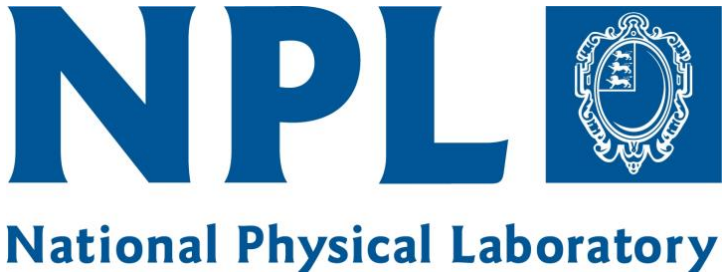
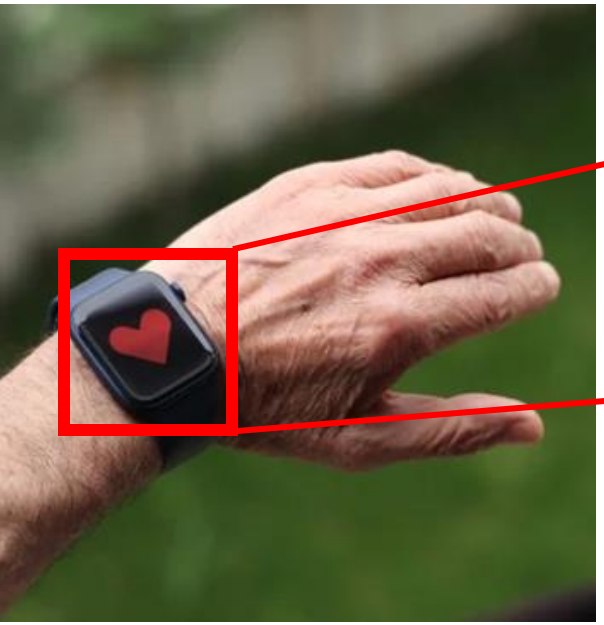
Evaluating uncertainty calibration in deep learning

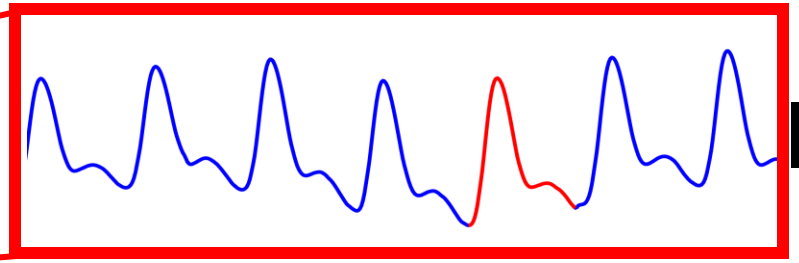
Vivek Desai, Ciaran Bench



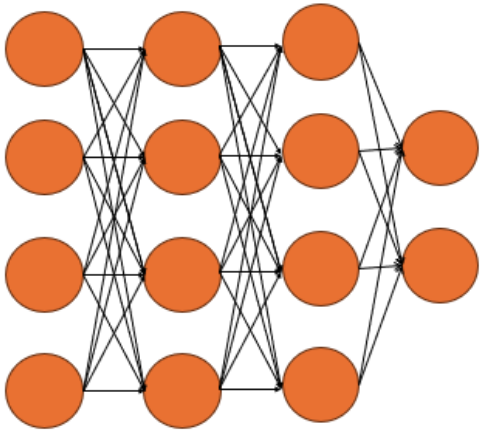
Case study: wearable photoplethysmography (PPG)

Deep learning + wearable Photoplethysmography (PPG) can enable home monitoring of blood pressure or other physiological parameters





Continuous daily monitoring



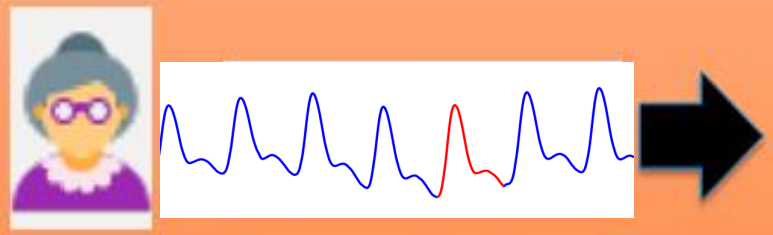
Blood pressure

120 mmHg

Risk of poor performance on unseen data slows commercial model deployment

Training patient population

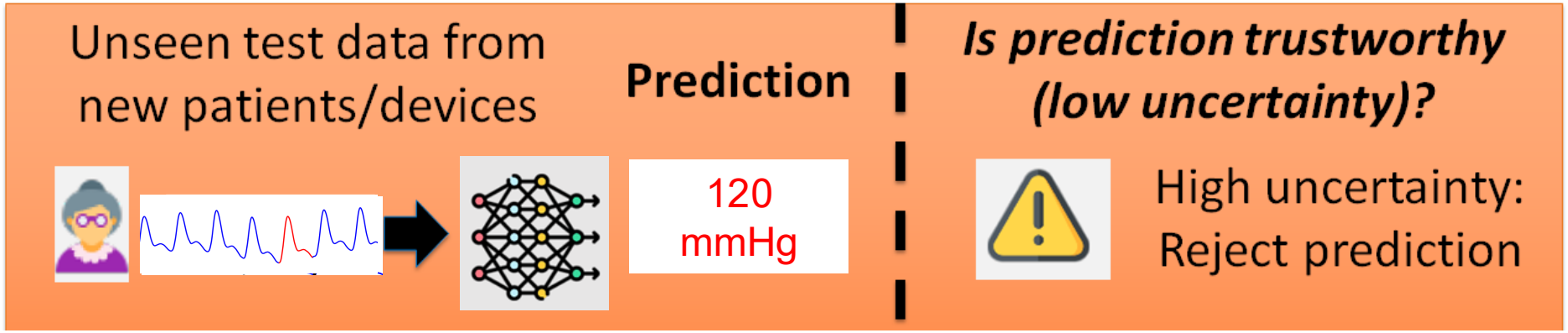
Unseen test data from new patients/devices

