Evaluating Reliability in Medical DNNs: A Critical Analysis of Feature and Confidence-based Out-of-Distribution Detection

Harry Anthony, Konstantinos Kamnitsas

Department of Engineering Science, University of Oxford, Oxford, UK.

harry.anthony@eng.ox.ac.uk

HarryAnthony

@HarryEJAnthony

MICCAI 2024 Marrakesh Morocco · UNSURE · Kamnitsas Lab for AI in Medicine · University of Oxford · EPSRC

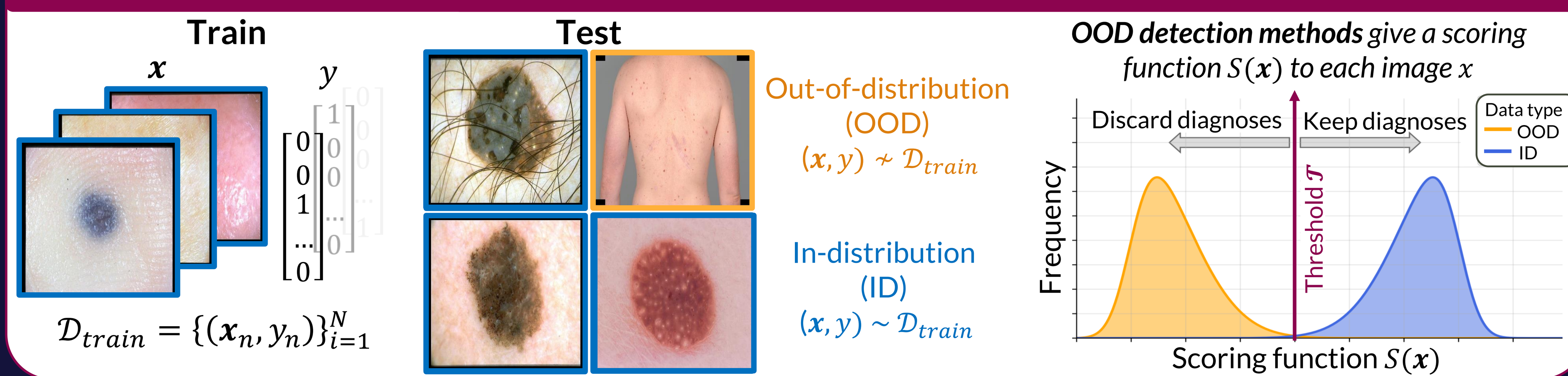
Introduction

Reliable neural networks must detect inputs that are **out-of-distribution (OOD)**.

- Individual OOD detection methods have strengths and weaknesses.
- We created new OOD counterfactual datasets to analyse these weaknesses.
- Combining complementary methods can mitigate against their weaknesses.

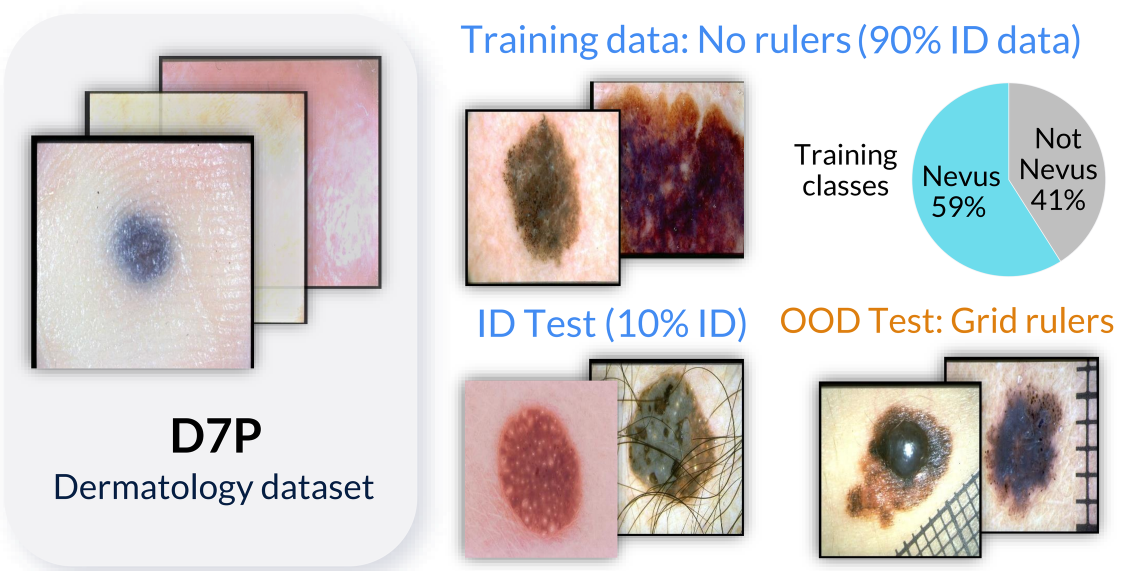
Paper available at: Code available at:

