The Invisible Gorilla Effect in Out-of-distribution Detection

Supplementary Material

1. Computational Resources

All computations were performed on a system equipped with NVIDIA RTX A5000 GPUs (24 GB VRAM), an Intel(R) Xeon(R) W-2275 CPU @ 3.30GHz and 256 GB of system RAM. This hardware configuration provided sufficient computational capacity for training and evaluating the models used in this study. Experiments were implemented using PyTorch 2.6.0 with CUDA 12.4 and cuDNN 9.1.0. To ensure reproducibility during evaluation, deterministic GPU execution was enabled by setting `torch.backends.cudnn.deterministic = True`.

2. Information on Model Training

All primary models were optimised using AdamW [19] with an initial learning rate of 1e-4 and weight decay 0.01. Network weights were initialised using He initialisation, appropriate for ReLU-based architectures. Models were trained using the cross-entropy loss for 600 optimisation steps. A multi-step learning rate schedule was applied via PyTorch's `MultiStepLR`, with decay milestones at [150,300,450] epochs. Models were trained with batch size 256.

We employed five repetitions of five-fold cross-validation to ensure robust evaluation across all experiments. Medical imaging datasets often contain repeated images or multiple images from the same patient, which can lead to data leakage. This can be an issue for evaluating generative-based methods such as DDPM-MSE, which have been shown to exhibit memorisation behaviour [5]. To address this, we enforced strict patient-level separation in the CheXpert dataset, ensuring that all images from the same patient were assigned exclusively to the training set. For the ISIC dataset, where generative models were evaluated, we removed all duplicate images prior to training in order to avoid inflating performance due to memorisation. A list of images in the ISIC dataset used for training are given in the Code Appendix.

During training, a series of image augmentations were applied to improve model generalisation. Input images were first resized to 224×224 pixels and then centre-cropped. Spatial augmentations included random rotations of up to 45°, random cropping with up to 25 pixels of padding, horizontal flipping with a probability of 0.5, and random perspective transformations with a distortion scale of 0.2. All images were converted to PyTorch tensors and normalized to float precision. Finally, Channel-wise normalisation was performed using dataset-specific means, which can be accessed in the Code Appendix. These augmentations were applied only during training. During evaluation, only resizing to 224×224 and channel-wise normalisation were applied, with no additional augmentations. Note that no colour-perturbing augmentations were used during training, as this would impact the colour of the model's regions of interest, which was a focus of this study.

For internal ad-hoc methods, the training regime of the primary models were different to increase the OOD-awareness of the model. Bayes by Backprop [7] is a variational inference method for training Bayesian neural networks. A Gaussian prior $\mathcal{N}(0,1)$ was placed over each weight. Posterior distributions were initialised with mean $\mu = 0$ and scale parameter $\rho = -3$, where the standard deviation was computed as $\sigma = \log(1+\exp(\rho))$. During each training epoch, we performed five Monte Carlo forward passes and optimised the model using the evidence lower bound (ELBO) [6], which consisted of a cross-entropy loss and a KL divergence loss. CIDER [21] was trained using supervised contrastive learning with a regularisation term based on class prototypes. Prototypes were updated using an exponential moving average (EMA). Two augmented views of each input were passed through the network, and the resulting features were normalised before computing the loss. For CIDER, the total loss was a combination of a discriminative loss and a compactness loss [21], which we gave equal weighting ($w = 1$). For Outlier Exposure [13], we applied a KL-divergence loss on the OOD samples, weighted by 0.5, and combined it with the cross-entropy loss for the main image classification task. For Reject Class [8], we optimised the model using only cross-entropy loss, assigning all OOD samples to the additional class label. For both Outlier Exposure and Reject Class, we used CIFAR-10 [18] as the auxiliary OOD dataset, following the setup from the original Outlier Exposure paper [13]. Finally, for Rotation Prediction [14], an additional head was added to the primary model, to predict the rotation of a training image from the set { 0°, 90°, 180°, 270°}. A cross-entropy loss was applied to the rotation prediction task, weighted by 0.25, and combined with the cross-entropy loss for the main image classification task.

For external methods, a model was trained external to the primary model trained for the image classification task. For DeepSVDD [23], we first trained a WideResNet-based [28] autoencoder to reconstruct the input images using mean squared error for 200 epochs. We then used the WideResNet encoder as the feature extractor for anomaly detection. After pre-training, we initialised the centre of the DeepSVDD hypersphere $c \in \mathbb{R}^{32}$ as the mean of the feature embeddings

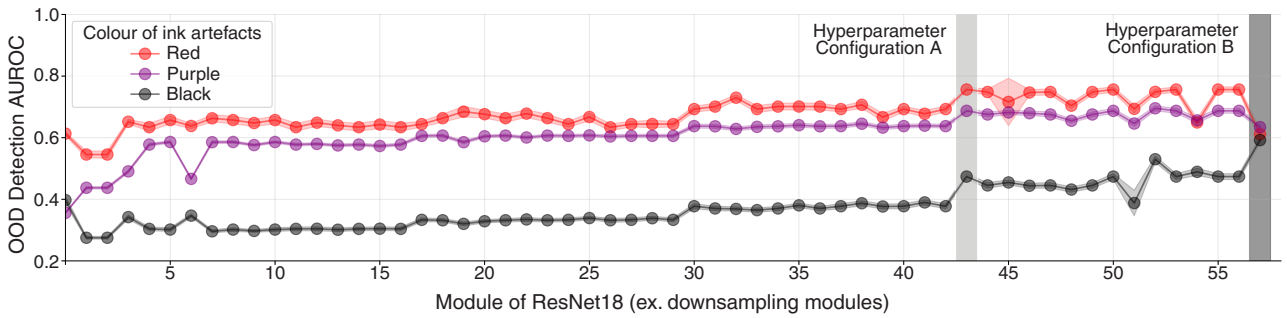

Figure 1. OOD detection AUROC for Mahalanobis score across ResNet18 modules (excluding downsampling layers) for red, purple and black ink annotations in ISIC. Light (hyperparameter configuration A) and dark grey (hyperparameter configuration B) bands highlight selected modules for comparison. The results show that the observed OOD detection performance varies significantly across layers, demonstrating that conclusions drawn using Mahalanobis score are dependent on the choice of hyperparameter. The Invisible Gorilla Effect would not be observed if only the dark grey hyperparameter is evaluated, highlighting the benefit of evaluating a wide range of hyperparameter settings. Each point represents the mean over 25 seeds, with coloured shaded regions indicating the 95% confidence interval.

from the training data, where 32 is the dimensionality of the encoder's output. To avoid trivial solutions, dimensions of c with absolute value below 1×10^{-5} were set a value of 1×10^{-5}. During the main DeepSVDD training phase, we optimised the encoder to minimise the average squared Euclidean distance between the feature representations of the inputs and the centre c. For training the DDPM [11], we followed the original setup by defining a linear noise schedule over $T = 1000$ timesteps. The variance schedule $\{\beta_t\}_{t=1}^{T}$ was linearly spaced between a minimum value of $\beta_{\text{min}} = 10^{-4}$ and a maximum value of $\beta_{\text{max}} = 0.02$. At each training step, noise was added to input images, and the model learned to predict this noise using a U-Net [22] conditioned on the diffusion timestep. The network was optimised using a mean squared error loss. For Foreign Patch Interpolation [26], each batch consisted of paired image patches extracted from both original and shuffled input images. The model was trained to distinguish whether a patch originated from the same image or from a different one. Patch centres were sampled using a core percent value of 0.99, meaning that only a narrow margin around the image borders was excluded to avoid placing patches near the edges. Training used a pixel-wise binary cross-entropy loss, where the loss was weighted based on the proportion of positive pixels in each batch [26]. Finally, for RealNVP [16], the model consisted of 6 affine coupling layers, each with a separate scale and translation network. These networks are feedforward MLPs that operate on a masked portion of the input to compute transformations for the remaining features. After every two coupling layers, a learnable permutation layer is applied to shuffle the feature dimension [16]. For training the RealNVP model, we extracted feature representations from a selected layer of a pre-trained primary model and computed per-class mean and precision statistics. We use these statistics to remove correlations between features, a process known as whitening. For each class, a separate RealNVP model was trained on the whitened features using maximum likelihood estimation.

3. Method Hyperparameters

For each method, we explored a wide range of hyperparameter settings. The specific hyperparameter configurations evaluated in our study are detailed in the following tables: feature-based methods (Table 1), internal ad-hoc methods (Table 2), confidence-based methods (Table 3), and external methods (Table 4). To highlight the importance of evaluating a broad range of hyperparameters, we report OOD detection AUROC for the Mahalanobis score across all layers of a ResNet18 on the ISIC benchmark with ink annotations (Figure 1). While red ink annotations consistently yield higher AUROC scores across most layers, our results show that specific hyperparameter choices can reverse this trend - making red ink annotations no more detectable than purple or black. This illustrates that conclusions about OOD performance can be highly sensitive to the choice of hyperparameters. Therefore, a full evaluation of method performance must consider a comprehensive hyperparameter search to ensure robust and fair comparisons across methods and settings. For each OOD artefact type, we selected the hyperparameter setting that achieved the highest mean performance (e.g. AUROC) for that specific artefact, averaged over 25 random seeds. This procedure was carried out independently for each artefact category, ensuring that the reported results reflect the best-performing configuration for the corresponding artefact.

