On the use of Mahalanobis distance for out-of-distribution detection with neural networks for medical imaging

Harry Anthony[1], Konstantinos Kamnitsas[1,2,3]

[1] Department of Engineering Science, University of Oxford, Oxford, UK.
[2] Department of Computing, Imperial College London, London, UK
[3] School of Computer Science, University of Birmingham, Birmingham, UK

Harry.Anthony@eng.ox.ac.uk

HarryAnthony
@HarryEJAnthony

MICCAI 2023 Vancouver CANADA

UNSURE

EPSRC Engineering and Physical Sciences Research Council

Introduction

Reliable neural networks must detect inputs that are **out-of-distribution (OOD)** compared to the training data.

Performance of OOD detection method **Mahalanobis score** is mixed in literature.

Research into its best practises is required.

Paper available at:

Code available at:

1. Out-of-distribution detection

Train

x y

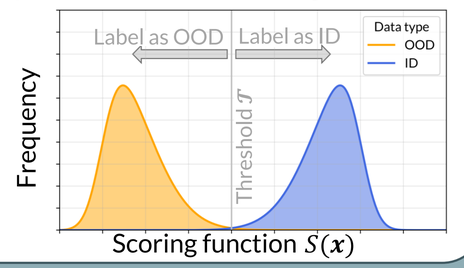
$\mathcal{D}_{train} = \{(\boldsymbol{x}_n, y_n)\}_{i=1}^{N}$

Test

Out-of-distribution (OOD)
$(\boldsymbol{x}, y) \nsim \mathcal{D}_{train}$

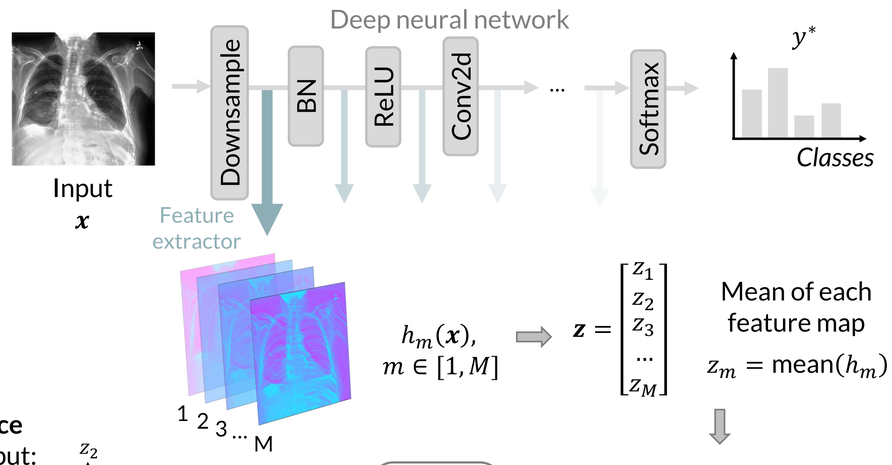
In-distribution (ID)
$(\boldsymbol{x}, y) \sim \mathcal{D}_{train}$

OOD detection is a binary classification problem, we want a scoring function S:



2. Mahalanobis distance-based OOD detection

Mean of each feature map extracted after a module in the neural network defines an embedding vector ***z****.*



Calculate **Mahalanobis distance** to class centroid c for a test input:

$$\mathcal{D}_{\mathcal{M}_c} = \sum_{i=1}^{M}(\boldsymbol{z_{i_c}} - \boldsymbol{\mu_c})\boldsymbol{\Sigma_c^{-1}}(\boldsymbol{z_{i_c}} - \boldsymbol{\mu_c})^T$$

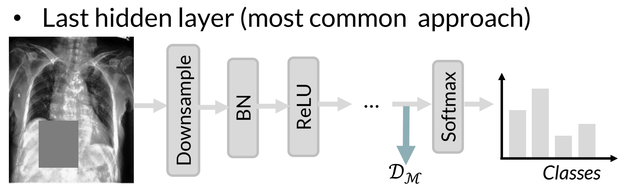
Use distance to closest class as OOD scoring function:

$$\mathcal{D}_{\mathcal{M}} = S_{Mahal.score} = \min_c(\mathcal{D}_{\mathcal{M}c})$$

Called the **Mahalanobis score**.

