FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

UAntwerp Symposium — 05/09/2025

Uncertainty Quantification with Neural Networks

From Selective Prediction to Active Learning

Drs. Arthur Thuy
Prof. dr. Dries F. Benoit

GHENT UNIVERSITY

fwo

FLANDERS MAKE

UAntwerp Uncertainty 05/09/2025

About me & us

Research areas

- Uncertainty quantification in neural networks
- Data-efficient natural language processing
- Applications in learning analytics (FWO)

Figure: Infeasible CAD design

Flanders Make projects in research group:

- ManProCAD: estimating manufacturing effort for 3D CAD designs
- REMASC: supporting decision-making and estimating cost for remanufacturing
- CostLeap: cost-sensitive methods for predictive maintenance

UAntwerp Uncertainty 05/09/2025

Overview

1 Neural networks in the wild

2 Uncertainty Quantification

3 Uncertainty in NNs

4 Selective Prediction

5 Active Learning

6 Conclusion

UAntwerp Uncertainty 05/09/2025

NEURAL NETWORKS IN THE WILD

UAntwerp
Uncertainty 05/09/2025

GHENT
UNIVERSITY

Deploying neural networks

- NNs are deployed widely in manufacturing
  - Quality Control (image data, classification)
  - Predictive Maintenance (tabular data, regression)
  - Defect detection from warranty claims (text data, classification)
- We assume in-distribution (ID) data, but often get out-of-distribution (OOD) data

ID	$p_{\text{TRAIN}}(\mathbf{X}, \mathbf{Y}) = p_{\text{TEST}}(\mathbf{X}, \mathbf{Y})$
OOD	$p_{\text{TRAIN}}(\mathbf{X}, \mathbf{Y}) \neq p_{\text{TEST}}(\mathbf{X}, \mathbf{Y})$

- Problem: NNs are confidently wrong on OOD data (Ovadia et al. 2019)

NN performance silently degrades $\rightarrow$ OOD data is dangerous

UAI tutorial Uncertainty 05/09/2025

Example case: Quality Control

Setup: Image data, classification

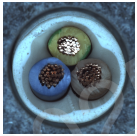
Shifted data

- Lighting: LED vs natural light
- Camera: different resolution or lens distortion
- Environment: clean lab vs production line with dust, vibrations, or thermal distortions

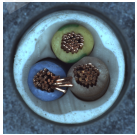
OOD data

- Novel defect types

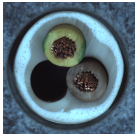
Image source: Bergmann et al. (2019)



(a) Good



(b) Bent wire



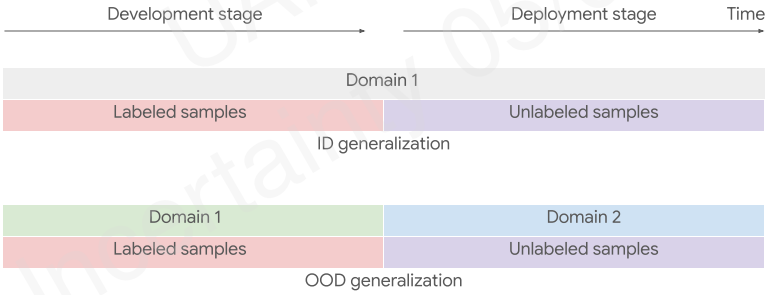
(c) Missing cable

Antwerp Uncertainty 05/09/2025

OOD Generalization

Objective

Construct robust models that perform accurately even on data not represented in the training set

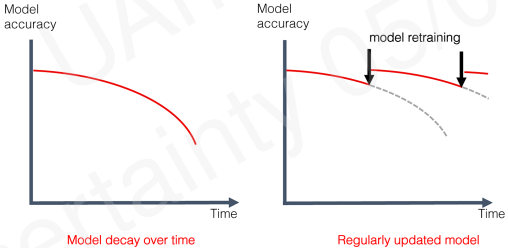


Taken from Mucsányi et al. (2023)

UAntwerp Uncertainty 05/09/2025

Why we care

- Typical solution is model retraining
- Requires: data collection, data annotation, model retraining on new data
- Expensive: MLOps (manpower, GPU, data management) and data curation



Taken from Mucsányi et al. (2023)

