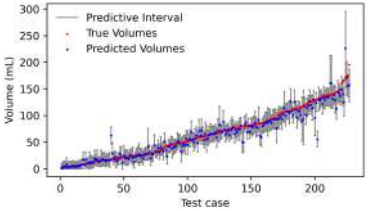
Necrosis:

- MAE: 3.10 ± 0.46mL
- Coverage: $90.78 \pm 2.71\%$



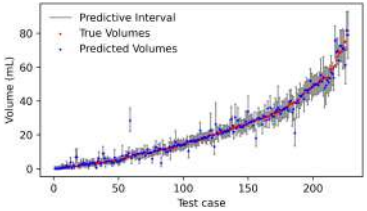
Edema:

- MAE: 8.22 ± 0.57mL
- Coverage: $90.76 \pm 2.70\%$



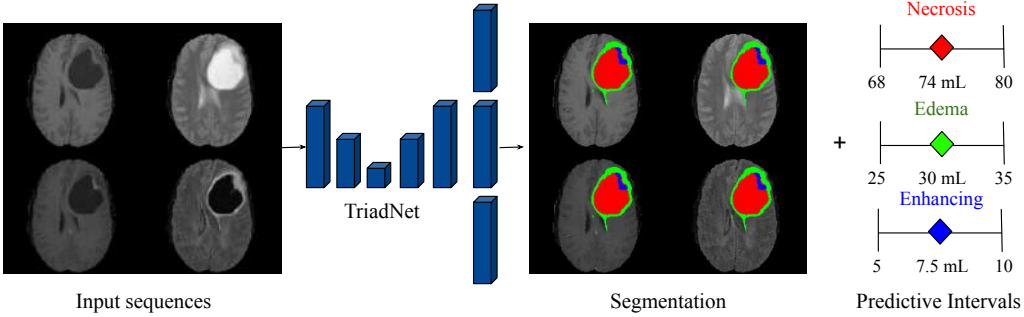
Enhancing tumor:

- MAE: 1.73 ± 0.19mL
- Coverage: $90.79 \pm 2.71\%$



Study summary - Predictive Intervals

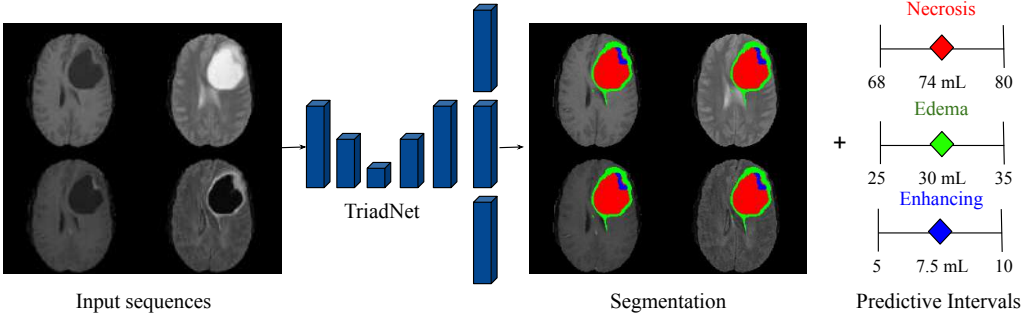
- We propose a direct approach for PI estimation based on a multi-head segmentation model, **TriadNet**[†].



[†] B. Lambert et al. (2023b). "TriadNet: Sampling-Free Predictive Intervals for Lesional Volume in 3D Brain MR Images". In: *UNSURE 2023, Held in Conjunction with MICCAI 2023*. LNCS 14291, pp. 32–41

Study summary - Predictive Intervals

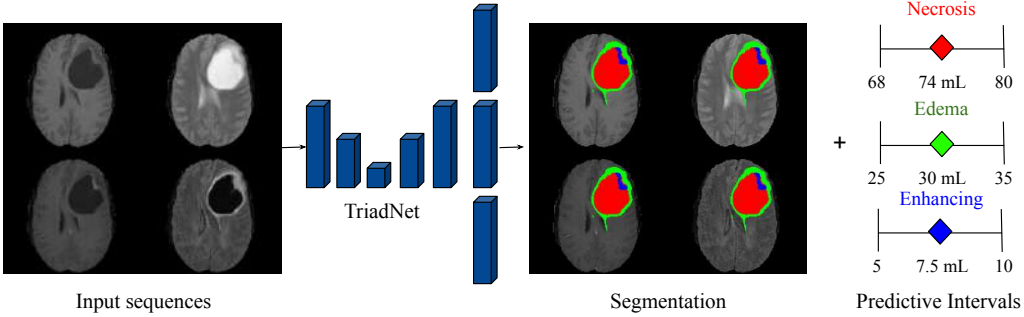
- We propose a direct approach for PI estimation based on a multi-head segmentation model, **TriadNet**[†].
- We leverage the **Conformal framework** to calibrate intervals.



[†] B. Lambert et al. (2023b). "TriadNet: Sampling-Free Predictive Intervals for Lesional Volume in 3D Brain MR Images". In: *UNSURE 2023, Held in Conjunction with MICCAI 2023*. LNCS 14291, pp. 32–41

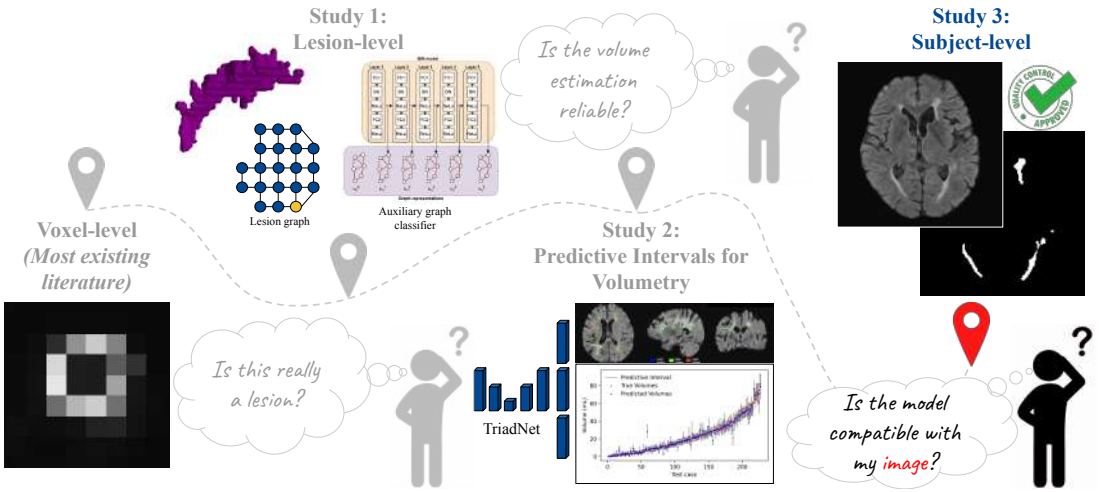
Study summary - Predictive Intervals

- We propose a direct approach for PI estimation based on a multi-head segmentation model, **TriadNet**[†].
- We leverage the **Conformal framework** to calibrate intervals.
- Intervals are **fast** to compute ($\leq$ 1s), as no sampling is required.



[†] B. Lambert et al. (2023b). "TriadNet: Sampling-Free Predictive Intervals for Lesional Volume in 3D Brain MR Images". In: *UNSURE 2023, Held in Conjunction with MICCAI 2023*. LNCS 14291, pp. 32–41

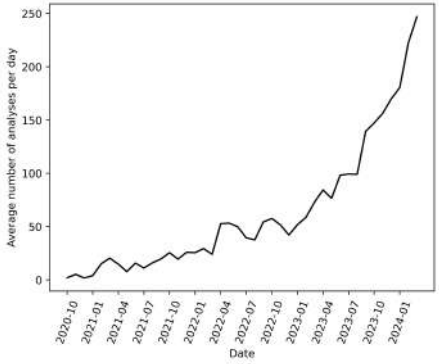
Outline


Study 1:
Lesion-level
Is the volume estimation reliable?
Lesion graph
Auxiliary graph classifier
Study 3:
Subject-level
QUALITY CONTROL
APPROVED
Voxel-level
(Most existing literature)
Study 2:
Predictive Intervals for Volumetry
Is this really a lesion?
TriadNet
Is the model compatible with my image?

Objective - Automatic Quality Control

Paradigm

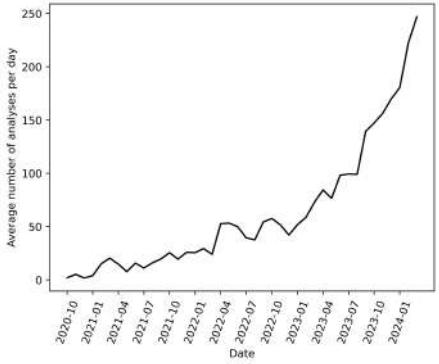
- The number of automatic analyses at Pixyl is steadily increasing.
- A fraction of the input images does not meet the defined quality criteria.

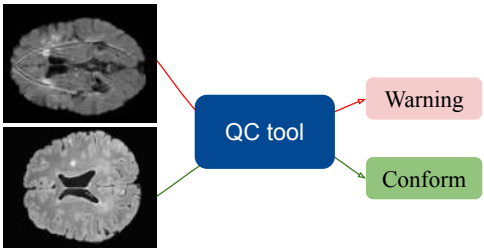


Objective - Automatic Quality Control

Paradigm

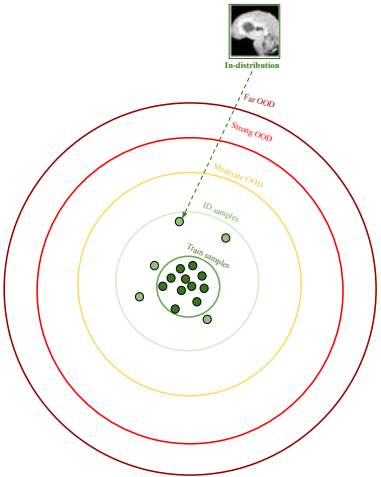
- The number of automatic analyses at Pixyl is steadily increasing.
- A fraction of the input images does not meet the defined quality criteria.
- An automatic Quality Control (QC) tool is desired to flag non-conform input images.





