TRAIL Summer Workshop' 25
Project Proposal

Full Name of Team Leader	Bertrand Braeckeveldt, Nicolas Sournac
Project Title	**Trustworthy medical diagnosis via uncertainty-driven adversarial training**
Profile of the Team Leader(s)	Bertrand Braeckeveldt and Nicolas Sournac are research scientists in the Fundamental AI team at Multitel led by Emmanuel Jean. Their research focuses on developing trustworthy deep learning models. They currently contribute to the ARIAC project. Bertrand is working on advanced methods for uncertainty quantification in deep learning systems. Nicolas is working on state-of-the art methods for robustness training and evaluation of AI models.
Abstract	As AI becomes increasingly integrated into high-stakes domains like healthcare, ensuring that models are not only accurate but also reliable and robust is essential. This project focuses on advancing trustworthy AI for medical image segmentation by exploring the synergy between uncertainty quantification and adversarial training. Specifically, it proposes to leverage evidential deep learning, which provides single-pass uncertainty estimates, to guide the generation of adversarial examples during training. By prioritizing high-uncertainty regions, we aim to improve both the model's robustness and the reliability of its confidence estimates. The approach will be applied to binary lung infection segmentation using the QaTa-COV19 chest X-ray dataset—a clinically relevant and well-annotated resource. This project focuses on evaluating models across three key dimensions: segmentation performance, robustness to perturbations, and quality of uncertainty estimates. The ultimate goal is to demonstrate how uncertainty-aware training strategies can make medical AI systems more dependable and aligned with emerging regulatory standards such as the European AI Act.
Project Objectives	With the increasing adoption of AI systems across diverse application domains, regulatory frameworks such as the European AI Act have emerged to ensure the trustworthiness of these technologies. Trustworthy AI goes beyond achieving high performance—it must also be explainable, robust, and uncertainty-aware. These requirements are especially critical in high-stakes applications like medical imaging, where the consequences of model errors can directly impact human lives. In this context, AI models used for diagnosis support must be capable of more than just making predictions. They should also recognize when they are uncertain and defer decisions to human experts rather than output potentially misleading results. Furthermore, they must maintain performance in the face of variability in data, such as differences in medical devices, acquisition protocols, and patient-specific conditions. This calls for models that are both robust to distributional shifts and adversarial perturbations, and that can provide reliable confidence estimates.

	This project focuses on the task of medical image segmentation, a foundational component in numerous clinical workflows including tumor delineation, lung pathology analysis, and cell structure identification. Accurate segmentation helps medical professionals localize regions of interest and make more informed decisions. However, due to the sensitivity of the task, segmentation models must be resilient to variations in imaging conditions and able to signal uncertainty in ambiguous or unfamiliar scenarios. To address these challenges, the project will pursue the following core objectives: 1) Develop a robust segmentation model through adversarial training. Adversarial training involves generating carefully crafted perturbations (adversarial examples) during training that can potentially mislead the model. By learning to resist such perturbations, the model becomes more robust to both natural and adversarial noise that may occur in real-world settings. 2) Design an uncertainty-aware segmentation model based on evidential deep learning. Unlike standard neural networks that output point estimates, evidential models output parameterized distributions over possible predictions, providing a natural way to estimate model uncertainty. These models accumulate evidence during training and express their confidence in the form of a belief distribution, enabling them to recognize and quantify uncertainty—particularly in out-of-distribution or ambiguous cases. 3) Combine adversarial training with uncertainty-guided example selection. The project will investigate how uncertainty estimates can be leveraged to drive adversarial training more effectively. Instead of generating perturbations uniformly or randomly, the training will prioritize samples with high predictive uncertainty, which are typically the most error-prone or fragile. Targeting the most uncertain examples when applying adversarial attacks should lower the required number of adversarial examples used during training, improving stability without sacrificing performance. This targeted strategy aims to both improve model robustness in vulnerable regions and enhance the reliability of uncertainty estimates. The project will evaluate these approaches along three essential dimensions: segmentation performance, robustness and uncertainty quality. By jointly addressing these criteria, the project seeks to demonstrate how state-of-the-art techniques in adversarial training and uncertainty modeling can be synergistically applied to build more trustworthy AI systems for medical imaging.
Project Dataset	This project aims to use the open dataset “QaTa-COV19 Dataset” available on Kaggle: https://www.kaggle.com/datasets/aysendegerli/qatacov19-dataset/data The QaTa-COV19 Dataset is a curated collection of chest X-ray (CXR) images specifically designed to support research in COVID-19 detection and infection region segmentation. It provides a high-quality and well-structured resource for training and evaluating AI models in medical imaging, with a focus on segmentation tasks. Developed through collaboration between Qatar University and Tampere University, the dataset includes both COVID-19 positive cases and control samples. The latest version

