FREE KIT / FOR RESEARCHERS WHO USE DEEP LEARNING

Uncertainty quantification, explained for researchers.

Accuracy tells you how often a model is right. Uncertainty tells you which answers to trust. This kit is the short version: the words you need, the main methods and their cost, how to test them, and what to write in your paper.



Dr. Ilhame Ait Lbachir

Computer science professor and AI researcher. Uncertainty quantification for breast cancer detection in mammography.

50+ peer reviews
Published in IEEE and Springer

INSIDE THIS KIT

01 The words you need
02 Two questions not to mix
03 Choose a method
04 Test it before you claim it
05 What to write in your paper
06 Questions reviewers ask
07 Further reading

> A good uncertainty estimate is high when the model is likely to be wrong, and low when it is likely to be right. Everything in this kit comes back to that one test.

01 The words you need

Uncertainty quantification	Methods that attach a measure of doubt to each prediction: a score, an interval, or a set of possible labels.
Aleatoric uncertainty	Noise in the data itself, such as an ambiguous image. More data of the same kind does not remove it.
Epistemic uncertainty	What the model has not learned. High on unfamiliar inputs. It can shrink with more or better data.
Calibration	Whether confidence matches reality. Of all cases predicted with 80 percent confidence, about 80 percent should be correct.
Reliability diagram	A plot of observed accuracy against confidence. A calibrated model follows the diagonal.
Expected calibration error	ECE. The average gap between confidence and accuracy across confidence bins, weighted by bin size. Lower is better.
Selective prediction	The model may abstain on cases it is unsure about and pass them to a person.
Coverage	The share of cases the model answers. Always report it next to the accuracy on those cases.
Confidence interval	The range around a metric such as AUC that reflects the size of your test set. It is about your evaluation, not about one prediction.

02 Two questions not to mix

UNCERTAINTY OF A PREDICTION

How sure is the model about this case?

Estimated by the methods on the next page. Used to flag, defer, or double check single cases.

UNCERTAINTY OF YOUR RESULTS

How sure are you about your reported metric?

A confidence interval, for example from the bootstrap. Resample patients, not images, when one patient has several images.

A paper needs both. An AUC of 0.91 on a small test set means little without its interval.

03 Choose a method

METHOD	WHAT IT GIVES	COST	USE IT WHEN
Softmax confidence	A baseline confidence score	None	Always. It is the baseline every other method must beat.
Temperature scaling	Better calibrated probabilities	One number, fitted on validation data	Probabilities will be read as risk. Accuracy does not change.
Monte Carlo dropout	Epistemic uncertainty	Many forward passes per input	Your network already uses dropout.
Deep ensembles	Epistemic uncertainty from disagreement	Several trained models	You can afford the training. Often a strong reference.
Noise output	Aleatoric uncertainty	An extra output and a changed loss	Your data are noisy or ambiguous.
Conformal prediction	Sets of labels with a coverage rate you choose	A held out calibration set	You need a guarantee, such as 90 percent coverage on average.

A sensible start: report softmax confidence, add temperature scaling, then compare with an ensemble or MC dropout if your budget allows. Conformal guarantees assume new data come from the same source as the calibration data.

04 Test it before you claim it

- [ ] **Calibration is shown, not just stated.** A reliability diagram plus ECE, with the number and type of bins written down.
- [ ] **Uncertainty separates errors from correct answers.** For example the AUROC for detecting misclassified cases.
- [ ] **The risk coverage curve is reported.** Error on the answered cases as the model abstains more. Not one hand picked threshold.
- [ ] **It is tested on new data.** Another hospital, scanner, or site. Epistemic uncertainty should rise.
- [ ] **Proper scores sit next to accuracy.** Negative log likelihood or the Brier score, plus accuracy or AUC.
- [ ] **Nothing was tuned on the test set.** Temperature, thresholds, and the conformal calibration all use separate data.

05 What to write in your paper

Uncertainty was estimated with `[method]`, using `[number of passes or models]`. The temperature was fitted on the validation set. On the test set we report ECE with `[number]` equal width bins, the Brier score, and the risk coverage curve, with 95 percent bootstrap confidence intervals computed over `[patients]`.

- [ ] **Every accuracy comes with its coverage.** 98 percent at 40 percent coverage is a different claim from 98 percent on all cases.
- [ ] **Who reads the deferred cases is described.** And how much extra work that creates.
- [ ] **Results are checked by subgroup.** Abstention should not pile up on one group of patients.

06 Questions reviewers ask

- [ ] **Is the uncertainty better than the softmax baseline?** Show the comparison.
- [ ] **Does it hold on external data?** Or is it only tested where the model was trained?
- [ ] **Is a well calibrated model also a useful one?** A model that predicts the disease rate for everyone is calibrated and tells you nothing.

07 Further reading

Guo et al., On calibration of modern neural networks, ICML 2017.

Gal and Ghahramani, Dropout as a Bayesian approximation, ICML 2016.

Lakshminarayanan et al., Simple and scalable predictive uncertainty estimation using deep ensembles, arXiv:1612.01474.

Kendall and Gal, What uncertainties do we need in Bayesian deep learning for computer vision?, arXiv:1703.04977.

Angelopoulos and Bates, A gentle introduction to conformal prediction and distribution free uncertainty quantification, arXiv:2107.07511.

Abdar et al., A review of uncertainty quantification in deep learning, Information Fusion, 2021.

A model earns trust by knowing when it might be wrong. *Measure that, test it, and report it with the same care as accuracy.*

The full guide, with links to calibration, the two kinds of uncertainty, and selective prediction:

`calibratedresearcher.com/field-notes/uncertainty-quantification-explained`