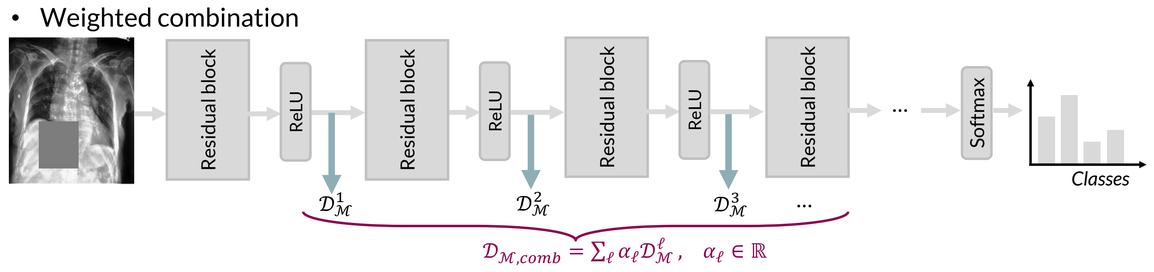
3. Feature extraction and combination

- Last hidden layer (most common approach)

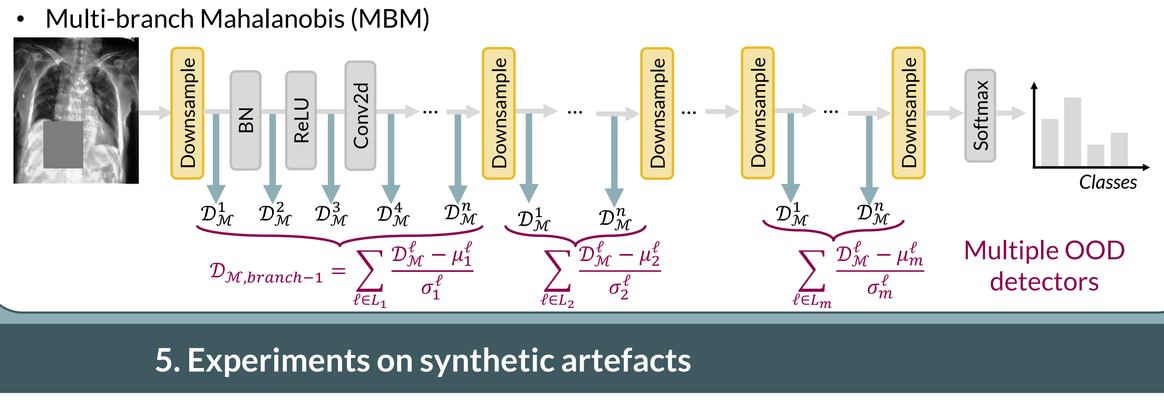


Visualisation of different techniques for extracting embeddings and combining distances for Mahalanobis score.

- Weighted combination



- Multi-branch Mahalanobis (MBM)

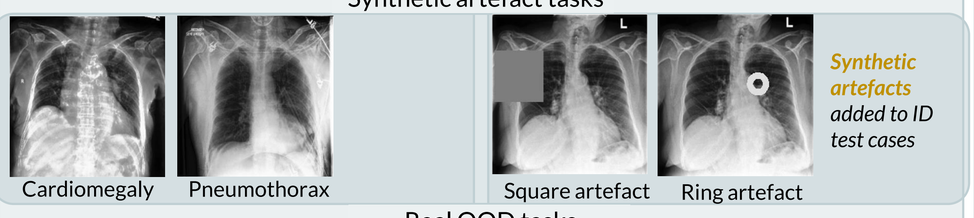


4. OOD detection tasks

In-distribution (train & test set)