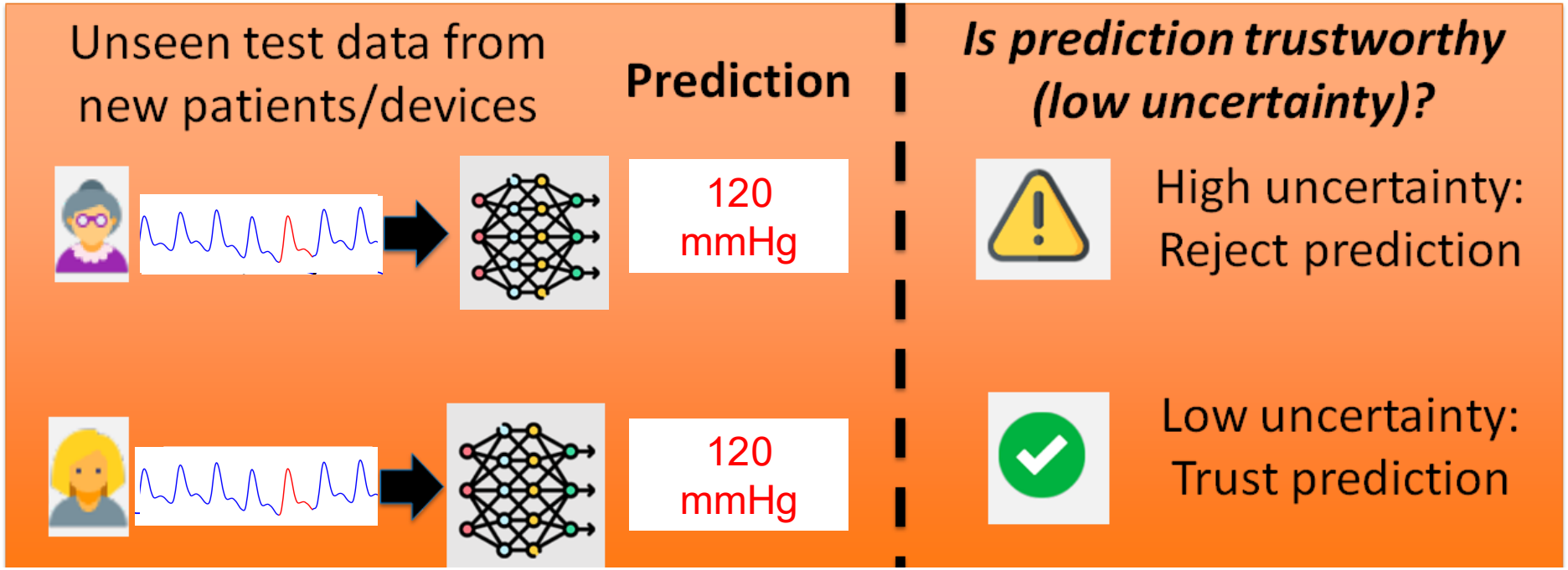


120 mmHg

How can we trust model's prediction for a new patient?

Uncertainty quantification has potential to streamline model development

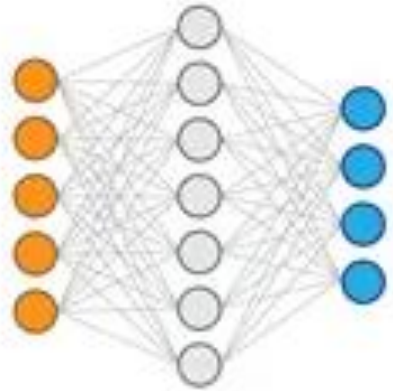

Unseen test data from new patients/devices
Prediction
120 mmHg
Is prediction trustworthy (low uncertainty)?
High uncertainty: Reject prediction


Unseen test data from new patients/devices
Prediction
Is prediction trustworthy (low uncertainty)?
120 mmHg
High uncertainty: Reject prediction
120 mmHg
Low uncertainty: Trust prediction

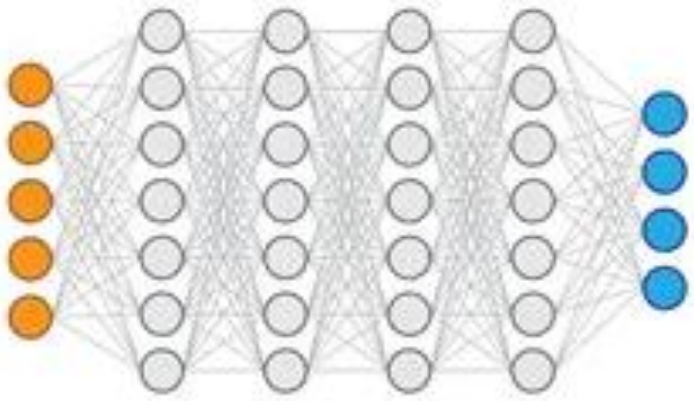
UQ in deep learning

→ Classical UQ techniques (e.g. Bayesian optimisation) don’t scale well with number of parameters

Smaller Model



Larger Model

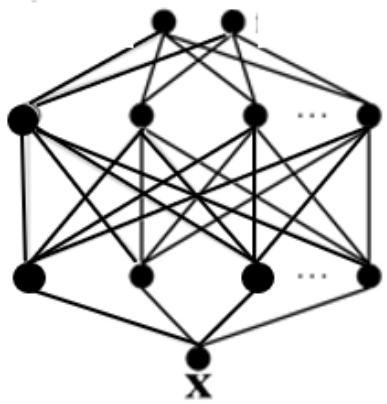


Scalable UQ

- Uncertainty in deep learning = variance of multiple answers you get for a given input
- Applying UQ? -> how to get a model to give you multiple answers instead of one?