UAntwerp Uncertainty 05/09/2025

“Solving” OOD Generalization

→ ML model will work well under new situations

Positive consequences

Lower cost

+ Less MLOps needed
+ Small businesses can adopt ML more easily

Real-world applicability

+ Reliable models are trustworthy
+ ML can be extended to high-stakes applications

UAntwerp Uncertainty 05/09/2025

Uncertainty Quantification

UAntwerp
Uncertainty 05/09/2025

GHENT
UNIVERSITY

Uncertainty is everywhere

- Perfectly understanding a system can only happen in simple environments
  $\rightarrow$ The world around us is not such an environment
- In classical ML, the main goal was to find methods that work well (standalone components)
- ML is increasingly integrated in decision-making
- Accuracy is not the biggest concern

Quantifying uncertainty in ML models makes them *know when they do not know*

UAntwerp
Uncertainty 05/09/2025

Dog-cat classifier

- NN trained on images of dogs and cats
- It outputs $p(y = \text{cat} \mid x)$ and $p(y = \text{dog} \mid x)$
- During deployment, it gets the following images
  $\rightarrow$ what would the NN predict?

UAntwerp
Uncertainty 05/09/2025

Dog-cat classifier

- NN trained on images of dogs and cats
- It outputs $p(y = \text{cat} \mid x)$ and $p(y = \text{dog} \mid x)$
- During deployment, it gets the following images
  $\rightarrow$ what would the NN predict?

(a) Dog and cat

UAntwerp
Uncertainty 05/09/2025

Dog-cat classifier

- NN trained on images of dogs and cats
- It outputs $p(y = \text{cat} \mid x)$ and $p(y = \text{dog} \mid x)$
- During deployment, it gets the following images
  $\rightarrow$ what would the NN predict?

(a) Dog and cat

(b) Dog with cat features

UAntwerp
Uncertainty 05/09/2025

Dog-cat classifier

- NN trained on images of dogs and cats
- It outputs $p(y = \text{cat} \mid x)$ and $p(y = \text{dog} \mid x)$
- During deployment, it gets the following images
  $\rightarrow$ what would the NN predict?

(a) Dog and cat

(b) Dog with cat features

(c) Horse

UAntwerp
Uncertainty 05/09/2025

Two sources of uncertainty (1)

Aleatoric uncertainty

- *"I am not sure because there are multiple plausible answers"*
- Uncertainty due to the inherent variability or randomness in the data
- Cannot be reduced by collecting more data
- Practice: observations where humans would struggle (ambiguity in the question or in the observation)

→ We expect 50%/50%

(a) Dog and cat

(b) Dog with cat features

UAntwerp Uncertainty 05/09/2025

Two sources of uncertainty (2)

Epistemic uncertainty

- *"I am not sure because I have not seen it before"*
- Uncertainty due to the lack of knowledge or information (underexplored data regions)
- Can be reduced by collecting more data
- Practice: shifted and OOD data

(a) Horse

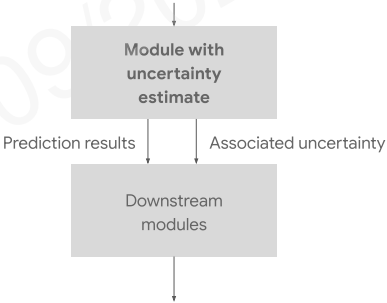
→ NNs are often confidently wrong (e.g., 95%/5%, 10%/90%) (Guo et al. 2017)

Predictive uncertainty: the sum of aleatoric and epistemic uncertainty

UAntwerp Uncertainty 05/09/2025

Critical building block

- ML model as part of a bigger pipeline
- Careful when using the model's prediction in downstream modules
  $\rightarrow$ people do not (and should not) trust every prediction
- Two model outputs: prediction and associated uncertainty estimates
- Uncertainty as XAI (Thuy and Benoit 2024a)
  - Local
  - Model-specific



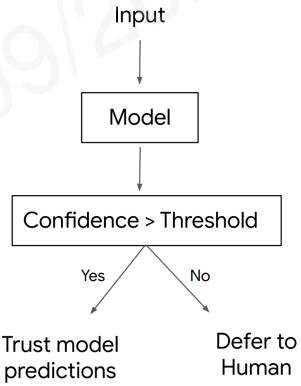
Taken from Mucsányi et al. (2023)