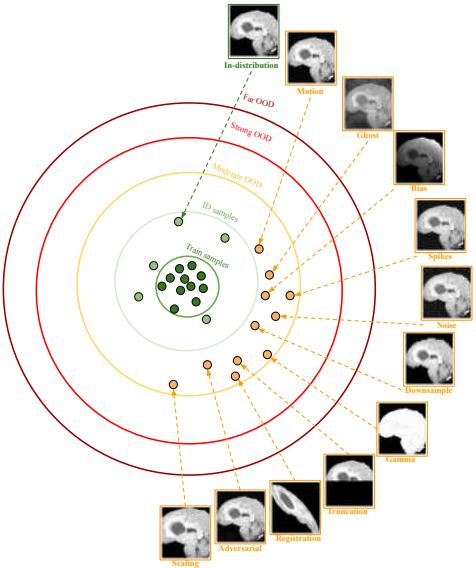
Target image-level module.

What makes a medical image out-of-distribution?



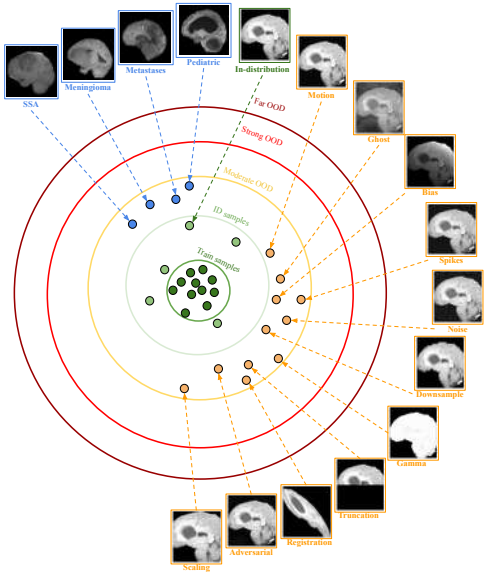
- In-distribution $\leftrightarrow$ training distribution (T1 MRI of Adult glioblastoma patients)

What makes a medical image out-of-distribution?



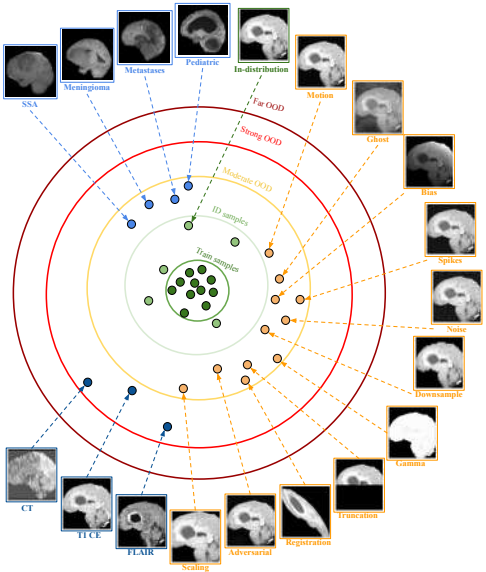
- In-distribution $\leftrightarrow$ training distribution (T1 MRI of Adult glioblastoma patients)
- Are **out-of-distribution**:
  - Images corrupted with artifacts.

What makes a medical image out-of-distribution?



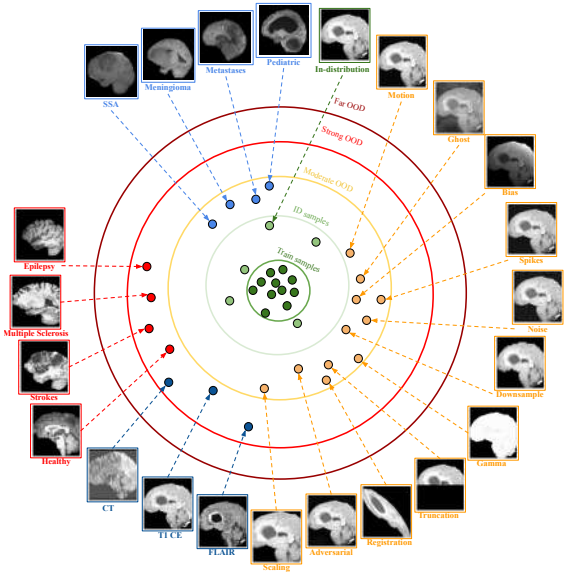
- In-distribution $\leftrightarrow$ training distribution (T1 MRI of Adult glioblastoma patients)
- Are **out-of-distribution**:
  - Images corrupted with artifacts.
  - Shifts in the imaged population.

What makes a medical image out-of-distribution?



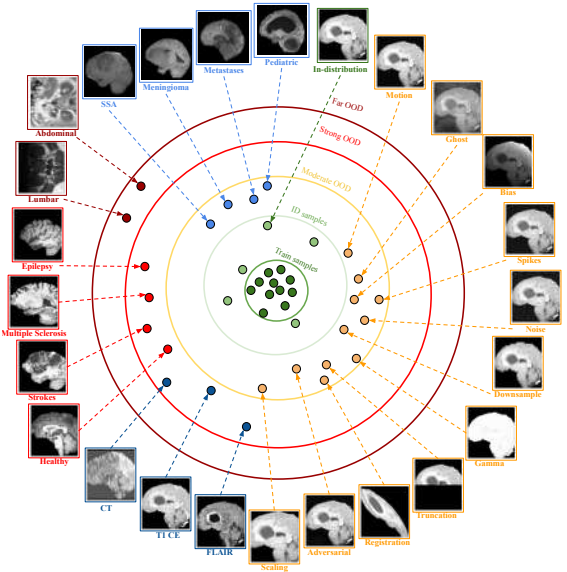
- In-distribution $\leftrightarrow$ training distribution (T1 MRI of Adult glioblastoma patients)
- Are **out-of-distribution**:
  - Images corrupted with artifacts.
  - Shifts in the imaged population.
  - Shifts in image modality.

What makes a medical image out-of-distribution?



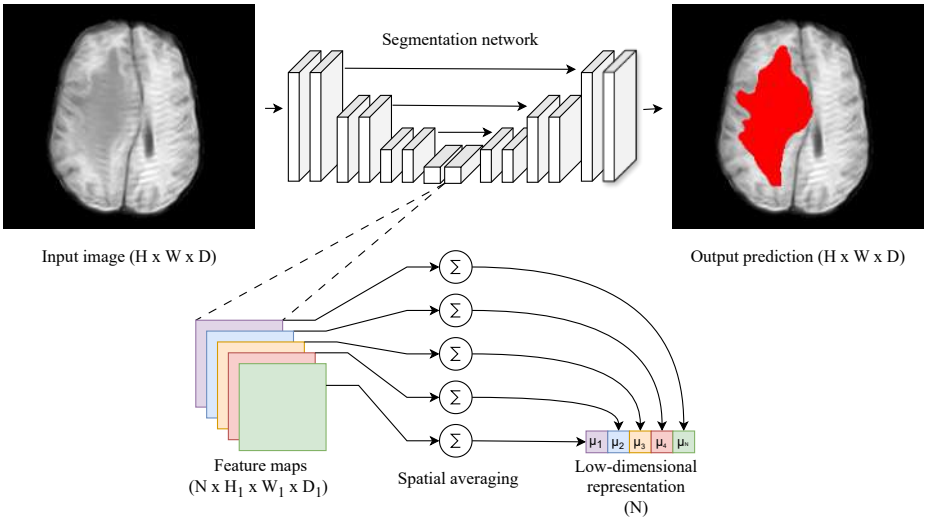
- In-distribution $\leftrightarrow$ training distribution (T1 MRI of Adult glioblastoma patients)
- Are **out-of-distribution**:
  - Images corrupted with artifacts.
  - Shifts in the imaged population.
  - Shifts in image modality.
  - Diseases not present in the training set.

What makes a medical image out-of-distribution?



- In-distribution $\leftrightarrow$ training distribution (T1 MRI of Adult glioblastoma patients)
- Are **out-of-distribution**:
  - Images corrupted with artifacts.
  - Shifts in the imaged population.
  - Shifts in image modality.
  - Diseases not present in the training set.
  - Incorrect organs.

Proposed solution: latent-space detection



Outlier detection using the Mahalanobis distance in latent space

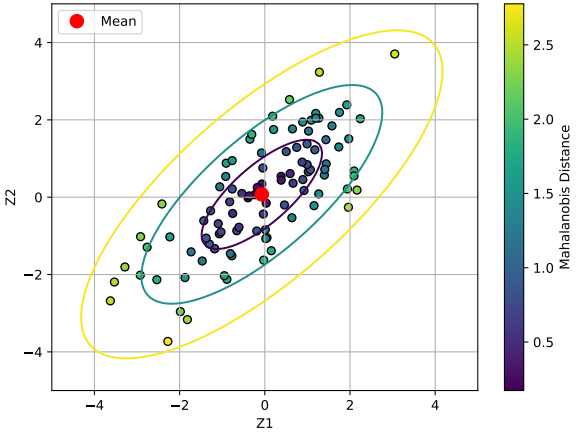
Mathematical formulation

Fit a multivariate Gaussian distribution from a training dataset of **in-distribution** latent representations $\{x_i\}_{i=1}^{N}$:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T$$

$$\mathrm{MD}(x_{test}; \mu, \Sigma) = (x_{test} - \mu)^T \Sigma^{-1} (x_{test} - \mu)$$



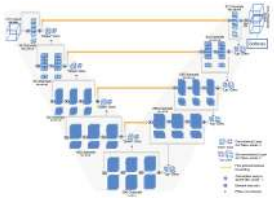
Visualization of the Mahalanobis distance in a 2-dimensional setting.