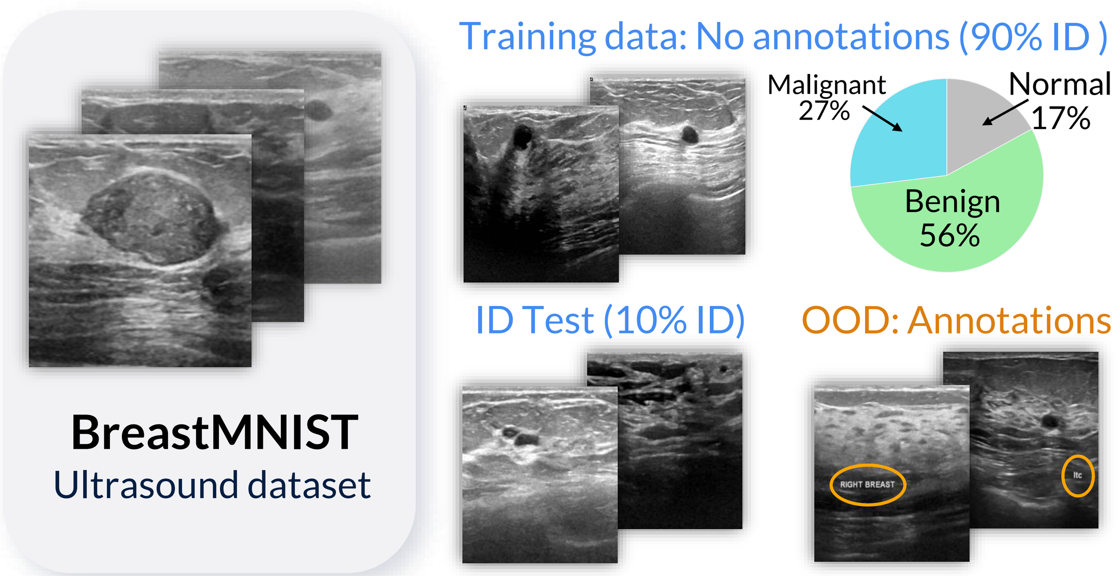
1. What is Out-Of-Distribution (OOD) detection?