UAntwerp Uncertainty 05/09/2025

HITL use case 1: Selective Prediction

Human-in-the-loop (HITL)

- When the model's uncertainty is high, it can say *"I'm not sure about the results"*
- Uncertain observations are deferred to a human expert



UIAntwerp Uncertainty 05/09/2025

HITL use case 2: Active Learning

- Decrease the labeling work for human annotators
- Request a label only for the *most informative* unlabeled observations

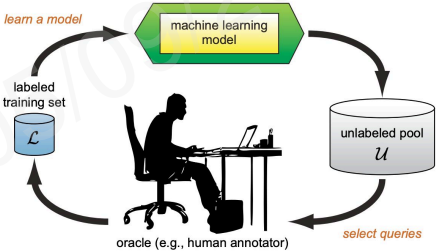


Image source: Active Learning Literature Survey, Settles 2010

UAntwerp Uncertainty 07/06/2025

Uncertainty in NNs

UAntwerp
Uncertainty 05/09/2025

GHENT
UNIVERSITY

Representing aleatoric uncertainty (1)

Ingredients

- Architecture: probabilistic output
- Loss function: proper scoring rule (match output distribution to ground truth)

UAntwerp
Uncertainty 05/09/2025

Representing aleatoric uncertainty (1)

Ingredients

- Architecture: probabilistic output
- Loss function: proper scoring rule (match output distribution to ground truth)

Standard NN classifier ✓

- Architecture: Softmax output over the classes, $p(y \mid x)$
- Loss: Negative log-likelihood (NLL) (proper scoring rule)

$$\text{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}(y_i = k) \log p(y = k \mid x_i, \theta)$$

UAntwerp Uncertainty 05/09/2025

Representing aleatoric uncertainty (2)

Standard NN regressor ✗

- y is a continuous random variable and follows a normal distribution + homoscedastic
- Architecture: single output unit for mean $\hat{y}$
- Loss: MSE

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

UAntwerp Uncertainty 05/09/2025

Representing aleatoric uncertainty (2)

Standard NN regressor ✗

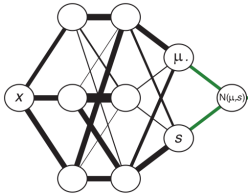
- y is a continuous random variable and follows a normal distribution + homoscedastic
- Architecture: single output unit for mean $\hat{y}$
- Loss: MSE

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Adapting the NN regressor ✓

- We allow the standard deviation to vary with x, i.e., heteroscedastic
- Architecture: two output units, for mean $\hat{y}$ and standard deviation $\hat{\sigma}$
- Loss: learned loss attenuation (Kendall et al. 2017)

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2\hat{\sigma}_i^2}(y_i - \hat{y}_i)^2 + \frac{1}{2}\log\hat{\sigma}_i^2\right]$$



Taken from Dürr et al. (2020)

UAntwerp Uncertainty 05/09/2025

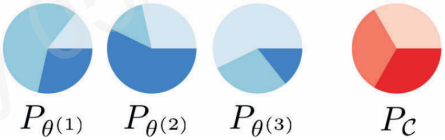
Representing epistemic uncertainty

Standard NN classifier ✗
Standard NN regressor ✗

Ingredient: a set of models → check the disagreement between models

Two model families

- Bayesian ML
- Ensembling



Taken from Settles (2009)

Most popular method in each family: Monte Carlo Dropout and Deep Ensembles

UAntwerp Uncertainty 05/09/2025

Representing epistemic uncertainty: Deep Ensembles (1)

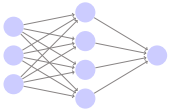
Ensemble of T NNs, trained on the same dataset and input features (Lakshminarayanan et al. 2017)

At test time, each of the T NNs performs one forward pass

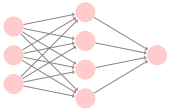
Diversity

- Different parameter initializations
- Noise in the stochastic gradient descent (SGD) training process
- Highly non-convex loss landscape

(a) Sample 1



(b) Sample 2



(c) Sample T

UAnTwerp Uncertainty 05/09/2025

Representing epistemic uncertainty: Deep Ensembles (1)

Ensemble of T NNs, trained on the same dataset and input features (Lakshminarayanan et al. 2017)

At test time, each of the T NNs performs one forward pass

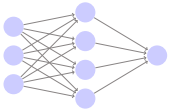
Diversity

- Different parameter initializations
- Noise in the stochastic gradient descent (SGD) training process
- Highly non-convex loss landscape

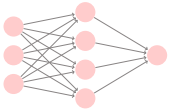
Already using ensembles? Get your UQ for free!