Challenges: segmentation architecture and layer selection

U-Net (Ronneberger et al. 2015)

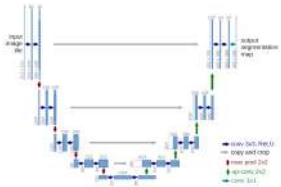
Challenges: segmentation architecture and layer selection

U-Net (Ronneberger et al. 2015)

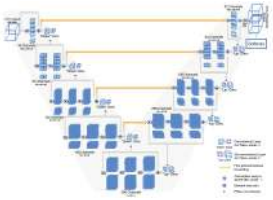


V-Net (Milletari et al. 2016)

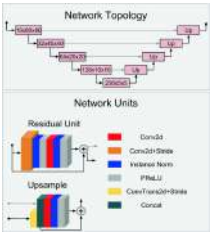
Challenges: segmentation architecture and layer selection



U-Net (Ronneberger et al. 2015)



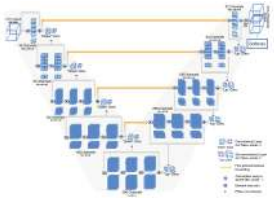
V-Net (Milletari et al. 2016)



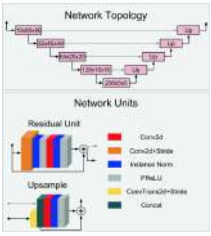
Residual U-Net (Kerfoot et al. 2018)

Challenges: segmentation architecture and layer selection

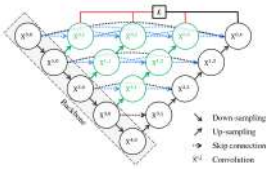
U-Net (Ronneberger et al. 2015)



V-Net (Milletari et al. 2016)



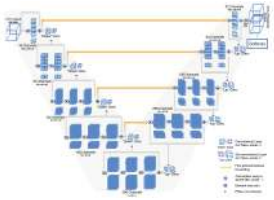
Residual U-Net (Kerfoot et al. 2018)



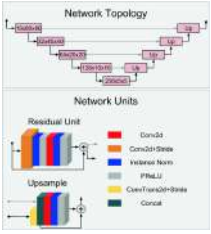
U-Net ++ (Zhou et al. 2018)

Challenges: segmentation architecture and layer selection

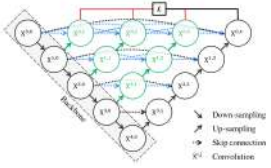
U-Net (Ronneberger et al. 2015)



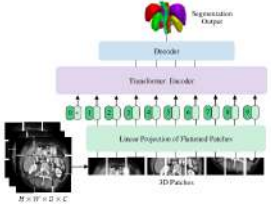
V-Net (Milletari et al. 2016)



Residual U-Net (Kerfoot et al. 2018)



U-Net ++ (Zhou et al. 2018)



U-Net Transformer (Hatamizadeh et al. 2022)

Challenges: segmentation architecture and layer selection

U-Net (Ronneberger et al.

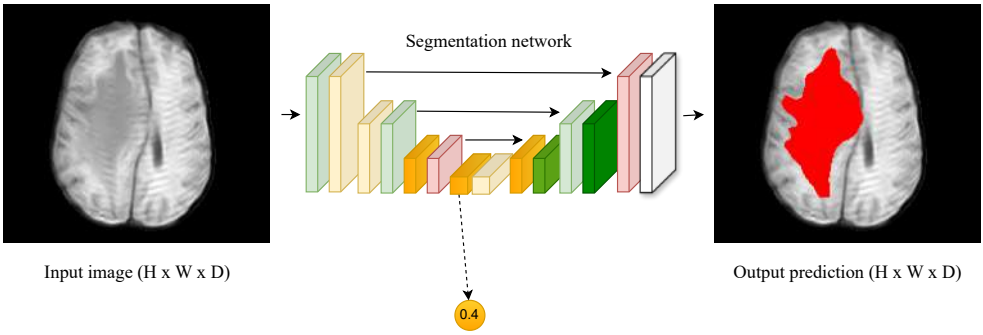
V-Net (Milletari et al. 2016)

Residual U-Net (Kerfoot et al.

- Architectures are diverse and may impact latent representations.
- The choice of a layer to extract latent representations seems crucial for outlier detection, but previous work focuses on single-layer approaches (Karimi et al. 2022, González et al. 2022, Diao et al. 2022)
- We adopt a multi-layer approach to circumvent these challenges.

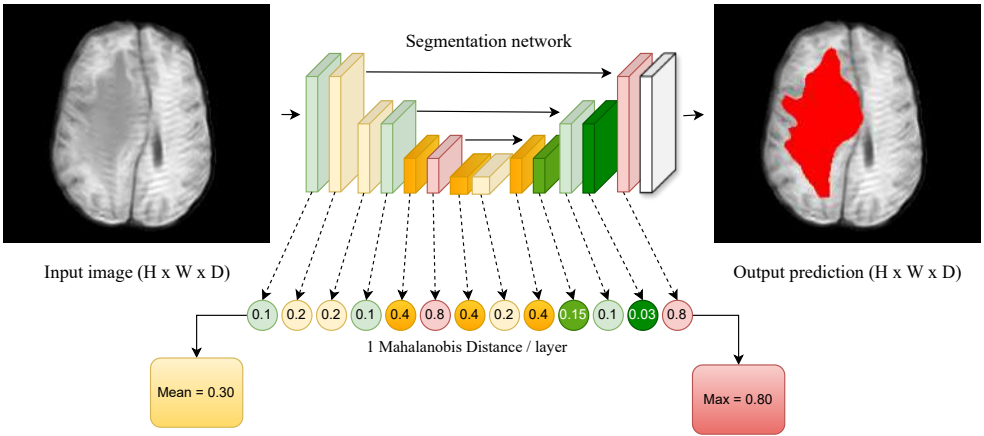
(Hatamizadeh et al. 2022)

Multi-layer aggregation of Mahalanobis distances



Standard single-layer approach.

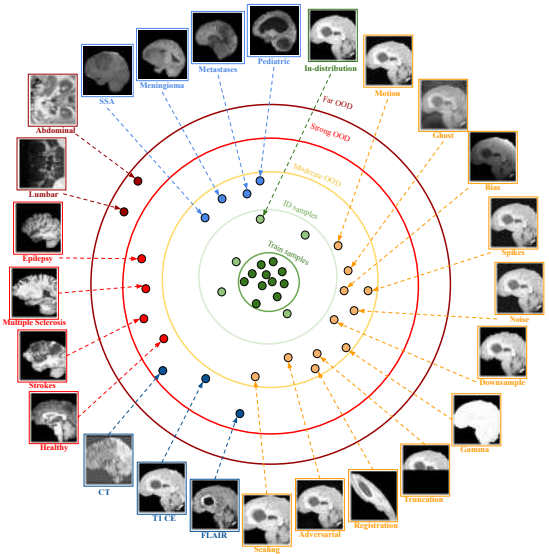
Multi-layer aggregation of Mahalanobis distances



Proposed multi-layer approach.[a]

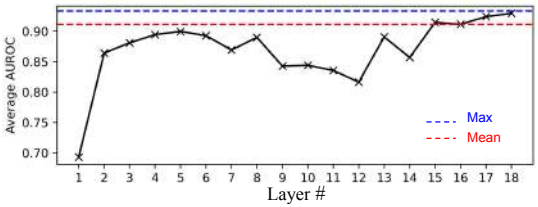
[a] B. Lambert et al. (2023a). "Multi-layer Aggregation as a key to feature-based OOD detection". In: *UNSURE 2023, Held in Conjunction with MICCAI 2023*. LNCS 14291, pp. 104–114

Experiments - Whole tumor segmentation in brain T1-weighted MRI

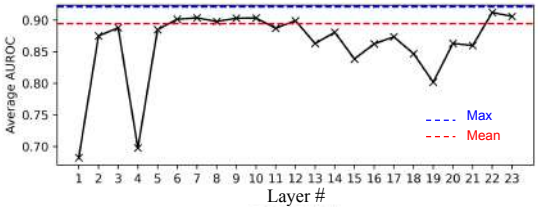


- Whole tumor segmentation in T1-weighted brain MRI as pretext task.
- 876 subjects for training, 30 for validation, 227 for in-distribution testing (BraTS 2023 dataset, Menze et al. 2014).
- 4 different segmentation models: Dynamic U-Net, Residual U-Net, V-Net, Attention U-Net.

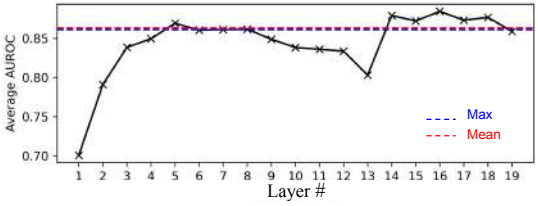
Results - Sensitivity to architecture and layer selection



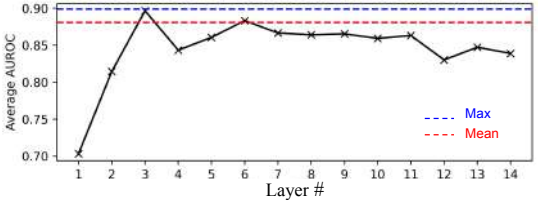
Dynamic U-Net



Residual U-Net



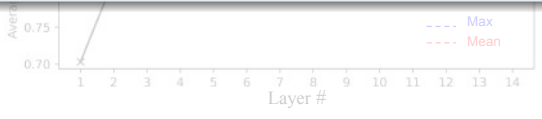
V-Net



Attention U-Net

Results - Sensitivity to architecture and layer selection

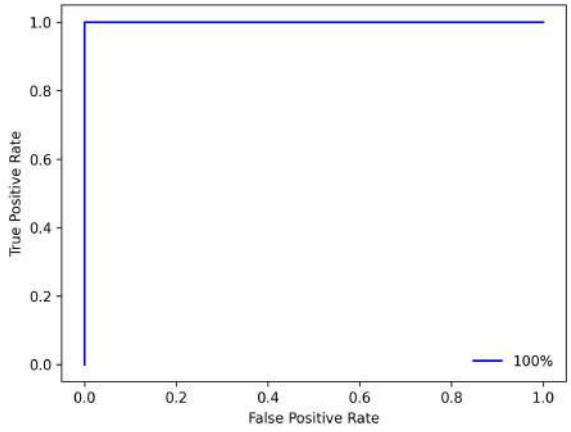
- The optimal layer for OOD detection depends on the segmentation architecture.
- The multi-layer scores (Mean and Max) provides high detection accuracy for each architecture.
- Overall it alleviates the cumbersome optimal layer selection.



