
Uncertainty and reproducibility in a practical machine learning model

Max Taylor, Dosimetry Services Department, UKHSA

February 19th, 2026

EURADOS Winter School on Artificial Intelligence and Machine Learning
in ionizing radiation dosimetry

Helsinki, Finland



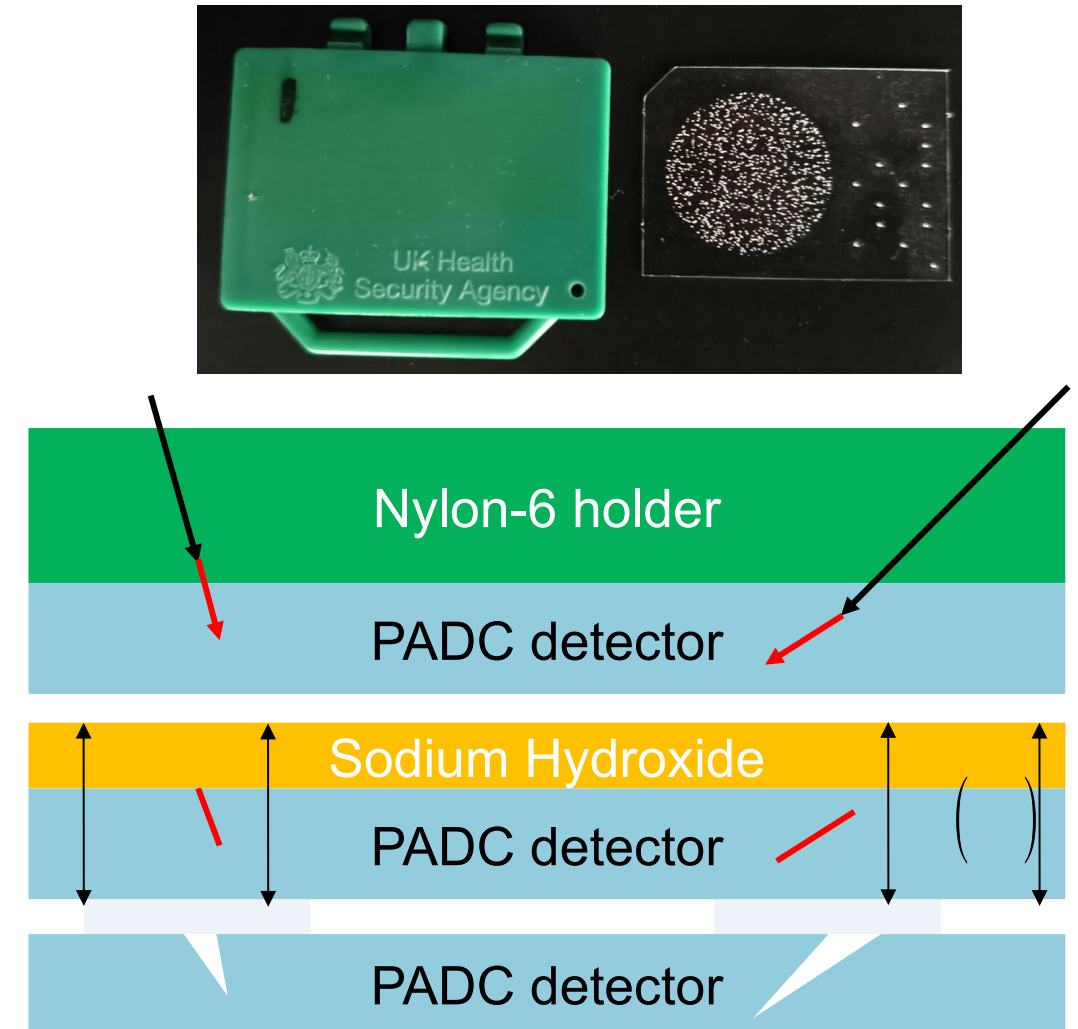
UK Health
Security
Agency

Introduction

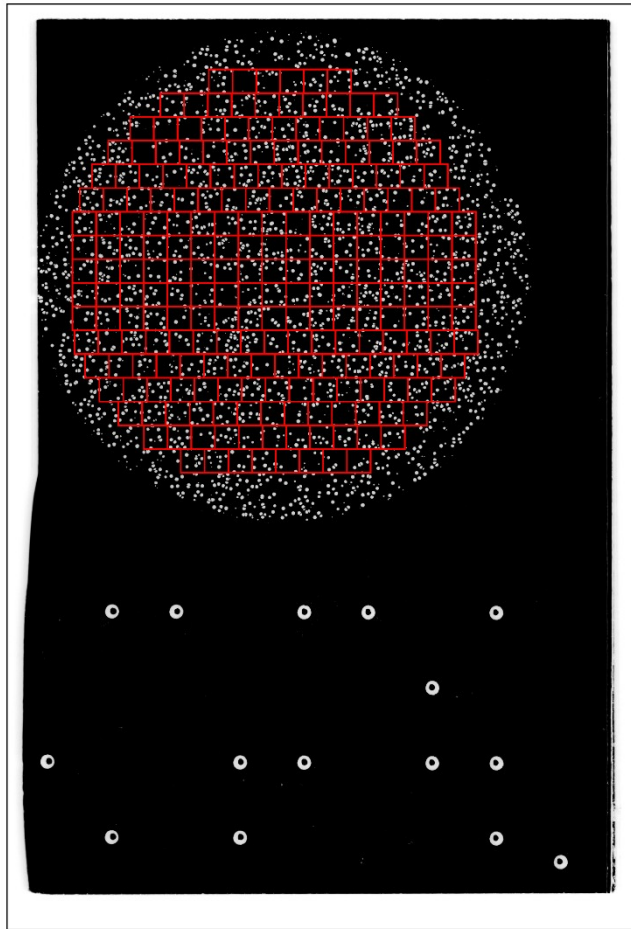
- **Uncertainty estimation** is required when reporting any measured value.
- This applies to **machine learning (ML) models** as much as any other instrument.
- Here, I will discuss uncertainty estimation in a homemade **regression neural network**, and the problems that can arise.

Context: Neutron dosimetry (and beyond!)

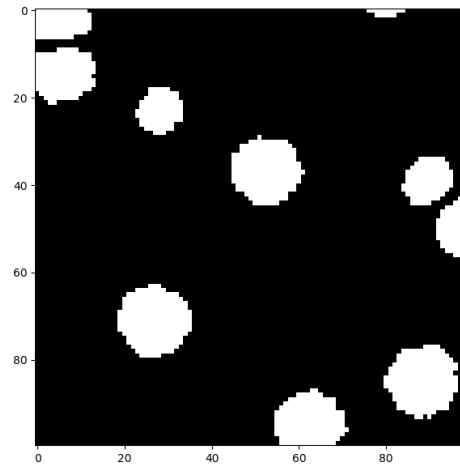
- PADC/CR-39 commonly used for neutron dosimetry.
- Tracks are **(electro)chemically etched** and counted to assess dose.
- ML methods currently being explored for **improved track analysis**.