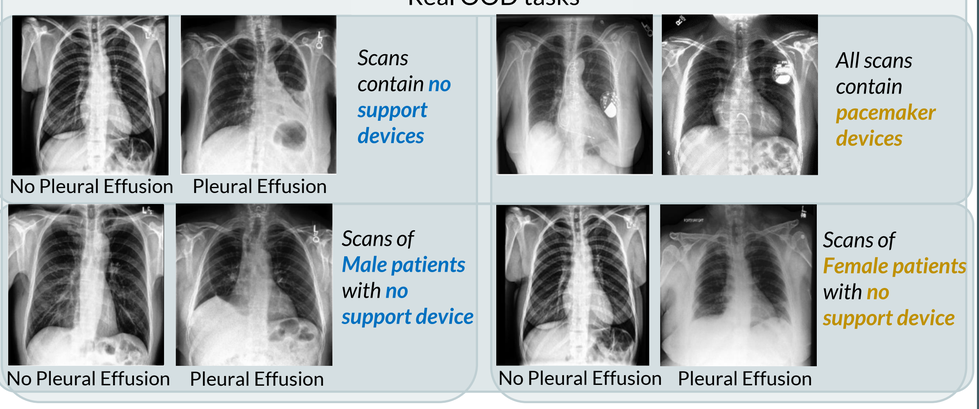
Out-of-distribution

Synthetic artefact tasks



Synthetic artefacts added to ID test cases

Real OOD tasks



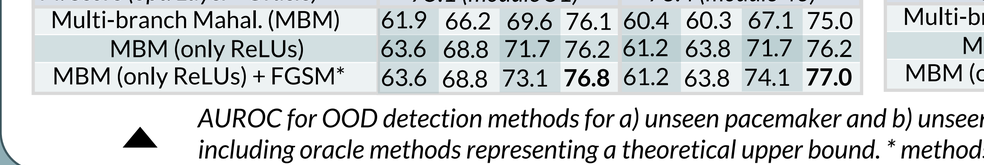
Scans contain ***no support devices***

All scans contain ***pacemaker devices***

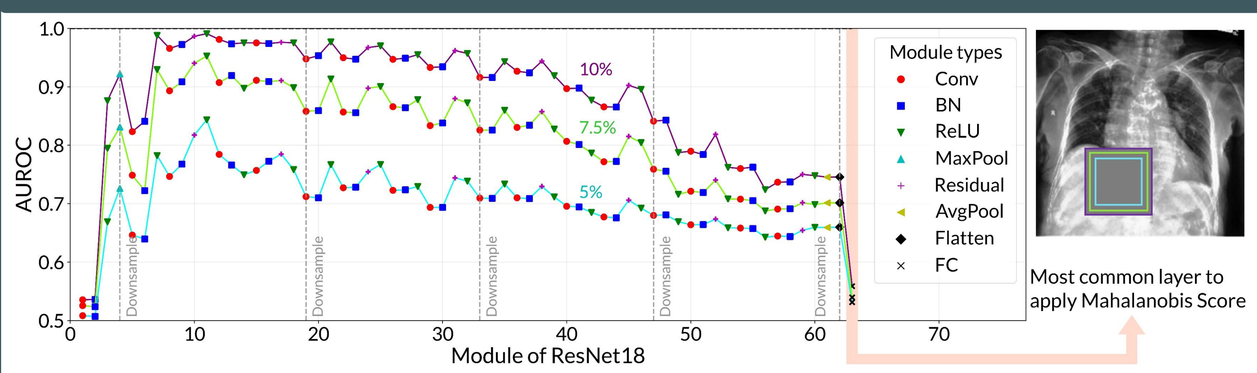
Scans of ***Male patients*** *with* ***no support device***

Scans of ***Female patients*** *with* ***no support device***

New test-bed for OOD detection: new public manual annotations of pacemakers & support devices in CheXpert

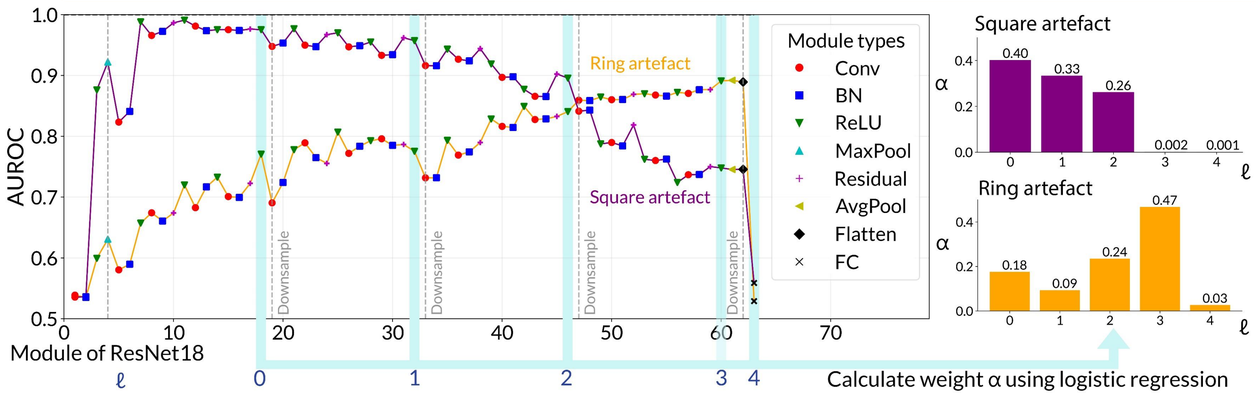


5. Experiments on synthetic artefacts



AUROC for Mahalanobis over the modules of ResNet18 for synthetic square of size 10% (purple), 7.5% (green) and 5% (blue) of the image.