	contains 9,258 CXR images of COVID-19-infected patients, each accompanied by a manually annotated ground-truth segmentation mask highlighting infected lung regions. Additionally, the dataset includes 12,544 images from a control group with no visible signs of lung infection, enabling the development of models that can differentiate between infected and healthy lungs. The infected-lung subset is already divided into training and test sets, with 7,145 images allocated for training and 2,113 for testing, representing a well-balanced split (~23% for testing). All images and segmentation masks are provided in PNG format, making them straightforward to preprocess, visualize, and integrate into standard deep learning workflows. Due to its binary segmentation nature (infected vs. non-infected regions), this dataset is particularly well-suited for a focused two-week workshop on medical image segmentation. Its manageable complexity allows for the exploration of advanced topics such as uncertainty quantification and adversarial robustness, without the overhead of multi-class segmentation challenges. The dataset's size, structure, and clinical relevance make it an ideal platform for investigating trustworthy AI methods in critical healthcare applications.
Background Information	Machine learning and deep learning models are increasingly deployed across a wide range of domains—from natural language processing to autonomous driving and medical diagnostics—due to their remarkable performance. However, in high-stakes applications such as healthcare, performance alone is not enough [1-3]. Models must also be trustworthy, which means they must be robust, explainable, and uncertainty aware. This need for trustworthy AI is further underscored by upcoming regulatory frameworks like the European AI Act [4], which imposes stringent requirements on AI systems, especially in critical domains such as medicine. In this context, ensuring that models can reliably express their uncertainty and remain robust against adversarial perturbations or natural noise becomes a necessity rather than a luxury. Moreover, the ability of a model to self-assess its uncertainty is very important, as it enables the system to identify cases where it is unsure and defer the decision to a human expert [5], rather than risking an incorrect or potentially harmful prediction. **Uncertainty quantification** Uncertainty quantification in machine and deep learning seeks to model the predictive uncertainty of AI models. This uncertainty can be decomposed into two primary categories: aleatoric uncertainty and epistemic uncertainty [5,6]. Aleatoric uncertainty refers to the inherent variability in real-world situations. This type of uncertainty can arise from noise during data collection, insufficient feature selection (where the available data does not contain enough information to adequately capture the

data generation process), or poor resolution (e.g., low sampling frequency or image resolution). Aleatoric uncertainty is irreducible by adding more data, as it is rooted in the intrinsic variations of the environment or system.

Epistemic uncertainty, on the other hand, relates to the model's limitations in understanding or representing the problem. It includes issues such as the representativeness of the data, shifts in data distributions, or out-of-distribution data. Epistemic uncertainty is also influenced by model choices, including the architecture, optimizer, loss function, and hyperparameters. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced by gathering more data or refining the model.

Traditional machine/deep learning models are typically deterministic in their predictions, providing only a point estimate that satisfies the optimization criteria, such as the mean or mode of the predictive distribution. To estimate uncertainty, various methods aim to recover the full predictive distribution. Some approaches, like Bayesian methods, estimate a posterior distribution over model parameters [5,7], using techniques such as local approximations [8-10], variational inference [11], and Monte Carlo dropout [12]. Other methods, such as ensemble methods [13], combine predictions from multiple models to capture variability across different potential solutions. These methods require multiple forward passes through the network, which can be computationally expensive, particularly for real-time applications [14].

To address these limitations, deterministic methods have been proposed to estimate uncertainty with a single forward pass. One such approach involves using an external model to capture the uncertainty of the primary model, often by measuring the distance in the representation space [15,16]. However, this can be prone to issues like feature collapse or inconsistencies in distance preservation between the input and representation spaces [17].

A promising deterministic approach is evidential learning, which is grounded in the Dempster-Shafer theory of evidence [14,18,19]. Instead of directly predicting the target, evidential methods model a belief distribution over possible target. During training, the model accumulates evidence from the data, refining its belief distribution. Regions where the model has gathered more evidence will have a more peaked belief distribution, while regions with less evidence will result in a flatter distribution. This method can also be used to detect out-of-distribution data through regularization techniques.

Evidential deep learning was initially introduced for classification problems [18] and has since been extended to regression tasks [14,19]. Various modifications have been proposed to enhance its ability to capture uncertainty. A key advantage of evidential methods is that the uncertainty estimates have an analytical closed form, allowing for

efficient computation in a single pass, alongside the prediction, which is typically represented as the expectation of the belief distribution.