Deep neural network for pit density estimates



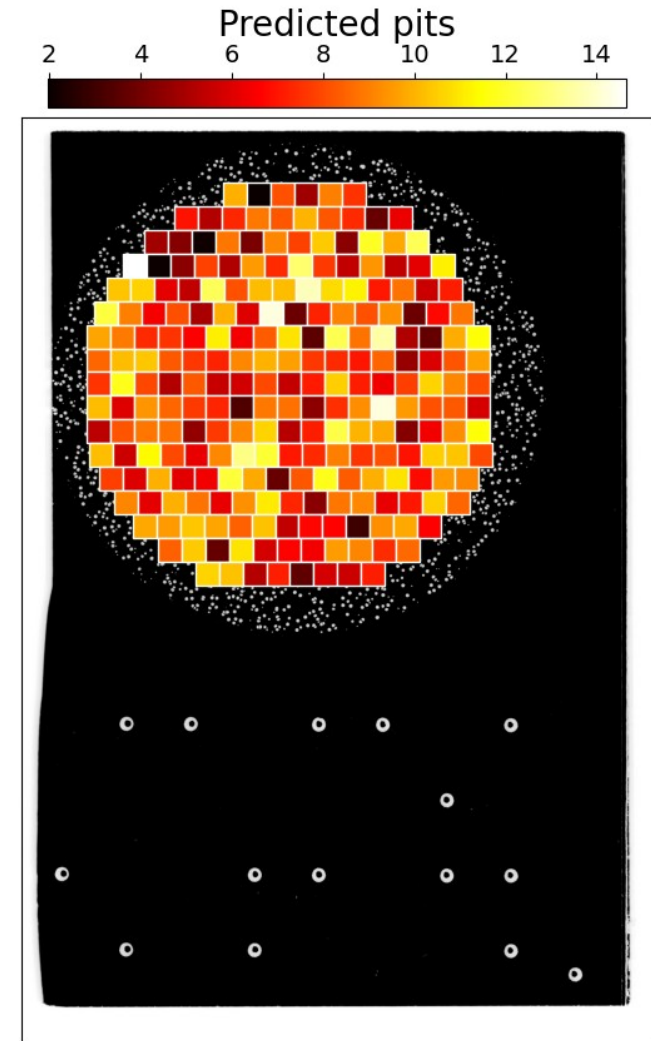
Split dosemeter
images into subareas



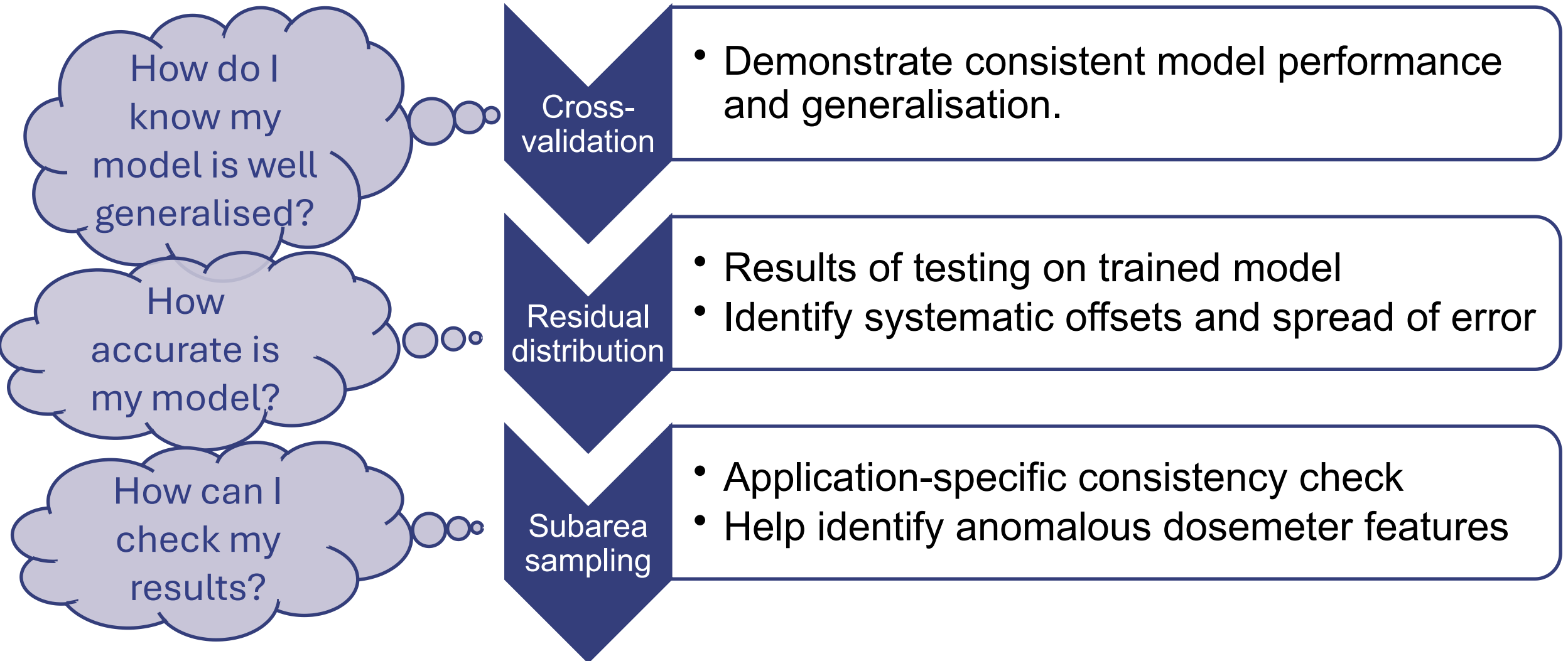
Machine Learning
model predicts pit count