Using Mahalanobis on last layer (most common practise) can be very suboptimal

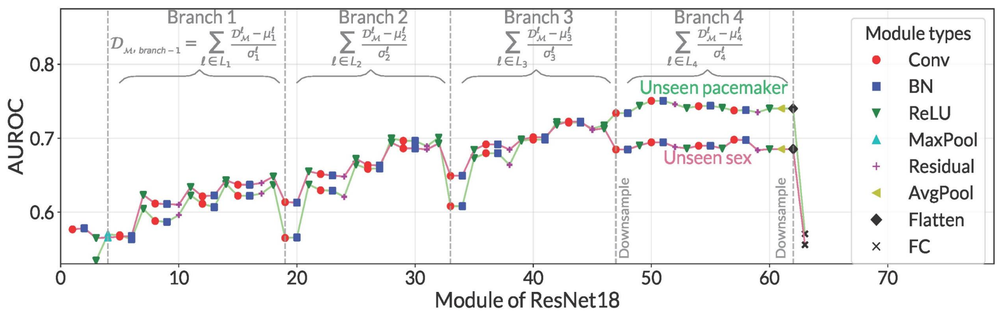


AUROC for Mahalanobis over the modules of ResNet18 for synthetic grey squares (purple) and white rings (orange).

The layers used for weighted combination (section 3) are highlighted in blue, and the weightings α_l for each layer are shown on the right.

Different OOD patterns are optimally detected at different depths of a network

6. Experiments on real OOD data



AUROC for Mahalanobis over the modules of ResNet18 for unseen pacemaker (green) and unseen sex (pink). MBM branches shown with grey brackets.

Multiple OOD detectors can be used to detect different OOD artefacts

a) Unseen pacemaker OOD	ResNet18 (AUROC ↑)				VGG16 (AUROC ↑)			
Maximum class Probability	58.4				58.3			
Monte Carlo Dropout	58.4				58.4			
Deep Ensemble	59.7				60.0			
ODIN *	66.1				70.3			
Mahal. Score (LHL)	57.1				55.8			
Mahal. Score (LHL+FGSM)*	57.4				57.5			
Mahal. Score (weight. comb)	64.5				66.0			
Mah. Score (w. Comb w/o LHL)	71.4				67.4			
*M. Score (opt. Layer - Oracle)**	*75.1 (module 51)*				*76.4 (module 40)*			
Multi-branch Mahal. (MBM)	61.9	66.2	69.6	76.1	60.4	60.3	67.1	75.0
MBM (only ReLUs)	63.6	68.8	71.7	76.2	61.2	63.8	71.7	76.2
MBM (only ReLUs) + FGSM*	63.6	68.8	73.1	**76.8**	61.2	63.8	74.1	**77.0**

b) Unseen sex OOD	ResNet18 (AUROC ↑)				VGG16 (AUROC ↑)			
Maximum class Probability	57.0				56.6			
Monte Carlo Dropout	57.0				56.7			
Deep Ensemble	58.3				57.7			
ODIN *	60.4				64.4			
Mahal. Score (LHL)	55.6				55.2			
Mahal. Score (LHL+FGSM)*	55.8				57.0			
Mahal. Score (weight. comb)	64.3				63.0			
Mah. Score (w. Comb w/o LHL)	70.3				66.7			
*M. Score (opt. Layer - Oracle)**	*72.2 (module 44)*				*76.3 (module 43)*			
Multi-branch Mahal. (MBM)	63.4	67.5	70.8	70.6	62.7	64.2	67.8	74.7
MBM (only ReLUs)	64.9	69.3	71.8	70.2	63.8	66.2	69.7	76.4
MBM (only ReLUs) + FGSM*	64.9	69.3	**72.1**	71.4	63.8	66.2	70.4	**78.0**

*AUROC for OOD detection methods for a) unseen pacemaker and b) unseen sex OOD tasks. Bold highlights the best result of methods, not including oracle methods representing a theoretical upper bound. * methods with hyperparameters optimised on OOD data.*

7. Finding thresholds for multiple detectors

Using multiple OOD detectors means finding multiple thresholds is required, which adds complexity.

Grid search to find thresholds for MBM's multiple detectors. If one detector flags OOD, the input is marked as such. Show it is feasible and can improve upon single detectors.

OOD detection method	Balanced accuracy		
	Both tasks	Unseen Sex	Pacemakers
Mahal score. (equally weighted comb w/o LHL)	67.64	65.63	70.37
Mahal. score (weighted comb with optimised α_l)	68.14	64.89	70.80
Multi-branch Mahal. (ReLU only)	**71.40**	**67.26**	**75.16**

Balanced accuracy for simultaneous detection of 2 OOD patterns, showing a multi-detector system can improve OOD detection over optimised single-detector systems.