FREE

(a) Sample 1



(b) Sample 2

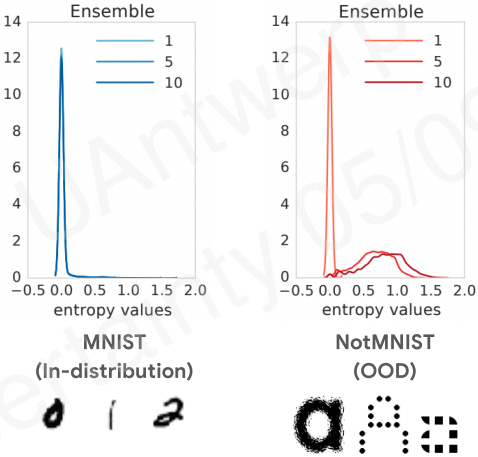


(c) Sample T

UAnwerp Uncertainty 05/09/2025

Representing epistemic uncertainty: Deep Ensembles (2)

Convolutional NN classifier trained on the MNIST dataset (digits 0–9)



Taken from Lakshminarayanan et al. (2017)

UAntwerp Uncertainty 05/09/2025

Representing epistemic uncertainty: MC Dropout

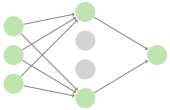
Apply dropout not only at training time but also at test time (Gal et al. 2016)

At test time, obtain T thinned dropout variants of the NN and perform one forward pass in each

Other interpretations

- Variational inference with Bernoulli distributions
- Bayesian approximation of the Gaussian process

(a) Sample 1



(b) Sample 2

(c) Sample T

UAntwerp Uncertainty 05/09/2025

Measuring uncertainty: classification

It is hard to disentangle predictive uncertainty into aleatoric and epistemic sources

Model output: $p(\hat{y} \mid x, \theta_t)$, for T samples (Depeweg et al. 2018)

$$u_{pred}(x) \approx \mathbb{H}\left[\frac{1}{T}\sum_{t=1}^{T} p(\hat{y} \mid x, \theta_t)\right]$$

$$u_{al}(x) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{H}\left[p(\hat{y} \mid x, \theta_t)\right]$$

$$u_{ep}(x) = u_{pred}(x) - u_{al}(x)$$

UAntwerp Uncertainty 05/09/2025

Measuring uncertainty: regression

Model output: mean prediction $\hat{y}_t$ and standard deviation $\hat{\sigma}_t$, for T samples (Kendall et al. 2017)

$$u_{al}(x) = \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_t^2 \approx \mathbb{E}[\hat{\sigma}_t^2]$$

$$u_{ep}(x) = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t\right) \approx Var[\hat{y}_t]$$

$$u_{pred}(x) = u_{al}(x) + u_{ep}(x)$$

UAntwerp
Uncertainty 25/09/2025

Practical: torch-uncertainty Python package

https://github.com/ENSTA-U2IS-AI/torch-uncertainty

UAntwerp Uncertainty 05/09/2025

Selective Prediction

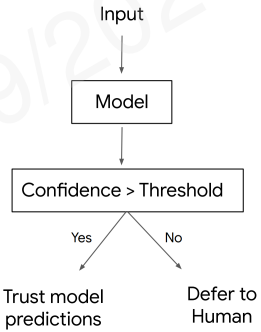
UAntwerp
Uncertainty 05/09/2025

GHENT UNIVERSITY

Selective Prediction

Reduce the error rate by abstaining from prediction when in doubt

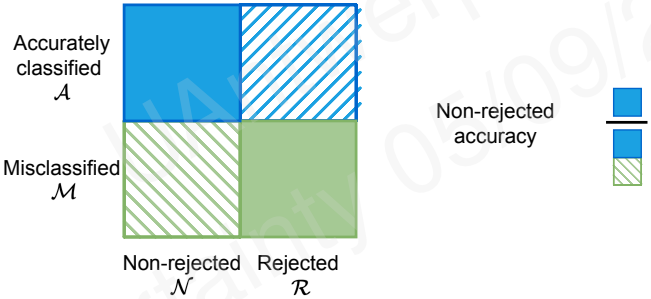
- Find uncertainty level at which to abstain
  $\rightarrow$ use in deployment
- Relevant if making an incorrect prediction costs more than consulting a human expert.
  (think far enough about effect on company! E.g., quality control)



UAntwerp Uncertainty 05/09/2025

METRIC

Non-rejected accuracy (NRA): ability to classify non-rejected samples accurately.