Table 1. List of hyperparameters for *feature-based OOD detection* methods. The table presents the hyperparameter configurations evaluated in our study, along with the total number of settings used per primary network architecture for each method.

Method	**Method Hyperparameter(s)**	**Number of Settings per Model**		
		ResNet18	**VGG16**	**ViT-b32**
CoP	• Dimensionality, $D_{\text{CoP}} \in [2, 10, 20]$	3	3	3
CoRP	• Dimensionality, $D_{\text{CoRP}} \in [2, 10, 20]$ • Gaussian Kernel, $\gamma_{\text{CoRP}} = [0.1]$ • Number of random Fourier features, $n_{rff} = [20]$	3	3	3
FeatureNorm	• Layer selection	66	42	139
GRAM	• Powers, $p_{GRAM} \in [[1], [10], [2, 4, 6]]$	3	3	3
KDE (Gaussian)	• Layer selection	66	42	139
KNN	• Layer selection • Number of neighbours, $k_n \in [1, 3, 5, 10, 20]$	330	210	695
LOF	• Layer selection • Number of neighbours, $k_n \in [1, 3, 5, 10, 20]$	330	210	695
Mahalanobis	• Layer selection	66	42	139
MBM	• Bracket, $B \in [1, 2, 3, 4]$	4	4	4
NAN	• Layer selection	66	42	139
NAC	• Layer selection, $l_{\text{NAC}} =$ [Avgpool] • Sigmoid alpha, $\alpha_{\text{NAC}} = [3]$ • Number of bins, $n_{\text{bin-NAC}} = [100]$	1	1	1
NMD	• Layer selection	66	42	139
NuSA	• Linear layer, $l_{NuSA} =$ [Avgpool]	1	1	1
PCX	• Layer selection	21	16	26
Residual	• Layer selection • Dimension, $D_{\text{residual}} \in [2, 10, 20]$	198	126	417
TAPUUD	• Number of clusters, $n_{\text{cluster}} \in [[3, 5, 7]]$	1	1	1
XOOD-M	• Layer selection • Covariance matrix scaling, $C \in [0.1, 1.0, 10^4]$	198	126	417

Table 2. List of hyperparameters for *internal ad-hoc OOD detection* methods. The table presents the hyperparameter configurations evaluated in our study.

Method	**Method Hyperparameter(s)**	**Number of Settings**
Bayes By Backprop	• Monte Carlo Samples, $n_{MC} = [30]$ • Scoring function, $\mathcal{S}_{\text{BNN}} =$ [Mutual Information]	1
CIDER	• Distance function, $d_{\text{CIDER}} =$[Cosine distance] • Feature extractor layer, $l_{\text{CIDER}} =$ [Output layer]	1
Outlier Exposure	• Scoring Function, $\mathcal{S}_{\text{OE}} =$ [MCP]	1
Reject Class		1
Rotation Prediction	• Scoring Function, $\mathcal{S}_{\text{RP}} =$ [Cross Entropy Loss]	1

4. Annotation Summary

For the ISIC dataset, we used two out-of-distribution benchmarks: colour charts and ink annotations. We manually annotated 8,964 instances of colour charts and 2,358 instances of ink annotations. To enable our analysis, we manually annotated each of these images by colour. For colour charts, we first annotated each image containing a chart where a specific colour was present (Table 5). To evaluate OOD detection performance by colour, we also created a subset where each image contained only a single annotated colour (Table 6). All images in this subset were used to generate synthetic, colour-swapped counterfactuals, as described in the Main Paper. For our final analysis, we restricted evaluation to colour charts occupying less than 10% of the image area (Table 7), as large charts are easily detected (achieving near-perfect OOD AUROC across all colours) and thus offer limited insight into the effect of colour on detection performance. Similarly, for ink annotations, we manually annotated images containing any ink color (Table 8), as well

Table 3. The table presents the hyperparameter configurations evaluated in our study for *confidence-based methods*, along with the total number of settings used per primary network architecture for each method. The table presents the hyperparameter configurations evaluated in our study, along with the total number of settings used per primary network architecture for each method.

Method	**Method Hyperparameter(s)**	**Number of Settings per Model**		
		ResNet18	**VGG16**	**ViT-b32**
ASH	• ASH function, $g_{\text{ASH}} \in$[ASH-b, ASH-p, ASH-s] • Percentile, $p_{\text{ASH}} \in [0.6, 0.7, 0.8, 0.9]$	12	12	12
Deep Ensemble	• Number of Ensemble members, $n_{\text{DE}} = [5]$ • Scoring function, $S_{\text{DE}} =$ [MCP]	1	1	1
DICE	• Sparsity parameter, $p_{\text{DICE}} \in [0.6, 0.7, 0.8, 0.9]$ • Truncation layer, $l_{\text{DICE}} =$ [Penultimate layer]	4	4	4
GAIA-A		1	1	1
GradNorm	• Parameter selection • Summation method, $\ell_n \in [\ell_1, \ell_2]$	132	64	304
GradOrth	• Epsilon threshold, $\epsilon_{\text{GradOrth}} \in [0.1, 0.5, 0.9]$	3	3	3
MCP		1	1	1
MC-Dropout	• Number of samples, $n_{\text{repeats}} = [30]$ • Dropout probability, $p_{\text{dropout}} \in [0.1, 0.2, 0.3, 0.4]$	4	4	4
ODIN	• Temperature Scaling, $T \in [1, 10, 100]$ • Preprocessing Magnitude, $\epsilon_{\text{ODIN}} \in \{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$	15	15	15
ReAct	• Percentile, $p_{\text{ReAct}} \in [0.6, 0.7, 0.8, 0.9]$	4	4	4
SHE		1	1	1
ViM	• Weighting coefficient, $\alpha \in [0.25, 0.5, 0.75]$ • Dimensionality, $D_{\text{ViM}} \in [2, 5, 10, 20]$	12	12	12
WeiPer	• Scoring function, $\mathcal{S}_{\text{WeiPer}} \in$ [MCP, KL-div] • Perturbations, $\delta_{\text{WeiPer}} \in [0.1, 1.0, 10]$ • Normalising factor, $\epsilon_{\text{WeiPer}} \in [0.01, 0.2]$ • Contribution of term 1, $\lambda_1 \in [0.0, 1.0, 2.5]$ • Contribution of term 2, $\lambda_2 \in [0.0, 0.1, 1.0]$ • Number of bins, $n_{WeiPer} \in [10, 50, 100]$ • Number of perturbations, $r_{\text{WeiPer}} = [30]$	324	324	324

Table 4. List of hyperparameters for *external OOD detection* methods. The table presents the hyperparameter configurations evaluated in our study, along with the total number of settings used per primary network architecture for each method.

Method	**Method Hyperparameter(s)**	**Number of Settings**
DeepSVDD		1
DDPM-MSE	• Number of steps, $n_{\text{DDPM-step}} = [1000]$ • Number of reconstructions, $n_{\text{recon}} = [100]$ • ID comparator, $\mathcal{D}_{\text{DDPM}} =$[Training Data]	1
DDPM-LPIPS	• Number of steps, $n_{\text{DDPM-step}} = [1000]$ • Number of reconstructions, $n_{\text{recon}} = [100]$ • ID comparator, $\mathcal{D}_{\text{DDPM}} =$[Training Data] • Feature extractor, $\phi =$[Pretrained AlexNet]	1
Foreign Patch Interpolation	• Threshold, $\tau_{\text{FPI}} \in [0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$	6
RealNVP	• Layer selection	• 66 for ResNet18 • 42 for VGG16 • 139 for Vit-b32

as a subset of images containing only a single ink colour (Table 9). Finally, a summary for the annotations for the MVTec-AD benchmarks is given in (Table 10).

Table 5. Number of images containing a colour chart in ISIC where the specified colour is present. For images with multiple colours, each colour is counted independently (i.e. counts are not mutually exclusive).

Colour of chart	Number of Images
Blue	4 203
Yellow	2 502
Green	2 315
Orange	2 193
Red	1 462
Grey / white	380
Black	47

Table 6. Number of images containing a colour chart in ISIC with a single, uniquely assigned colour. Each image is associated with exactly one colour category.

Colour of chart	Number of Images
Blue	2 334
Yellow	1 202
Green	648
Orange	571
Red	805
Grey / white	365
Black	42

Table 7. Number of images containing a colour chart in ISIC with a uniquely assigned colour, where the size of the colour chart is **less than 10%** of the image area. Each image is associated with exactly one colour category.

Colour of chart	Number of Images
Blue	904
Yellow	419
Green	185
Orange	207
Red	321
Grey / white	92
Black	5

Table 8. Number of images containing an ink annotation in ISIC with a uniquely assigned colour where the specified colour is present. For images with multiple colours, each colour is counted independently (i.e. counts are not mutually exclusive).

Colour of Ink	Number of Images
Purple	1 558
Black	887
Green	26
Red	25

Table 9. Number of images containing an ink annotation in ISIC with a single, uniquely assigned colour. Each image is associated with exactly one colour category.