V-Net

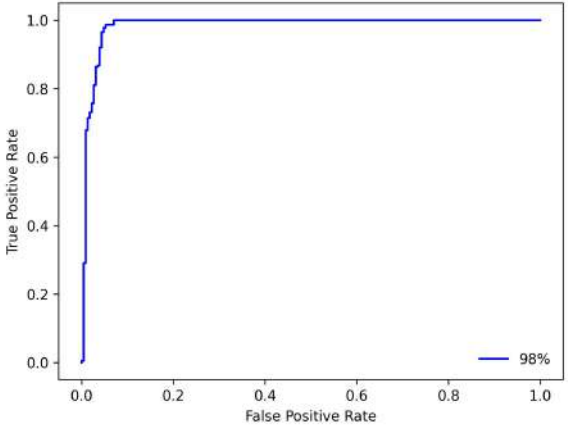
Attention U-Net

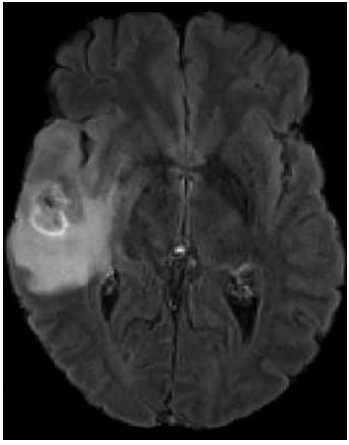
Success and failure cases - Dynamic U-Net



Incorrect organ (Lumbar, N=250)

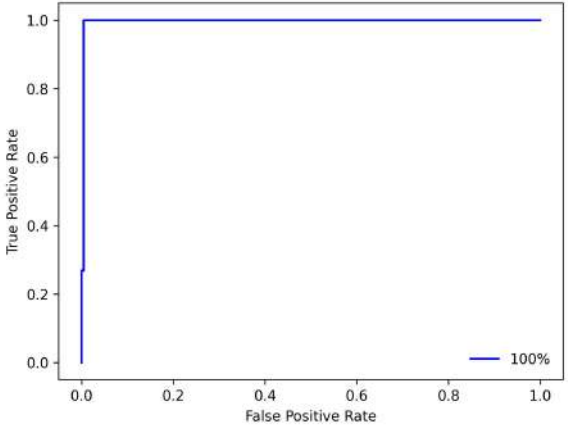
Success and failure cases - Dynamic U-Net

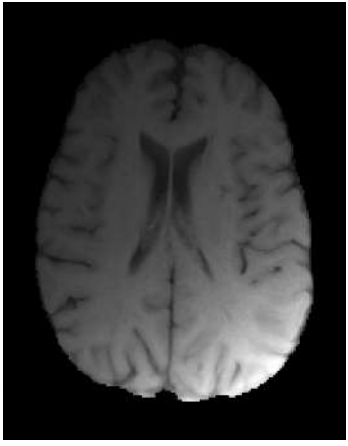




Incorrect modality (FLAIR MRI, N=227)

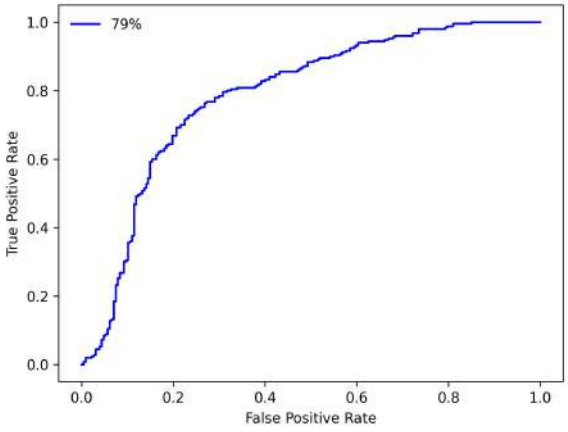
Success and failure cases - Dynamic U-Net





Strong bias artifact (N=227)

Success and failure cases - Dynamic U-Net



Different tumor subtype (Meningioma, N=250)

From input QC to output QC

Limits of input QC

- Latent-space distances are efficient in detecting images **far from the training distribution**.

From input QC to output QC

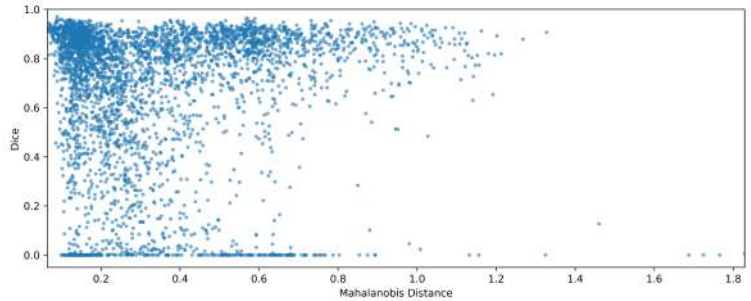
Limits of input QC

- Latent-space distances are efficient in detecting images **far from the training distribution**.
- What about detecting poor-quality predictions?

From input QC to output QC

Limits of input QC

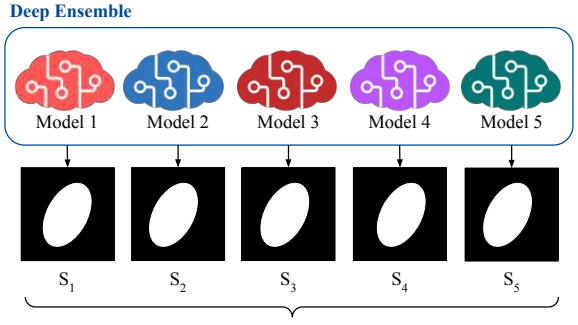
- Latent-space distances are efficient in detecting images **far from the training distribution**.
- What about detecting poor-quality predictions?
  - Pearson's correlation between segmentation quality (Dice) and the Mahalanobis distance: $\rho = 0.065$ (N=3825)

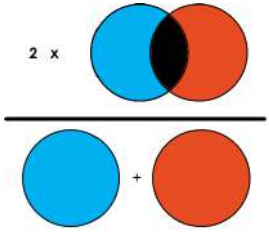


A proxy score for segmentation quality assessment (Hann et al. 2021)

Paradigm

- **Goal**: estimate the true segmentation accuracy.
- **How?** Measure the segmentation variability among an ensemble of models.
- We note S_k the individual segmentations and MV the majority vote segmentation
- Ensemble Prediction Agreement (EPA): $\text{EPA} = \frac{1}{K}\sum_{i=1}^{K} \text{Dice}(S_k, MV)$



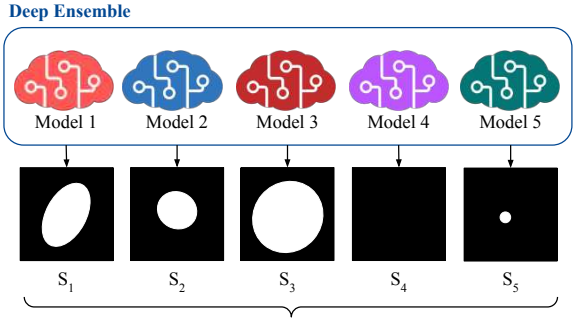


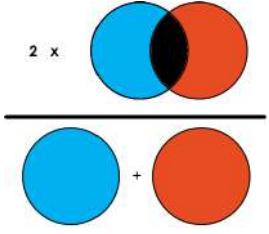
Dice score

A proxy score for segmentation quality assessment (Hann et al. 2021)

Paradigm

- **Goal**: estimate the true segmentation accuracy.
- **How?** Measure the segmentation variability among an ensemble of models.
- We note S_k the individual segmentations and MV the majority vote segmentation
- Ensemble Prediction Agreement (EPA): $\text{EPA} = \frac{1}{K}\sum_{i=1}^{K} \text{Dice}(S_k, MV)$



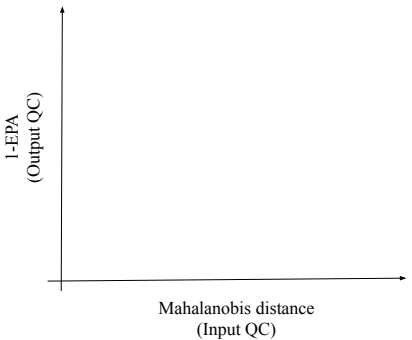


Dice score

Unified Input/Output Quality Control

- The Mahalanobis distance is used as input QC score.
- The Ensemble Prediction Agreement is used as output QC score[a].

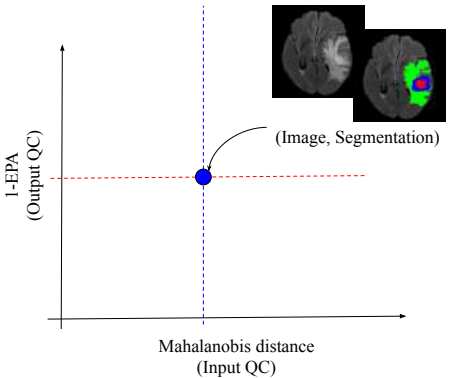
[a] B. Lambert et al. (in prep.). "From Out-of-distribution detection to Quality Control". In: *Trustworthy AI in Medical Imaging, MICCAI book series*



Unified Input/Output Quality Control

- The Mahalanobis distance is used as input QC score.
- The Ensemble Prediction Agreement is used as output QC score[a].