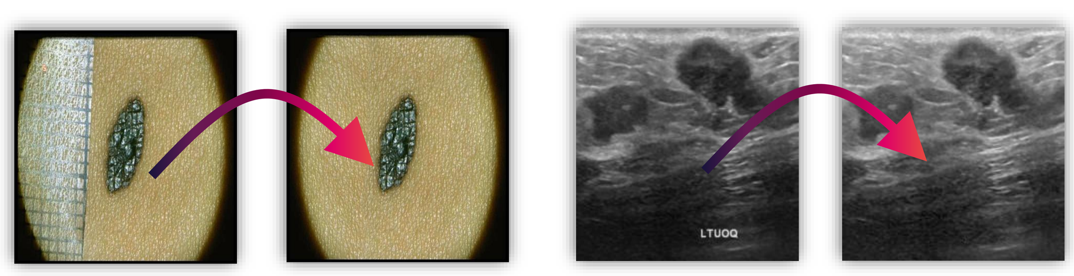
2. OOD Detection benchmarks

a) New out-of-distribution benchmarks

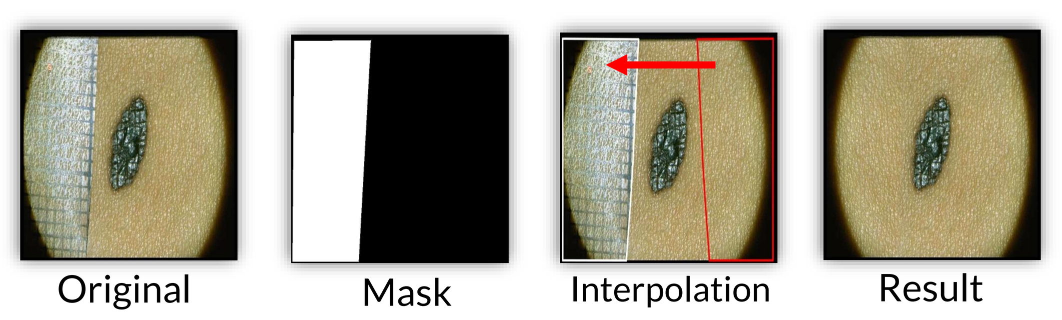




b) New Counterfactual Datasets



*Dataset was created with **inter-image interpolation**, using a patch from the same image to remove OOD artefacts.*

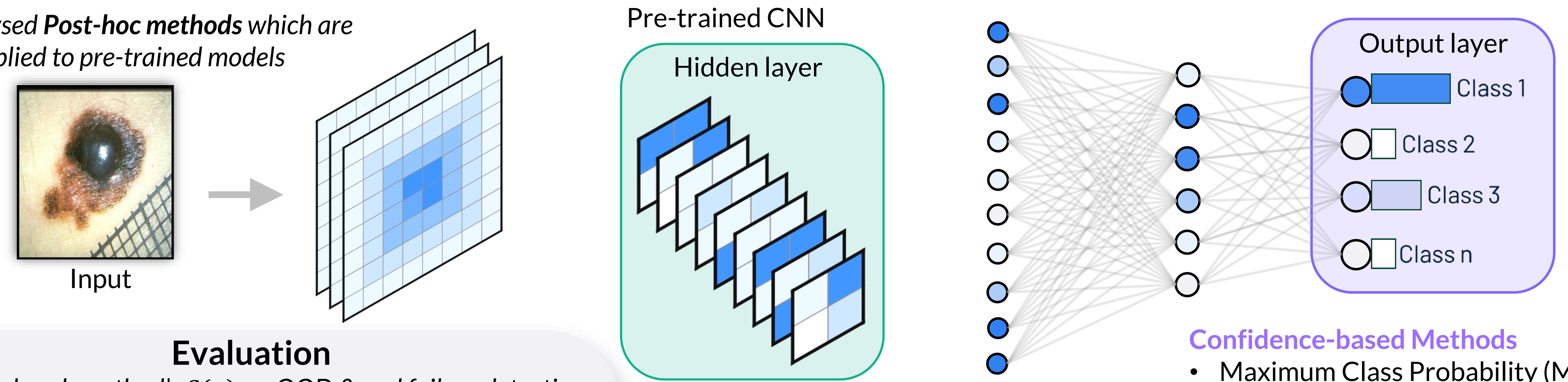


Access this new data

- Annotations for 2 new OOD benchmarks
- Pixel-wise artefact masks
- 478 image counterfactual datasets