Colour of Ink	Number of Images
Purple	1 481
Black	839
Green	22
Red	22

Table 10. Number of images containing an ink annotation with a single, uniquely assigned colour in the MVTec benchmarks, separated by a) Metal Nuts and b) Pills.

Colour of Ink	Number of Images
a) Metal Nut	
Black	8
Blue	8
b) Pill	
Red	12
Yellow	6

5. Evaluating Statistical Significance

To assess the statistical significance of the OOD detection performance difference between similar and dissimilar artefacts, we conducted a two-sided Wilcoxon signed-rank test across 25 random seeds (e.g. 25 models). This non-parametric test was chosen as it does not assume a Gaussian distribution. We compared AUROC scores from red ink annotations (visually similar to skin lesions) and green ink annotations (visually dissimilar), using the best-performing feature layer of the method Mahalanobis Score on the full OOD dataset. Green was selected as the representative dissimilar artefact because Main Paper Figure 1 shows it yielded the highest AUROC among non-similar colours. A histogram of the OOD detection AUROCs per experiment are plotted in Figure 2. The resulting p-value was 3.28×10^{-6}, indicating a statistically significant difference in detection performance between the two OOD artefact colours.

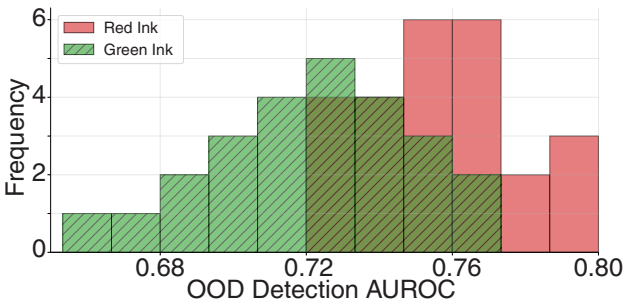

Figure 2. Distribution of AUROC scores across 25 random seeds for Mahalanobis Score on red (visually similar) and green (visually dissimilar) ink artefacts in the ISIC benchmark. Results are shown for the best-performing feature layer over the OOD dataset. A statistically significant difference was observed ($p = 3.28 \times 10^{-6}$, Wilcoxon signed-rank test).

6. Quantifying colour similarity with RGB distance

To quantify colour similarity, we used linear Euclidean RGB distance. While metrics such as CIEDE2000 more closely reflect human colour perception [20], our goal in this study is to analyse how the neural network processes colour similarity within its input representation, which is defined in RGB. Therefore, RGB Euclidean distance was used because it enables a more direct measure of similarity in the space in which the network operates.

We constructed OOD detection benchmarks to evaluate the Invisible Gorilla Effect. To obtain ROI masks for each dataset, we used the Segment Anything Model (SAM) [17]. We define the ROI as the part of the training image where the task-relevant visual features are expected to reside. For ISIC, we defined the ROI as the lesion area, as this is where features for malignant-benign classification are expected to reside. As classifiers can sometimes use contextual shortcuts (Clever Hans), we verified the model focuses on the lesion region using saliency maps from seven explainable AI methods: Expected Gradients [10], LayerCAM [15], HiResCAM [9], LRP [4], CRP [1], Integrated Gradients [25] and GradCAM [24] (Figure 3). For MVTec-AD, we defined the ROI as the entire industrial tool (e.g. metal nut).

A human annotator iteratively provided prompts to SAM until the generated segmentation closely matched the target object (example segmentations shown in Figure 4). Utilising these segmentation masks, the mean RGB values of the model's ROI in the data were calculated over the training (ID) dataset (summarised in Table 11). Note that for ISIC, a sample of 100 images was used to calculate the mean RGB, list of images used provided upon paper acceptance.

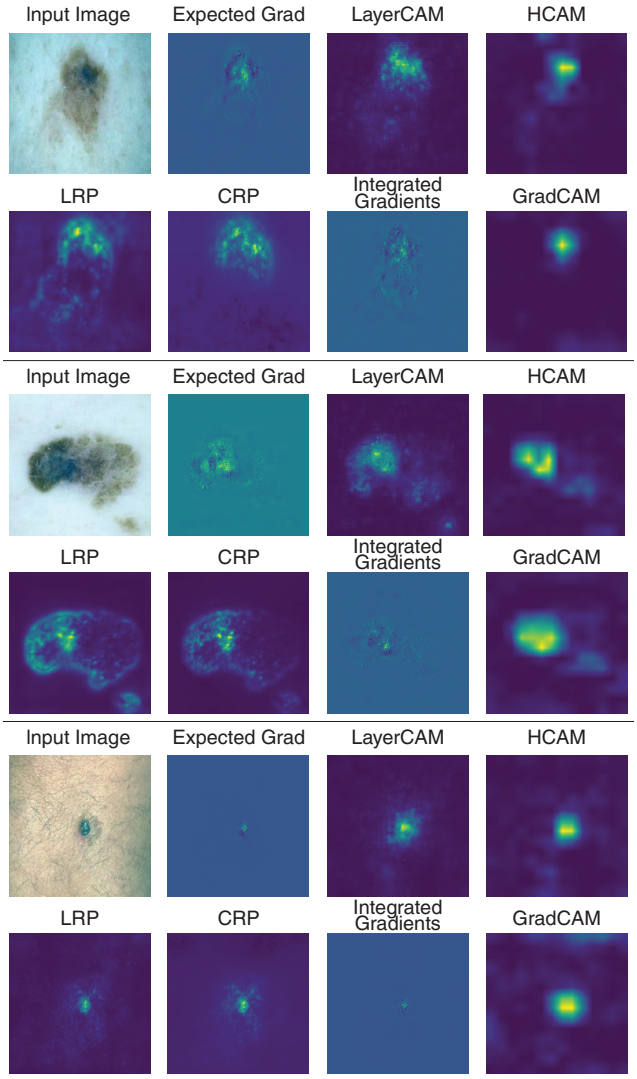

Figure 3. Explainability analysis of model focus on ISIC lesion regions. Three dermoscopic images from ISIC (top left) and their corresponding saliency maps generated using seven explainable AI methods (Expected Gradients [10], LayerCAM [15], HiResCAM [9], LRP [4], CRP [1], Integrated Gradients [25] and GradCAM [24]) for a VGG16 primary model trained for classifying between malignant versus benign lesions. cross all methods, the highlighted regions consistently localise on the lesion, providing evidence that the primary models's region of interest is the lesion.

Table 11. Mean RGB for the model's ROI across the training/ID dataset (or a sample of 100 images for ISIC), using segmentation masks created with SAM using human annotator-guided prompts.

Dataset	Model ROI	Mask	Mean RGB
ISIC	Skin Lesion	SAM	(176, 116, 77)
MVTec-AD	Metal Nut	SAM	(77, 86, 83)

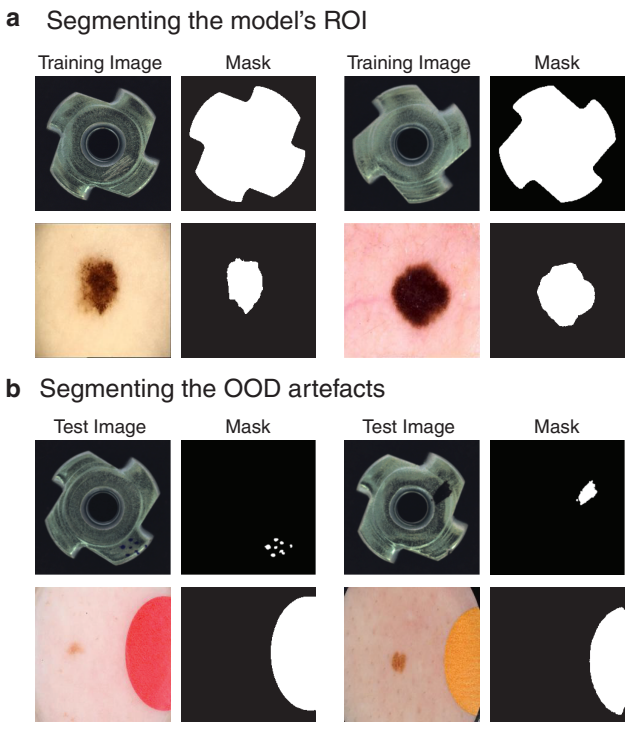

Figure 4. Examples of segmentation masks used to calculate the mean RGB of regions of interest (ROIs) for in-distribution data and OOD artefacts. A human annotator iteratively prompted the Segment Anything Model (SAM) until the produced mask closely matched the target region. Panel (a) shows segmentations of the model's ROI and panel (b) shows segmentations the OOD artefacts for the ISIC and MVTec benchmarks.

This process was then repeated to segment the naturally occurring OOD artefacts in these datasets. The segmentation masks were either created using SAM with a human annotator, or using provided ground truth masks (for MVTec covariate OOD artefacts). The mean RGB for each artefact was calculated for the artefacts in ISIC (Table 12) and MVTec (Table 13). For each dataset, we then calculated the Euclidean RGB distance between the artefact colour and the model's ROI colour and applied a dataset-specific threshold to categorise artefacts as similar (distance below threshold) or dissimilar (distance above threshold). This thresholding serves as an operational definition of similarity, as our goal was not to determine an absolute boundary but to analyse how OOD detection performance varies with colour distance. This allows us to examine the counterintuitive phenomenon where OOD detection may improve as artefact colours become more similar to those found in the training data.