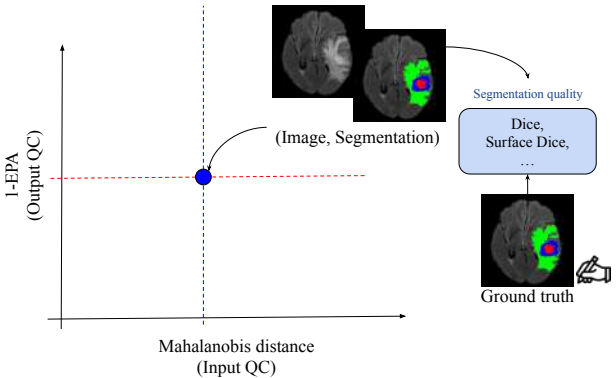
[a] B. Lambert et al. (in prep.). "From Out-of-distribution detection to Quality Control". In: *Trustworthy AI in Medical Imaging, MICCAI book series*



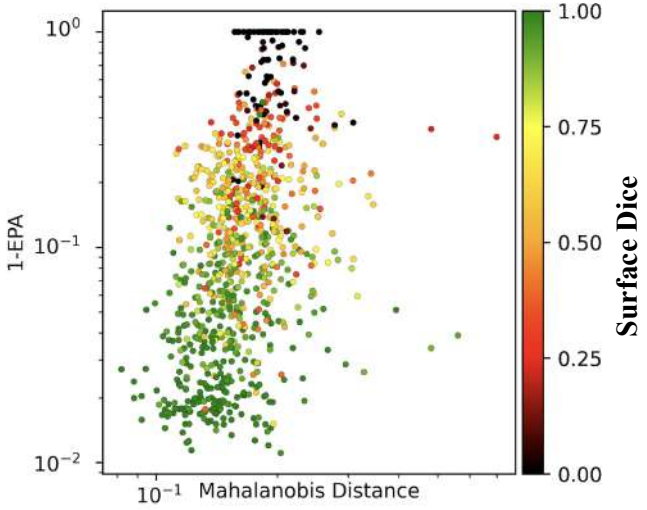
Unified Input/Output Quality Control

- The Mahalanobis distance is used as input QC score.
- The Ensemble Prediction Agreement is used as output QC score[a].

[a] B. Lambert et al. (in prep.). "From Out-of-distribution detection to Quality Control". In: *Trustworthy AI in Medical Imaging, MICCAI book series*



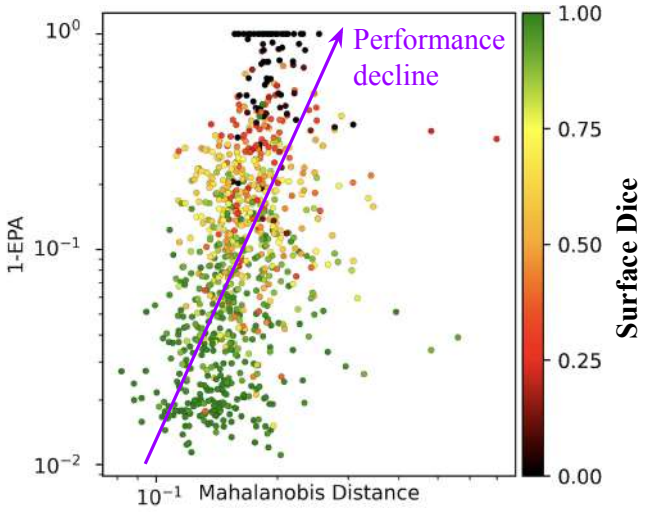
Prediction space stratification for multi-class brain tumor segmentation



Experimental Setting

- 5 Dynamic U-Nets are trained to segment gliomas.
- QC scores are computed for 874 test subjects with **variable difficulty**.

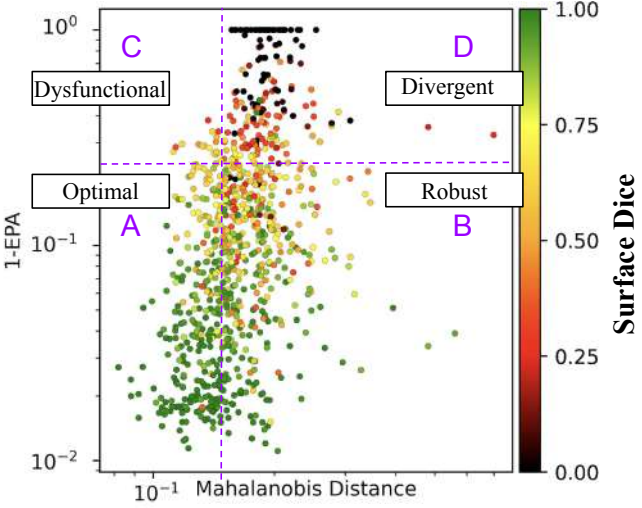
Prediction space stratification for multi-class brain tumor segmentation



Experimental Setting

- 5 Dynamic U-Nets are trained to segment gliomas.
- QC scores are computed for 874 test subjects with **variable difficulty**.

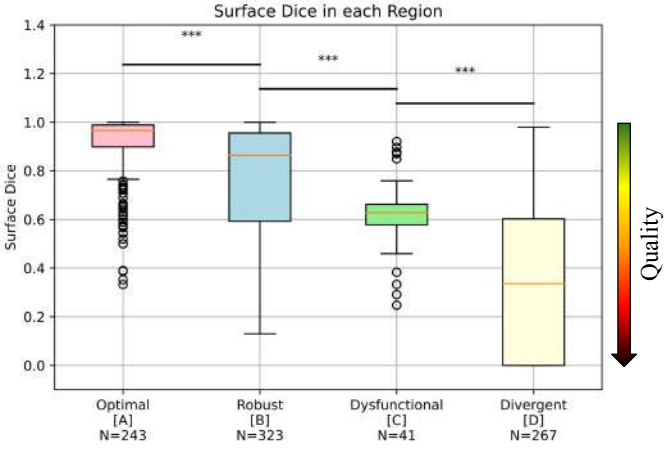
Prediction space stratification for multi-class brain tumor segmentation



Experimental Setting

- 5 Dynamic U-Nets are trained to segment gliomas.
- QC scores are computed for 874 test subjects with **variable difficulty**.
- 4 regimes (A-B-C-D) identified by fitting thresholds on a validation dataset (N=30).

Prediction space stratification for multi-class brain tumor segmentation



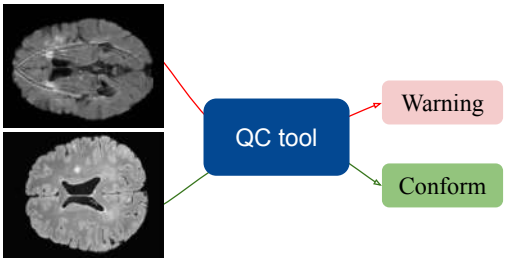
Surface Dice in each predictive region. ***: p-value ≤ 0.001 (Student's t-test).

Experimental Setting

- 5 Dynamic U-Nets are trained to segment gliomas.
- QC scores are computed for 874 test subjects with **variable difficulty**.
- 4 regimes (A-B-C-D) identified by fitting thresholds on a validation dataset (N=30).

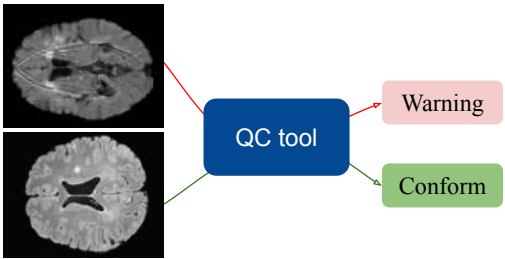
Study summary - Quality Control

- The **Mahalanobis distance** in the latent space is efficient in detecting images far from the training distribution.

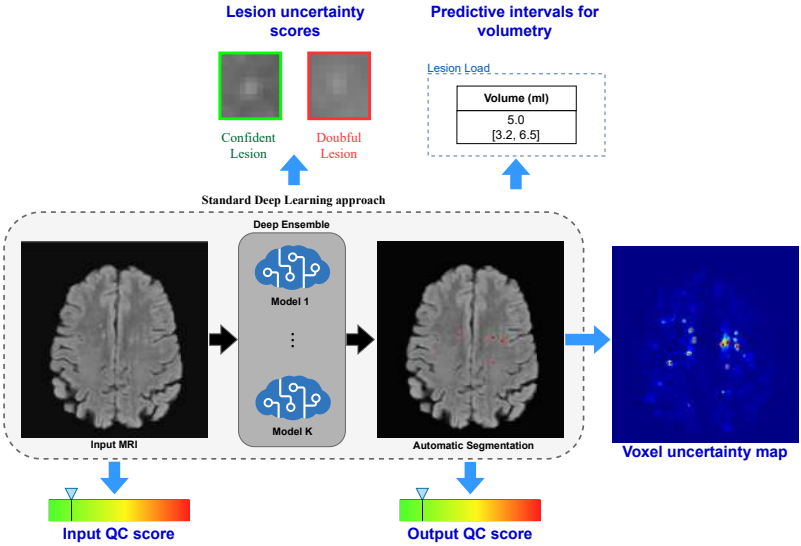


Study summary - Quality Control

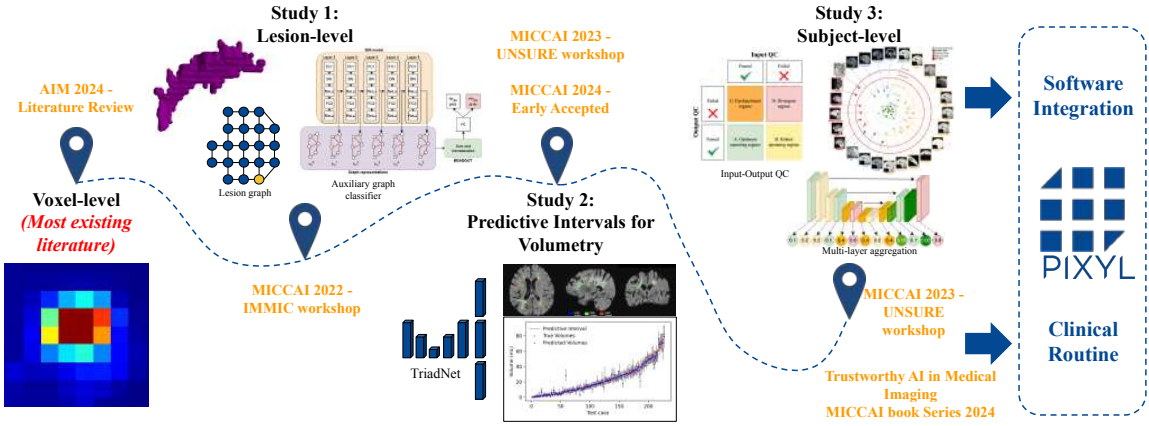
- The **Mahalanobis distance** in the latent space is efficient in detecting images far from the training distribution.
- It can be completed with the **Ensemble Prediction Agreement** score to assess the quality of the segmentation.