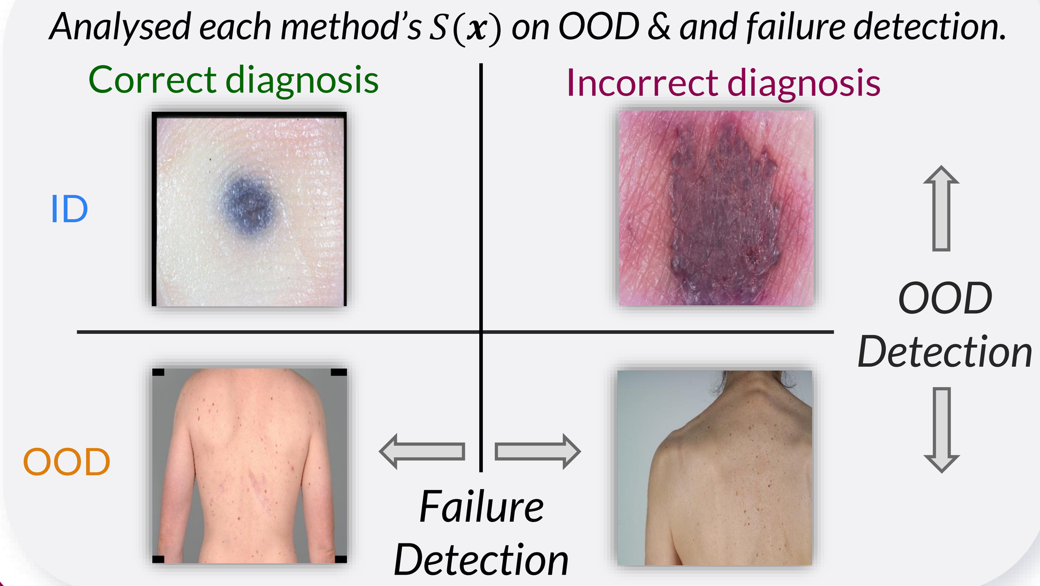
3. Out-of-distribution Detection Methods: Confidence-based & Feature-based

*We analysed **Post-hoc methods** which are applied to pre-trained models*



Evaluation

Analysed each method's $S(x)$ on OOD & and failure detection.