Table 12. Mean RGB values of the colour-chart artefact colours used to make the OOD benchmark for ISIC. Segmentation masks were created using SAM with human annotator-guided prompts. The Euclidean distance between the mean artefact and model's ROI is given (ℓ_2), along with the categorisation of similar and dissimilar colours under a threshold of $\ell_2 = 90$.

Colour	Mask	Mean RGB	ℓ_2 dist	Label
Red	SAM	(222, 52, 57)	81.3	Sim.
Orange	SAM	(207, 123, 48)	43.0	Sim.
Yellow	SAM	(207, 191, 48)	86.2	Sim.
Blue	SAM	(65, 52, 57)	129.7	Diss.
Green	SAM	(53, 152, 69)	128.4	Diss.
White	SAM	(144, 158, 162)	100.1	Diss.
Black	SAM	(66, 61, 60)	124.2	Diss.

Table 13. Mean RGB values of the ink artefact colours used to make the OOD benchmark for ISIC. Segmentation masks were created using the ground truth (GT) masks provided by MVTec-AD. The Euclidean distance between the mean artefact and model's ROI is given (ℓ_2), along with the categorisation of similar and dissimilar colours under a threshold of $\ell_2 = 42$.

Colour	Mask	Mean RGB	ℓ_2 dist	Label
Blue	GT	(57, 59, 59)	40.9	Sim.
Black	GT	(49, 53, 68)	45.1	Diss.

7. On the Trade-Off Between OOD Generalisation and OOD Detection

A debated question in OOD detection is which inputs a model should be expected to generalise to and which should instead be rejected by an OOD detector. Some works argue that models ought to generalise to all covariate-shifted inputs, and OOD detection should exclusively be used to detect semantic shifts [27, 30]. Within this framing, robustness to every form of covariate shift is the desired goal, and evaluating OOD detection methods on covariate-shifted data is considered misaligned with this objective.

We argue that this framing is insufficient for real-world deployment, particularly in high-risk domains such as the focus of this work. In real-world scenarios, covariate shifts can produce inputs that the primary model should *not* generalise to. For example, covariate shifts can produce inputs for which the main classification task becomes ill-posed, even when the label space itself has not changed. Such shifts can hide, distort, or remove the visual cues needed for a reliable clinical decision. As a result, the model may produce high-confidence predictions even though the input no longer contains enough usable information to justify them. This applies even when the model happens to output the correct label: correctness by coincidence does not imply that the prediction is trustworthy, and such cases remain important to detect [2]. In these cases, enforcing generalisation

is undesirable: the desirable behaviour is to recognise that the input falls outside the model's scope [12] and to flag it as OOD. In such cases, the optimal behaviour is not to encourage the model to generalise to this covariate-shifted data, but instead to label these inputs as OOD and discard the predictions as unreliable. This is why many prior OOD detection works treat covariate shifts as OOD [3, 29]. We note that covariate shifts span a continuum from benign corruptions to task-invalidating artefacts, and we recognise that the boundary between robustness and covariate OOD is not universally defined. Therefore, in Section 2.1 in the main paper, we explicitly operationalise the covariate OOD settings to remove ambiguity about which artefacts we aim to detect.

For these reasons, we evaluate OOD detection on covariate-shift benchmarks that produce inputs outside the model's reliable operating conditions. In our experiments, these naturally occurring artefacts led to a substantial degradation in classification accuracy (main paper, Fig. 4), confirming that they meaningfully impact task performance. In medical imaging, collecting training data that covers the full range of real-world artefacts is often impractical due to privacy and acquisition constraints. Consequently, detecting previously unseen covariate shifts that compromise the reliability of model predictions is essential for safe deployment.

8. Supplementary Experimental Results for Section 4.2

In addition to the main results, we conducted supplementary experiments using VGG16 and ViT-B/32 architectures to assess the generality of our findings across different primary model architectures. We report results using three metrics: OOD detection AUROC, OOD detection AURC (Area Under the Risk–Coverage Curve), and FPR@TPR80. These metrics were chosen because they are unaffected by class imbalance between in-distribution and OOD samples, enabling fair comparison of OOD detection performance across artefact colours with differing numbers of images. The metric FPR@TPR80 places a high threshold (many diagnoses discarded) on the scoring function whereas the other metrics (AUROC and AURC) are threshold-independent, highlighting that threshold adjustment does not eliminate the bias. We report the post-hoc method results for the ISIC benchmarks with a ResNet primary model, with metrics AUROC (Table 2 main paper), AURC (Table 14) and FPR@TRP80 (Table 15). Post-hoc method results for the ISIC benchmarks are shown for both a ViT-B/32 primary model (Table 16) and a VGG16 primary model (Table 17). In addition, post-hoc method results for MVTec benchmarks are shown for a ResNet18 primary model (Table 18), ViT-B/32 primary model (Table 19) and VGG16 primary model (Table 20).

9. Supplementary Experimental Results for Section 4.3

In addition to the main results, we conducted supplementary experiments for the mitigation strategies across different primary model architectures and mitigation method hyperparameters.

Colour jitter augmentation: We report the OOD detection AUROC results for all the post-hoc methods for a ResNet18 primary model (Table 21). In addition, we repeated the analysis for ViT-B/32 primary model using the same colour jitter augmentation, with results given in table 22.

Subspace Projection: We first evaluate the inference latency for a single image across three feature-based OOD detection methods, both with and without subspace projection, on the ISIC Ink benchmark (Table 23). The results show that incorporating the subspace projection introduces only a negligible increase in inference time. In contrast, the latency remains substantially lower than that of external generative approaches such as DDPM-MSE.

We then evaluated how setting the number of PC's in the nuisance subspace (k) impacted the OOD detection performance of method Mahalanobis (see Sec 3.4). We evaluated k$\in [2, 5, 10, 20]$, with results plotted in Table 24. The results show that $k = 5$ resulted in the highest performing OOD detection performance for both similar and dissimilar artefacts.

Finally, we applied the subspace projection method on a number of feature-based methods for a ResNet18 primary model (Table 25). Some feature-based methods, such as NuSA and PCX, would not be impacted by projecting out the nuisance subspace due to the design of the method. Our results show that projected features can reduce the Invisible Gorilla Effect across several feature-based methods (reducing the gap in OOD detection performance between similar and dissimilar artefacts), but not for all methods (e.g. XOOD-M).

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. Number: 9 Publisher: Nature Publishing Group. 6

[2] Harry Anthony and Konstantinos Kamnitsas. Evaluating Reliability in Medical DNNs: A Critical Analysis of Feature and Confidence-Based OOD Detection. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 160–170, Cham, 2025. Springer Nature Switzerland. 7

[3] Reza Averly and Wei-Lun Chao. Unified out-of-distribution detection: A model-specific perspective. In *Proceedings of*

Table 14. OOD detection **AURC** (%) results for the ISIC benchmark, using a **ResNet18** primary model for internal methods. Detection was evaluated on artefacts that are either visually similar to skin lesions (red for ink artefacts; red, orange, yellow for colour charts) or visually dissimilar (green, purple, black for ink artefacts; green, blue, black, grey for colour charts). Methods are grouped into two groups: feature-based and confidence-based. Each entry shows the best-performing hyperparameter setting, reported as the mean AURC over 25 seeds, with 95% confidence intervals in brackets. For AURC, lower values indicate better performance.