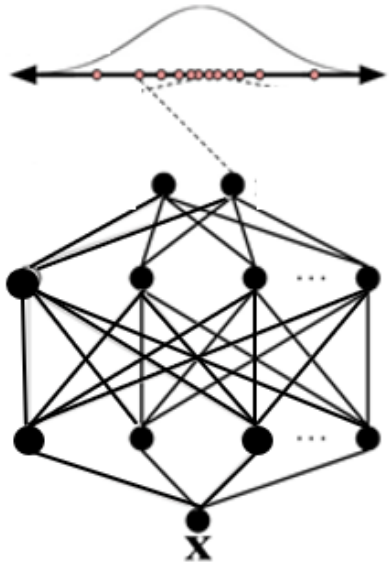
Typical Network

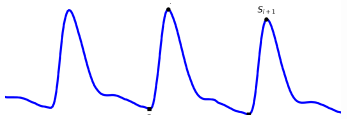
Singular outputs, ***no*** uncertainty

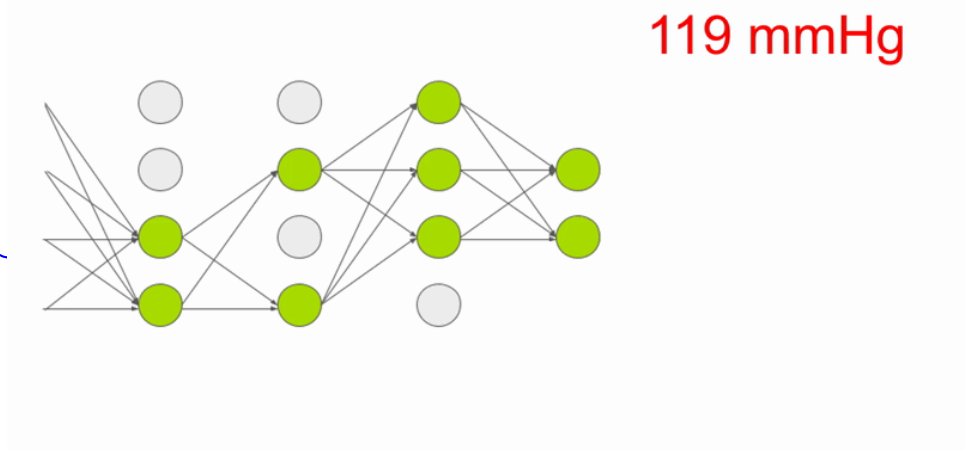


Ideal Network

Distribution of outputs, ***can*** compute uncertainty

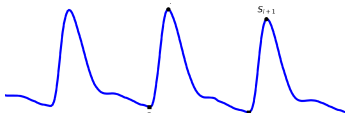


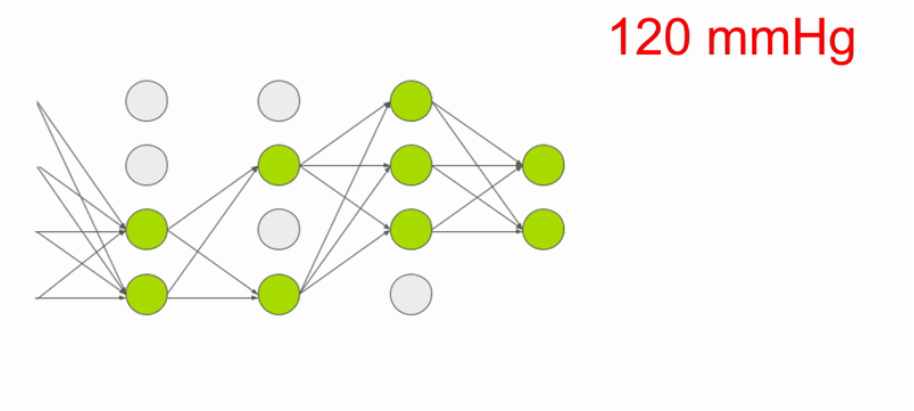




Large difference
across multiple BP predictions indicates high uncertainty

Reject Prediction





Small difference
across multiple BP predictions indicates low uncertainty

Accept Prediction

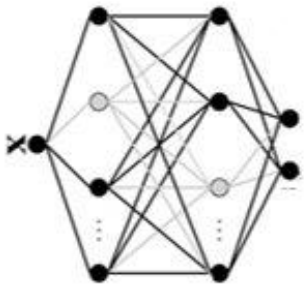
How do we trust our uncertainties?

CALIBRATION

Practically useful uncertainty quantification

1. Uncertainty should encode underlying doubt in a prediction

- Various ways to define this….



$Prediction = \mu$

$Uncertainty = \sigma^2$

y is ground truth

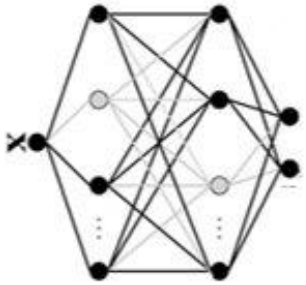
$|\mu - y| \approx \sigma$

- **Calibration**: Uncertainty magnitude σ should correspond with prediction error magnitude $|\mu - y|$

Practically useful uncertainty quantification

1. Uncertainty should encode underlying doubt in a prediction

- Various ways to define this….



$Prediction = \mu$

$Uncertainty = \sigma^2$

y is ground truth

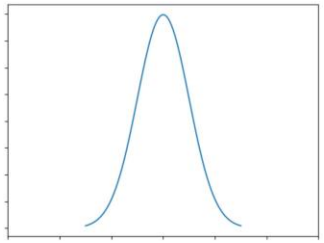
$|\mu - y| \approx \boldsymbol{\sigma}$

- **Calibration**: Uncertainty magnitude σ should correspond with prediction error magnitude $|\mu - y|$

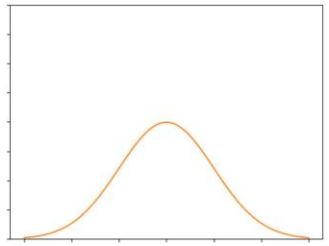
2. Calibration caters to ultimate use case of the model

- E.g. if few predictions are used to make a decision, predicted uncertainties should be calibrated at the level of individual predictions
- **Sharpness**: Width of prediction interval should be small, and centred at GT

Sharp



Not Sharp



Key concepts in uncertainty evaluation

Scale of calibration

Global calibration

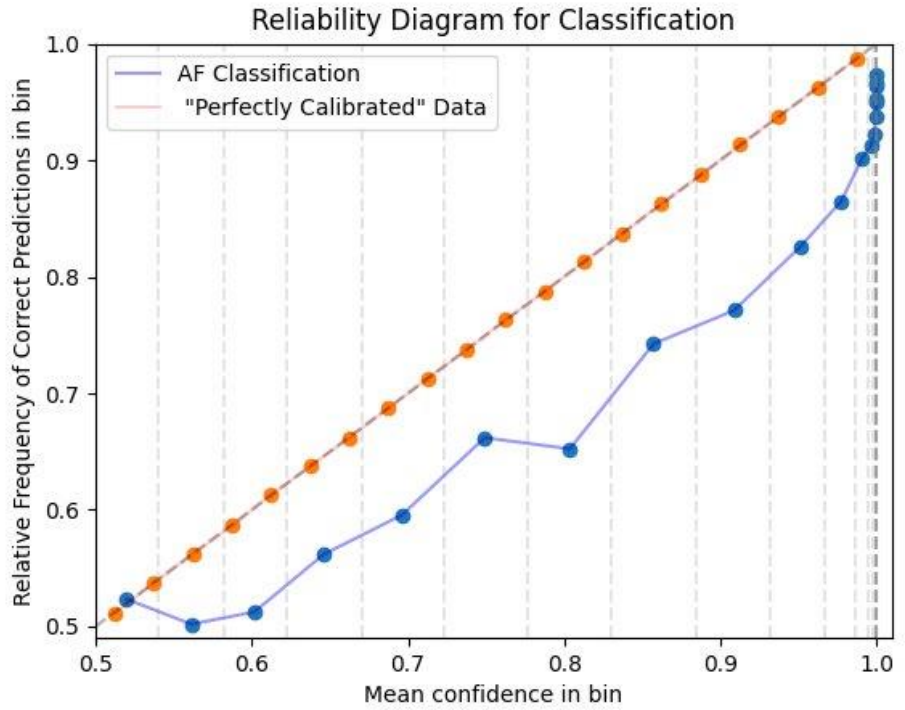
- Average uncertainty of **all** test examples ≈ average prediction error
- Scalar value…

Global calibration

- Average uncertainty of **all** test examples ≈ average prediction error
- Scalar value…