Feature-based Methods

- Mahalanobis Score
- Multi-Branch Mahal. (MBM)
- RMS
- GRAM Matrices

Confidence-based Methods

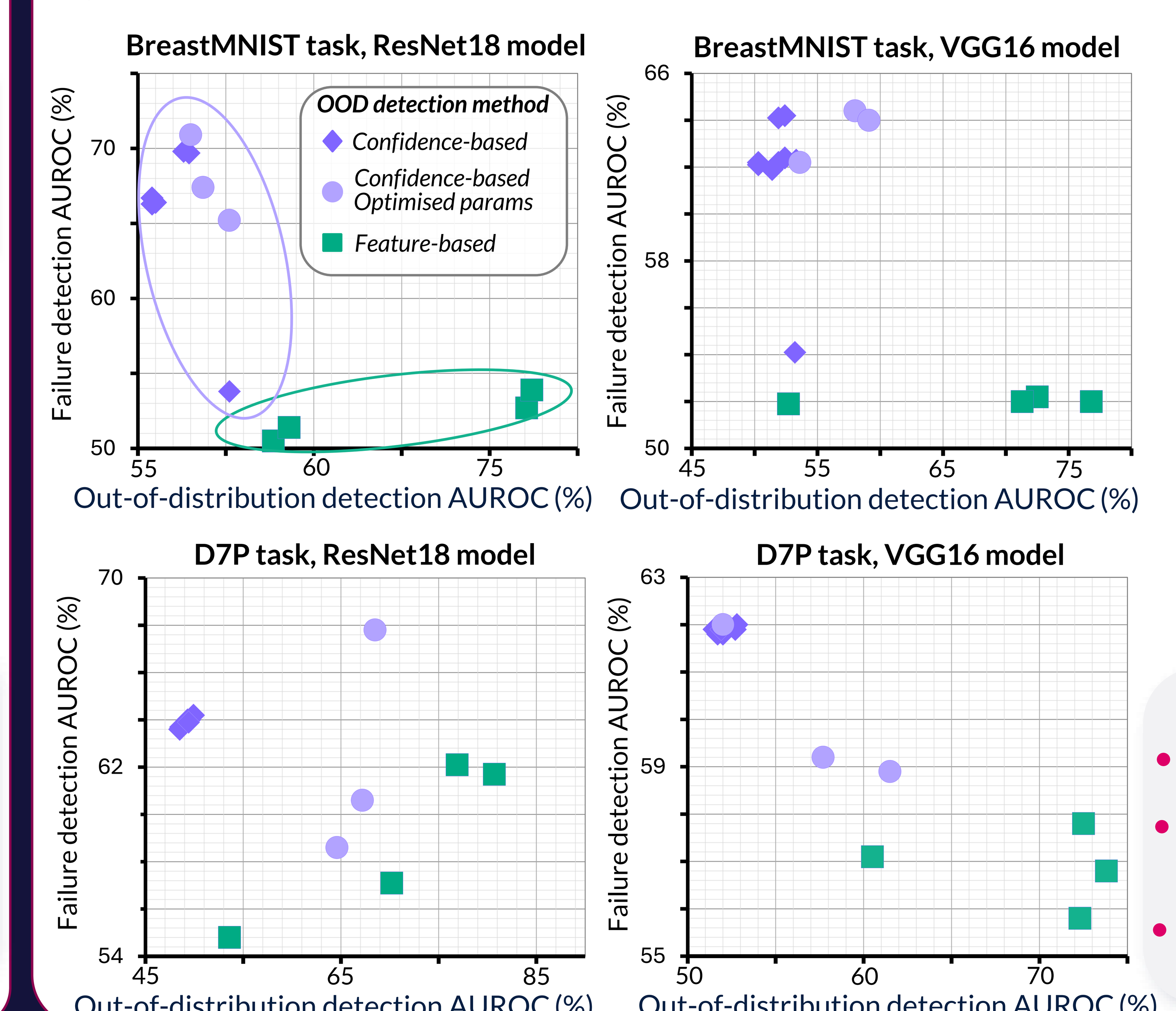
- Maximum Class Probability (MCP)
- Maximum Logit Score (MLS)
- Shannon Entropy (SE)
- Energy Score
- MC Dropout – MCP
- MC Dropout – Predictive Entropy
- MC Dropout – Mutual Information
- Deep Ensembles – MCP
- GradNorm