Recombine
subareas

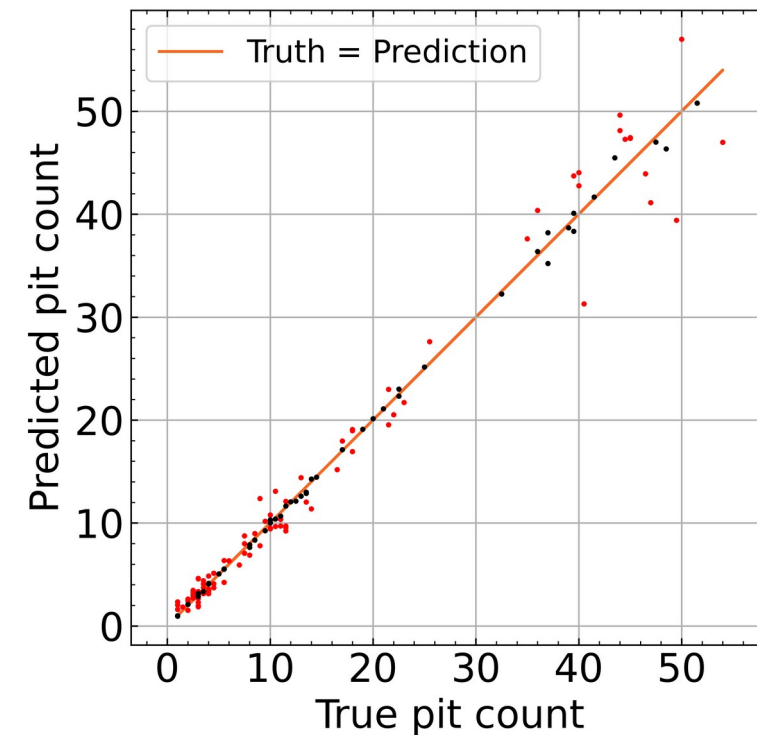
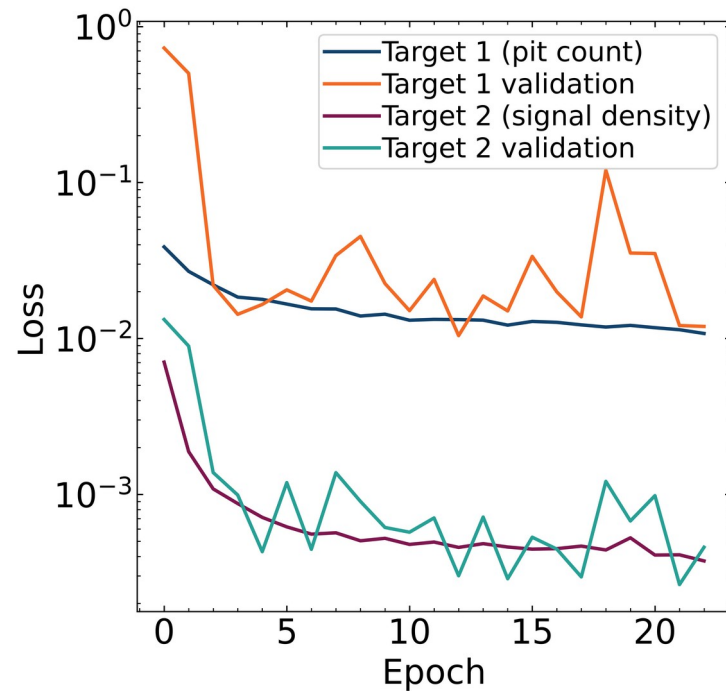