Local calibration

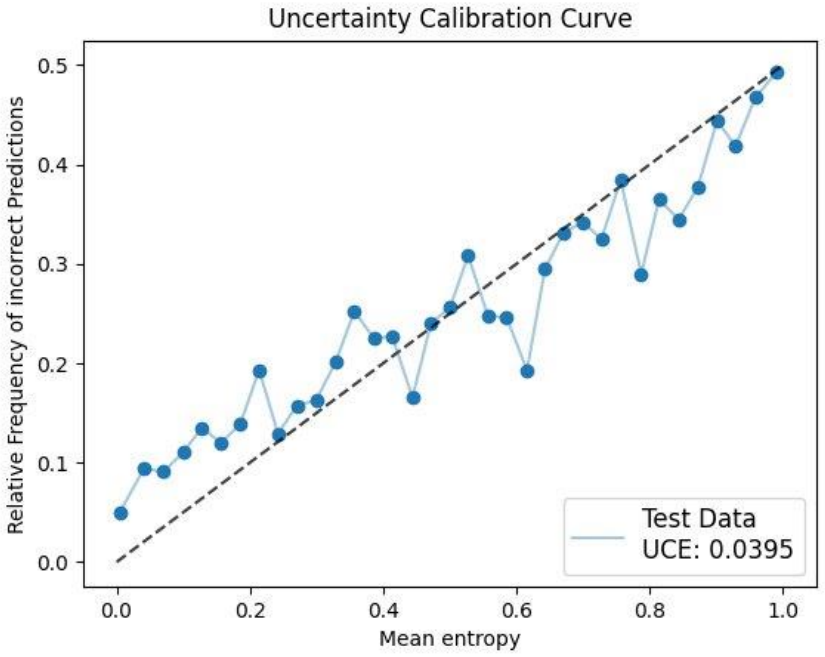
- Bin test predictions by uncertainty magnitude
- Average uncertainty in bin ≈ average prediction error in bin
- Reliability diagrams



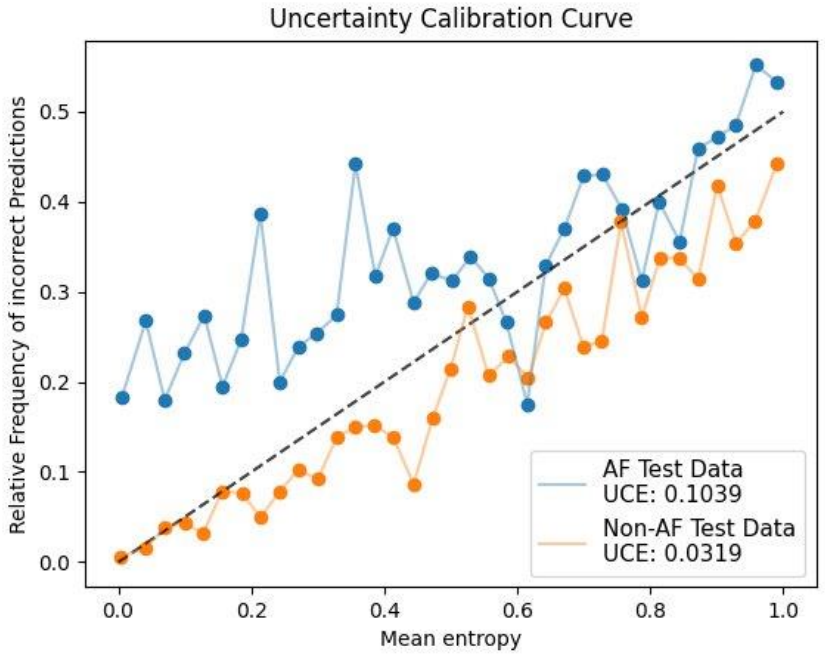
Adaptivity

- Expanding idea of local calibration
  - Bin predictions by something ***other than*** uncertainty magnitude
  - Average uncertainty *in bin* ≈ average prediction error *in bin*
- E.g. per class uncertainty calibration – often overlooked

Non-adaptive



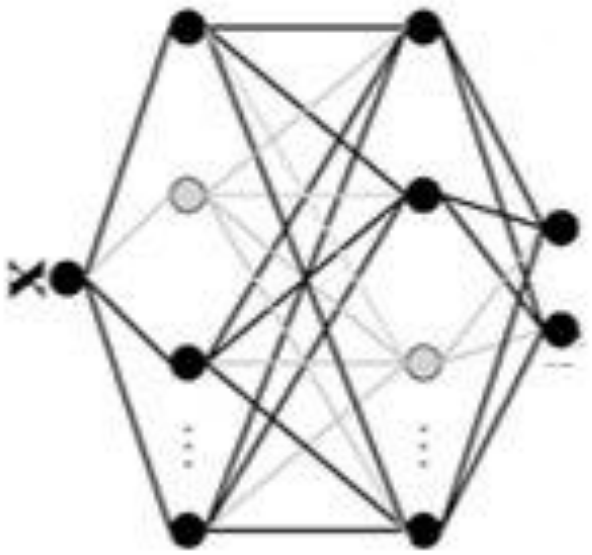
Adaptive



Regression

What is our uncertainty for regression tasks?