+ with optimised hyperparameters

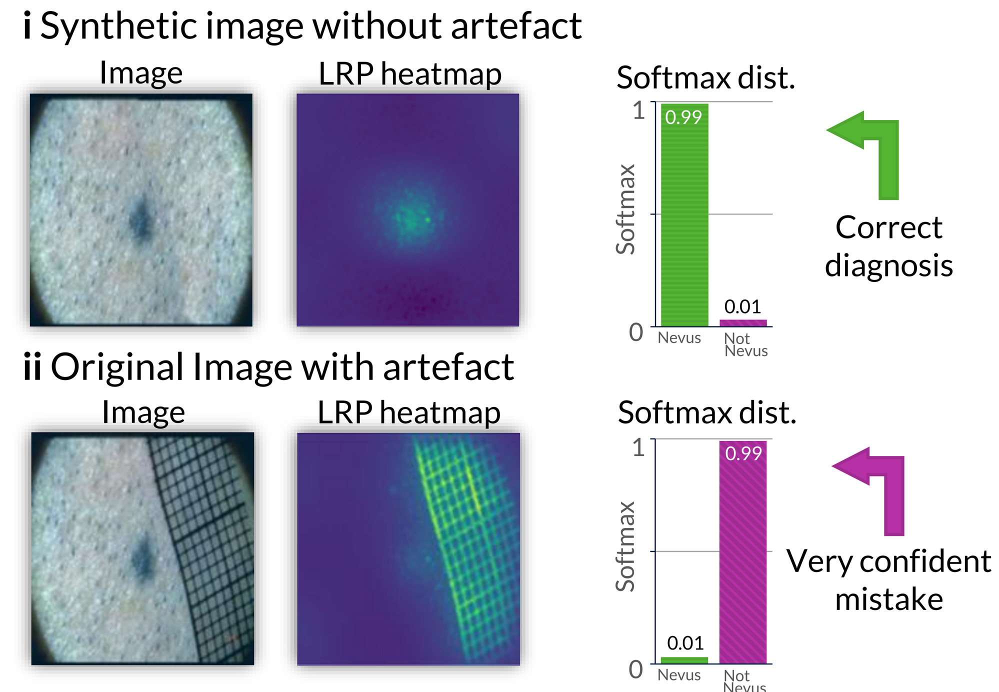
- ODIN
- ReAct
- DICE

4. Experiments on OOD detection tasks

a) OOD detection and failure detection evaluation



b) Why do confidence-based methods have poor OOD detection?



→ OOD artefacts can lead to high confidence predictions which confidence-based methods won't detect!

Key Takeaways

- Out-of-distribution detection ≠ Failure detection
- **Confidence-based** methods typically outperform **feature-based** methods at **Failure Detection**
- **Feature-based** methods typically outperform **confidence-based** methods at **OOD Detection**

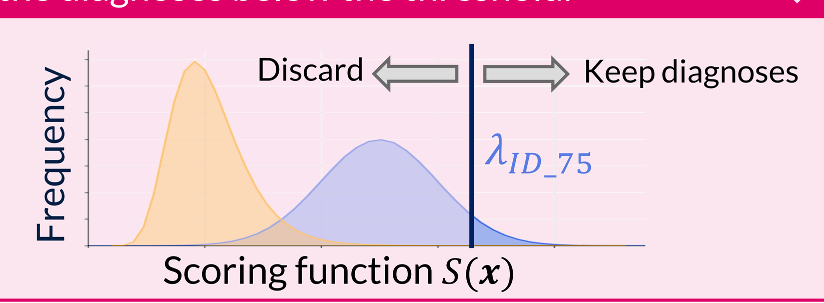
5. Combining OOD detection methods to mitigate their weaknesses