Taken from Thuy and Benoit (2024a)

UAnvers Uncertainty 05/09/2025

SETUP

Different evaluation sets

- Only ID data
- ID data & shifted data (shifted data has original labels)
- ID data & fully OOD data (fully OOD data has other labels $\rightarrow$ always incorrect)

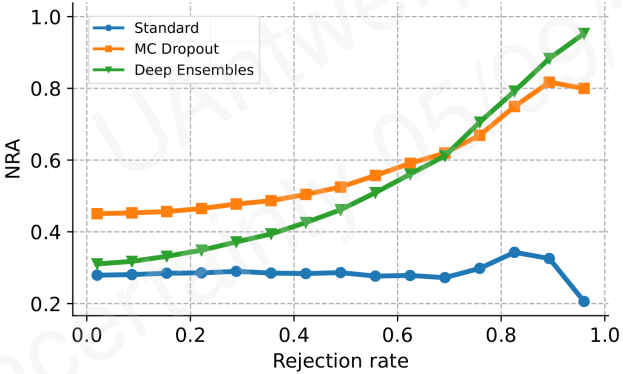
The model should exhibit higher uncertainty for incorrect predictions, and reject these first

UAntwerp Uncertainty 05/09/2025

Plot

What we want to see: Monotonically increasing accuracy when increasing the threshold

Example output on ID & shifted data



Taken from Thuy and Benoit (2024a)

UAntwerp Uncertainty 05/09/2025

Active Learning

GHENT UNIVERSITY

UAntwerp Uncertainty 05/09/2025

Active learning

Active learning

- Supervised human-in-the-loop approach
- Minimize labeled data
- Match performance of fully supervised models

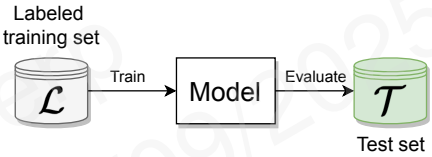
Prerequisites for relevancy

- Unlabeled data points can be easily collected
- You need a lot of labeled data to train an accurate ML model (NNs!)
- An oracle can provide labels (might be expensive, e.g., an expert)

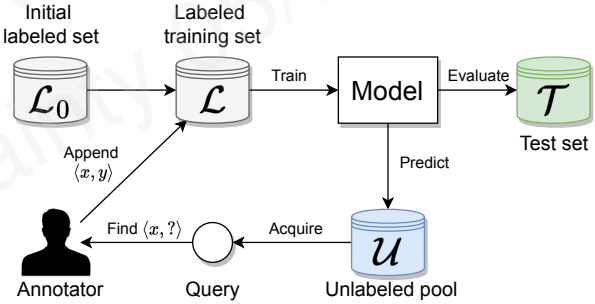
UAntwerp Uncertainty 05/09/2025

WORKFLOW

Supervised learning



Active learning
(pool-based sampling)



UAntwerp
Uncertainty 25/09/2025

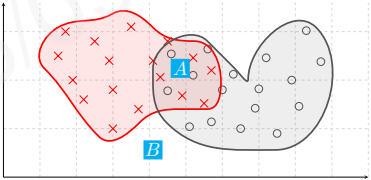
Acquisition functions

Query by committee

Any disagreement-based approach that uses a committee (or ensemble) of models

Translates to epistemic uncertainty

- Method: Deep Ensembles, MC Dropout
- High where few training observations
  $\rightarrow$ observations we want to label B



Adapted from Hüllermeier et al. (2021)

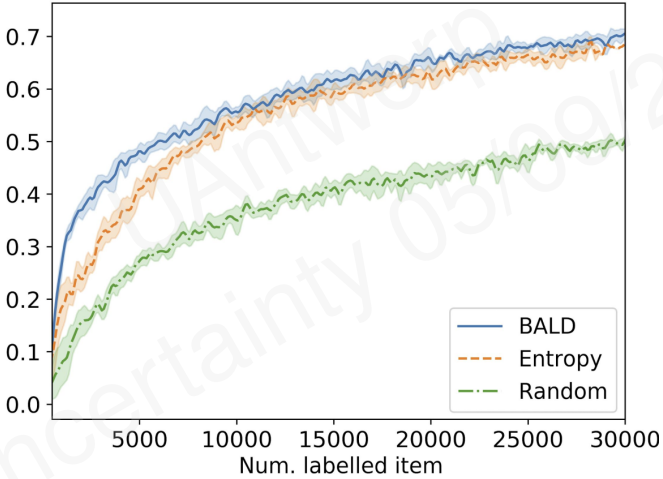
Acquisition functions

- Classification: Bayesian Active Learning by Disagreement (BALD) (Houlsby et al. 2011)
- Regression: Variance of $\hat{y}$

UAntwerp Uncertainty 05/09/2025

Plot

What we want to see: accuracy is higher



Taken from Atighehchian et al. (2020)

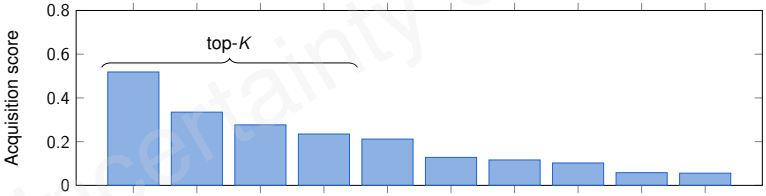
UAntwerpen Uncertainty 05/09/2025

PRACTICAL: BATCH-MODE (1)

In practice, select batches of data points

Common heuristic: top-K acquisition
= Select top-K highest single-acquisition scores (i.e., top-K BALD, top-K Variance)

+ Computationally cheap
− Overlooks interactions between points of a batch



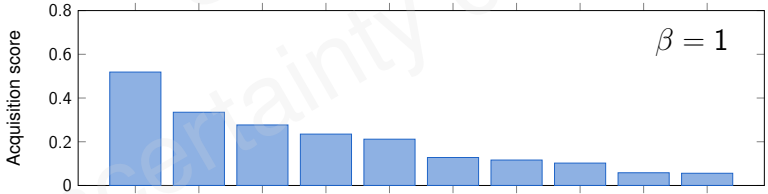
UAntwerp Uncertainty 05/09/2025

PRACTICAL: BATCH-MODE (2)

Simple heuristic: PowerBALD (Kirsch, Farquhar, et al. 2021)
= stochastically select highest single-acquisition scores, following a Power distribution $s(i)^\beta$

+ as cheap as top-K acquisition
+ comparable performance to BatchBALD (infeasible) (Kirsch, Van Amersfoort, et al. 2019)

→ Analogously for regression: PowerVariance (Thuy, Loginova, et al. 2024)



UAntwerp Uncertainty 05/09/2025

Practical: Baal Python package

https://github.com/baal-org/baal

UAntwerp Uncertainty 05/09/2025

Conclusion

UAntwerp
Uncertainty 05/09/2025

Conclusion

Motivation

- Shifted and OOD data is everywhere
- Standard NNs tend to be overconfident in such situations

UQ offers tools for NNs to *know when they do not know*

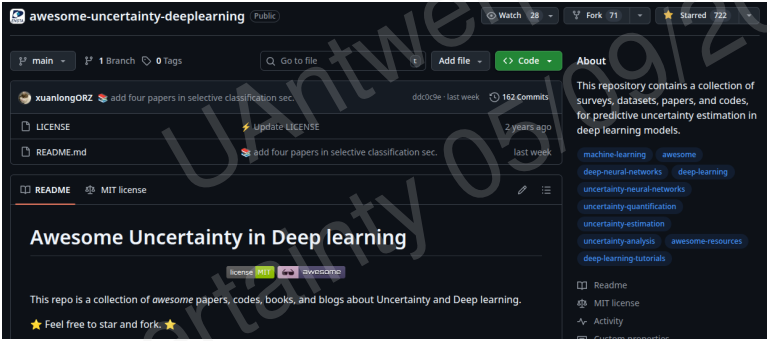
Enabling

- Trustworthy and reliable NNs, ultimately benefiting user adoption
- HITL: selective prediction, active learning