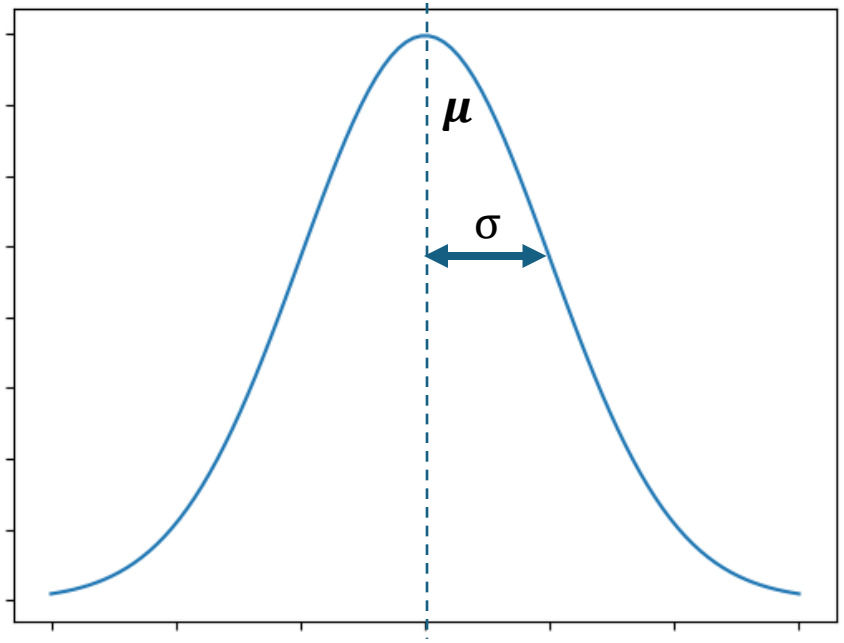
Model predicts a distribution



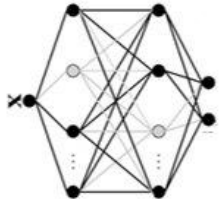
$Prediction = \mu$

$Uncertainty = \sigma^2$

Parameterise Gaussian



Variance/coverage based calibration metrics



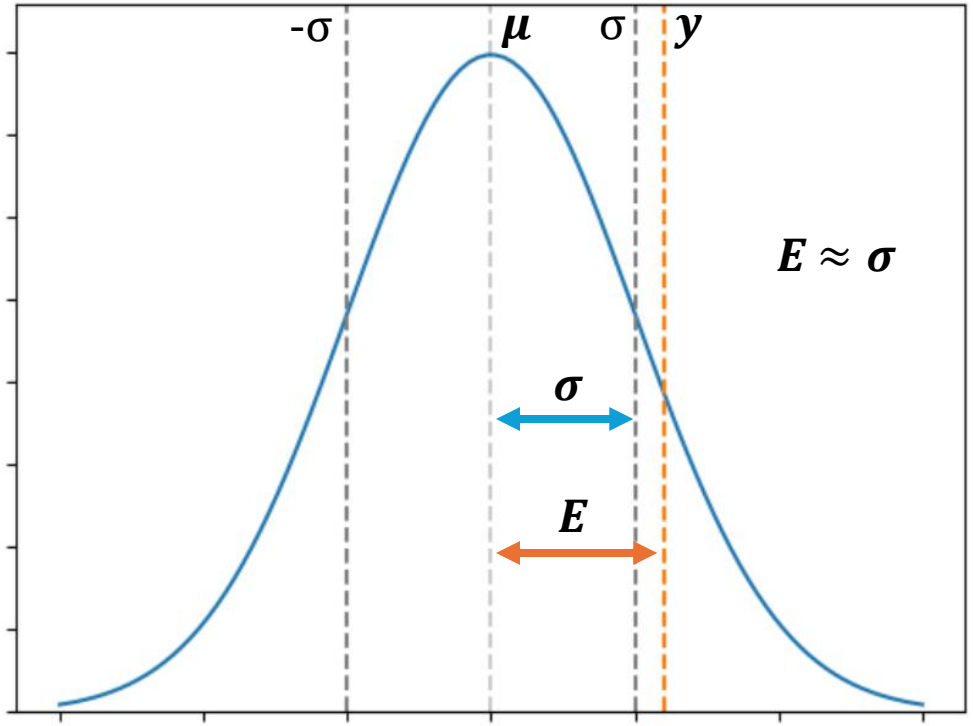
$\mu = Prediction$

$\sigma^2 = Uncertainty$

$y = Ground\ truth$

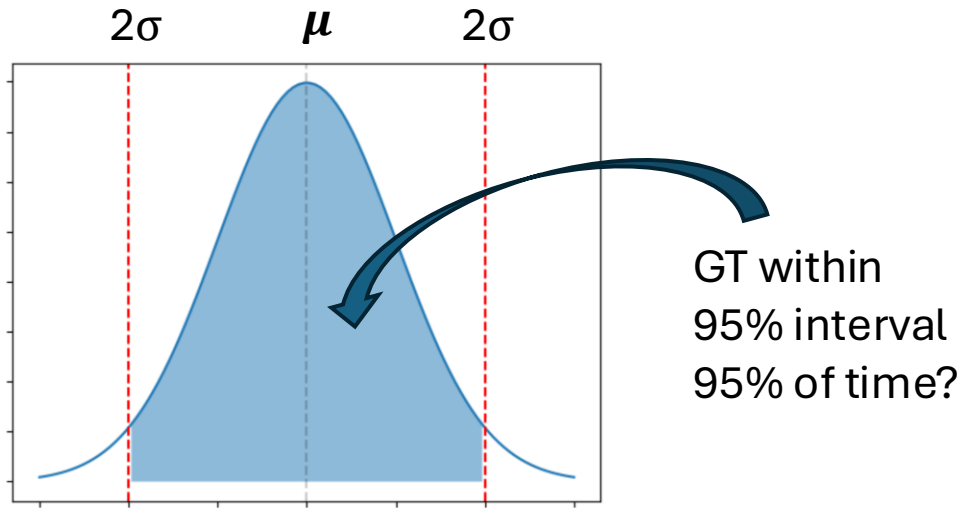
Variance

- Magnitude of uncertainty ≈ magnitude of prediction error



Coverage

- Define Gaussian/distribution based on prediction + uncertainty
- Do GTs in test set fall within X% confidence interval X% of the time?

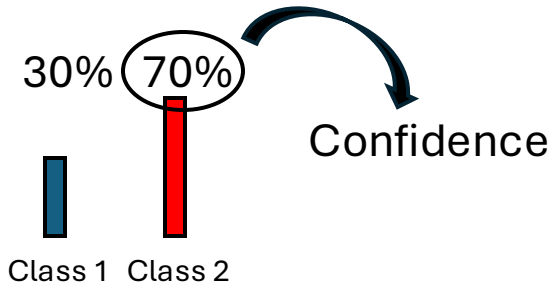


Classification

Expressions of uncertainty

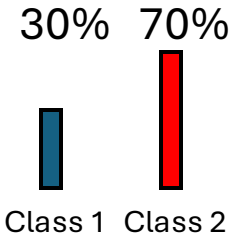
Class probabilities

- Confidence
- Some metrics that cater to this expression:
  - **Expected Calibration Error** (ECE)
  - Smooth Expected Calibration Error (smECE)
  - Adaptive Calibration Error (ACE)