OOD Method	**Ink Artefacts AURC (↓)**		**Colour Chart Artefacts AURC (↓)**	
	Similar	Dissimilar	Similar	Dissimilar
Feature-based Methods				
CoP	0.11 (0.01)	17.34 (0.01)	2.73 (0.01)	4.28 (0.01)
CoRP	0.14 (0.03)	17.39 (0.01)	3.03 (0.01)	4.71 (0.01)
FeatureNorm	0.15 (0.01)	26.01 (0.03)	1.53 (0.01)	2.11 (0.02)
GRAM	0.12 (0.01)	14.91 (0.01)	8.78 (0.01)	11.79 (0.01)
KDE (Gaussian)	0.18 (0.01)	18.74 (0.01)	2.12 (0.01)	3.14 (0.01)
KNN	0.06 (0.01)	14.76 (0.01)	0.08 (0.01)	0.13 (0.01)
LOF	0.11 (0.01)	19.35 (0.01)	1.70 (0.01)	2.73 (0.01)
Mahalanobis	0.10 (0.01)	17.51 (0.01)	1.23 (0.01)	1.98 (0.01)
MBM	0.09 (0.01)	17.01 (0.01)	1.12 (0.01)	1.82 (0.01)
NAN	0.16 (0.01)	26.14 (0.01)	6.50 (0.02)	9.39 (0.02)
NAC	0.44 (0.01)	32.94 (0.01)	6.52 (0.02)	8.43 (0.03)
NMD	0.42 (0.01)	31.17 (0.01)	7.87 (0.08)	9.40 (0.08)
NuSA	0.12 (0.01)	12.85 (0.03)	10.02 (0.02)	13.71 (0.03)
PCX	0.11 (0.01)	19.41 (0.01)	0.58 (0.01)	0.86 (0.01)
Residual	0.35 (0.01)	28.13 (0.01)	1.78 (0.01)	2.40 (0.01)
TAPUUD	0.14 (0.01)	14.36 (0.01)	5.40 (0.08)	6.27 (0.08)
XOOD-M	0.13 (0.01)	17.06 (0.04)	3.15 (0.01)	4.94 (0.01)
Confidence-based Methods				
ASH	0.17 (0.01)	14.43 (0.01)	5.22 (0.01)	7.27 (0.01)
Deep Ensemble	0.14 (0.01)	12.99 (0.01)	7.50 (0.01)	10.06 (0.01)
DICE	0.15 (0.01)	14.29 (0.01)	9.03 (0.01)	12.09 (0.01)
GAIA-A	0.21 (0.01)	19.80 (0.01)	13.10 (0.01)	17.54 (0.01)
GradNorm	0.14 (0.01)	14.74 (0.01)	10.82 (0.08)	13.36 (0.07)
GradOrth	0.13 (0.01)	14.04 (0.01)	8.98 (0.01)	12.47 (0.01)
MCP	0.17 (0.01)	15.74 (0.01)	9.09 (0.01)	12.27 (0.01)
MC-Dropout	0.31 (0.01)	21.24 (0.01)	10.07 (0.01)	13.26 (0.01)
ODIN	0.11 (0.01)	13.82 (0.01)	2.47 (0.01)	4.09 (0.01)
ReAct	0.21 (0.01)	20.94 (0.04)	11.99 (0.02)	14.91 (0.02)
SHE	0.14 (0.01)	14.01 (0.01)	9.31 (0.01)	12.77 (0.01)
ViM	0.11 (0.01)	12.92 (0.01)	2.31 (0.01)	2.41 (0.01)
WeiPer	0.12 (0.01)	12.30 (0.01)	2.32 (0.01)	2.44 (0.01)

the IEEE/CVF International Conference on Computer Vision, pages 1453–1463, 2023. 8

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10 (7):e0130140, 2015. 6

[5] Ricardo Baptista, Agnimitra Dasgupta, Nikola B. Kovachki, Assad Oberai, and Andrew M. Stuart. Memorization and Regularization in Generative Diffusion Models, 2025. arXiv:2501.15785 [cs]. 1

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. 1

[7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 1

[8] C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 2003. 1

[9] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020. 6

[10] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M

Table 15. OOD detection **FPR@TPR80** (%) results for the ISIC benchmark, using a **ResNet18** primary model for internal methods. Detection was evaluated on artefacts that are either visually similar to skin lesions (red for ink artefacts; red, orange, yellow for colour charts) or visually dissimilar (green, purple, black for ink artefacts; green, blue, black, grey for colour charts). Methods are grouped into two groups: feature-based and confidence-based. Each entry shows the best-performing hyperparameter setting, reported as the mean FPR@TPR80 over 25 seeds, with 95% confidence intervals in brackets. For FPR@TPR80, lower values indicate better performance.

OOD Method	**Ink Artefacts FPR@TPR80 (↓)**		**Colour Chart Artefacts FPR@TPR80 (↓)**	
	Similar	Dissimilar	Similar	Dissimilar
Feature-based Methods				
CoP	55.30 (0.01)	61.85 (0.01)	22.06 (0.02)	28.34 (0.03)
CoRP	56.91 (0.03)	61.91 (0.01)	26.16 (0.03)	32.39 (0.03)
FeatureNorm	34.91 (0.02)	77.85 (0.03)	60.44 (0.02)	64.73 (0.04)
GRAM	49.98 (0.01)	52.08 (0.01)	73.44 (0.01)	76.19 (0.01)
KDE (Gaussian)	45.45 (0.01)	65.60 (0.01)	41.79 (0.01)	49.13 (0.01)
KNN	24.05 (0.01)	56.67 (0.01)	10.07 (0.01)	11.44 (0.01)
LOF	36.18 (0.01)	64.98 (0.01)	8.70 (0.01)	12.85 (0.01)
Mahalanobis	54.18 (0.01)	63.61 (0.01)	0.54 (0.01)	1.35 (0.01)
MBM	52.91 (0.01)	59.46 (0.01)	0.53 (0.01)	1.24 (0.01)
NAN	31.64 (0.02)	84.62 (0.03)	46.42 (0.07)	54.48 (0.07)
NAC	87.09 (0.01)	90.70 (0.01)	80.65 (0.01)	81.40 (0.01)
NMD	83.33 (0.02)	87.69 (0.03)	33.79 (0.03)	34.77 (0.03)
NuSA	47.64 (0.05)	47.95 (0.03)	82.99 (0.02)	84.43 (0.02)
PCX	55.09 (0.01)	59.17 (0.01)	3.29 (0.01)	4.53 (0.01)
Residual	84.55 (0.01)	83.27 (0.01)	5.73 (0.01)	6.29 (0.01)
TAPUUD	52.08 (0.01)	54.41 (0.01)	3.57 (0.01)	7.40 (0.01)
XOOD-M	30.00 (0.03)	54.20 (0.04)	18.26 (0.03)	23.92 (0.06)
Confidence-based Methods				
ASH	54.55 (0.01)	55.64 (0.01)	44.46 (0.05)	47.51 (0.04)
Deep Ensemble	48.18 (0.01)	49.19 (0.01)	72.93 (0.01)	73.86 (0.01)
DICE	61.46 (0.01)	61.50 (0.01)	73.79 (0.01)	75.69 (0.01)
GAIA-A	55.27 (0.02)	63.54 (0.02)	82.29 (0.03)	83.70 (0.03)
GradNorm	52.91 (0.03)	53.97 (0.02)	41.49 (0.02)	47.94 (0.02)
GradOrth	51.64 (0.01)	52.36 (0.01)	72.96 (0.03)	74.75 (0.02)
MCP	52.00 (0.01)	52.31 (0.01)	73.01 (0.03)	74.76 (0.02)
MC-Dropout	61.82 (0.01)	68.93 (0.01)	75.89 (0.01)	76.98 (0.01)
ODIN	48.00 (0.01)	52.30 (0.02)	17.45 (0.01)	24.59 (0.01)
ReAct	63.45 (0.02)	70.45 (0.03)	78.32 (0.05)	78.47 (0.05)
SHE	51.64 (0.02)	53.55 (0.01)	75.54 (0.03)	77.00 (0.02)
ViM	50.45 (0.01)	51.55 (0.01)	62.33 (0.01)	64.81 (0.01)
WeiPer	50.65 (0.01)	50.64 (0.01)	61.64 (0.01)	62.12 (0.01)

Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021. 6

[11] Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2948–2957, 2023. 2

[12] Joris Guérin, Kevin Delmas, Raul Sena Ferreira, and Jérémie Guiochet. Out-Of-Distribution Detection Is Not All You Need, 2023. arXiv:2211.16158 [cs, eess]. 8

[13] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018. 1

[14] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. 1

[15] Peng-Tao Jiang, Changlin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 6

[16] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*,

Table 16. OOD detection AUROC (%) results for the ISIC benchmark, using a **ViT-B/32** primary model for internal methods. Detection was evaluated on artefacts that are either visually similar to skin lesions (red for ink artefacts; red, orange, yellow for colour charts) or visually dissimilar (green, purple, black for ink artefacts; green, blue, black, grey for colour charts). Methods are grouped into two groups: feature-based and confidence-based. Each entry shows the best-performing hyperparameter setting, reported as the mean AUROC over 25 seeds, with 95% confidence intervals in brackets.

