TRAIL

TRUSTED AI LABS

Summer Workshop 25' London

Trustworthy medical diagnosis via uncertainty-driven adversarial training

Project n°10

KING'S College LONDON

KING'S INSTITUTE FOR ARTIFICIAL INTELLIGENCE

Wallonie - Bruxelles International.be

ncp WALLONIE

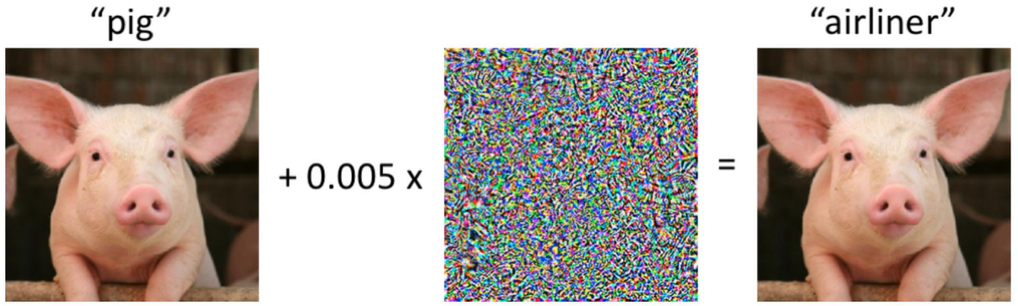
Wallonie service public SPW

S3

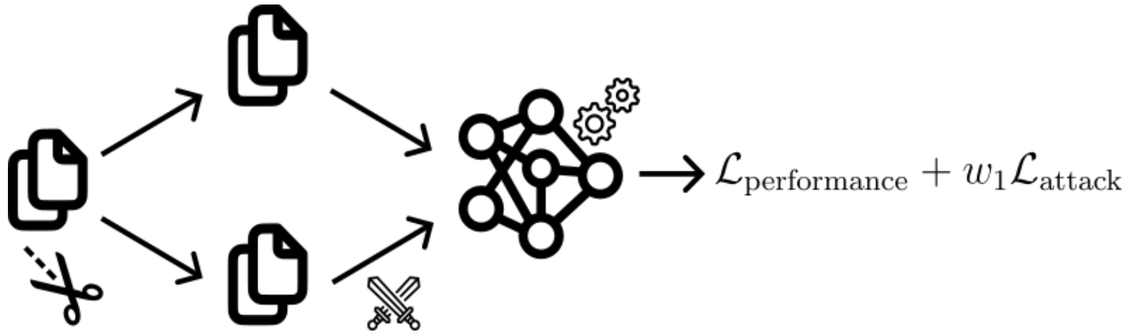
TRAIL TRUSTED AI LABS

Adversarial training

Motivation: Models can't be fooled in critical domains



Solution: Train with adversarial examples



Challenges:

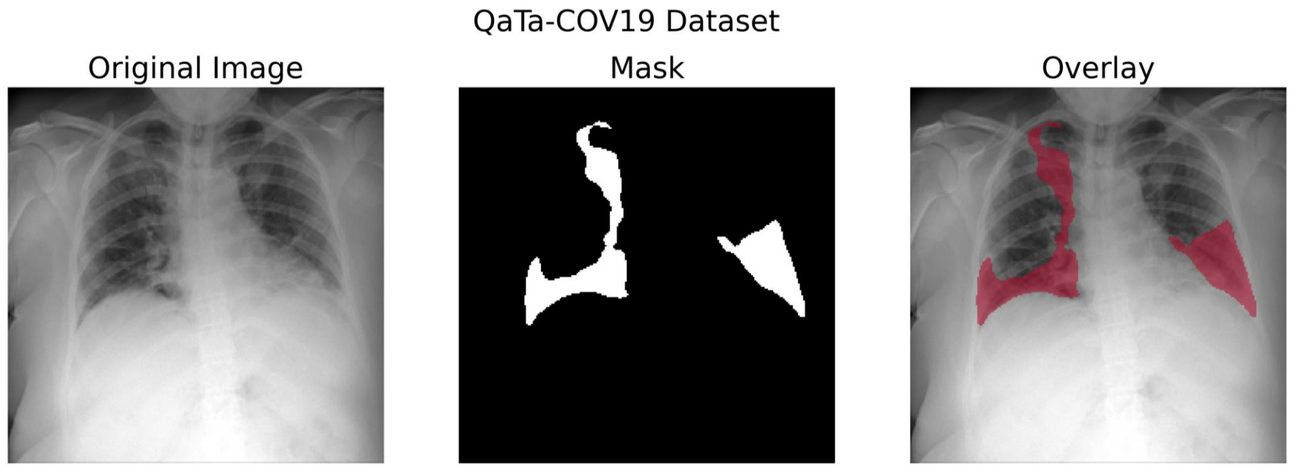
- Computational cost
- Sample selection for adversarial attacks