Entropy

- "Width" of distribution

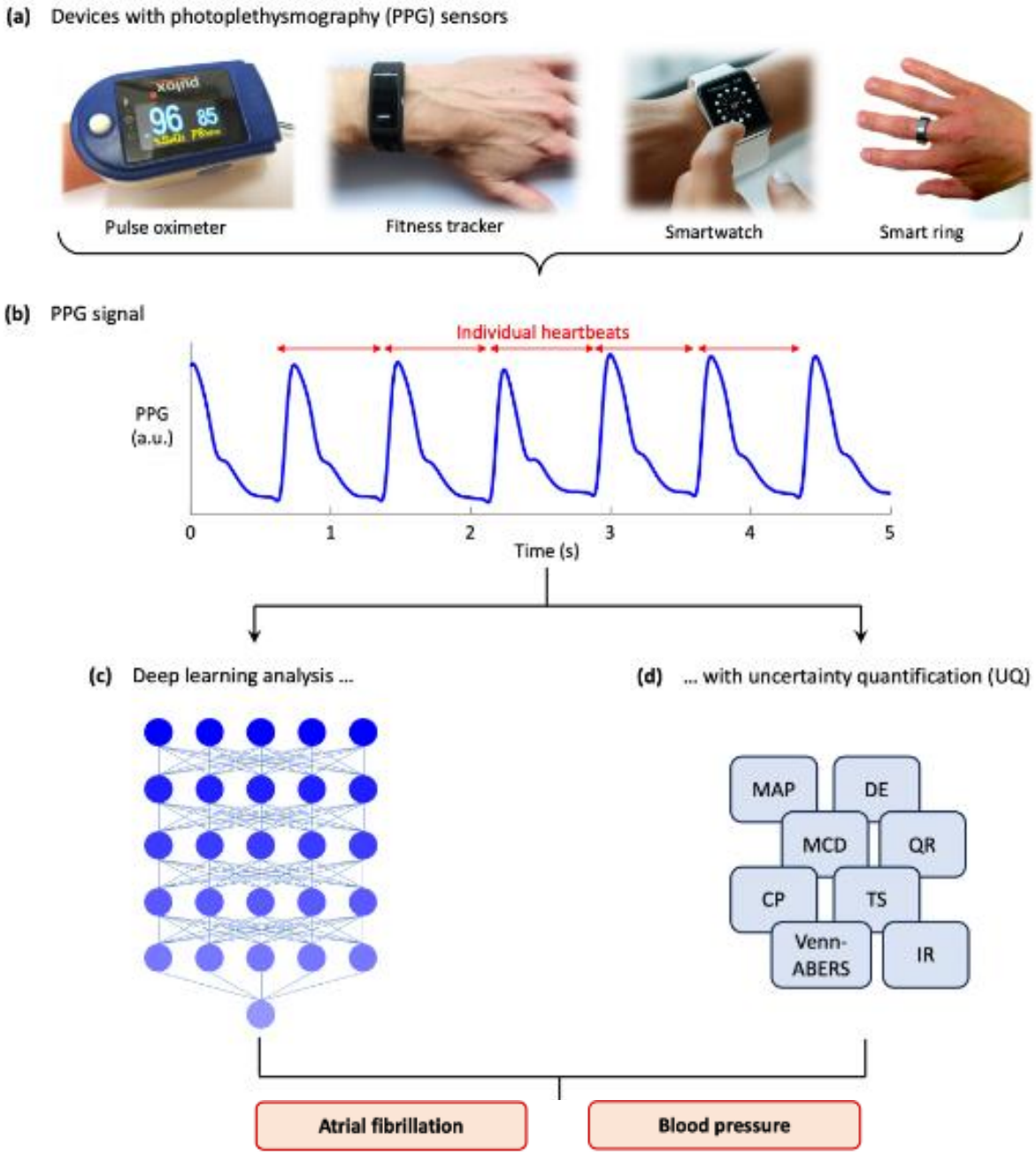


$$H = -\sum_{i=1}^{2} p_i \log_2 p_i$$

$$H \approx 0.88$$

- Metrics:
  - Uncertainty Calibration Error (UCE)
  - "Better" for multi-class case

Insight from case study in PPG



- Effectiveness of UQ methods on real large scale problems
- Think about uncertainty evaluation in practical decision making.

Optimal UQ depends on expression of uncertainty/metric/parameterisation

UQ type	Performance Metrics						
	ECE↓	ACE↓	smECE↓	UCE↓	VCE↓	NLL↓	AUC↑
MAP	0.098	0.100	0.086	0.062	0.306	1.169	0.84
MCD	0.087	0.084	0.076	0.055	0.269	1.053	0.85
DE	0.074	0.076	0.064	**0.054**	0.234	0.973	**0.86**
MCD+TS	0.078	0.078	0.073	0.158	0.110	0.691	0.85
MCD+IR	0.055	0.075	**0.044**	0.133	0.095	0.712	0.85
DE+TS	0.075	0.084	0.070	0.149	0.102	0.692	**0.86**
DE+IR	**0.050**	**0.059**	**0.044**	0.129	**0.090**	**0.682**	0.85
Venn-ABERS	0.055	0.063	0.048	0.143	0.103	0.691	0.84

Confidence based calibration metrics
Optimal UQ method: Deep Ensembles + Isotonic regression

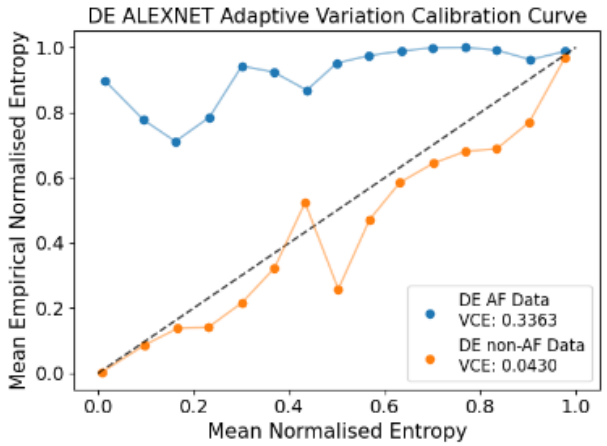
Optimal UQ depends on expression of uncertainty/metric/parameterisation

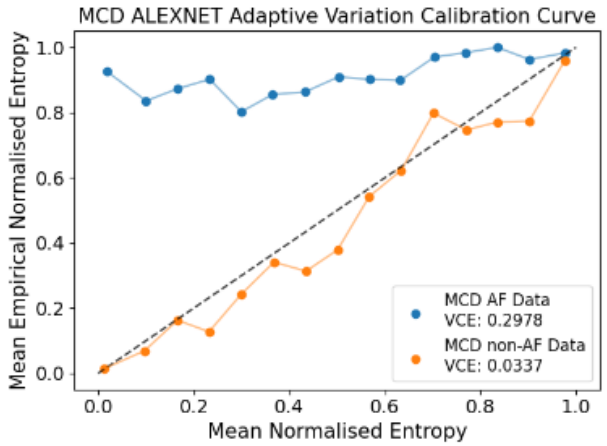
UQ type	Performance Metrics						
	ECE↓	ACE↓	smECE↓	UCE↓	VCE↓	NLL↓	AUC↑
MAP	0.098	0.100	0.086	0.062	0.306	1.169	0.84
MCD	0.087	0.084	0.076	0.055	0.269	1.053	0.85
DE	0.074	0.076	0.064	**0.054**	0.234	0.973	**0.86**
MCD+TS	0.078	0.078	0.073	0.158	0.110	0.691	0.85
MCD+IR	0.055	0.075	**0.044**	0.133	0.095	0.712	0.85
DE+TS	0.075	0.084	0.070	0.149	0.102	0.692	**0.86**
DE+IR	**0.050**	**0.059**	**0.044**	0.129	**0.090**	**0.682**	0.85
Venn-ABERS	0.055	0.063	0.048	0.143	0.103	0.691	0.84