Building a neural network with uncertainty in mind



Standard ML model characterisation and testing

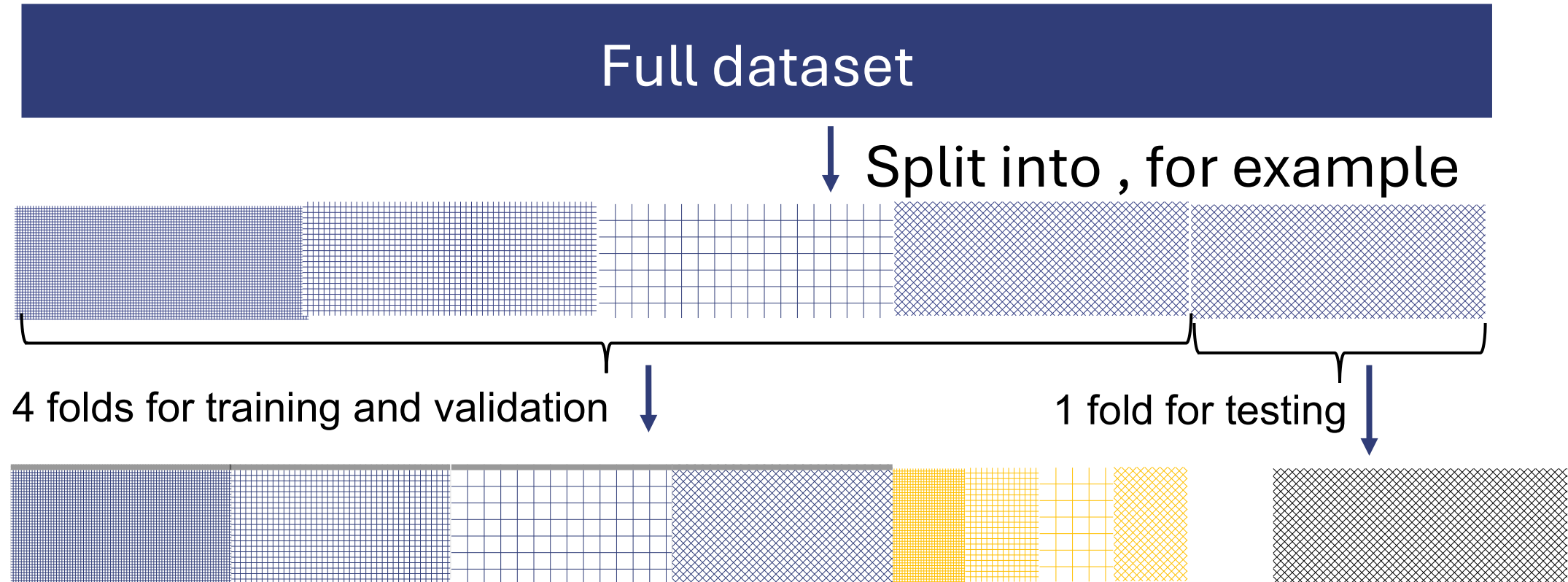
- Loss function trends during training.
- Early indication of fitting quality.
- Accuracy of predictions on independent test data.
- Determine model error statistics.