a) Method for studying diagnoses above $S(x)$ threshold

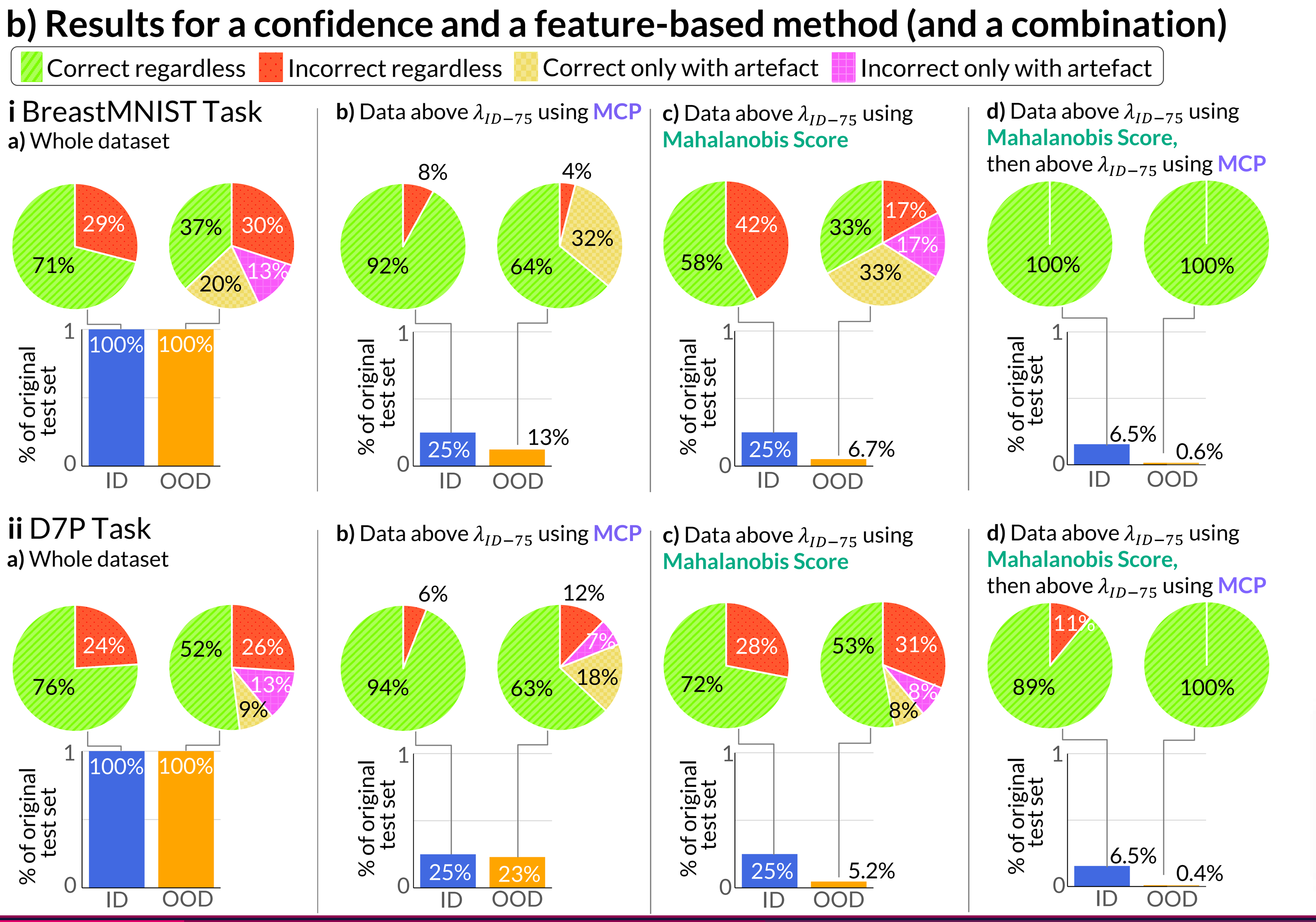
1. Sort all OOD images into one of four categories

Label	Image with artefact	Counterfactual Image without artefact
Correct regardless	Correct	Correct
Incorrect regardless	Incorrect	Incorrect
Correct only with artefact	Correct	Incorrect
Incorrect only with artefact	Incorrect	Correct

2. Calculate Scoring function for OOD method.
3. Calculate 75 percentile on ID test data λ_{ID_75}.
4. Remove the diagnoses below the threshold.

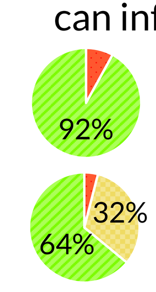


b) Results for a confidence and a feature-based method (and a combination)



c) In-depth insights

- **Correct only with artefact** diagnoses can inflate failure detection AUROC.



ID
- Accuracy: 92%

OOD
- Accuracy: 64+32=96% > 92%
- Correct regardless: 64% < 92%

→ *Could give false sense of security*

- OOD detection methods with high OOD AUROC, but which also cause a higher risk of incorrect diagnoses, may not make neural nets more trustworthy.
- Using multiple detectors leads to more discarded diagnoses, but the remaining diagnoses are more trustworthy.

Key Takeaways

- **Combining** confidence & feature-based methods in a pipeline can **mitigate their respective weaknesses.**