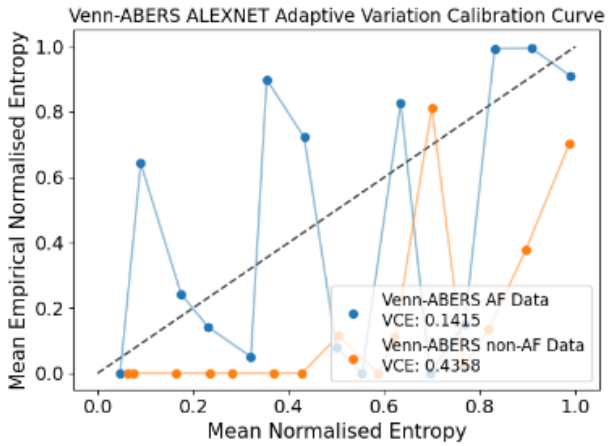
Confidence based calibration metrics
Optimal UQ method: Deep Ensembles + Isotonic regression

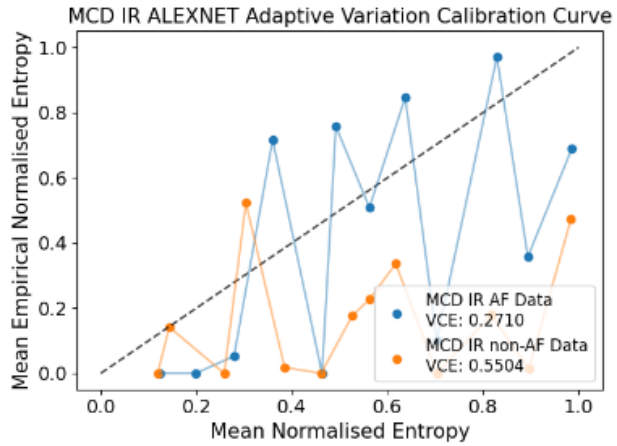
Entropy based calibration metric
Optimal UQ method: Deep Ensembles + Isotonic regression

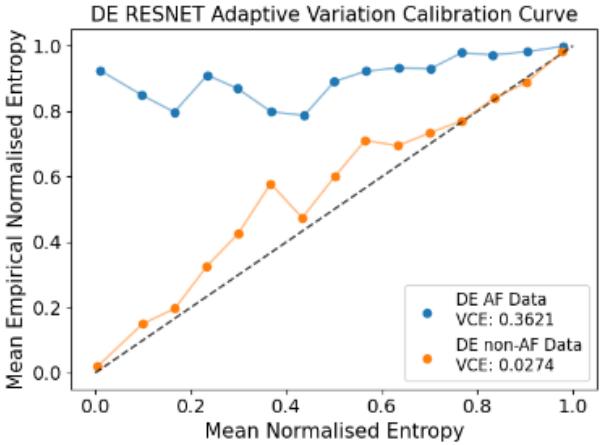
UQ methods do not optimise for adaptivity