UAntwerp Uncertainty 05/09/2025

LEARN MORE

https://github.com/ENSTA-U2IS-AI/awesome-uncertainty-deeplearning



Uncertainty 05/19/2025

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

Arthur Thuy

PhD Researcher

Ghent University (Belgium)

arthur.thuy@ugent.be

Socials

arthur-thuy

Arthur Thuy

arthurthuy.com

Slides available at https://arthurthuy.com/uantwerp_uncertainty_2025

UAntwerp Uncertainty 05/09/2025

GHENT UNIVERSITY

FWO

FLANDERS MAKE

References I

[1] Burr Settles. “Active learning literature survey”. In: (2009).

[2] Neil Houlsby et al. “Bayesian active learning for classification and preference learning”. In: *arXiv preprint arXiv:1112.5745* (2011).

[3] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[4] Chuan Guo et al. “On calibration of modern neural networks”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1321–1330.

[5] Gao Huang et al. “Snapshot ensembles: Train 1, get m for free”. In: *arXiv preprint arXiv:1704.00109* (2017).

[6] Alex Kendall and Yarin Gal. “What uncertainties do we need in bayesian deep learning for computer vision?” In: *Advances in neural information processing systems* 30 (2017).

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in neural information processing systems* 30 (2017).

UAntwerp Uncertainty 05/09/2025

References II

[8] Stefan Depeweg et al. “Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning”. In: *International conference on machine learning*. PMLR. 2018, pp. 1184–1193.

[9] Paul Bergmann et al. “MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9592–9600.

[10] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. “Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning”. In: *Advances in neural information processing systems* 32 (2019).

[11] Wouter Kool, Herke Van Hoof, and Max Welling. “Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3499–3508.

[12] Yaniv Ovadia et al. “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift”. In: *Advances in neural information processing systems* 32 (2019).

UHAntwerpen Uncertainty 15/09/2025

References III

[13] Parmida Atighehchian, Frédéric Branchaud-Charron, and Alexandre Lacoste. "Bayesian active learning for production, a systematic study and a reusable library". In: *arXiv preprint arXiv:2006.09916* (2020).

[14] Oliver Dürr, Beate Sick, and Elvis Murina. *Probabilistic deep learning: With python, keras and tensorflow probability*. Manning, 2020.

[15] Marton Havasi et al. "Training independent subnetworks for robust prediction". In: *arXiv preprint arXiv:2010.06610* (2020).

[16] Yeming Wen, Dustin Tran, and Jimmy Ba. "Batchensemble: an alternative approach to efficient ensemble and lifelong learning". In: *arXiv preprint arXiv:2002.06715* (2020).

[17] Florian Wenzel et al. "Hyperparameter ensembles for robustness and uncertainty quantification". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6514–6527.

[18] Eyke Hüllermeier and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine Learning* 110.3 (2021), pp. 457–506. DOI: 10.1007/s10994-021-05946-3.

UAntwerp Uncertainty 05/09/2025

References IV

[19] Andreas Kirsch, Sebastian Farquhar, et al. “Stochastic Batch Acquisition: A Simple Baseline for Deep Active Learning”. In: *arXiv preprint arXiv:2106.12059* (2021). DOI: https://doi.org/10.48550/arXiv.2106.12059.

[20] Bálint Mucsányi et al. “Trustworthy machine learning”. In: *arXiv preprint arXiv:2310.08215* (2023).

[21] Arthur Thuy and Dries F Benoit. “Explainability through uncertainty: Trustworthy decision-making with neural networks”. In: *European Journal of Operational Research* 317.2 (2024), pp. 330–340. DOI: 10.1016/j.ejor.2023.09.009.

[22] Arthur Thuy and Dries F Benoit. “Fast and reliable uncertainty quantification with neural network ensembles for industrial image classification”. In: *Annals of Operations Research* (2024), pp. 1–27. DOI: 10.1007/s10479-024-06440-4.

[23] Arthur Thuy, Ekaterina Loginova, and Dries F Benoit. “Active Learning to Guide Labeling Efforts for Question Difficulty Estimation”. In: *arXiv preprint arXiv:2409.09258* (2024). DOI: 10.48550/arXiv.2409.09258.

Uploaded on 05/09/2025