Thesis framework summary



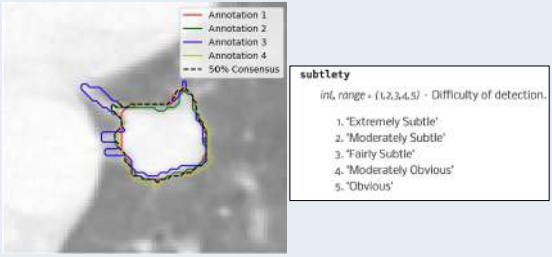
Methodological contributions



Limits and Future Directions

Evaluation of uncertainty

- Generally restricted to detecting errors
- "Ground truth" uncertainty labels are promising but costly and subjective.

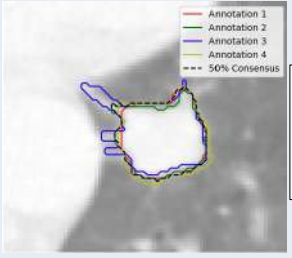


Annotations from the LIDC-IDRI dataset for lung cancer (Armato III et al. 2011)

Limits and Future Directions

Evaluation of uncertainty

- Generally restricted to detecting errors
- "Ground truth" uncertainty labels are promising but costly and subjective.



subtlety

int, range = [1,2,3,4,5] - Difficulty of detection.

1. 'Extremely Subtle'
2. 'Moderately Subtle'
3. 'Fairly Subtle'
4. 'Moderately Obvious'
5. 'Obvious'

Annotations from the LIDC-IDRI dataset for lung cancer (Armato III et al. 2011)

Added value in clinical routine

Measuring the benefit of uncertainty in AI-assisted clinical routine:

- Increased trust and acceptability?
- Faster reviewing time?
- Better decision-making?

PIXYL

AI-Powered Patient Care

GIN Grenoble Institut Neurosciences

Inria

Thank you !

Michel Dojat

Florence Forbes

Senan Doyle

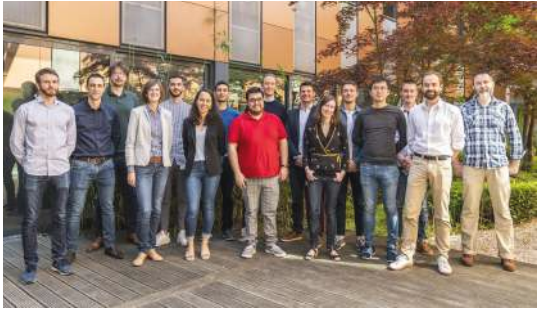
Alan Tucholka

Julien Perrin

Team Pixyl

Benjamin Lemasson

Team NIPC

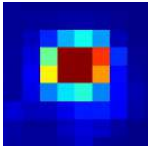


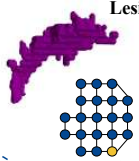
Methodological contributions

Study 1: Lesion-level

AIM 2024 - Literature Review

Voxel-level *(Most existing literature)*



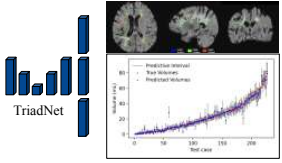


MICCAI 2022 - IMMIC workshop

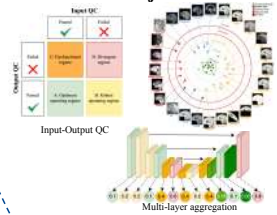
MICCAI 2023 - UNSURE workshop

MICCAI 2024 - Early Accepted

Study 2: Predictive Intervals for Volumetry



Study 3: Subject-level



MICCAI 2023 - UNSURE workshop

Trustworthy AI in Medical Imaging MICCAI book Series 2024

Software Integration

PIXYL

Clinical Routine

References I

Angelopoulos, A. N., S. Bates, A. Fisch, L. Lei, and T. Schuster (2022). "Conformal risk control". In: *arXiv preprint arXiv:2208.02814*.

Angelopoulos, A. N. and S. Bates (2023). "Conformal Prediction: A Gentle Introduction". In: *Foundations and Trends in Machine Learning* 16.4, pp. 494–591.

Armato III, S. G., G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. (2011). "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans". In: *Medical physics* 38.2, pp. 915–931.

Barber, R. F., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2023). "Conformal prediction beyond exchangeability". In: *The Annals of Statistics* 51.2, pp. 816–845.

Commowick, O., M. Kain, R. Casey, R. Ameli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, S. Camarasu-Pop, T. Glatard, et al. (2021). "Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset". In: *Neuroimage* 244, p. 118589.

Diao, Z., H. Jiang, and T. Shi (2022). "A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity". In: *Knowledge-Based Systems* 246, p. 108739.

González, C., K. Gotkowski, M. Fuchs, A. Bucher, A. Dadras, R. Fischbach, I. J. Kaltenborn, and A. Mukhopadhyay (2022). "Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation". In: *Medical Image Analysis* 82, p. 102596.

Grech, N., T. Dalli, S. Mizzi, L. Meilak, N. Calleja, and A. Zrinzo (2020). "Rising incidence of glioblastoma multiforme in a well-defined population". In: *Cureus* 12.5.

References II

Hann, E., R. A. Gonzales, I. A. Popescu, Q. Zhang, V. M. Ferreira, and S. K. Piechnik (2021). "Ensemble of deep convolutional neural networks with monte carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets". In: *Annual Conference on Medical Image Understanding and Analysis*, pp. 280–293.

Hatamizadeh, A., Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu (2022). "Unetr: Transformers for 3d medical image segmentation". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584.

Kamraoui, R. A., B. Mansencal, J. V. Manjon, and P. Coupé (2022). "Longitudinal detection of new MS lesions using deep learning". In: *Frontiers in Neuroimaging* 1, p. 948235.

Karimi, D. and A. Gholipour (2022). "Improving calibration and out-of-distribution detection in deep models for medical image segmentation". In: *IEEE Transactions on Artificial Intelligence*.

Kerfoot, E., J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel (2018). "Left-ventricle quantification using residual U-Net". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 371–380.

Lambert, B., F. Forbes, S. Doyle, H. Dehaene, and M. Dojat (2024). "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis". In: *Artificial Intelligence in Medicine*, p. 102830.

Lambert, B., F. Forbes, S. Doyle, and M. Dojat (2023a). "Multi-layer Aggregation as a key to feature-based OOD detection". In: *UNSURE 2023, Held in Conjunction with MICCAI 2023*. LNCS 14291, pp. 104–114.

— (2023b). "TriadNet: Sampling-Free Predictive Intervals for Lesional Volume in 3D Brain MR Images". In: *UNSURE 2023, Held in Conjunction with MICCAI 2023*. LNCS 14291, pp. 32–41.

References III

Lambert, B., F. Forbes, S. Doyle, and M. Dojat (in prep.). "From Out-of-distribution detection to Quality Control". In: *Trustworthy AI in Medical Imaging, MICCAI book series*.

Menze, B. H., A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. (2014). "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10, pp. 1993–2024.

Milletari, F., N. Navab, and S.-A. Ahmadi (2016). "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision*, pp. 565–571.

Ronneberger, O., P. Fischer, and T. Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241.

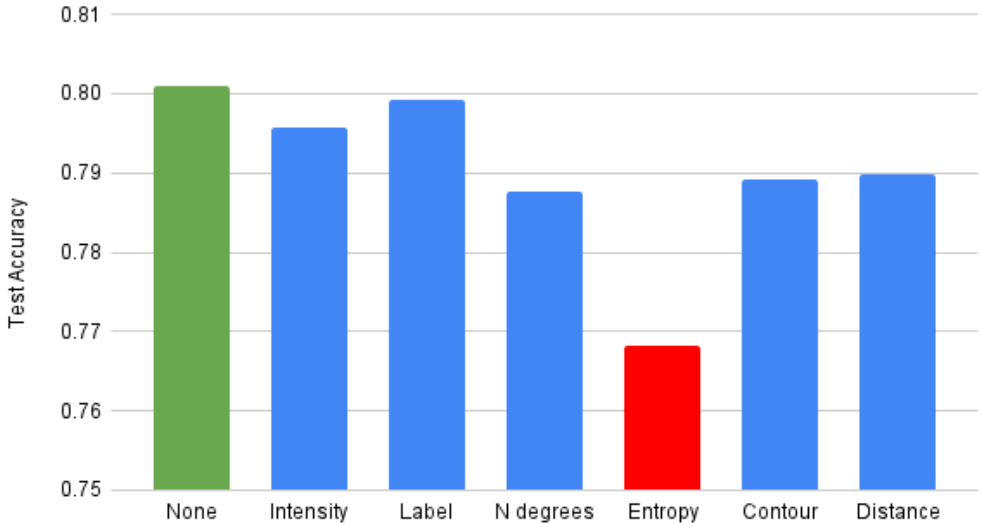
Salehi, S. S. M., D. Erdogmus, and A. Gholipour (2017). "Tversky loss function for image segmentation using 3D fully convolutional deep networks". In: *International Workshop on Machine Learning in Medical Imaging*, pp. 379–387.

Vovk, V. (2012). "Conditional validity of inductive conformal predictors". In: *Asian conference on machine learning*, pp. 475–490.