OOD Method	**Ink Artefacts**		**Colour Chart Artefacts**	
	Similar	Dissimilar	Similar	Dissimilar
Feature-based Methods				
CoP	77.43 (0.02)	72.08 (0.01)	91.09 (0.01)	87.99 (0.02)
CoRP	77.51 (0.03)	70.05 (0.01)	89.35 (0.01)	85.52 (0.02)
FeatureNorm	68.37 (0.03)	52.33 (0.03)	86.06 (0.01)	76.39 (0.04)
GRAM	63.54 (0.03)	52.68 (0.01)	78.90 (0.03)	77.92 (0.03)
KDE (Gaussian)	73.89 (0.04)	64.14 (0.03)	85.72 (0.03)	84.27 (0.01)
KNN	82.22 (0.01)	70.33 (0.01)	91.48 (0.02)	91.00 (0.01)
LOF	91.29 (0.01)	76.50 (0.01)	92.86 (0.02)	92.71 (0.01)
Mahalanobis	79.49 (0.02)	69.39 (0.02)	91.46 (0.02)	91.18 (0.02)
MBM	81.20 (0.01)	68.50 (0.01)	91.58 (0.02)	91.48 (0.02)
NAN	53.07 (0.03)	52.02 (0.01)	55.12 (0.02)	54.61 (0.02)
NAC	45.86 (0.02)	45.32 (0.02)	57.25 (0.02)	56.13 (0.02)
NMD	57.88 (0.05)	57.53 (0.02)	56.68 (0.08)	56.26 (0.03)
NuSA	69.97 (0.04)	68.36 (0.02)	67.14 (0.04)	67.11 (0.01)
PCX	60.48 (0.04)	56.79 (0.01)	55.81 (0.02)	55.47 (0.02)
Residual	80.72 (0.02)	73.86 (0.01)	91.50 (0.01)	90.67 (0.01)
TAPUUD	57.70 (0.02)	55.27 (0.02)	68.64 (0.02)	67.93 (0.03)
XOOD-M	65.35 (0.04)	56.71 (0.01)	86.79 (0.02)	86.09 (0.04)
Confidence-based Methods				
ASH	69.49 (0.03)	70.15 (0.01)	61.68 (0.02)	61.74 (0.02)
Deep Ensemble	75.48 (0.02)	72.58 (0.01)	72.02 (0.01)	70.83 (0.02)
DICE	67.92 (0.01)	67.71 (0.01)	60.81 (0.02)	60.30 (0.02)
GAIA-A	34.78 (0.06)	31.19 (0.01)	34.20 (0.02)	37.18 (0.02)
GradNorm	71.19 (0.03)	70.70 (0.01)	64.94 (0.03)	63.53 (0.04)
GradOrth	68.28 (0.02)	70.20 (0.01)	59.82 (0.03)	62.29 (0.02)
MCP	69.18 (0.02)	69.08 (0.01)	58.72 (0.02)	58.64 (0.02)
MC-Dropout	72.32 (0.02)	72.14 (0.01)	60.23 (0.03)	59.34 (0.02)
ODIN	71.33 (0.03)	70.35 (0.01)	61.29 (0.03)	61.40 (0.02)
ReAct	67.89 (0.04)	66.16 (0.03)	58.96 (0.04)	57.17 (0.05)
SHE	69.98 (0.03)	69.63 (0.01)	58.67 (0.03)	58.87 (0.02)
ViM	70.65 (0.03)	70.14 (0.01)	59.91 (0.03)	60.74 (0.02)
WeiPer	69.98 (0.02)	66.31 (0.02)	61.26 (0.02)	60.57 (0.02)

33:20578–20589, 2020. 2

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. pages 4015–4026, 2023. 6

[18] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 1

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 1

[20] M.R. Luo, G. Cui, and B. Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26(5):340–350, 2001. 6

[21] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022. 1

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[23] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. 1

Table 17. OOD detection AUROC (%) results for the ISIC benchmark, using a **VGG16** primary model for internal methods. Detection was evaluated on artefacts that are either visually similar to skin lesions (red for ink artefacts; red, orange, yellow for colour charts) or visually dissimilar (green, purple, black for ink artefacts; green, blue, black, grey for colour charts). Methods are grouped into two groups: feature-based and confidence-based. Each entry shows the best-performing hyperparameter setting, reported as the mean AUROC over 25 seeds, with 95% confidence intervals in brackets.

OOD Method	**Ink Artefacts**		**Colour Chart Artefacts**	
	Similar	Dissimilar	Similar	Dissimilar
Feature-based Methods				
CoP	64.07 (0.01)	57.30 (0.01)	54.70 (0.02)	54.38 (0.02)
CoRP	63.63 (0.01)	56.07 (0.01)	54.63 (0.02)	54.16 (0.02)
FeatureNorm	82.48 (0.01)	56.18 (0.01)	60.49 (0.02)	59.88 (0.03)
GRAM	68.25 (0.04)	55.25 (0.03)	59.32 (0.02)	59.27 (0.02)
KDE (Gaussian)	80.71 (0.01)	55.74 (0.02)	60.10 (0.02)	58.37 (0.02)
KNN	83.67 (0.01)	62.11 (0.01)	81.32 (0.01)	74.24 (0.03)
LOF	85.28 (0.01)	64.28 (0.01)	82.38 (0.04)	82.11 (0.02)
Mahalanobis	80.02 (0.01)	52.41 (0.01)	74.38 (0.03)	74.10 (0.02)
MBM	81.33 (0.01)	52.06 (0.01)	75.39 (0.02)	75.01 (0.02)
NAN	77.25 (0.01)	48.08 (0.01)	45.54 (0.03)	44.71 (0.03)
NAC	51.52 (0.02)	45.23 (0.01)	55.63 (0.02)	55.34 (0.02)
NMD	60.13 (0.04)	58.24 (0.05)	37.50 (0.02)	37.29 (0.02)
NuSA	60.27 (0.06)	58.03 (0.05)	82.91 (0.02)	82.82 (0.02)
PCX	78.33 (0.04)	63.15 (0.05)	71.23 (0.02)	70.99 (0.02)
Residual	42.64 (0.01)	42.75 (0.01)	60.19 (0.02)	59.13 (0.02)
TAPUUD	62.47 (0.03)	53.79 (0.03)	61.17 (0.01)	60.69 (0.01)
XOOD-M	80.77 (0.01)	58.62 (0.01)	78.94 (0.02)	76.37 (0.02)
Confidence-based Methods				
ASH	65.03 (0.01)	65.87 (0.01)	41.90 (0.02)	41.88 (0.01)
Deep Ensemble	65.72 (0.06)	65.12 (0.04)	47.70 (0.01)	47.75 (0.02)
DICE	72.92 (0.01)	72.22 (0.01)	40.94 (0.02)	40.44 (0.01)
GAIA-A	48.85 (0.01)	34.86 (0.03)	42.75 (0.01)	42.29 (0.02)
GradNorm	62.24 (0.01)	62.13 (0.01)	47.83 (0.02)	46.86 (0.02)
GradOrth	65.76 (0.01)	65.99 (0.01)	43.26 (0.02)	42.35 (0.02)
MCP	60.12 (0.01)	59.82 (0.01)	45.45 (0.01)	45.04 (0.01)
MC-Dropout	61.34 (0.01)	61.42 (0.01)	46.17 (0.01)	46.09 (0.02)
ODIN	67.89 (0.01)	66.03 (0.01)	58.94 (0.02)	57.88 (0.02)
ReAct	72.81 (0.01)	72.60 (0.01)	50.95 (0.03)	50.78 (0.04)
SHE	65.37 (0.01)	64.64 (0.01)	42.20 (0.02)	41.43 (0.02)
ViM	66.43 (0.01)	66.12 (0.02)	42.47 (0.02)	42.60 (0.02)
WeiPer	59.30 (0.01)	52.34 (0.01)	53.47 (0.02)	53.19 (0.02)

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 6

[26] Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, et al. Detecting outliers with foreign patch interpolation. *Machine Learning for Biomedical Imaging*, 1 (April 2022 issue):1–27, 2022. 2

[27] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023. 7

[28] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Procedings of the British Machine Vision Conference 2016*, pages 87–1. British Machine Vision Association, 2016. 1

[29] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023. 8

[30] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE*

Table 18. OOD detection AUROC (%) results for the MVTec benchmark, using a **ResNet18** primary model for internal methods. Detection was evaluated on artefacts that are either visually similar to the model's ROI (red for pill; black for metal nut) or visually dissimilar (yellow for ink artefacts; blue for metal nut). Methods are grouped into two groups: feature-based and confidence-based. Each entry shows the best-performing hyperparameter setting, reported as the mean AUROC over 25 seeds, with 95% confidence intervals in brackets.

OOD Method	Pill		Metal Nut	
	Similar	Dissimilar	Similar	Dissimilar
Feature-based Methods				
CoP	88.27 (0.01)	81.15 (0.02)	53.07 (0.01)	39.20 (0.03)
CoRP	81.92 (0.01)	71.41 (0.02)	48.86 (0.01)	39.09 (0.03)
FeatureNorm	71.67 (0.02)	56.41 (0.02)	55.00 (0.01)	52.69 (0.02)
GRAM	74.76 (0.01)	69.83 (0.01)	68.94 (0.01)	64.70 (0.01)
KDE (Gaussian)	81.35 (0.02)	73.59 (0.02)	63.52 (0.01)	43.98 (0.01)
KNN	93.33 (0.01)	86.15 (0.01)	71.02 (0.01)	36.93 (0.01)
LOF	83.78 (0.01)	62.18 (0.01)	69.30 (0.02)	42.95 (0.02)
Mahalanobis	71.86 (0.01)	68.72 (0.01)	69.77 (0.01)	58.30 (0.01)
MBM	73.18 (0.01)	71.36 (0.01)	70.32 (0.02)	58.12 (0.01)
NAN	64.10 (0.02)	56.92 (0.01)	54.43 (0.01)	52.39 (0.03)
NAC	68.27 (0.02)	67.88 (0.01)	52.38 (0.01)	51.93 (0.02)
NMD	85.26 (0.01)	82.31 (0.01)	68.75 (0.01)	68.94 (0.01)
NuSA	80.64 (0.01)	73.21 (0.01)	58.41 (0.01)	44.09 (0.02)
PCX	64.10 (0.02)	56.41 (0.02)	51.42 (0.01)	49.15 (0.01)
Residual	86.09 (0.01)	68.97 (0.02)	78.30 (0.03)	70.91 (0.03)
TAPUUD	53.14 (0.03)	49.87 (0.04)	59.27 (0.01)	58.30 (0.02)
XOOD-M	75.71 (0.02)	66.67 (0.01)	67.16 (0.01)	66.93 (0.01)
Confidence-based Methods				
ASH	81.35 (0.02)	80.64 (0.01)	57.84 (0.01)	58.75 (0.02)
Deep Ensemble	79.54 (0.01)	79.51 (0.01)	64.72 (0.01)	62.75 (0.01)
DICE	87.63 (0.01)	82.56 (0.01)	63.52 (0.01)	48.98 (0.02)
GAIA-A	36.47 (0.01)	32.18 (0.01)	59.43 (0.02)	42.61 (0.02)
GradNorm	80.13 (0.01)	79.07 (0.01)	60.34 (0.01)	59.83 (0.01)
GradOrth	86.86 (0.01)	78.59 (0.01)	65.45 (0.01)	53.00 (0.02)
MCP	78.46 (0.02)	78.33 (0.01)	58.75 (0.01)	45.34 (0.01)
MC-Dropout	79.57 (0.02)	79.04 (0.01)	60.74 (0.02)	60.83 (0.02)
ODIN	82.69 (0.01)	80.77 (0.01)	58.75 (0.01)	53.18 (0.02)
ReAct	83.72 (0.01)	80.00 (0.01)	59.43 (0.01)	45.91 (0.03)
SHE	81.09 (0.01)	80.02 (0.01)	57.50 (0.01)	46.02 (0.01)
ViM	83.21 (0.01)	80.77 (0.01)	57.27 (0.02)	43.75 (0.02)
WeiPer	80.77 (0.02)	80.29 (0.01)	67.91 (0.01)	66.02 (0.01)