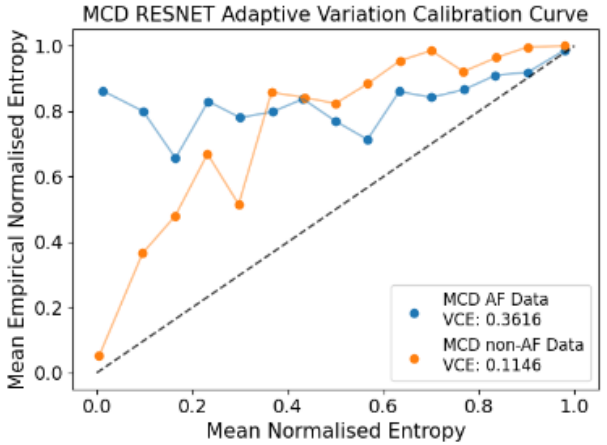


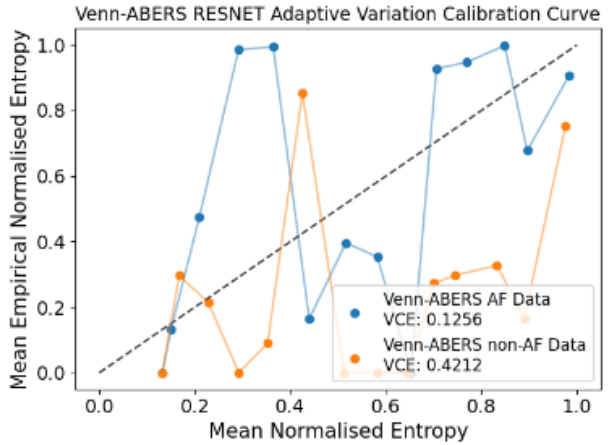


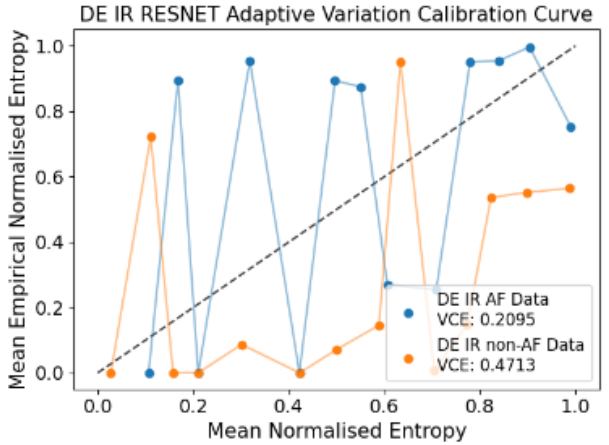




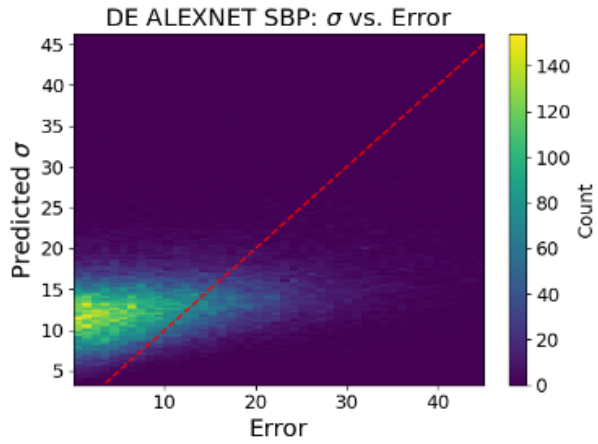


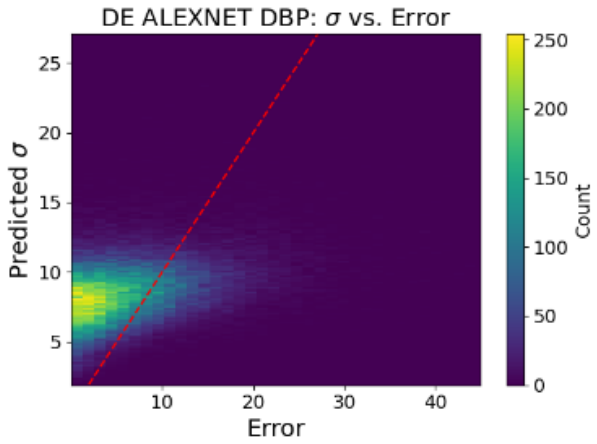


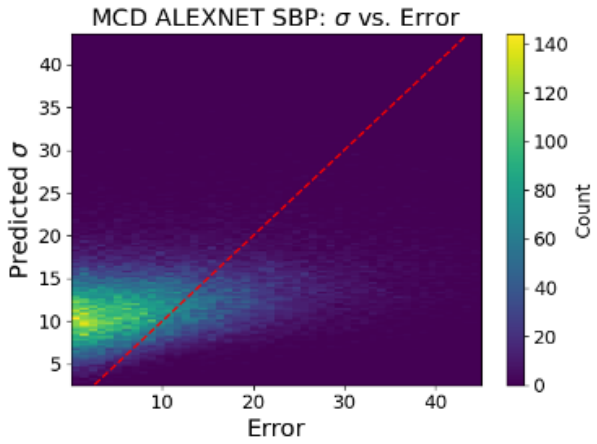


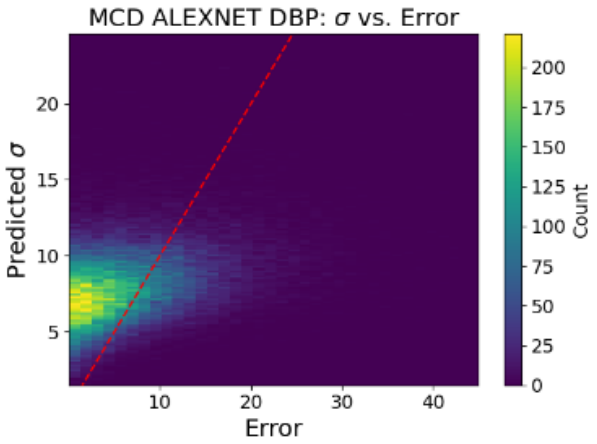


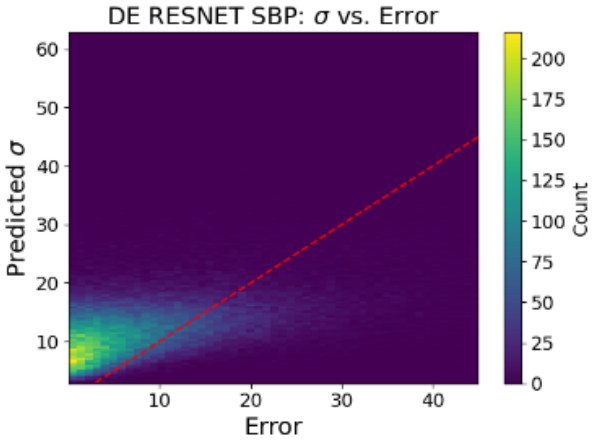
UQ methods do not optimise for small scale reliability

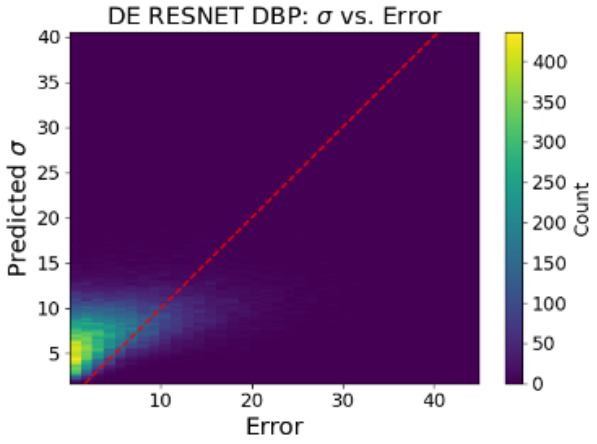


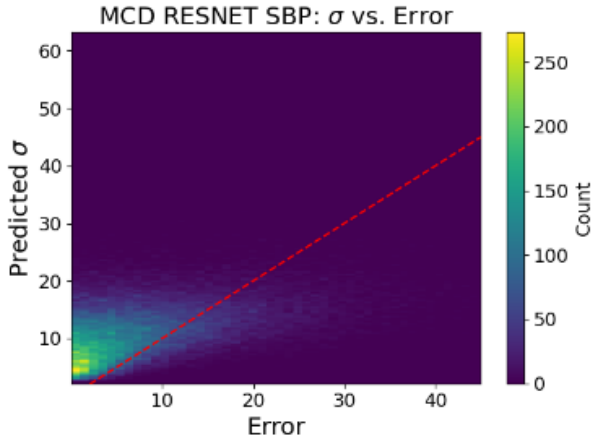


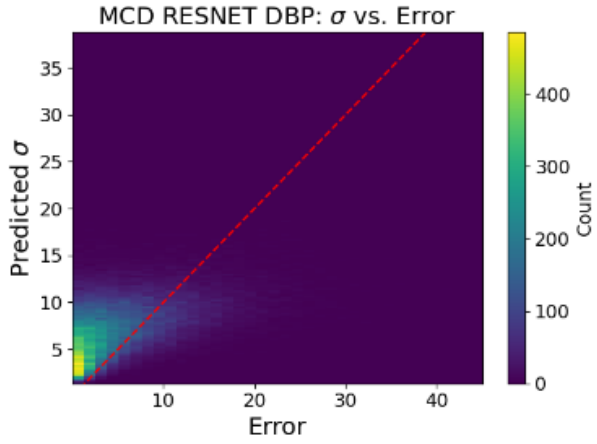












Conclusion

- Practically useful uncertainties in medicine should exhibit
  - Local calibration
  - Adaptivity
  - Small scale-reliability (few measurements used per decision)
- Concepts are not new but perhaps need more regular use/reference in DL.
- Next wave of UQ methods should incorporate more of these aspects
  - Current UQ techniques do not account for them.

Related Outputs

- Publications:
  - Bench, Ciaran, et al. "A systematic evaluation of uncertainty quantification techniques in deep learning: a case study in photoplethysmography signal analysis." *Machine Learning: Health* 2.1 (2026): 015011.
  - Bench, Ciaran, et al. "Uncertainty quantification with approximate variational learning for wearable photoplethysmography prediction tasks." *Machine Learning: Health* 1.1 (2025): 015013.

- Preprints:
  - Thompson, Andrew, and Vivek Desai. "Extending confidence calibration to generalised measures of variation." *arXiv preprint arXiv:2602.12975* (2026).
  - Bench, Ciaran. "Uncertainty quantification in deep learning is unsatisfactory for clinical applications and complex decision making." *Authorea Preprints* (2026).