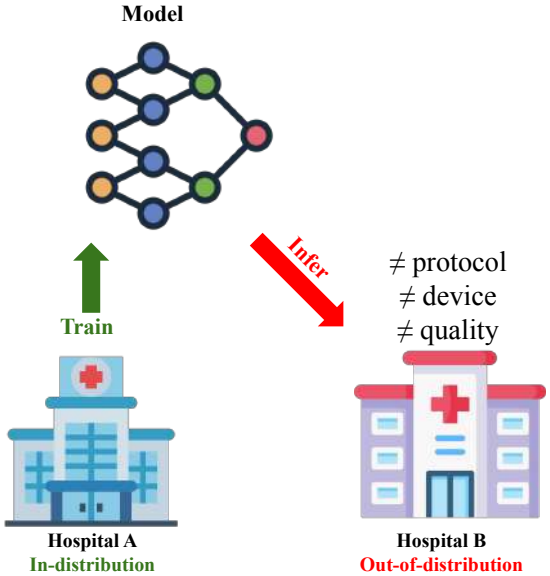
Wang, X. and M. Zhang (2022). "How Powerful are Spectral Graph Neural Networks". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Proceedings of Machine Learning Research 162, pp. 23341–23362.

Zhou, Z., M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang (2018). "Unet++: A nested u-net architecture for medical image segmentation". In: pp. 3–11.

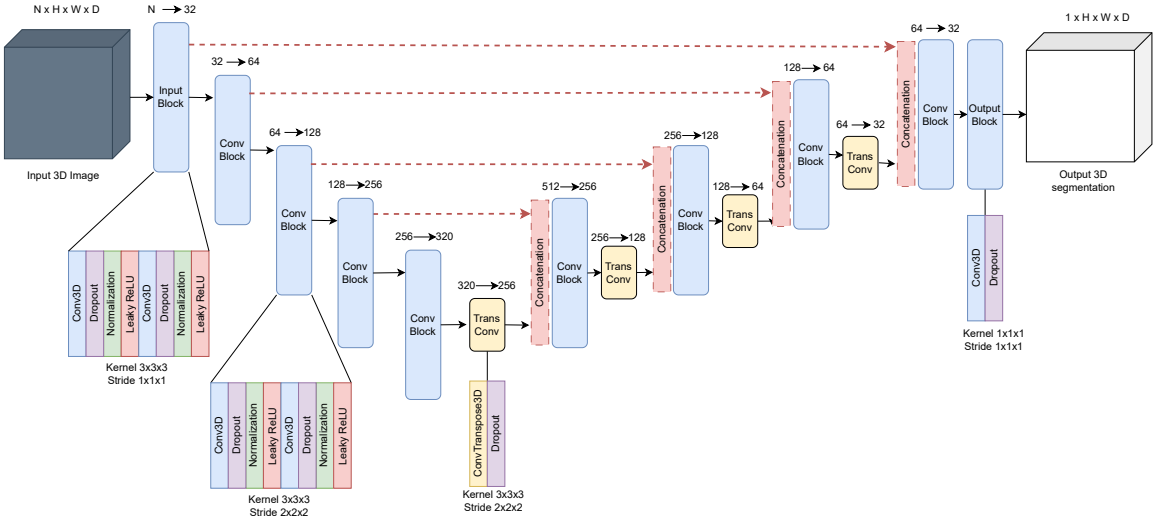
Feature importance - GIN model



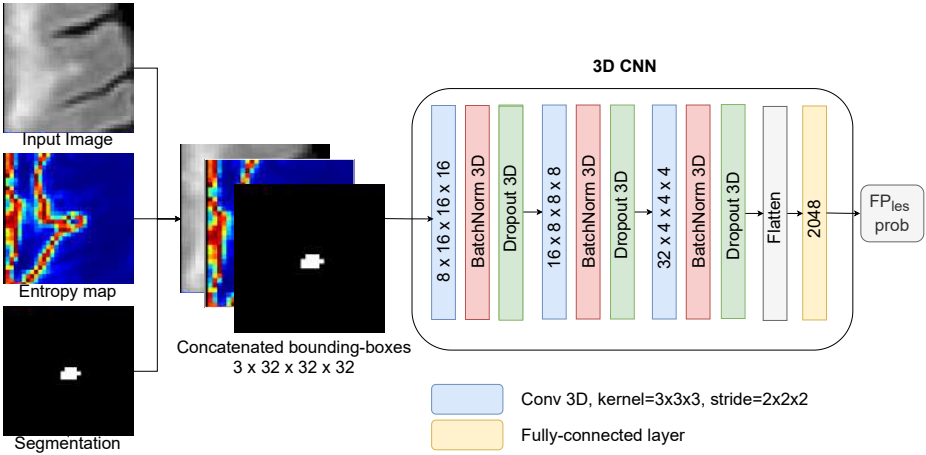
Domain-shift in medical-image analysis



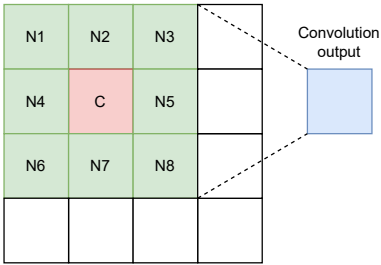
Dynamic U-Net model (16.5 million trainable parameters



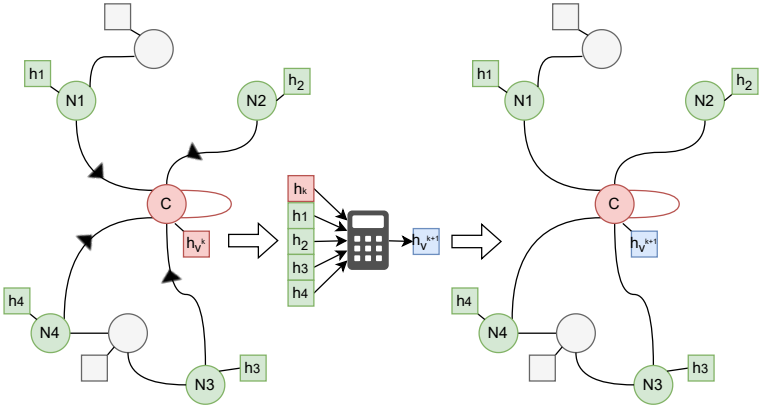
CNN-based lesion uncertainty quantification



Analogy between CNN and GNN



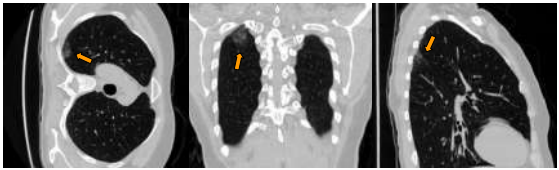
Convolutional Neural Network
Grid-like data (euclidean)



Graph Neural Network
Any structured data (can be non-euclidean)

Aggregate Combine

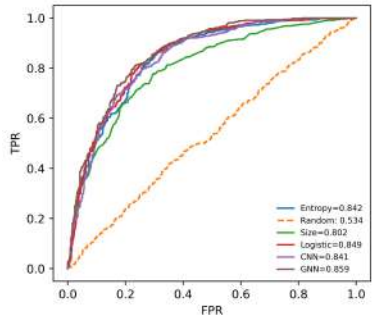
Lesion uncertainty - Application to lung nodule detection



Lung CT scan presenting a lung nodule.

Dataset

LIDC-IDRI dataset (Armato III et al. 2011) with 710 subjects for training, 50 for validation, 250 for testing.

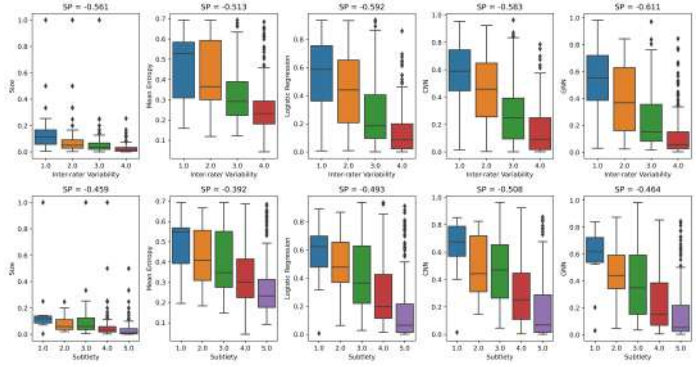


Lung nodules - Correlation with human uncertainty

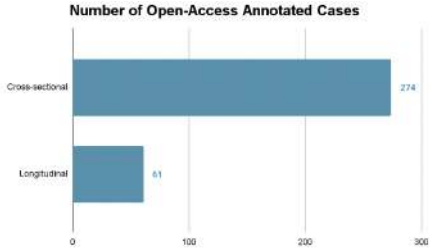
Nodule-level uncertainty ground truth scores

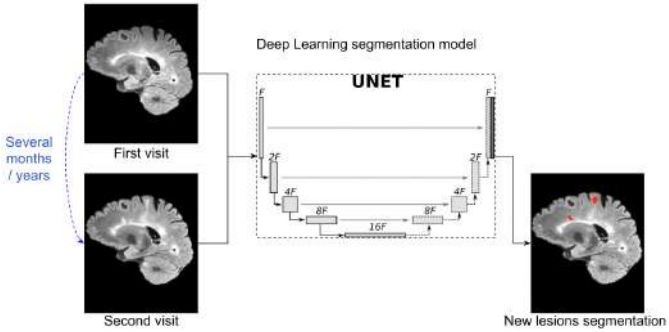
For each lung nodule, we have access to:

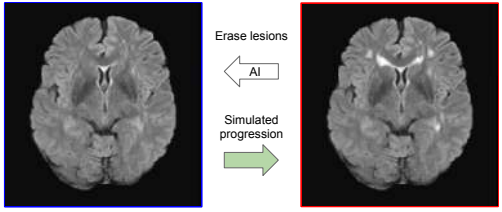
- the number of experts that marked the finding as a nodule (inter-rater variability)
- the subjective difficulty of detection (subtlety score).