transactions on pattern analysis and machine intelligence, 45(4):4396–4415, 2022. 7

Table 19. OOD detection AUROC (%) results for the MVTec benchmark, using a **ViT-B/32** primary model for internal methods. Detection was evaluated on artefacts that are either visually similar to the model's ROI (red for pill; black for metal nut) or visually dissimilar (yellow for ink artefacts; blue for metal nut). Methods are grouped into two groups: feature-based and confidence-based. Each entry shows the best-performing hyperparameter setting, reported as the mean AUROC over 25 seeds, with 95% confidence intervals in brackets.

OOD Method	**Pill**		**Metal Nut**	
	Similar	Dissimilar	Similar	Dissimilar
Feature-based Methods				
CoP	72.76 (0.01)	66.92 (0.01)	78.79 (0.02)	42.33 (0.02)
CoRP	69.87 (0.01)	64.38 (0.01)	75.19 (0.03)	41.86 (0.02)
FeatureNorm	66.03 (0.02)	62.49 (0.02)	85.23 (0.02)	68.75 (0.01)
GRAM	57.55 (0.03)	54.83 (0.02)	78.23 (0.02)	52.54 (0.02)
KDE (Gaussian)	66.35 (0.01)	48.85 (0.01)	82.39 (0.02)	70.80 (0.02)
KNN	77.95 (0.02)	71.28 (0.02)	83.18 (0.02)	78.75 (0.01)
LOF	63.01 (0.01)	57.69 (0.01)	81.59 (0.01)	79.66 (0.01)
Mahalanobis	70.28 (0.02)	65.22 (0.02)	77.70 (0.01)	67.84 (0.02)
MBM	70.09 (0.03)	66.27 (0.02)	78.21 (0.01)	66.90 (0.01)
NAN	60.90 (0.01)	58.85 (0.02)	68.64 (0.01)	57.39 (0.02)
NAC	48.29 (0.02)	43.21 (0.02)	54.54 (0.02)	52.13 (0.03)
NMD	67.37 (0.01)	64.87 (0.02)	70.57 (0.03)	64.43 (0.02)
NuSA	70.64 (0.01)	59.10 (0.02)	58.07 (0.01)	40.23 (0.01)
PCX	65.87 (0.02)	63.15 (0.02)	64.02 (0.03)	57.20 (0.03)
Residual	66.03 (0.01)	63.72 (0.01)	74.77 (0.04)	70.11 (0.02)
TAPUUD	68.90 (0.02)	65.64 (0.02)	65.11 (0.02)	53.64 (0.01)
XOOD-M	70.08 (0.03)	68.73 (0.02)	79.97 (0.03)	53.27 (0.02)
Confidence-based Methods				
ASH	61.25 (0.02)	60.51 (0.01)	62.84 (0.01)	61.70 (0.01)
Deep Ensemble	66.43 (0.02)	64.75 (0.02)	70.14 (0.04)	63.48 (0.02)
DICE	58.21 (0.02)	51.54 (0.02)	71.82 (0.04)	66.93 (0.05)
GAIA-A	52.05 (0.02)	44.29 (0.02)	37.16 (0.02)	35.80 (0.01)
GradNorm	65.94 (0.03)	62.46 (0.04)	71.75 (0.03)	69.39 (0.02)
GradOrth	59.46 (0.03)	59.02 (0.02)	69.03 (0.02)	67.99 (0.03)
MCP	63.72 (0.01)	55.26 (0.01)	66.14 (0.01)	62.16 (0.02)
MC-Dropout	66.27 (0.02)	54.96 (0.02)	68.56 (0.02)	66.47 (0.02)
ODIN	63.97 (0.01)	63.91 (0.01)	71.25 (0.04)	62.16 (0.02)
ReAct	54.36 (0.02)	34.42 (0.02)	63.18 (0.01)	56.59 (0.01)
SHE	70.26 (0.02)	59.04 (0.02)	58.30 (0.01)	44.09 (0.02)
ViM	55.45 (0.02)	37.44 (0.02)	63.86 (0.01)	57.61 (0.01)
WeiPer	64.04 (0.02)	62.18 (0.02)	66.02 (0.01)	62.27 (0.02)

Table 20. OOD detection AUROC (%) results for the MVTec benchmark, using a **VGG16** primary model for internal methods. Detection was evaluated on artefacts that are either visually similar to the model's ROI (red for pill; black for metal nut) or visually dissimilar (yellow for ink artefacts; blue for metal nut). Methods are grouped into two groups: feature-based and confidence-based. Each entry shows the best-performing hyperparameter setting, reported as the mean AUROC over 25 seeds, with 95% confidence intervals in brackets.

OOD Method	**Pill**		**Metal Nut**	
	Similar	Dissimilar	Similar	Dissimilar
Feature-based Methods				
CoP	76.68 (0.04)	69.07 (0.03)	68.18 (0.02)	53.52 (0.02)
CoRP	69.42 (0.02)	65.77 (0.02)	63.63 (0.02)	50.90 (0.04)
FeatureNorm	52.28 (0.02)	51.15 (0.01)	57.95 (0.02)	45.45 (0.04)
GRAM	64.90 (0.03)	60.19 (0.03)	72.46 (0.04)	65.32 (0.03)
KDE (Gaussian)	63.85 (0.03)	54.74 (0.02)	69.89 (0.04)	58.75 (0.01)
KNN	91.28 (0.03)	82.82 (0.02)	80.57 (0.03)	74.89 (0.03)
LOF	71.79 (0.01)	68.72 (0.02)	70.00 (0.02)	63.52 (0.02)
Mahalanobis	71.15 (0.01)	69.78 (0.02)	60.80 (0.01)	54.89 (0.04)
MBM	72.03 (0.01)	69.58 (0.02)	61.28 (0.02)	55.76 (0.02)
NAN	51.77 (0.06)	50.26 (0.02)	53.98 (0.04)	46.48 (0.03)
NAC	52.10 (0.03)	50.08 (0.02)	43.53 (0.04)	41.42 (0.03)
NMD	60.51 (0.02)	57.95 (0.01)	82.16 (0.03)	74.09 (0.02)
NuSA	55.64 (0.02)	49.10 (0.04)	61.16 (0.02)	58.92 (0.03)
PCX	68.56 (0.02)	66.29 (0.03)	74.97 (0.02)	64.52 (0.02)
Residual	79.87 (0.01)	76.09 (0.04)	85.11 (0.02)	77.95 (0.02)
TAPUUD	59.36 (0.02)	58.46 (0.02)	52.36 (0.02)	48.30 (0.02)
XOOD-M	74.08 (0.03)	68.84 (0.03)	74.20 (0.04)	54.89 (0.2)
Confidence-based Methods				
ASH	60.45 (0.02)	60.13 (0.02)	70.80 (0.02)	66.36 (0.03)
Deep Ensemble	60.32 (0.02)	60.62 (0.03)	76.77 (0.03)	75.34 (0.03)
DICE	57.05 (0.03)	53.72 (0.02)	75.91 (0.02)	69.66 (0.01)
GAIA-A	48.33 (0.01)	41.73 (0.02)	55.57 (0.02)	45.57 (0.02)
GradNorm	61.94 (0.03)	58.42 (0.02)	63.14 (0.02)	59.98 (0.02)
GradOrth	64.22 (0.02)	60.18 (0.03)	59.09 (0.03)	51.70 (0.03)
MCP	57.95 (0.02)	53.59 (0.02)	73.98 (0.04)	71.07 (0.03)
MC-Dropout	58.32 (0.03)	55.31 (0.02)	72.89 (0.02)	71.30 (0.02)
ODIN	58.46 (0.01)	53.78 (0.02)	78.07 (0.01)	73.98 (0.03)
ReAct	54.87 (0.02)	51.79 (0.02)	75.23 (0.04)	66.48 (0.01)
SHE	56.54 (0.02)	54.10 (0.02)	67.05 (0.02)	52.84 (0.02)
ViM	85.90 (0.01)	76.18 (0.03)	75.91 (0.03)	69.89 (0.01)
WeiPer	58.85 (0.01)	53.59 (0.02)	75.34 (0.03)	74.32 (0.04)

Table 21. OOD AUROC (%) on the ISIC ink benchmark using 10 **ResNet18** primary models, trained with either light or heavy **colour jitter** augmentations. Brackets show the OOD AUROC difference compared to the primary model's trained without colour jitter augmentations (Table 2, main paper).