Checking model consistency

- Models can show variation originating from the specific subsets of data chosen for training, validation, and testing.
- Particularly an issue if working with **small datasets**.
- One way to test this is with **K-fold cross validation**.
- Standard model testing is performed on **different combinations** of the dataset to determine consistency

K-fold cross validation



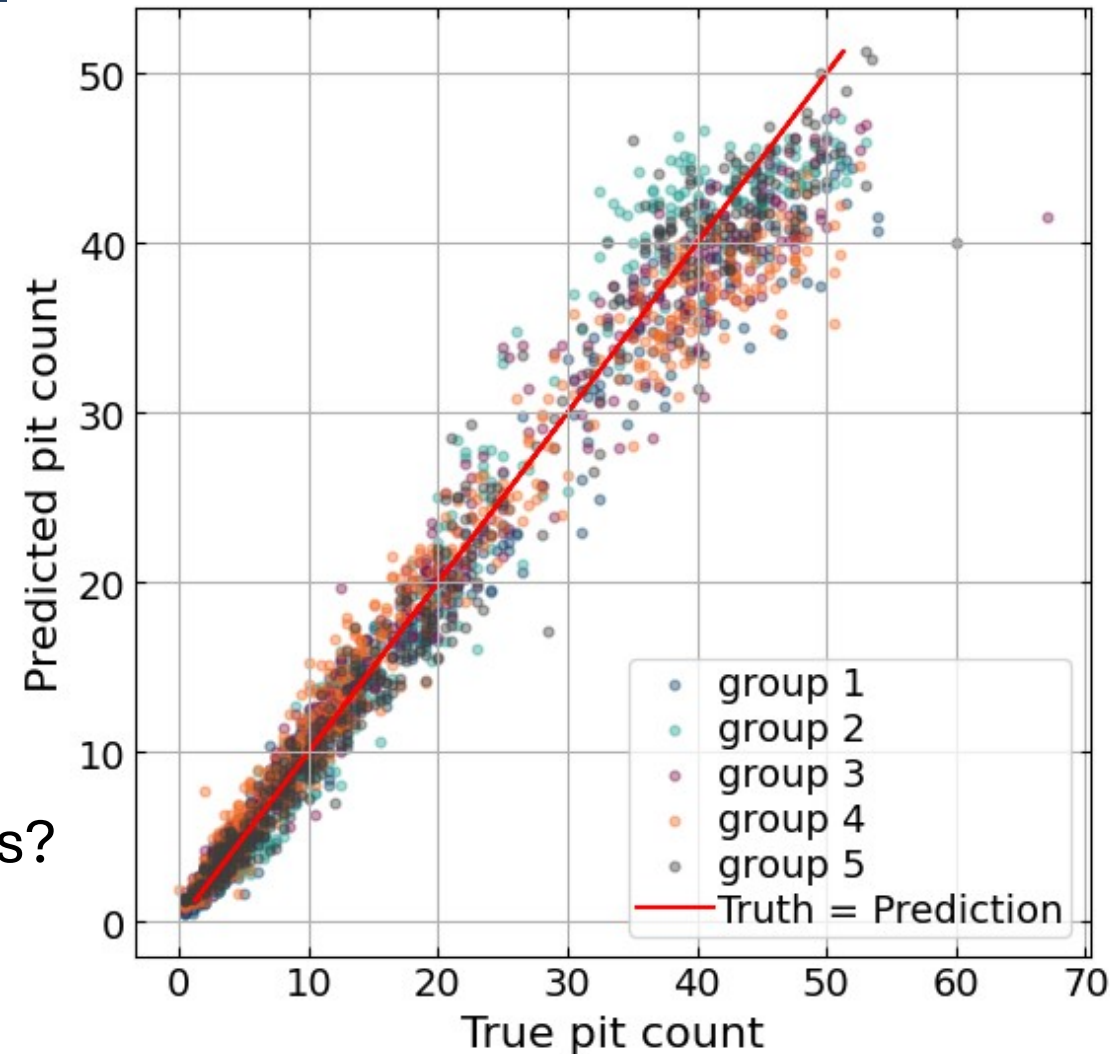
Gives different training and testing sets using different fold combinations

Comparing cross validation groups

- Primarily interested in the error from the test dataset.
- Inconsistency between groups increasingly visible at high pit count (more spread).
- Can assess this with metrics such as **Mean Absolute Error**

$$= \frac{\sum_{i=1}^n |e_i|}{n}, \quad \forall i$$

- How much does the MAE vary between groups?
- Important – looking for **similar error values regardless of size.**



Cross validation: Binned MAE

Bin	Group 1	Group 2	Group 3	Group 4	Group 5	Mean	Std	CoV (%)
0-10	0.73	0.74	0.83	1.02	0.68	0.80	0.13	16
11-20	1.12	1.42	1.22	1.80	1.46	1.40	0.26	19
21-30	2.36	3.22	2.33	1.83	2.54	2.46	0.50	20
31-40	2.52	4.24	2.10	2.60	3.15	2.92	0.83	28
>40	5.29	2.63	4.26	5.99	2.93	4.22	1.46	35

- Binning to account for the dependence on pit count
- Compute MAE for each group, for each bin.
- Compare consistency across groups.

Cross validation: binned MAE

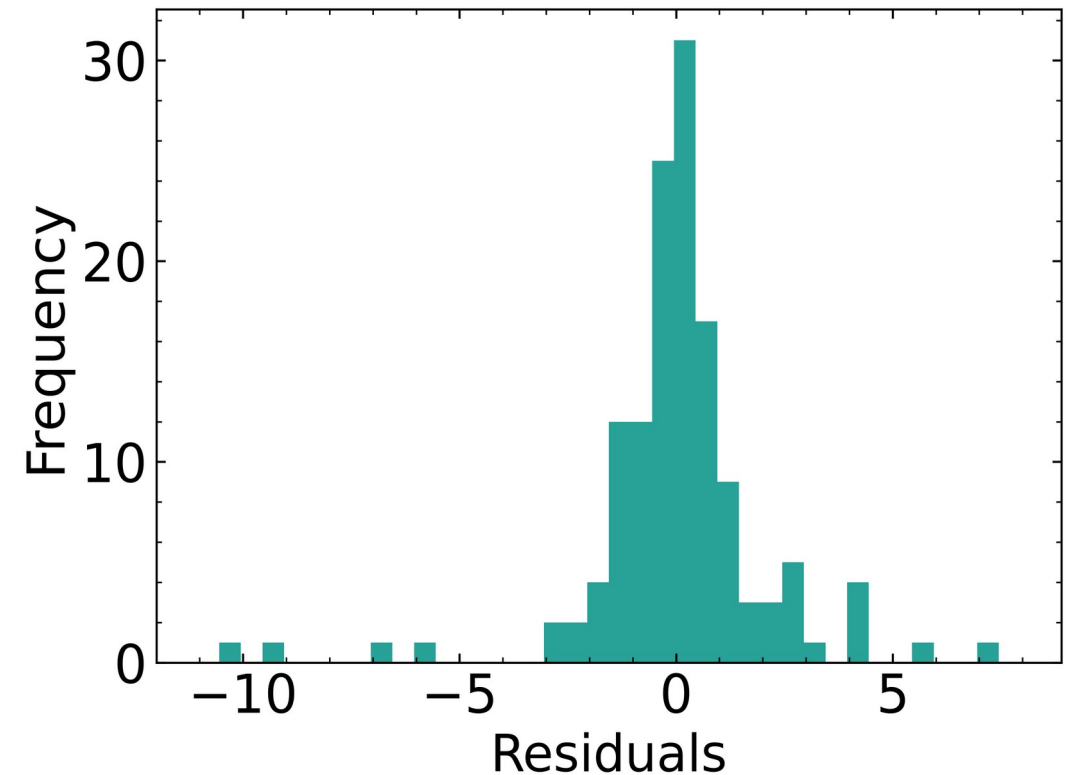
Bin	Group 1	Group 2	Group 3	Group 4	Group 5	Mean	Std	CoV (%)
0-10	0.73	0.74	0.83	1.02	0.68	0.80	0.13	16
11-20	1.12	1.42	1.22	1.80	1.46	1.40	0.26	19
21-30	2.36	3.22	2.33	1.83	2.54	2.46	0.50	20
31-40	2.52	4.24	2.10	2.60	3.15	2.92	0.83	28
>40	5.29	2.63	4.26	5.99	2.93	4.22	1.46	35