Evidential learning can be integrated into existing architectures with minimal adjustments, such as modifying the loss function or replacing the output layer and activation function. This flexibility means that any existing model can be adapted to gain insights into its uncertainty estimates. Once uncertainty estimates are obtained, they can be compared to a threshold, determined by the specific use case, to decide whether to accept or reject a prediction. If the uncertainty is too high, the prediction can be flagged for human expert verification [5].

Evidential deep learning has already been successfully applied to various domains, including image segmentation. For instance, it has been used for out-of-distribution obstacle detection in autonomous driving applications [20] and for tumor segmentation in the medical domain [21]. These applications demonstrate the potential of evidential learning to handle uncertainty effectively in critical tasks, where both the model's prediction and its uncertainty are essential for decision-making.

Robustness

Robustness in machine learning and deep learning refers to evaluating model's ability to maintain consistent predictions when exposed to input perturbations. This typically involves assessing whether the model's output remains stable or satisfies some properties under various modifications to the input.

Both the nature of input perturbations and the definition of the output property being evaluated may vary depending on the task or application domain. Nonetheless, the concept of stability—defined as a model's ability to preserve its performance level under such perturbations—is widely recognized as a fundamental aspect of robustness. This notion is formalized in standards such as ISO/IEC TR 24028-1 and ISO/IEC TR 24029-1.

Deep learning models are known to be vulnerable to two main types of perturbations: adversarial examples, which are carefully crafted and imperceptible modifications designed to manipulate the model's output [22]; and natural perturbations, which arise from real-world conditions such as lighting variations, sensor noise, or other environmental factors. These vulnerabilities are significant obstacles to the deployment of AI models in safety-critical domains and are respectively categorized under adversarial robustness and non-adversarial robustness.

Robustness is generally evaluated locally, in the neighborhood of specific input examples. This involves assessing whether perturbations—whether adversarial or

natural—can lead to violations of the desired model property when applied to input examples.

Adversarial robustness studies the existence of adversarial examples, which are generated using adversarial attacks (also known as evasion attacks) that aim to compromise a model's robustness. Some of these attacks, such as the Fast Gradient Sign Method (FGSM) [23] and Projected Gradient Descent (PGD) [24], leverage gradient information from the model to construct perturbations. Others, like DeepFool [25] and the Carlini & Wagner (C&W) attack [26], rely on geometric approaches (e.g., orthogonal projections to estimated decision boundaries) or formulate explicit optimization problems to craft adversarial examples. Adversarial examples are typically constrained by a perturbation budget or radius, which bounds their magnitude according to a specified norm.

Adversarial training

Adversarial training has naturally emerged as a prominent defense mechanism against adversarial attacks. It involves optimizing the model with respect to the worst-case loss over a perturbation region, often referred to as the robustness loss. Since this loss is intractable, adversarial attacks are commonly used to approximate it, providing a lower bound loss during training [24].

Adversarial training presents several challenges. It is computationally intensive [27], and it implies a trade-off between standard accuracy (on regular inputs) and robustness (on perturbed inputs) [28,29]. A widely adopted approach to make the learning procedure efficient and scalable is to group both normal and adversarial examples together before each training step [30]. This is usually done by randomly selecting data from the batch to generate adversarial samples. In this setup, carefully selecting which examples to attack based on an uncertainty measure could improve the stability of the training process and help mitigate the trade-off between robustness and standard accuracy.

Although adversarial training and uncertainty estimation are both active research areas, to the best of our knowledge, the integration of uncertainty estimation to improve adversarial training has not yet been thoroughly explored. Previous studies have shown that adversarial examples can fool uncertainty estimation techniques [31,32], highlighting the vulnerability of models that attempt to estimate uncertainty. However, the reverse—using uncertainty to enhance adversarial robustness—remains an open question in the literature.

In this proposal, we aim to address this gap by introducing evidential learning as the uncertainty estimation technique in the context of adversarial training. Evidential learning offers a key advantage: it provides uncertainty estimates in a single forward

pass, making it computationally efficient and scalable to large datasets and complex models. This efficiency is crucial in adversarial training, where multiple training iterations are required, and speed is essential to manage the computational cost.

Furthermore, by directly applying adversarial training to an uncertainty-aware model, we hypothesize that the model's robustness will improve. This is because the model not only becomes more resistant to adversarial attacks on its predictions but also to adversarial manipulation of its uncertainty estimates. In this way, the model should become more challenging to deceive, improving both the reliability of its predictions and the trustworthiness of its uncertainty estimates.