An adversarial approach to longitudinal Multiple Sclerosis case synthesis



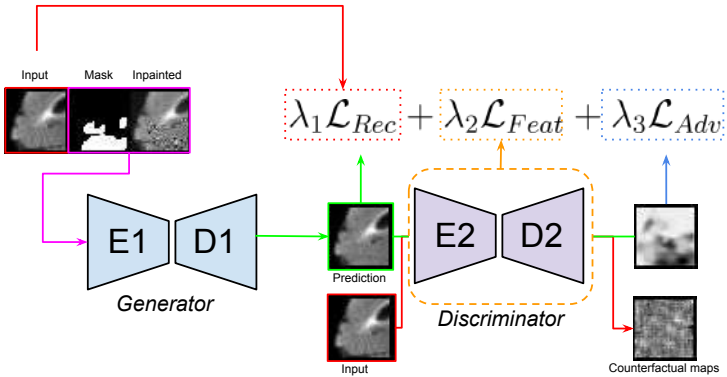




An adversarial approach to longitudinal Multiple Sclerosis case synthesis

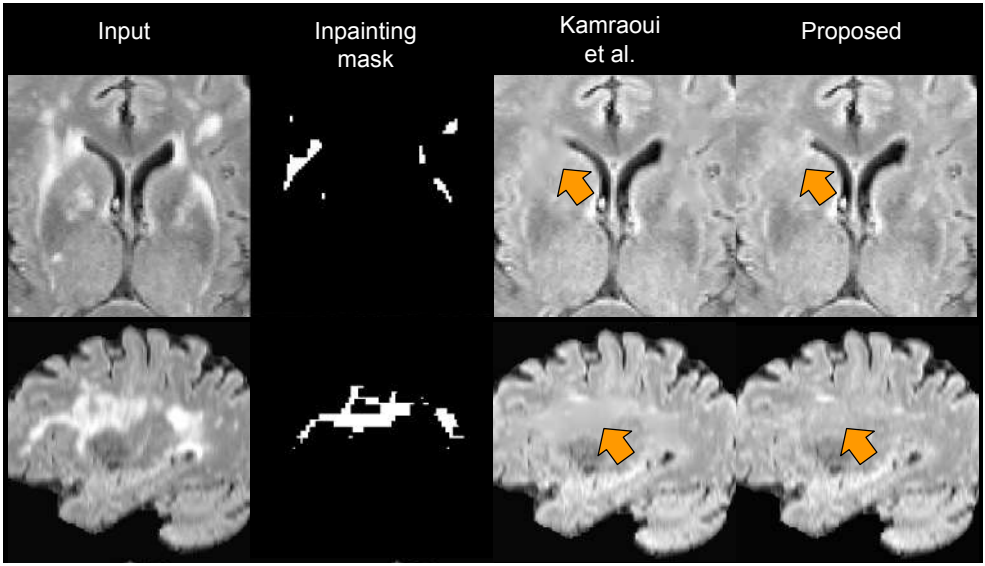
Inspiration: lesion inpainting (Kamraoui et al. 2022)

- Train an autoencoder model to erase lesions inside cross-sectional images.
- This yields a longitudinal case (2 visits) and a ground truth mask of new lesions.

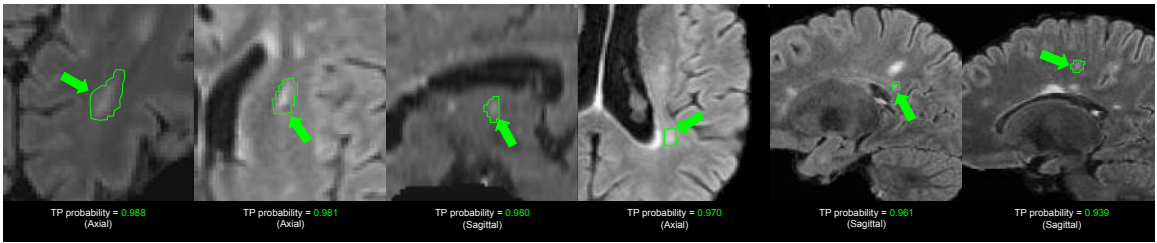


Proposed extension: improved realism using Generative Adversarial Networks

Synthetic longitudinal cases



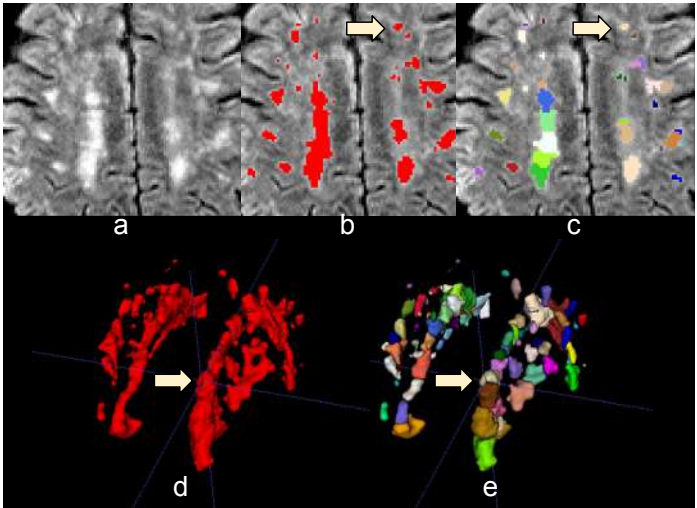
Human or Machine: Who is right?



False positive lesions associated with low uncertainty scores.

The lesion division algorithm

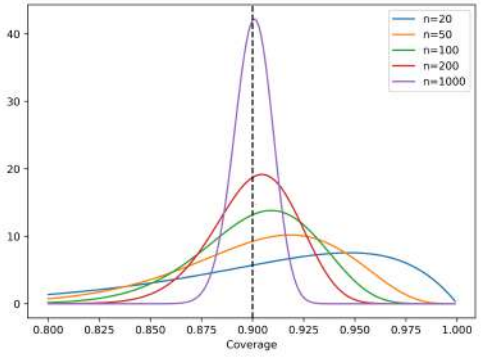
- Binary segmentation ⇒ lesion instances by identifying peaks in the probability map.



Conformal Prediction: the size of the calibration dataset

Analytic distribution of coverages (Vovk 2012)

$$\mathcal{P}(Y_{test} \in \Gamma_\alpha(X_{test}) | \{(X_i, Y_i)_{i=1}^N\}) \sim \text{Beta}(N + 1 - k, k)$$
with $k = \lfloor (N + 1) \times \alpha \rfloor$



Non-exchangeable Conformal Prediction

Main assumption of Conformal Prediction

- Calibration and test data should be exchangeable.
- In other words, there should be no **domain-shift**.
- Unrealistic for real-world medical applications.

Solution: reweight datapoints to make them exchangeable (Barber et al. 2023)

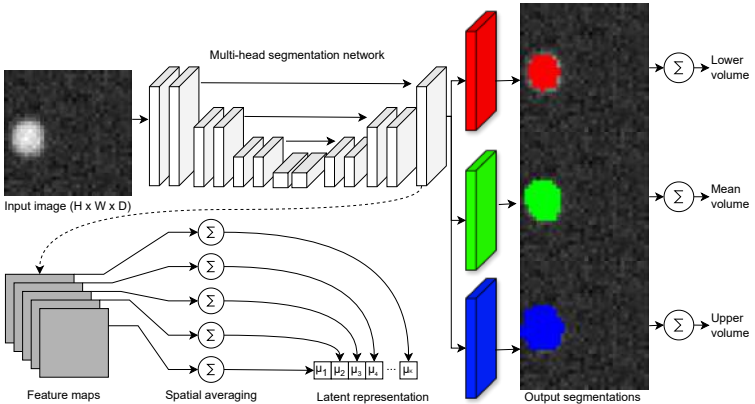
Writing $(X_1, ..., X_n)$ the n calibration sample, $1 - \alpha$ the desired coverage, s the score function:

- Estimate the density ratio: $w = dP_{\text{test}}/dP_{\text{train}}$
- Reweight calibration samples: $p_i^w(x) = \frac{w(X_i)}{\sum_{i=1}^{N} w(X_j) + w(x)}$

Practical estimation of the density ratio for high-dimensional medical images

Issues with Weighted Conformal Prediction

- The calibration and test distributions should not be too far apart, otherwise the density ratio is undefined.
- Estimating the density ratio is intractable in very high dimensions (3D MRIs).



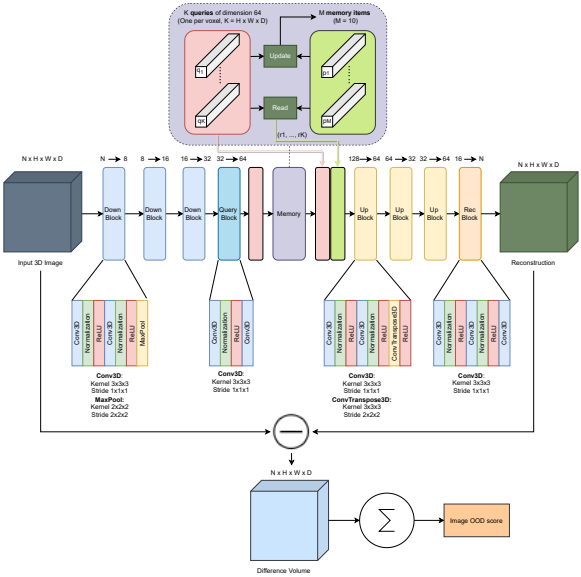
Classification-based density ratio estimation (Angelopoulos et al. 2022)

Definition

- Writing $X_1, ..., X_n$ the calibration points and $X_{n+1}, ..., X_{n+m}$ the test points.
- We set $C_i = 0$ for $i = 1, ..., n$ and $C_i = 1$ for $i = n + 1, ..., n + m$.
- Writing $\widehat{p}(x) = \mathcal{P}(C = 1|X = x)$ the probability predicted by a classifier trained on the $\{X_i, C_i\}$ dataset that the input sample x belongs to the test distribution.
- The density ratio is estimated as:

$$\widehat{w}(x) = \frac{\widehat{p}(x)}{1 - \widehat{p}(x)} \tag{6}$$

Reconstruction-based OOD detection



Pledge of integrity

In the presence of my peers.

With the completion of my doctorate, in my quest for knowledge, I have carried out demanding research, demonstrated intellectual rigor, ethical reflection, and respect for the principles of research integrity.

As I pursue my professional career, whatever my chosen field, I pledge, to the greatest of my ability, to continue to maintain integrity in my relationship to knowledge, in my methods, and in my results.

Drinks / Pot de thèse

Drinks and buffet on the fourth floor, library room!