- Variation in prediction quality depending on training and testing subsets.
- Final model could exhibit this **variation in predictions on real data**.
- Clear diagnosis – need to **increase dataset size** for higher pit counts.

Determining an uncertainty

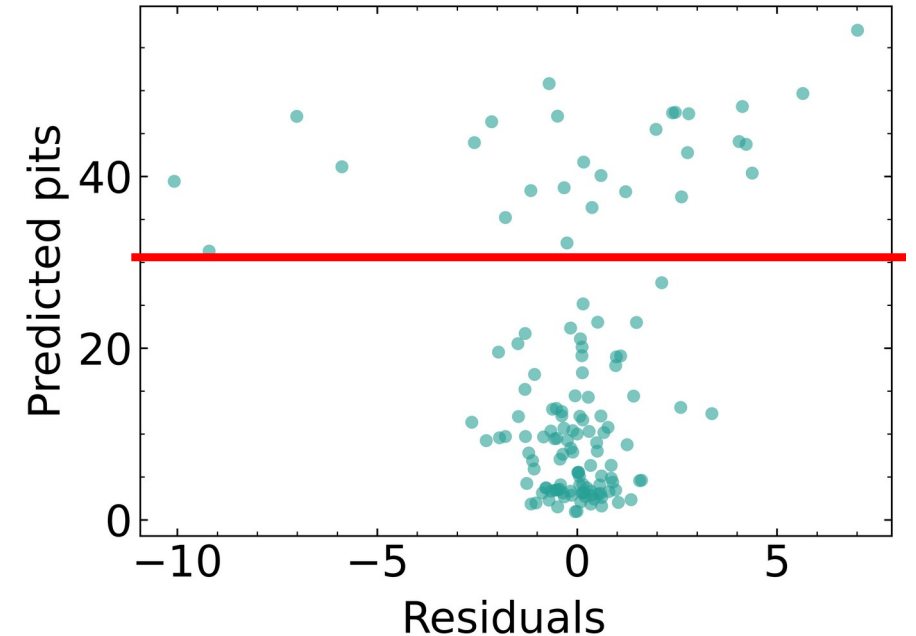
- Once a final model is trained, can **assess for uncertainty**.
- Use trained model to **predict results for test dataset**.
- Residuals are the **difference between prediction and truth**.
- Roughly **symmetric** – no systematic offset
- **Centred around 0** – could use standard deviation as an uncertainty estimate.

Distribution of test residuals



Extra layer of complexity

- As before, residuals more spread at high pit count
- Unfair to use same uncertainty at all levels
- Sub distributions can be used to assess appropriate value for each range.
- Example, split at 30 pit level:
 - Slightly different centres
 - Significantly different std



⋮

Combined uncertainty for a dosemeter

- Each **subarea** now **has an individual uncertainty** from the model testing dataset – but **no uncertainty in empty subareas!**
- Pits on each subareas are summed for the total, so **uncertainty contributions are summed in quadrature:**
- However, results of **cross-validation** may suggest an **additional component** should be included.

More to consider

- Different cross validation testing metrics and values can reveal more information.
- In this application, positioning of the subareas could lead to slightly different results.
- How does the **ML model uncertainty compare to other sources**, like variation in dosimeter sensitivity and knowledge of the radiation field?

4. Summary

- The ML model is an important source of uncertainty to consider.
- Cross validation is a useful approach to check model generalisation.
- Ongoing testing can help to reveal weaknesses in the model.
- Next steps: increase dataset and keep testing!