Task and dataset

Lung infected regions segmentation with QaTa-COV19

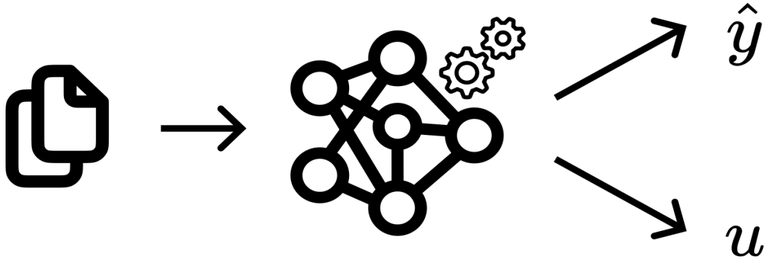
Binary segmentation from chest X-Ray images

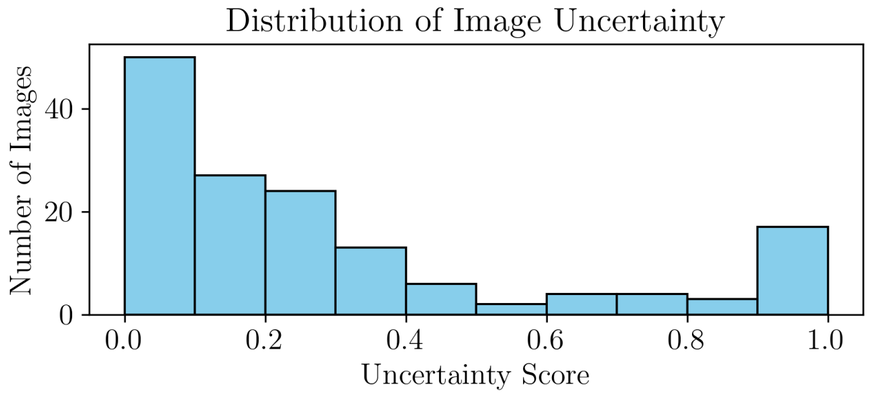
- Kaggle dataset
- 9258 infected patients
- 12544 healthy patients



Uncertainty estimation

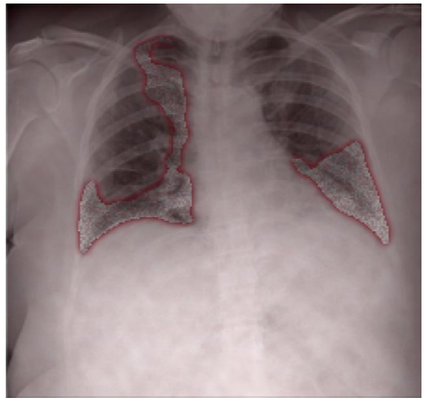
Motivation





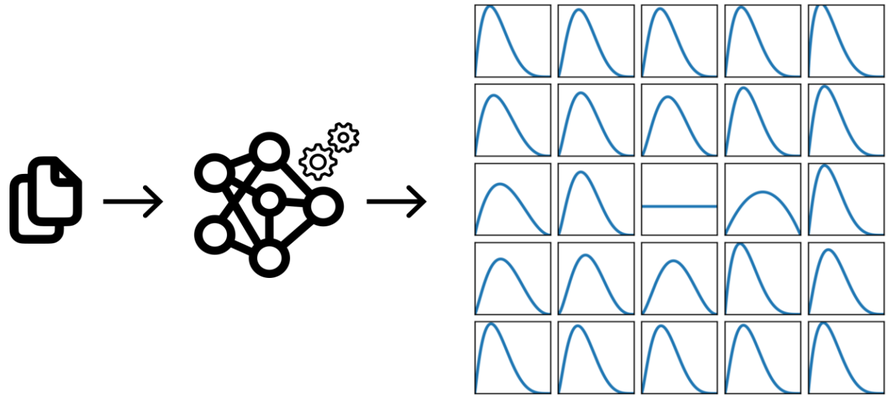
Pixel class probability

0.0	0.0	0.0	0.1	0.0
0.1	0.8	0.9	0.5	0.2
0.2	0.9	1.0	0.3	0.0
0.0	0.1	0.5	0.3	0.1
0.3	0.2	0.0	0.0	0.1