Bibliographic References

1. Rasheed et al., *Explainable, trustworthy, and ethical machine learning for healthcare: A survey*, Comput. Biol. Med. **106043** (2024), doi: 10.1016/j.compbiomed.2022.106043.
2. Markus et al., *Explainable, trustworthy, and ethical machine learning for healthcare: A survey*, J. Biomed. Inform. 113, **103655** (2021), doi: 10.1016/j.jbi.2020.103655
3. Zou et al., *A Review of Uncertainty Estimation and its Application in Medical Imaging*, arXiv:2302.08119.
4. European Parliament, *EU AI Act: first regulation on artificial intelligence*, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed Apr. 10, 2025).
5. Gawlikowski et al., *A Review of Uncertainty Estimation and its Application in Medical Imaging*, arXiv:2107.03342.
6. Abdar et al., *A review of uncertainty quantification in deep learning: Techniques, applications and challenges*, Inf. Fusion **76**, 243 (2021), doi: 10.1016/j.inffus.2021.05.008
7. Mena et al., *A Survey on Uncertainty Estimation in Deep Learning Classification Systems from a Bayesian Perspective*, ACM Comput. Surv. **54**, 193:1 (2021), doi: https://doi.org/10.1145/3477140
8. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
9. Maddox et al., *A Simple Baseline for Bayesian Uncertainty in Deep Learning*, arXiv:1902.02476.
10. Kristiadi et al., *Learnable Uncertainty under Laplace Approximations*, arXiv:2010.02720.
11. Rezende et al., *Variational Inference with Normalizing Flows*, arXiv:1505.05770.
12. Gal et al., *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, arXiv:1506.02142.
13. Lakshminarayanan et al., *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*, arXiv:1612.01474.
14. Gao et al., *A Comprehensive Survey on Evidential Deep Learning and Its Applications*, arXiv:2409.04720.
15. van Amersfoort et al., *Uncertainty Estimation Using a Single Deep Deterministic Neural Network*, arXiv:2003.02037.

16. Ramalho et al., *Density estimation in representation space to predict model uncertainty*, arXiv:1908.07235.
17. Liu et al., *A Simple Approach to Improve Single-Model Deep Uncertainty via Distance-Awareness*, arXiv:2205.00403.
18. Sensoy et al., *Evidential Deep Learning to Quantify Classification Uncertainty*, arXiv:1806.01768.
19. Amini et al., *Deep Evidential Regression,* arXiv:1910.02600.
20. Ancha et al., in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), **6943** (2024).
21. Zou et al., *TBraTS: Trusted Brain Tumor Segmentation*, arXiv:2206.09309.
22. Szegedy et al., *Intriguing properties of neural networks*, arXiv:1312.6199.
23. Goodfellow et al., *Explaining and Harnessing Adversarial Examples*, arXiv:1412.6572.
24. Madry et al., *Towards Deep Learning Models Resistant to Adversarial Attacks*, arXiv:1706.06083.
25. Moosavi-Dezfooli et al., *DeepFool: a simple and accurate method to fool deep neural networks*, arXiv:1511.04599.
26. Carlini et al., *Towards Evaluating the Robustness of Neural Networks*, arXiv:1608.04644.
27. Wong et al., *Fast is better than free: Revisiting adversarial training*, arXiv:2001.03994.
28. Zhang et al., *Theoretically Principled Trade-off between Robustness and Accuracy*, arXiv:1901.08573.
29. Tsipras et al., *Robustness May Be at Odds with Accuracy*, arXiv:1805.12152.
30. Kurakin et al., *Adversarial Machine Learning at ScaleRobustness May Be at Odds with Accuracy*, arXiv:1611.01236.
31. Carlini et al., *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*, arXiv:1705.07263.
32. Ledda et al., *Adversarial Attacks Against Uncertainty Quantification*, arXiv:2309.10586.
33. Qubvel-org, *segmentation_models.pytorch*, https://github.com/qubvel-org/segmentation_models.pytorch (accessed Apr. 11, 2025).
34. Kristoffersson Lind et al., *Uncertainty Quantification Metrics for Deep Regression*, Pattern Recognit. Lett. **186**, 91 (2024).
35. Trusted-AI, *adversarial-robustness-toolbox*, https://github.com/Trusted-AI/adversarial-robustness-toolbox (accessed Apr. 14, 2025).

Detailed Work Plan

This project is dedicated to the implementation and evaluation of robust, uncertainty-aware image segmentation models, with a focus on medical applications. The work plan is structured around three core pillars: image segmentation, uncertainty estimation, and robustness through adversarial training. These components will be developed in parallel, with coordinated integration and benchmarking.

Image Segmentation

The core task of image segmentation is approached as a pixel-wise classification problem. Given its clinical relevance, we will adopt architectures well-established in medical imaging, such as U-Net, Fully Convolutional Networks (FCN) or DeepLabV3. These models are readily available through popular libraries such as PyTorch, Hugging Face, or segmentation_models.pytorch [33]. Pretrained weights can be leveraged to accelerate development and rapidly establish strong baselines.

A classical segmentation model will be trained on the QaTa-COV19 dataset, which provides binary segmentation masks indicating infected lung regions. This serves as the foundation for the uncertainty and robustness enhancements introduced in later stages.

Uncertainty Estimation

A key objective of the project is to incorporate evidential deep learning into the segmentation pipeline. Evidential models learn to produce a distribution over possible predictions, enabling the estimation of epistemic uncertainty in a single forward pass. This will require minor modifications to the model's final layer (e.g., from sigmoid to softplus activations for evidence output) and training with an evidential loss, such as the Bayesian Risk Loss or the Type-II Maximum Likelihood Loss tailored for the Beta distribution in binary segmentation.

Pixel-wise uncertainty can be derived from the evidential parameters (e.g., entropy or variance of the Beta distribution). We will also experiment with heuristic uncertainty surrogates in non-evidential models, such as the Bernoulli variance, to compare effectiveness.

At the image level, uncertainty aggregation strategies will be explored—ranging from simple statistics (mean, max) to threshold-based counts (e.g., number of pixels above a certain uncertainty level). Novel aggregation strategies may also emerge through discussion and experimentation during the workshop.

Robustness via Adversarial Training

To improve model reliability under input perturbations, adversarial training techniques will be implemented. This involves generating adversarial examples—deliberate, small perturbations of the input designed to fool the model—and incorporating them into training. We plan to explore fast and effective methods such as FGSM (Fast Gradient Sign Method), R+FGSM (Random-FGSM) and PGD (Projected Gradient Descent).

A unique aspect of this project is to guide adversarial training using uncertainty. Specifically, we will prioritize generating adversarial examples from high-uncertainty samples from the batch, based on either evidential or surrogate uncertainty estimates. This selective strategy is expected to target weak points of the model more effectively than random sampling. Three adversarial training regimes will be evaluated. Random sample-based adversarial training, uncertainty-guided adversarial training and a baseline training without adversarial examples. These will be applied to both classical and evidential models to evaluate the benefits of combining adversarial robustness with uncertainty awareness.

Evaluation and Benchmarking

Evaluation will cover three complementary aspects. Segmentation performance, uncertainty quality and robustness. Performance can be accessed via Dice Coefficient or Intersection over Union (IoU). Uncertainty quality involves the Area Under the Sparsification Error curve (AUSE) [34]. Robustness will be studied via performance degradation under adversarial perturbations (Dice/IoU vs. perturbation magnitude) and uncertainty shift under attack (e.g. measure of correlation).

	These metrics will allow us to benchmark models not only on their accuracy, but also on how well they know what they don't know—and how resilient they are to difficult or malicious inputs. **Project Management** The project welcomes contributors with backgrounds in machine/deep learning, uncertainty modeling, computer vision, or software engineering. Familiarity with image segmentation or PyTorch is beneficial but not required. Members with strong development experience will play a key role in building reliable workflows, experiment tracking, and reproducibility tools. The project can be separated into the following tasks: • Data preprocessing, augmentation, and loading • Training classical segmentation models • Adapting models for evidential learning • Implementing the adversarial training pipeline • Developing evaluation and visualization tools Roles and responsibilities will be assigned collaboratively during a pre-workshop online meeting. Key architectural and experimental design choices—such as model selection, loss functions, and adversarial attack methods—will also be discussed in advance to ensure an efficient start. Tasks like classical and evidential model training, uncertainty processing, and robustness evaluation will be developed in parallel to avoid bottlenecks and encourage iterative improvements. **Tools and Resources** • Programming Language: Python • Frameworks: PyTorch, PyTorch Lightning (optional), segmentation_models.pytorch [33], Torchmetrics • Visualization: Plotly, Matplotlib • Adversarial Tools: Adversarial Robustness Toolbox [35] • Version Control: Git • Environment Management: Virtual environment or Docker (based on team preferences and constraints) • Compute Resources: Access to the LUCIA supercomputer via ARIAC/Cenaero
Other Remarks	This project is directly related to the "Grand Challenge 6" (Trustworthy AI for Critical Systems) of the TRAIL initiative.