OOD Method	Light Augmentation		Heavy Augmentation	
	Similar	Dissimilar	Similar	Dissimilar
Feature-based Methods				
CoP	77.04 (+5.30)	70.34 (+4.58)	79.58 (+7.84)	69.81 (+4.05)
CoRP	74.60 (+3.44)	68.72 (+3.74)	74.33 (+3.17)	67.00 (+2.02)
FeatureNorm	75.92 (+0.80)	64.24 (+11.33)	75.66 (+0.54)	68.09 (+15.18)
GRAM	67.25 (-13.07)	59.24 (-13.47)	66.83 (-13.49)	61.21 (-11.50)
KDE (Gaussian)	89.70 (+4.15)	75.67 (+7.51)	87.43 (+1.88)	76.89 (+8.73)
KNN	90.13 (+4.45)	77.26 (+7.16)	87.88 (+2.20)	79.23 (+9.13)
LOF	84.12 (+1.77)	74.52 (+11.62)	88.84 (+6.49)	74.73 (+11.83)
Mahalanobis	85.55 (+8.57)	70.92 (+7.28)	87.32 (+10.34)	75.16 (+11.52)
MBM	85.52 (+8.12)	71.39 (+7.68)	87.17 (+9.77)	73.88 (+10.17)
NAN	76.03 (+0.45)	58.88 (+10.42)	73.57 (-2.01)	64.38 (+15.92)
NAC	39.67 (+0.05)	34.50 (-2.82)	36.18 (-3.44)	42.66 (+5.34)
NMD	72.78 (-6.53)	70.68 (-3.05)	75.08 (-4.23)	74.22 (+0.49)
NuSA	64.02 (-11.00)	71.26 (-3.71)	58.45 (-16.57)	72.52 (-2.45)
PCX	77.34 (+1.73)	68.93 (+4.18)	75.24 (-0.37)	66.18 (+1.43)
Residual	60.22 (-5.78)	58.69 (+0.42)	69.53 (+3.53)	64.29 (+6.02)
TAPUUD	79.66 (+8.87)	67.34 (+10.33)	80.22 (+9.43)	69.30 (+12.29)
XOOD-M	84.14 (+3.38)	76.19 (+12.28)	83.92 (+3.16)	76.12 (+12.21)
Confidence-based Methods				
ASH	53.29 (-19.12)	43.48 (-29.54)	57.18 (-15.23)	42.76 (-30.26)
Deep Ensemble	64.13 (-8.93)	73.38 (+0.96)	59.91 (-13.15)	72.19 (-0.23)
DICE	61.82 (-7.22)	71.59 (+0.29)	58.14 (-10.90)	71.56 (+0.26)
GAIA-A	49.24 (-19.69)	52.93 (-12.62)	47.23 (-21.70)	49.65 (-15.90)
GradNorm	62.81 (-12.82)	71.27 (-1.16)	69.09 (-6.54)	75.61 (+3.18)
GradOrth	63.76 (-9.04)	71.88 (-0.86)	58.43 (-14.37)	72.06 (-0.68)
MCP	63.50 (-6.33)	71.80 (+3.06)	58.42 (-11.41)	71.65 (+2.91)
MC-Dropout	68.34 (-2.08)	72.81 (+3.60)	58.23 (-12.19)	71.03 (+1.82)
ODIN	63.70 (-9.06)	72.24 (-0.12)	58.42 (-14.34)	71.65 (-0.71)
ReAct	62.84 (-1.33)	70.46 (+10.96)	55.62 (-8.55)	67.72 (+8.22)
SHE	62.68 (-9.52)	71.02 (-1.34)	57.79 (-14.41)	72.58 (+0.22)
ViM	63.36 (-11.93)	71.43 (-3.09)	61.28 (-14.01)	73.67 (-0.85)
WeiPer	65.74 (-8.85)	66.23 (-7.54)	72.31 (-2.28)	73.00 (-0.77)

Table 22. OOD AUROC (%) on the ISIC ink benchmark using 10 **ViT-B/32** primary models, trained with light **colour jitter** augmentations. Brackets show the OOD AUROC difference compared to the primary model's trained without colour jitter augmentations (Table 16).

OOD Method	Light Augmentation	
	Similar	Dissimilar
Feature-based Methods		
CoP	87.82 (+10.39)	72.23 (+0.15)
CoRP	80.95 (+3.44)	69.28 (-0.77)
FeatureNorm	82.09 (+13.72)	73.27 (+20.94)
GRAM	44.24 (-19.30)	44.65 (-8.03)
KDE (Gaussian)	85.33 (+11.44)	75.05 (+10.91)
KNN	86.76 (+4.54)	79.98 (+9.65)
LOF	93.95 (+2.66)	81.13 (+4.63)
Mahalanobis	85.06 (+5.57)	79.63 (+10.24)
MBM	84.56 (+3.36)	79.21 (+10.71)
NAN	56.24 (+3.17)	55.84 (+3.82)
NAC	43.19 (-2.67)	45.62 (+0.30)
NMD	80.99 (+23.11)	74.81 (+17.28)
NuSA	52.14 (-17.83)	71.97 (+3.61)
PCX	63.40 (+2.92)	59.76 (+2.97)
Residual	96.80 (+16.08)	86.46 (+12.60)
TAPUUD	90.20 (+32.50)	75.35 (+20.08)
XOOD-M	86.33 (+20.98)	68.82 (+12.11)
Confidence-based Methods		
ASH	51.61 (-17.88)	64.85 (-5.30)
Deep Ensemble	49.72 (-25.76)	70.04 (-2.54)
DICE	76.39 (+8.47)	66.25 (-1.46)
GAIA-A	49.46 (+14.68)	30.10 (-1.09)
GradNorm	57.27 (-13.92)	69.96 (-0.74)
GradOrth	77.28 (+9.00)	73.72 (+3.52)
MCP	48.82 (-20.36)	69.40 (+0.32)
MC-Dropout	58.28 (-14.04)	58.60 (-13.54)
ODIN	48.82 (-22.51)	70.43 (+0.08)
ReAct	38.80 (-29.09)	53.37 (-12.79)
SHE	57.67 (-12.31)	75.48 (+5.85)
ViM	39.83 (-30.82)	57.24 (-12.90)
WeiPer	73.18 (+3.20)	70.95 (+4.64)

Table 23. Average inference latency (in milliseconds) for a single image on the ISIC Ink benchmark on a ResNet18 primary model. Results are shown for three feature-based OOD detection methods with and without the proposed subspace projection, showing small increases in inference latency. The forward pass through the model is included in this value. The inference latency of the external method baseline (DDPM-MSE) is included to highlight that some OOD detection methods have significantly higher computational cost.

Method	**Inference Latency (ms)**
Mahalanobis	12.15
+ proj.	14.23
Featurenorm	8.63
+ proj.	10.32
NaN	11.04
+ proj.	13.09
DDPM-MSE	4767.90

Table 24. Average OOD detection AUROC (%) for the Mahalanobis method as a function of the nuisance subspace dimensionality k. Removing a small number of nuisance directions (up to $k = 5$) improves performance for both similar and dissimilar artefacts, whereas removing larger subspaces ($k \geq 10$) removes task-relevant structure and substantially degrades AUROC.

OOD Artefact	**k=2**	**k=5**	**k=10**	**k=20**
Similar	77.2	77.5	53.7	48.2
Dissimilar	67.0	75.8	59.4	47.0

Table 25. OOD performance on ISIC ink and colour chart benchmarks with and without nuisance subspace projection (equation 3), averaged over 25 ResNet18 models with random seeds. Using projected features can reduce the Invisible Gorilla Effect across several feature based OOD detection methods, although not for all methods (e.g. XOOD-M).

Method	**Ink Artefacts**		**Colour Charts**	
	Sim.	Diss.	Sim.	Diss.
Mahalanobis	77.0	63.6	96.7	95.4
+Proj.	77.5	75.8	95.8	97.7
MBM	77.4	63.7	97.0	95.5
+Proj.	77.6	75.3	96.1	97.2
FeatureNorm	75.1	52.9	62.4	58.1
+Proj.	75.3	74.5	75.9	76.3
NAN	75.6	48.5	72.5	68.1
+Proj.	75.3	76.8	77.9	77.4
XOOD-M	80.8	63.9	88.5	84.8
+Proj.	68.5	61.0	61.2	59.8