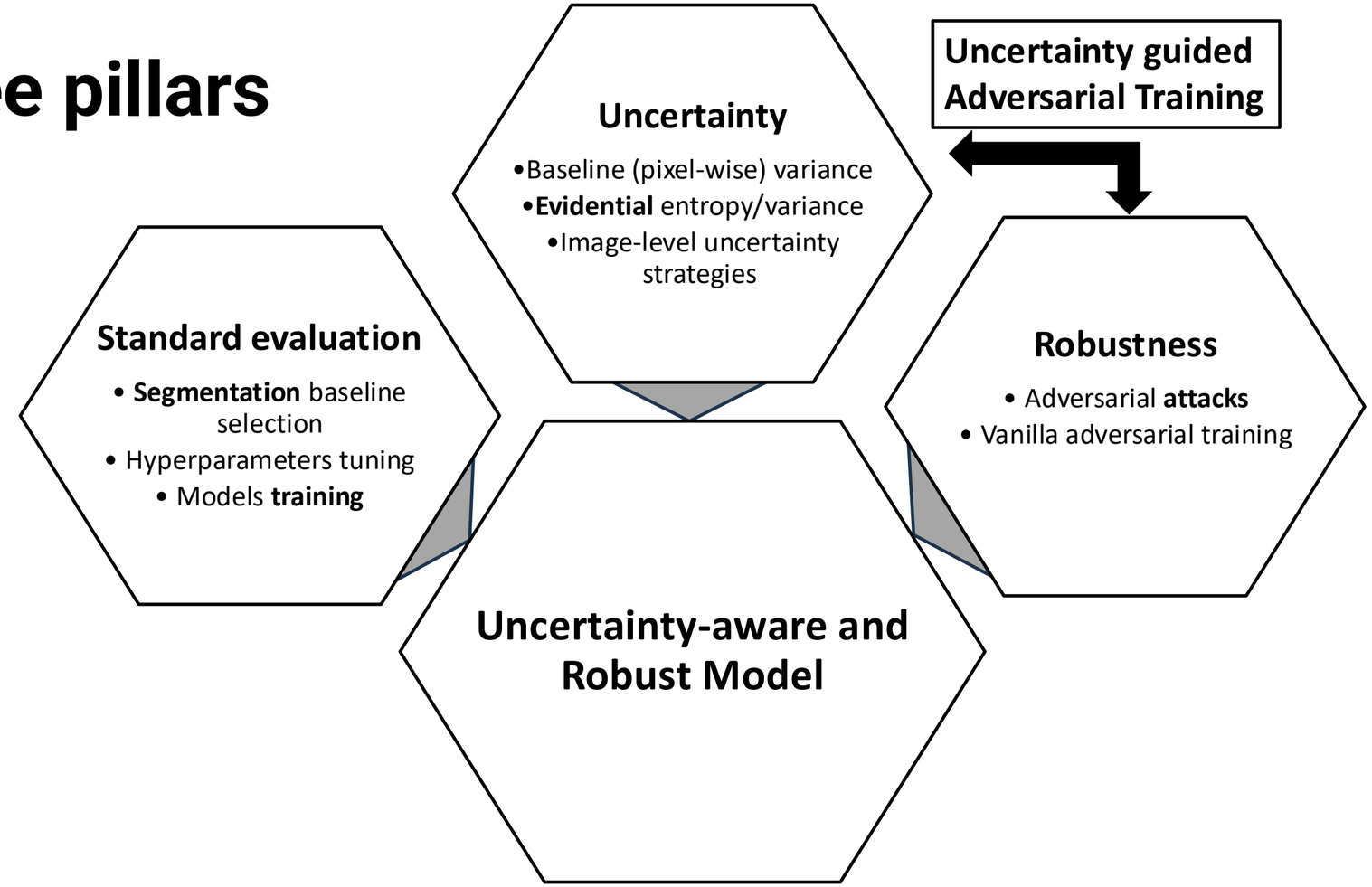
Evidential method

Output full distributions



Work Plan

Three pillars



Metrics

Performance: Dice, Intersection over Union, ...

Robustness: Attack success rate, Robust accuracy

Uncertainty: Performance against uncertainty filtering

Current members and expected outcome

Nicolas Sournac:
Expert in **robustness**

Bertrand Braeckeveldt:
Expert in **uncertainty** estimation

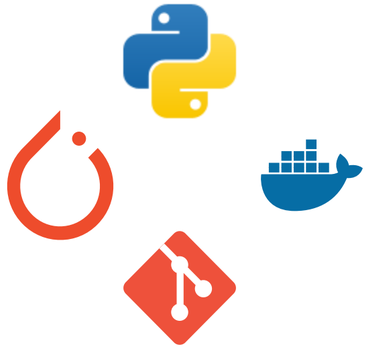
Part of GD6: Trustworthy AI for Critical Systems

Expected outcome:
Publication
Possibility to extend results after the TSW

Expertise Sought

Tools (minimal):

- PyTorch
- Git



Skills (minimal):

- Deep Learning fundamentals (training, monitoring, …)

Tools (optimal)

- Torchmetrics
- PyTorch-Lightning
- Matplotlib or Plotly

Skills (optimal)

- Experience in image segmentation
- Statistics (uncertainty and robustness)
- Structured deep learning workflows

TRAIL Summer Workshop 25’ London

TRUSTED AI LABS

Thank you for your attention !

Contact:
Sournac@multitel.be
Braeckeveldt@multitel.be

Full proposal: