OpenMIBOOD: Open Medical Imaging Benchmarks for Out-Of-Distribution Detection

Supplementary Material

A. Datasets

This section provides a detailed overview of all datasets used in this work. To facilitate reproducibility, we include preprocessing scripts for each dataset in our public GitHub repository, enabling the transformation of the downloaded datasets into the utilized ID and OOD datasets. For clarity, the fundamental steps executed by these scripts are outlined below.

A.1. MIDOG benchmark

MIDOG [5] The MIDOG dataset consists of 503 whole slide images stained with Hematoxylin & Eosin, a widely used stain for differentiating tissue components and evaluating tissue morphology. The dataset is divided into ten distinct domains, labeled as 1_a, 1_b, 1_c, 2, 3, 4, 5, 6_a, 6_b, and 7. Annotations are provided for mitotic cells and imposter cells. The preprocessing steps outlined in Sec. 3.1 yielded 451 mitotic cell crops, 724 imposter cell crops, and 1153 additional crops extracted from 50 whole slide images within the ID domain 1_a. Each domain's number corresponds to a semantic cell type shift, stemming from seven different cancer types and two species: human and canine. The cancer types include breast carcinoma, lung carcinoma, lymphosarcoma, cutaneous mast cell tumor, neuroendocrine tumor, soft tissue sarcoma, and melanoma. Furthermore, the domains display differing levels of covariate shift caused by variations in imaging hardware and staining protocols. Subscripts are used to indicate domains with multiple sources of covariate shift. Domains 2, 3, 4, 6_a, and 6_b exhibit covariate shifts, whereas domains 5 and 7 do not show apparent covariate shifts, as their images were generated using the same imaging hardware as the ID dataset and originate from the same institute, employing the same staining protocol as the ID set. The whole MIDOG dataset serves as ID (1_a), cs-ID (1_b, 1_c), and near-OOD (2 – 7) datasets.

CCAgT [2, 4] The CCAgT dataset comprises 15 tissue slides stained using the AgNOR technique, labeled alphabetically from 'A' to 'O'. The AgNOR stain specifically targets regions within the cell nucleus, providing insights into distinct cellular properties. Nuclei annotations are available for these slides and were used to generate the same type of image crops as those from the various domains of the ID dataset (Sec. 3.1). This process produced 29 675 crops which are utilized as the first far-OOD dataset.

FNAC 2019 [55] The FNAC 2019 dataset comprises 212 images of human breast tissue samples obtained via fine needle aspiration cytology. Of these, 113 images are classified as malignant, while 99 are labeled as benign. Due to the absence of cell-level annotations, we extract cell crops through a multi-step processing pipeline. First, each image is segmented using a binary threshold with a value of 100. Next, morphological opening with a kernel size of 5 and erosion with a kernel size of 3 are applied to isolate cell clusters. From the resulting processed images, the ten largest clusters are identified, and 50×50 px crops are generated around the centroids of these clusters. The resulting 2088 crops are subsequently utilized as the second far-OOD dataset.

A.2. PhaKIR benchmark

Acknowledgements We thank the creators of the PhaKIR dataset for granting permission to use their dataset ahead of the challenge results' publication.

PhaKIR [53] The PhaKIR dataset consists of eight endoscopic videos of cholecystectomy procedures, with annotations for 19 instrument classes provided as segmentation masks and keypoints for every 25th frame. For this study, only frames containing a single surgical instrument were selected. However, one instrument class, the trocar, is exclusively an access instrument and, therefore, frequently visible alongside other surgical instruments. Consequently, frames showing a trocar in conjunction with a single surgical instrument were also included and assigned the label of the accompanying surgical instrument.

To enhance the object-to-background ratio in the selected frames, frames were excluded if the instrument covered less than 0.5% of the image area or if the distance between the instrument's endpoint and tip was less than 150 px. In the PhaKIR dataset, the tip refers to the part of the instrument that directly contacts the organ, while the endpoint denotes the location where the instrument appears at the image border. Rueckert *et al*. [54] provided annotations for the first four videos of the PhaKIR-Challenge dataset. In this work, we extend these annotations to include Video 05 and Video 07. Following the previously established filtering process, each frame was categorized into three levels of smoke intensity by utilizing the respective annotations – none, medium, and heavy – using the corresponding annotations. The categorization criteria were as follows: None, if no smoke was perceptible; Medium, if

Table 2. Summary of available frames for each instrument class. Video 06 is employed as test data for the official challenge evaluation and therefore not publicly available.

	Video 01	Video 02	Video 03	Video 04	Video 05	Video 07	Sum
Clip-Applicator	63	151	53	22	26	0	315
Grasper	40	13	7	81	52	125	318
PE-Forceps	68	891	72	109	52	42	1234
Needle-Probe	20	27	6	29	12	31	125
Palpation-Probe	18	45	35	187	25	110	420
Suction-Rod	20	96	7	45	37	152	357
No-Instrument	198	483	323	442	166	279	1891

smoke was present but the instrument remained clearly distinguishable; and Heavy, if the instrument was no longer clearly distinguishable. Frames without visible smoke from the first six videos were designated as ID data, while frames containing medium or heavy smoke were used as cs-ID data.

Within the ID dataset, instrument classes with fewer than 80 available training images were excluded, resulting in a final dataset of 2769 frames across six instrument classes (Tab. 2). To prevent an unintended semantic shift, images from excluded instrument classes were also removed from the cs-ID sets.

Cholec80 [66] The Cholec80 dataset comprises 80 endoscopic videos of cholecystectomies, with annotations identifying instrument classes present in every 25th frame. Consistent with the methodology used for PhaKIR, only frames containing a single surgical instrument were selected for analysis.

This dataset's role within the MIB is to evaluate semantic shifts arising from variations in surgical instruments. Consequently, instrument classes that overlap with those in PhaKIR were excluded, yielding a total of 74 049 frames.

Most videos, except for videos 40, 60, 65, and 80, exhibit a pronounced black vignette. To avoid introducing unintended covariate shifts, a rectangular region within the vignette was cropped while maintaining the original aspect ratio of the ID images. The dataset is employed as a near-OOD dataset.

EndoSeg15 [10] The EndoSeg15 dataset from the EndoVis 2015 challenge comprises 160 training images, evenly distributed across four distinct laparoscopic surgeries. Similar to the Cholec80 dataset, the frames in EndoSeg15 often contain black vignettes or borders, which we exclude by extracting rectangular crops with the same aspect ratio as the ID dataset. This crop is carefully positioned within the vignette or usable image content, ensuring the exclusion of black borders. The resulting dataset is then used as a near-OOD dataset.

EndoSeg18 [1] The EndoSeg18 test dataset, part of the EndoVis Challenge 2018, consists of 1000 frames captured during four porcine surgical procedures featuring robotic instruments. No preprocessing was applied to this dataset and it was utilized for near-OOD detection evaluation.

Kvasir-SEG [31] The Kvasir-SEG dataset, an extension of the original Kvasir dataset introduced by Pogorelov *et al*. [50], comprises 1000 images of colorectal polyps, from which ten contain surgical instruments different from those in the ID dataset. Kvasir-SEG is employed as a far-OOD dataset.

CATARACTS [20] The CATARACTS dataset comprises videos of cataract surgeries and includes 21 ophthalmological instruments, which are completely distinct from those in the ID dataset. Given the large size of this dataset, we limit our analysis to the first five videos from the official test split. The initial 43, 203, 130, 29, and 159 frames were excluded from these videos, because these frames contain only black content. Afterwards, the dataset still contains 181 986 usable frames. CATARACTS is utilized as the second far-OOD dataset.

A.3. OASIS-3 benchmark

Acknowledgements Data were provided in part by OASIS-3: Longitudinal Multimodal Neuroimaging: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

OASIS-3 [38] The longitudinal OASIS-3 dataset comprises 2842 MRI scans from 1378 subjects, covering multiple modalities, including T1w and T2w scans. Each scan is labeled with the number of days since the subject's initial visit. Additionally, clinical diagnoses, also timestamped by

days since the initial visit, are provided. However, these timestamps do not precisely align between the clinical diagnoses and MRI scans, requiring the matching of clinical diagnoses to imaging visits as described in Section 3.1.

We categorize the dataset into subjects with clinical diagnoses of cognitively normal (CN) and Alzheimer's disease (AD), using the provided diagnostic labels. Subject 30753 was excluded due to the absence of a diagnosis, and subjects 30937 and 31357 were discarded as they lacked MRI scans. Additionally, MRI scans for which the skull-stripping process failed were excluded. Specifically, this affected T2w scans from subjects 30649, 30724, and 30815, as well as a T1w scan from subject 30339.

In cases where multiple MRI scans were available for a single acquisition timestamp, the scan with the highest index number was selected. MRI scans containing only the hippocampal region were excluded to avoid introducing an unintended domain shift.

For each MRI scan, we reviewed the associated metadata to identify the acquisition device used. One out of eight scanners, namely the *Siemens Vision* device, exhibited a unique orientation and axis configuration that differed from other devices. To ensure consistency with the default orientation of the broader dataset and prevent an unintended domain shift, these images were realigned to match the orientation used by other devices.

T1w MRI scans from all devices, except the Siemens Vision, are designated as ID data. The corresponding T2w MRI scans from these subjects, if available, are labeled as cs-ID data, due to the change in imaging modality. The withheld T1w MRI scans from the Siemens Vision device are labeled as cs-ID, as the covariate shift arises from differences in the acquisition device.

Following these preprocessing steps, the final dataset consists of 944 CN and 288 AD MRI scans, forming the ID dataset.

ATLAS [43] The ATLAS challenge dataset consists of 33 cohorts, each containing multiple subjects with brain lesions resulting from strokes. For this study, we exclusively utilize the T1w MRI scans from the official training split. Cohorts R027, R047, R049, and R050 were excluded from our analysis due to significant quality degradation compared to the remaining cohorts. Consequently, a total of 595 MRI scans were included in the analysis as the first near-OOD dataset.

All selected OASIS-3 and ATLAS MRI scans are preprocessed by resampling to an isotropic voxel spacing of 1 mm^3 and applying skull-stripping using HD-BET [29], version 2.0.1 (official release).

BraTS [6, 7, 47] The BraTS 2023 Glioma challenge dataset includes subjects with large gliomas in the brain. As the data was preprocessed prior to release, including steps such as resampling and skull-stripping, no further preprocessing was necessary. Therefore, the complete official training split, comprising 1251 T1w MRI scans, was used as the second near-OOD dataset.

CT from OASIS-3 [38] In addition to the MRI data, the OASIS-3 dataset includes 1472 low-dose CT scans, which were acquired to perform attenuation correction for PET scans [38]. To emphasize brain tissue, we clipped these scans to a range of 0 – 80 Hounsfield units and subsequently normalized them. The resulting dataset was used as the third near-OOD dataset.

MSD-H [3, 64] The MSD-H dataset consists of 30 MRI scans of the human heart, all acquired during a single cardiac phase using a 3D balanced steady-state free precession acquisition method. These scans were initially employed in a benchmark for left atrium segmentation [64]. The dataset encompasses images of varying quality, ranging from high-resolution scans to those with substantial noise. Utilized as the first far-OOD dataset.

CHAOS [32, 33] The official test split of the CHAOS challenge dataset consists of 20 MRI scans of the abdomen, originally designed for the task of segmenting abdominal organs. The dataset includes both in-phase and out-of-phase images from dual-echo MRI sequences. For the far-OOD evaluation, we use the in-phase scans, as they exhibit stronger visual alignment with the imaging characteristics of the ID dataset, while still maintaining significant anatomical differences.

A.4. Splits

MIDOG The following randomly selected whole-slide image identifiers from MIDOG's ID domain 1_a were utilized for each respective split:

Train: 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24, 28, 29, 30, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50

Validation: 20, 26, 31, 32, 33

Test: 3, 17, 25, 27, 49

The remaining splits of the MIDOG dataset are detailed in Tab. 3. For the CCAgT dataset, the 15 available slides were randomly divided into three validation slides and twelve test slides, corresponding to a 20 – 80 split.

Validation: E, L, O

Test: A, B, C, D, F, G, H, I, J, K, M, N

The FNAC 2019 dataset comprises images categorized as either benign or malignant, and we partitioned each category separately into 10 % validation and 90 % test subsets, resulting in:

Validation: Benign 90 – 99; Malignant 103 – 113

Table 3. Whole slide image identifiers utilized for test and validation. 1–7 denote the different domains of the MIDOG dataset. The last row describes the total number of extracted crops per domain.

	1_b	1_c	2	3	4	5	6_a	6_b	7
Test	51 – 95	101 – 145	200 – 240	245 – 294	300 – 344	350 – 399	405 – 481	490 – 503	505 – 549
Valid.	96 – 100	145 – 150	241 – 244	295 – 299	345 – 349	400 – 404	482 – 489	504	550 – 553
# Crops	3258	3174	3548	16059	7202	4743	6260	974	4084

Test: Benign 1 – 89; Malignant 1 – 102

PhaKIR From the six available ID videos, one was randomly selected for validation and another for testing.

Train: Video 02, 03, 04, 07

Validation: Video 05

Test: Video 01

Similarly to the ID split, we use Video 05 as the validation data for the cs-ID datasets Medium Smoke and Heavy Smoke, while the remaining five videos serve as the test datasets. For the Cholec80 dataset, we employ a 10 – 90 split, resulting in:

Validation: Videos 73 – 80

Test: Videos 1 – 72

For EndoSeg15 and EndoSeg18, the first three videos of each dataset are designated as test data, while the remaining one is used for validation. Similarly, in the CATARACTS dataset, the first four videos are allocated as test data, with the final video serving as validation data.

For the Kvasir-SEG dataset, a 10 – 90 validation–test split is applied. Given the dataset's size and the use of filenames resembling globally unique identifiers (GUIDs) individual images, it is impractical to list the exact split. Therefore, detailed information about the data split can be found in the accompanying public GitHub repository.

OASIS-3 After preprocessing all T1w and T2w MRI scans as outlined in Appendix A.3, the remaining CN and AD data were randomly divided into 70 % for training, 15 % for validation, and 15 % for testing. For the cs-ID datasets Modality and Scanner, a randomized 10 – 90 split was applied for validation and test sets.

The ATLAS dataset was partitioned by assigning 10 % of the MRI scans from each cohort to the validation set, with the remaining 90 % allocated to the test set. For the MSD-H dataset, instead of performing a random split, the official test split was utilized as validation data, while the official training split was used as test data.

For the other benchmark datasets, specifically BraTS, CT data from OASIS-3, and CHAOS, the data were randomly divided into 10 % for validation and 90 % for testing.

Detailed information on the dataset sizes and specific splits can be found in the associated public GitHub repository, which includes the exact subject identifiers and filenames for each partition.

Table 4. Table showing the number of ID test samples relative to the average number of cs-ID and OOD samples. The red number indicates the factor by which the average number of cs-ID and OOD samples exceeds that of the ID set.

Data source	MIDOG	PhaKIR	OASIS-3
ID test	251	427	181
cs-ID & OOD	6110 × 24.34	25 350 × 59.36	595 × 3.29

A.5. Metrics

In this work, we employ the AUPR_{IN} and AUPR_{OUT} metrics to assess OOD detection performance. However, interpreting these metrics can be challenging due to significant imbalances between ID and OOD data, which are inherent to many OOD detection tasks. Specifically, the number of OOD samples often greatly exceeds the number of ID samples, as shown in Tab. 4. This imbalance directly influences precision, defined as $\frac{TP}{TP+FP}$. A higher number of OOD samples increases the likelihood of false positives, leading to lower AUPR_{IN} values due to reduced precision for ID samples. Conversely, AUPR_{OUT} values are generally higher because the abundance of OOD samples skews precision favorably when OOD is treated as the positive class. Recall, defined as $\frac{TP}{TP+FN}$, is similarly affected by these imbalances.

This effect is particularly evident in the results from the Cholec80 and CATARACTS datasets within the PhaKIR benchmark, as presented in Tab. 14.

In clinical applications, however, machine learning models are more frequently exposed to ID data, where detecting rare occurrences of OOD inputs becomes crucial. Consequently, developers might prioritize highly sensitive OOD detection methods (high AUPR_{OUT} values) to ensure such inputs are reliably flagged. At the same time, it is equally important to minimize false positive OOD detections (high AUPR_{IN} values), as these can compromise the system's usability. To address this trade-off, we report the harmonic mean of AUPR_{IN} and AUPR_{OUT} in the main text (Tab. 1),

Table 5. Metadata for the training pipeline of each MIB classifier. LR stands for learning rate, WD for weight decay, and BS for batch size.

MIB	Split	Architecture	Optimizer	Seed	Epochs	LR	WD	BS
MIDOG	80-10-10 randomly	ResNet50 [21]	SGD with β of 0.9 [63]	0	300	5×10^{-4}	1×10^{-1}	128
PhaKIR	6 videos split into 4 train, 1 validation, 1 test	ResNet18 [21]	Adam [36]	0	500	1×10^{-4}	1×10^{-3}	48
OASIS-3	70-15-15 randomly	R(2+1)D [65]	Adam [36]	0	300	5×10^{-5}	1×10^{-6}	15

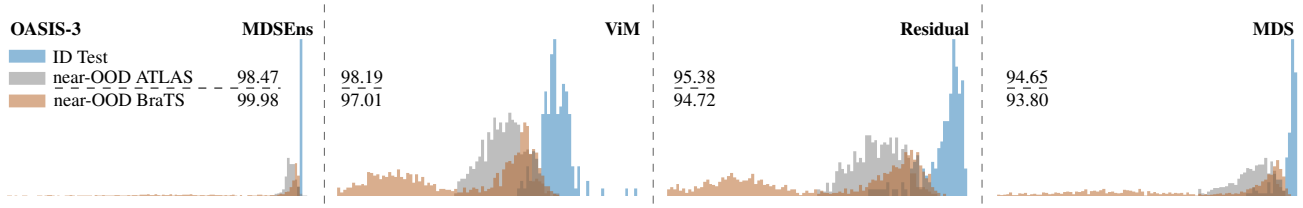

Figure 4. Distribution of OOD scores for the top four methods on two near-OOD datasets from the OASIS-3 benchmark, including AUROC values for each dataset and method.

calculated as:

$$\text{AUPR} = \frac{2 \cdot (\text{AUPR}_{\text{IN}} \cdot \text{AUPR}_{\text{OUT}})}{\text{AUPR}_{\text{IN}} + \text{AUPR}_{\text{OUT}}}$$

This approach prevents a weak performance in one metric from being overshadowed by strong results in the other, as can occur with the arithmetic mean.

By reporting a comprehensive set of metrics, including AUROC (overall OOD detection performance), FPR@95 (threshold-specific behavior), and AUPR_{IN}/AUPR_{OUT} (detailed insights into ID and OOD detection performance), we provide a nuanced evaluation of OOD detection performance.

B. Experiments

B.1. Classifier Training

Table 5 provides additional information regarding the training of each benchmark classifier. To be able to reuse existing model architectures, the final fully connected layer is replaced with a new one, where the output dimension corresponds to the number of classes in each respective classification task.

For training the MIDOG and PhaKIR classifiers, the CE loss function was weighted according to the inverse distribution of class frequencies. However, the PhaKIR training data was highly imbalanced, particularly with respect to the PE-Forceps class, which was overrepresented due to 891 images from Video 02 (Tab. 2), as well as the overall dominance of the No-Instrument class. To address this imbalance and stabilize the training process, 200 images from the PE-Forceps class in Video 02 and 400 images from all ID training videos were randomly sampled in each epoch, with the remaining images withheld for that epoch.

A similar imbalance is present in the OASIS-3 benchmark, with 949 CN and 288 AD MRI scans in the ID data. Following the 70–15–15 split, 660 MRI and 197 MRI scans are available, respectively. To address this imbalance, 100 scans were randomly selected from the available CN and AD MRI sessions per epoch.

Table 6. The employed mean and standard deviation (SD) values for each dataset. For MIDOG and PhaKIR, the values correspond to the red, green, and blue channel.

	MIDOG	PhaKIR	OASIS-3
Mean	0.712 / 0.496 / 0.756	0.517 / 0.361 / 0.336	z-Normalization
SD	0.167 / 0.167 / 0.110	0.166 / 0.143 / 0.137	

Augmentation To enhance classification performance on unseen data, each classifier was trained using additional data augmentations.

For the MIDOG classifier, TrivialAugment Wide [48] was chosen, as it encompasses a diverse range of augmentations. In contrast, for the PhaKIR classifier, a custom augmentation pipeline provided the best results.

Resize: size=(360, 640)
RandomHorizontalFlip: p=0.5
RandomPerspective: distortion_scale=0.2, p=0.5
ColorJitter: brightness=0.2, contrast=0.2, saturation=0.1, hue=0.1

Similarly, the OASIS-3 classifier was trained using a custom augmentation pipeline:

RandomFlip: axes='lr', p=0.5

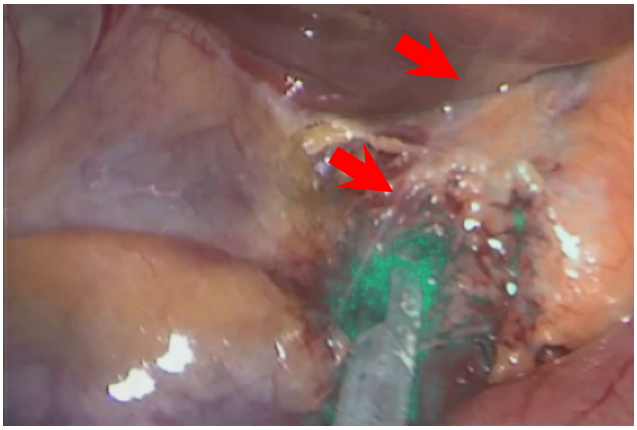

Figure 5. Image from the cs-ID Medium Smoke dataset. Classification attribution is visualized in turquoise using Integrated Gradients (Sundarajan *et al*. [62]), revealing the PhaKIR classifier's tendency to base decisions on regions containing instruments. Arrows indicate an area with localized smoke.

RandomAffine: scales=(0.9, 1.1), degrees=10, isotropic=True, default_pad_value='minimum', p=0.9
RandomMotion: degrees=5, translation=5, p=0.2
RandomNoise: std=(0, 0.1), p=0.9
RandomBlur: std=(0, 0.2), p=0.9
RandomBiasField: coefficients=(0.1, 0.3), p=0.8
RandomElasticDeformation: max_displacement=(5, 5, 5), p=0.1

The final step in each data transformation pipeline was normalization, with the corresponding values provided in Tab. 6.

B.2. Results

The AUROC scores of the four top-performing methods on the most challenging datasets from the OASIS-3 MIB, ATLAS, and BraTS are presented in Fig. 4.

The substantial gap in discriminative power between classification- and hybrid-based methods, compared to feature-based methods, is shown in Tab. 7.

To allow for an easier interpretation of the results, Tab. 8 – Tab. 10 include descriptions for all evaluated OOD methods.

Table 7. AUROC performance averaged across all OOD detection methods for each type. The red percentages indicate the performance degradation compared to the best performing type Feature.

Information source	MIDOG	PhaKIR	OASIS-3
Feature	66.08	70.68	92.08
Combined	57.18 -13 %	50.71 -28 %	69.86 -24 %
Classification	55.81 -16 %	49.45 -30 %	52.13 -43 %

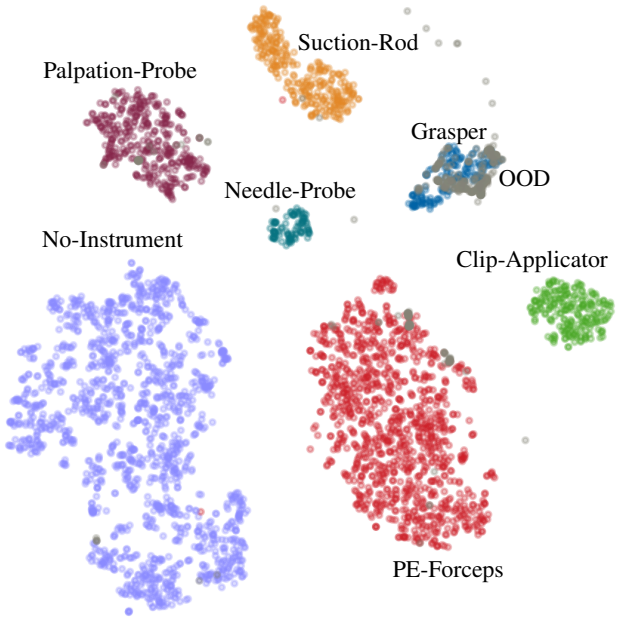

Figure 6. t-SNE [67] projection of the feature space generated by the classifier trained on the PhaKIR ID dataset. Features are from the train ID data (Class labels, colored) and the EndoSeg18 dataset (OOD, gray).

Figure 5 shows that the PhaKIR classifier predominantly bases its decisions on regions containing instruments. Thus, when smoke is located away from the instrument, it is likely that the feature embeddings are less influenced by the smoke.

In Fig. 6, the t-SNE [67] visualization of the PhaKIR classifier's feature space illustrates that OOD samples from the EndoSeg18 dataset are primarily clustered near the Grasper class.

Figure 7 presents success and failure cases for all OOD settings. For the MIDOG and PhaKIR benchmarks, these examples are derived from the two highest-ranked methods. However, for OASIS-3, due to the lack of misclassifications in several OOD categories for MDSEns and ViM, we selected the best performing methods that still exhibit failure cases in these scenarios: SHE and RMDS.

The remaining tables in this section (Tab. 11 – Tab. 18) provide detailed results for all MIBs and their corresponding datasets across all metrics as well as for the ImageNet1k benchmark from OpenOOD [72, 73]. Entries in each table are sorted according to the overall results presented in Tab. 1.

B.3. Employed Hyperparameters

Most OOD detection methods rely on one or more hyperparameters to optimize their performance by using the OOD validation set. The search space for each method's param-

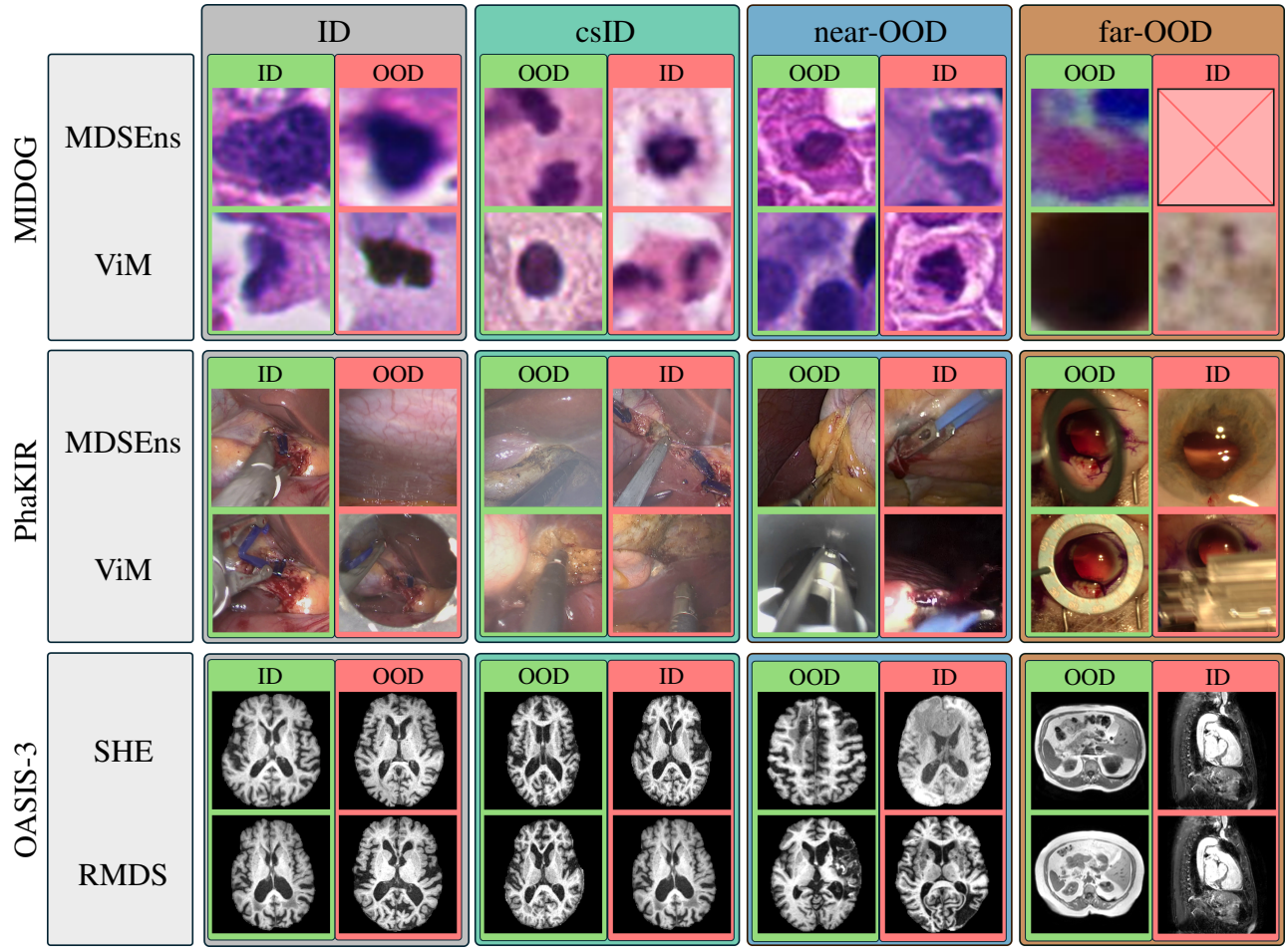

Figure 7. Example images illustrating success and failure cases across all OOD settings. We present the two top-performing methods from the MIDOG and PhaKIR benchmarks. As these methods exhibited no failure cases for several categories in the OASIS-3 benchmark, we instead selected the next two best-performing methods that exhibit failure cases: SHE and RMDS.

eters is outlined in Tab. 19. Methods not included in this table either do not have tunable hyperparameters or feature parameters that are not easily adjustable. The parameter ranges are based on the OpenOOD framework [72, 77], with minor adjustments. Table 20 provides a summary of all automatically selected hyperparameters for the OOD detection methods.

Table 8. Description of each evaluated classification-based approach.

Method	Description
EBO [45]	Motivated by energy-based learning [39], Liu *et al*. transform the final logits into a single scalar using the energy function $E(\mathbf{x}; f) = -T \cdot \log \sum_i^K e^{f_i(\mathbf{x})/T}$. This scalar is then used as the confidence score for OOD detection.
Dropout [17]	Based on the uncertainty estimation from Gal *et al*., this method repeatedly sets entire channels from the penultimate feature layer to zero at random. The softmax probabilities of the resulting logits' mean are used as the confidence score.
GEN [46]	Liu *et al*. take the *N* largest softmax probability and use the generalized entropy $G_\gamma(\mathbf{p}) = \sum_i p_i^\gamma (1 - p_i)^\gamma$ as confidence score.
KLM [24]	Hendrycks *et al*. compute the class-wise distribution of mean softmax probabilities. During inference the minimal Kullback-Leibler divergence between these mean distributions and the current sample distribution is used as confidence score.
MLS [26]	Instead of employing the maximum softmax probability, Hendrycks *et al*. use the maximum logit as confidence score.
MSP [23]	As one of the earliest baselines, Hendrycks *et al*. use the maximum softmax probability as confidence score.
ODIN [42]	Liang *et al*. employ input perturbation and subsequent temperature scaling on the logits. Subsequently, the maximum softmax probability from these logits is used as confidence score.
OpenMax [8]	Bendale and Boult first estimate Weibull distributions for all classes based on the top *k* distances to mean logits. These distributions are used to rescale the logits. Subsequently an additional pseudo-logit is added, while the total activation level remains constant, which serves as OOD class. The probability of this pseudo-class is used as confidence score.
TempScale [19]	Guo *et al*. learn a temperature scaling on the ID dataset and use the temperature-scaled softmax probabilities as confidence score.

Table 9. Description of each evaluated feature-based approach.

Method	Description
KNN [61]	Sun *et al*. compute the *k*-th nearest neighbor of a sample inside the set of normalized activations from the penultimate layer. The distance to this neighbor is used as confidence score.
MDS [40]	Lee *et al*. uses the Mahalanobis distance between a sample's penultimate layer activations and class conditional Gaussian distributions derived from ID data.
MDSEns [40]	Lee *et al*. extend MDS by aggregating these distances from all intermediate layers through weighted averaging. Additionally, input perturbation from ODIN is applied.
Residual [70]	Wang *et al*. project the activations from the penultimate layer to a low-variance subspace defined by the *N* smallest eigenvalues of the empirical covariance matrix, estimated from ID data and uses the norm of those activations as confidence score.
RMDS [52]	Ren *et al*. extend MDS by introducing an additional Mahalanobis distance, computed between the penultimate layer activations and the Gaussian distribution estimated from the entire ID dataset. The final confidence score is obtained by subtracting this new distance from the original MDS distances.
SHE [76]	Zhang *et al*. define confidence scores using the distance between a sample's penultimate-layer activations and its class-conditional mean, computed exclusively from correctly classified samples.

Table 10. Description of each evaluated hybrid-based approach.

Method	Description
ASH [15]	Djurisic *et al*. set the lowest *p*th-percentile of activations in the penultimate layer to zero. The remaining activations are then processed in one of three ways: they are either left unchanged, replaced with a positive constant, or scaled by a ratio derived from the activations before and after pruning. Subsequently, these adjusted activations are used as input to the energy-score from EBO to yield the confidence score. We follow the implementation from OpenOOD [77] and use the variant with positive constants.
DICE [59]	Sun *et al*. calculate the class-wise contribution of weights in the final fully-connected layer based on the empirically estimated ID mean of the ID dataset. By preserving only the *p*-th percentile of the most important weights, they calculate the final logits and use those as input to the energy-score from EBO to yield the confidence score.
fDBD [44]	Liu *et al*. estimate the distance of the penultimate layers' features to class-decision boundaries and regularize this distance by the distance between the activation and the mean of activations from the ID dataset.
NNGuide [49]	Park *et al*. use the average distance of a samples' activation from the penultimate layer to the ID distribution of activations and use it to scale the energy-score from EBO to yield the confidence score.
RankFeat [58]	Song *et al*. propose to remove the rank-1 matrix from activations from the two last feature layers. These matrices are composed by the largest singular value established from a Singular Value Decomposition. The adjusted activations are forwared to yield new logits, which are then averaged and used as input to the energy-score from EBO to yield the confidence score.
ReAct [60]	Sun *et al*. calculate a threshold from the *p*th-percentile of all ID activations in the penultimate layer and use this threshold to set activations above this threshold to zero. Subsequently, logits resulting from these activations serve as input to the energy-score from EBO, yielding the final confidence score.
Relation [35]	Kim *et al*. estimate the relational structure on the feature-space of the penultimate layer based on the activations and corresponding class labels. This structure allows to identify similar feature embeddings with different label information. The confidence score is then calculated by evaluating their proposed similarity functions on a subset of the ID data.
SCALE [71]	Motivated by ASH, Xu *et al*. omit the pruning step from ASH but keep the activation scaling based on the *p*th-percentile of activations. These activations are then used as input to the energy-score from EBO to yield the confidence score.
ViM [70]	Wang *et al*. create an additional virtual-logit based on the subspace from Residual and calculate the energy-score over this and the original logits.

Table 11. Results from the MIDOG benchmark for the AUROC and FPR@95 metrics.

	cs-ID			near-OOD								far-OOD		
	1_b	1_c	Avg	2	3	4	5	6_a	6_b	7	Avg	CCAgT	FNAC	Avg
AUROC↑														
MDSEns [40]	98.56	99.82	99.19	99.56	99.71	99.68	71.50	98.95	99.71	73.78	91.84	100.00	100.00	100.00
ViM [70]	58.97	60.48	59.73	69.06	68.22	65.87	61.97	61.78	61.13	50.65	62.67	89.80	79.76	84.78
Residual [70]	57.87	62.65	60.26	72.56	72.15	67.57	64.48	65.21	63.56	54.89	65.78	94.80	89.91	92.35
MDS [40]	56.89	60.31	58.60	68.94	68.20	65.60	63.67	62.64	60.96	52.45	63.21	93.50	88.31	90.91
KNN [61]	56.72	59.21	57.97	66.62	70.41	63.45	63.25	59.22	57.82	50.62	61.63	91.10	89.26	90.18
SHE [76]	57.06	58.72	57.89	66.92	68.49	61.20	62.96	60.73	61.27	51.03	61.80	91.12	91.06	91.09
RMDS [52]	49.06	49.87	49.46	50.35	50.69	51.44	56.05	52.53	49.61	54.97	52.23	59.98	61.39	60.68
Relation [35]	54.93	56.42	55.68	62.68	62.23	59.27	62.12	58.08	56.43	50.17	58.71	86.10	86.24	86.17
fDBD [44]	50.74	54.33	52.54	60.80	58.02	58.06	61.83	59.43	54.10	56.06	58.33	82.99	83.06	83.03
SCALE [71]	51.65	55.24	53.44	55.61	58.63	54.22	59.14	53.91	53.20	55.27	55.71	79.68	84.38	82.03
ReAct [60]	51.83	55.74	53.79	58.58	59.30	56.56	61.23	56.79	52.94	56.41	57.40	83.78	85.94	84.86
ASH [15]	52.31	55.55	53.93	56.14	59.47	54.63	59.47	53.94	53.65	54.86	56.02	80.47	84.93	82.70
RankFeat [58]	48.37	47.98	48.17	48.99	51.63	52.01	56.12	52.75	41.39	59.91	51.83	64.90	47.97	56.44
OpenMax [8]	48.80	50.96	49.88	53.78	53.29	53.11	56.77	52.08	48.45	51.80	52.75	67.44	64.30	65.87
ODIN [42]	52.82	59.94	56.38	64.40	71.77	58.36	59.69	60.36	62.31	52.69	61.37	79.61	90.56	85.08
GEN [46]	51.00	54.08	52.54	56.22	58.26	55.74	59.97	53.75	50.83	53.46	55.46	77.88	81.68	79.78
MSP [23]	51.71	55.02	53.37	56.70	58.86	56.01	60.16	54.09	51.80	53.71	55.90	78.83	83.00	80.91
Dropout [17]	51.56	54.95	53.25	56.57	58.70	55.84	60.00	53.95	51.65	53.60	55.76	78.78	82.96	80.87
TempScale [19]	51.91	55.23	53.57	56.84	59.08	56.10	60.25	54.21	52.00	53.83	56.05	79.24	83.52	81.38
NNGuide [49]	56.79	59.38	58.08	61.63	68.14	58.95	61.74	56.47	58.61	53.31	59.84	87.33	91.83	89.58
KLM [24]	47.92	48.76	48.34	50.92	47.45	51.34	56.05	53.14	48.70	53.22	51.54	71.49	75.57	73.53
EBO [45]	52.70	56.21	54.46	57.34	60.33	56.08	60.90	55.00	53.33	54.97	56.85	82.02	86.88	84.45
MLS [26]	52.50	55.94	54.22	57.17	59.88	56.13	60.62	54.70	53.02	54.56	56.58	80.59	85.37	82.98
DICE [59]	49.63	53.49	51.56	53.23	52.74	53.55	59.72	53.44	49.58	56.80	54.15	76.38	81.79	79.08
FPR@95↓														
MDSEns [40]	3.59	0.80	2.19	1.20	1.20	1.20	84.06	1.99	1.20	86.85	25.38	0.00	0.00	0.00
ViM [70]	89.64	85.66	87.65	79.68	79.68	79.68	86.45	84.86	87.25	90.04	83.95	40.24	54.58	47.41
Residual [70]	91.24	86.45	88.84	78.09	76.89	78.88	87.65	84.06	85.66	91.24	83.21	22.71	36.25	29.48
MDS [40]	89.24	86.85	88.05	78.49	78.49	79.68	86.85	84.06	86.85	92.83	83.89	32.67	41.83	37.25
KNN [61]	86.06	90.04	88.05	82.07	73.71	85.26	90.84	94.42	89.24	97.21	87.54	29.88	39.84	34.86
SHE [76]	84.86	87.25	86.06	80.08	76.10	85.26	89.64	94.02	84.86	96.41	86.62	28.29	34.66	31.47
RMDS [52]	96.02	95.62	95.82	96.81	96.81	95.62	95.62	95.62	96.81	91.63	95.56	99.20	99.20	99.20
Relation [35]	86.45	86.85	86.65	80.88	79.28	83.67	87.25	90.44	87.25	97.21	86.57	37.05	50.20	43.63
fDBD [44]	93.63	92.83	93.23	85.66	84.06	88.05	90.04	88.05	92.03	92.03	88.56	55.78	64.14	59.96
SCALE [71]	97.21	96.81	97.01	93.23	84.86	97.21	95.22	96.81	90.44	95.22	93.28	76.49	67.73	72.11
ReAct [60]	94.42	92.83	93.63	92.83	89.64	94.82	94.42	94.82	94.82	92.83	93.45	46.22	58.17	52.19
ASH [15]	96.02	95.62	95.82	92.43	85.26	96.41	94.42	95.62	90.44	95.62	92.89	74.50	65.34	69.92
RankFeat [58]	95.62	96.81	96.22	95.62	95.22	95.22	94.02	95.22	98.01	93.23	95.22	87.25	97.21	92.23
OpenMax [8]	94.02	92.03	93.03	90.44	86.06	93.63	92.43	94.02	89.64	94.82	91.58	62.55	66.53	64.54
ODIN [42]	95.22	93.23	94.22	89.24	78.49	94.02	93.63	92.03	88.84	96.41	90.38	86.85	54.18	70.52
GEN [46]	94.42	93.63	94.02	90.84	86.85	94.82	93.23	94.82	91.24	96.02	92.54	63.35	62.55	62.95
MSP [23]	94.02	93.63	93.82	90.84	86.45	94.42	93.63	94.82	91.24	96.02	92.49	63.75	61.75	62.75
Dropout [17]	94.82	94.02	94.42	91.24	86.85	94.82	92.83	95.22	91.24	96.41	92.66	64.14	62.95	63.55
TempScale [19]	94.02	93.63	93.82	90.84	86.45	94.42	93.63	94.82	91.24	96.02	92.49	63.75	60.96	62.35
NNGuide [49]	92.83	92.43	92.63	88.05	78.49	94.02	92.03	94.82	87.65	96.81	90.27	54.58	35.46	45.02
KLM [24]	96.41	94.02	95.22	94.42	97.21	94.82	93.63	93.63	92.43	93.63	94.25	75.70	91.63	83.67
EBO [45]	94.42	94.02	94.22	91.24	86.06	94.42	92.83	94.42	91.24	96.02	92.32	67.33	58.17	62.75
MLS [26]	94.42	94.02	94.22	91.24	86.45	94.42	93.23	94.42	91.24	96.02	92.43	66.93	58.96	62.95
DICE [59]	96.02	96.02	96.02	95.62	93.23	96.02	94.02	96.02	95.22	94.02	94.88	63.75	72.51	68.13

Table 12. Results from the MIDOG benchmark for the $AUPR_{IN}$ and $AUPR_{OUT}$ metrics.

	cs-ID			near-OOD								far-OOD		
	1_b	1_c	Avg	2	3	4	5	6_a	6_b	7	Avg	CCAgT	FNAC	Avg
$AUPR_{IN}$ ↑														
MDSEns [40]	95.64	99.28	97.46	98.85	98.83	98.86	13.30	97.90	99.55	13.16	74.35	100.00	99.99	100.00
ViM [70]	10.20	12.75	11.48	16.55	5.27	9.69	10.49	7.24	30.35	7.58	12.45	44.07	57.24	50.65
Residual [70]	10.35	13.69	12.02	19.55	6.85	12.35	10.88	8.92	34.45	8.00	14.43	55.84	72.06	63.95
MDS [40]	10.46	13.03	11.74	19.06	7.30	12.41	11.35	8.78	32.40	7.32	14.09	51.93	67.59	59.76
KNN [61]	11.81	11.15	11.48	15.55	11.42	9.60	10.33	6.15	31.42	6.33	12.97	28.37	65.82	47.10
SHE [76]	16.30	13.74	15.02	18.78	12.18	8.71	10.78	6.65	35.75	6.50	14.19	55.35	72.39	63.87
RMDS [52]	7.55	7.59	7.57	6.82	1.61	3.99	7.41	4.39	21.76	7.97	7.71	1.32	13.77	7.54
Relation [35]	15.31	13.28	14.29	17.75	10.19	8.70	10.96	6.70	32.53	6.50	13.33	47.93	62.33	55.13
fDBD [44]	8.45	9.15	8.80	12.10	3.50	6.35	10.03	6.96	26.48	8.41	10.55	29.44	51.14	40.29
SCALE [71]	8.48	9.06	8.77	8.81	3.12	4.39	8.62	4.63	25.63	7.82	9.00	4.27	44.60	24.44
ReAct [60]	8.54	9.41	8.98	9.24	2.80	4.89	8.98	5.07	23.87	8.60	9.06	14.11	47.90	31.00
ASH [15]	8.67	9.16	8.92	8.98	3.26	4.46	8.75	4.64	25.93	7.65	9.09	4.62	45.69	25.15
RankFeat [58]	7.34	7.17	7.26	6.83	1.71	4.14	7.65	4.42	18.55	8.49	7.40	2.08	10.66	6.37
OpenMax [8]	7.96	8.63	8.30	8.56	2.62	4.55	8.31	4.49	23.30	7.21	8.43	4.75	39.87	22.31
ODIN [42]	8.27	10.47	9.37	11.76	5.86	4.90	8.58	5.84	33.21	6.78	10.99	3.28	57.47	30.37
GEN [46]	8.78	9.26	9.02	9.50	3.61	4.80	8.98	5.03	24.56	6.99	9.07	5.65	46.10	25.88
MSP [23]	8.94	9.48	9.21	9.81	3.83	4.81	8.99	5.07	25.25	6.97	9.25	5.69	46.54	26.12
Dropout [17]	8.85	9.36	9.10	9.67	3.64	4.82	8.91	5.02	24.94	6.94	9.13	5.51	46.43	25.97
TempScale [19]	8.97	9.52	9.24	9.84	3.85	4.82	9.01	5.09	25.34	6.98	9.27	5.72	46.84	26.28
NNGuide [49]	12.98	11.51	12.24	13.28	8.84	5.83	9.73	5.09	31.69	6.62	11.58	11.56	71.19	41.38
KLM [24]	7.03	7.55	7.29	7.22	1.48	3.95	7.76	4.61	23.32	7.24	7.94	3.02	21.62	12.32
EBO [45]	9.08	9.66	9.37	9.73	3.78	4.76	9.00	5.12	25.99	7.16	9.37	5.52	49.42	27.47
MLS [26]	9.04	9.62	9.33	9.72	3.76	4.78	8.98	5.10	25.81	7.11	9.32	5.53	48.48	27.01
DICE [59]	8.02	8.54	8.28	8.24	2.49	4.44	8.90	4.91	22.98	8.36	8.62	4.94	33.88	19.41
$AUPR_{OUT}$ ↑														
MDSEns [40]	99.77	99.98	99.88	99.95	99.99	99.98	97.04	99.89	99.88	97.23	99.14	100.00	100.00	100.00
ViM [70]	94.72	94.76	94.74	96.52	99.19	97.68	95.31	97.22	85.78	93.33	95.00	99.85	95.81	97.83
Residual [70]	94.43	95.20	94.82	97.06	99.32	97.89	95.91	97.59	86.93	94.44	95.59	99.93	98.19	99.06
MDS [40]	94.22	94.82	94.52	96.55	99.20	97.72	95.74	97.38	85.87	94.13	95.23	99.91	97.90	98.91
KNN [61]	93.92	94.50	94.21	95.97	99.22	97.48	95.97	97.09	83.43	94.39	94.79	99.87	98.26	99.07
SHE [76]	93.94	94.49	94.21	96.01	99.16	97.33	95.94	97.20	84.46	94.43	94.93	99.87	98.59	99.23
RMDS [52]	92.10	92.54	92.32	93.25	98.42	96.30	94.69	96.17	77.70	94.38	92.99	99.25	92.70	95.97
Relation [35]	93.36	93.88	93.62	95.22	98.92	97.07	95.67	96.86	81.93	94.17	94.26	99.76	97.66	98.71
fDBD [44]	92.92	93.81	93.36	95.22	98.83	97.07	95.66	97.00	81.85	94.76	94.34	99.73	97.25	98.49
SCALE [71]	92.76	93.69	93.23	94.29	98.78	96.73	95.45	96.57	80.31	94.83	93.85	99.68	97.57	98.62
ReAct [60]	92.89	93.86	93.37	94.82	98.85	96.91	95.70	96.83	80.88	94.95	94.13	99.75	97.86	98.80
ASH [15]	92.92	93.76	93.34	94.41	98.82	96.77	95.50	96.58	80.61	94.79	93.93	99.69	97.67	98.68
RankFeat [58]	91.52	91.64	91.58	92.42	98.32	96.16	94.44	96.07	72.45	95.41	92.18	99.16	87.55	93.35
OpenMax [8]	91.49	91.81	91.65	93.15	98.25	96.12	94.10	95.77	76.13	93.03	92.37	99.01	87.31	93.16
ODIN [42]	92.98	94.52	93.75	95.85	99.28	97.14	95.57	97.27	84.79	94.39	94.90	99.69	98.66	99.18
GEN [46]	92.47	93.22	92.84	94.12	98.70	96.72	95.23	96.37	78.93	94.28	93.48	99.56	96.08	97.82
MSP [23]	92.64	93.48	93.06	94.30	98.75	96.79	95.37	96.48	79.44	94.46	93.66	99.62	96.99	98.30
Dropout [17]	92.63	93.48	93.06	94.29	98.75	96.78	95.36	96.47	79.37	94.46	93.64	99.62	96.99	98.30
TempScale [19]	92.70	93.55	93.13	94.35	98.77	96.81	95.42	96.51	79.68	94.53	93.72	99.63	97.17	98.40
NNGuide [49]	93.77	94.44	94.11	95.31	99.14	97.18	95.89	96.89	82.89	94.79	94.58	99.82	98.78	99.30
KLM [24]	92.77	92.85	92.81	93.68	98.49	96.57	95.01	96.46	78.91	94.38	93.36	99.53	96.24	97.89
EBO [45]	92.97	93.89	93.43	94.63	98.87	96.87	95.71	96.72	80.71	94.91	94.06	99.73	98.05	98.89
MLS [26]	92.90	93.78	93.34	94.51	98.83	96.86	95.59	96.63	80.38	94.75	93.94	99.68	97.70	98.69
DICE [59]	91.68	93.04	92.36	93.39	98.36	96.33	95.39	96.31	76.77	94.94	93.07	99.57	97.13	98.35

Table 13. Results from the PhaKIR benchmark for the AUROC and FPR@95 metrics. M. Smoke and H. Smoke stands for Medium and Heavy Smoke. CAT. stands for CATARACTS.

	cs-ID			near-OOD				far-OOD		
	M. Smoke	H. Smoke	Avg	Cholec80	EndoSeg15	EndoSeg18	Avg	Kvasir	CAT.	Avg
AUROC↑										
MDSEns [40]	45.67	84.43	65.05	96.65	94.81	99.87	97.11	99.98	97.02	98.50
ViM [70]	63.64	81.14	72.39	68.44	83.25	91.73	81.14	50.04	60.64	55.34
Residual [70]	39.11	75.13	57.12	59.53	75.98	95.46	76.99	48.33	66.30	57.31
MDS [40]	38.25	73.84	56.04	58.27	76.37	94.81	76.48	40.20	62.75	51.47
KNN [61]	24.01	44.09	34.05	64.17	61.22	40.92	55.44	31.22	44.28	37.75
SHE [76]	25.89	46.47	36.18	62.40	55.13	33.48	50.34	54.30	40.50	47.40
RMDS [52]	31.67	44.76	38.22	60.86	69.85	72.48	67.73	24.98	45.99	35.49
Relation [35]	24.41	37.72	31.06	62.98	63.50	56.57	61.02	26.55	34.90	30.72
fDBD [44]	26.38	33.78	30.08	58.13	54.72	37.56	50.13	18.10	36.98	27.54
SCALE [71]	27.35	44.46	35.91	61.19	42.47	13.26	38.97	47.75	45.91	46.83
ReAct [60]	24.15	36.52	30.34	58.08	52.64	34.41	48.38	16.63	35.16	25.89
ASH [15]	36.95	55.04	45.99	60.14	42.62	17.76	40.17	73.39	56.78	65.08
RankFeat [58]	54.27	49.75	52.01	41.96	51.71	42.10	45.26	14.61	40.09	27.35
OpenMax [8]	23.95	39.85	31.90	64.55	69.04	64.51	66.03	31.21	35.90	33.56
ODIN [42]	31.16	37.39	34.28	63.17	43.58	18.59	41.78	83.53	60.10	71.82
GEN [46]	24.16	41.14	32.65	61.29	55.97	37.40	51.55	29.33	36.02	32.68
MSP [23]	23.77	40.30	32.04	61.67	54.20	34.61	50.16	28.45	36.56	32.51
Dropout [17]	23.78	40.13	31.96	61.41	54.27	34.60	50.10	28.59	36.52	32.56
TempScale [19]	23.56	39.90	31.73	61.73	52.96	31.49	48.73	27.87	37.12	32.50
NNGuide [49]	21.87	36.43	29.15	60.93	42.19	10.84	37.98	34.56	47.32	40.94
KLM [24]	52.39	57.91	55.15	57.33	58.25	53.45	56.34	29.09	42.64	35.87
EBO [45]	23.74	40.25	31.99	60.77	45.38	14.38	40.18	27.36	41.32	34.34
MLS [26]	23.74	40.25	32.00	60.76	45.36	14.39	40.17	27.35	41.31	34.33
DICE [59]	30.22	41.70	35.96	60.62	57.94	41.43	53.33	22.47	22.82	22.65
FPR@95↓										
MDSEns [40]	100.00	56.67	78.34	14.99	15.93	0.70	10.54	0.00	8.90	4.45
ViM [70]	95.78	78.69	87.24	84.54	52.22	28.34	55.04	95.08	93.21	94.15
Residual [70]	99.53	86.42	92.97	88.52	62.06	16.63	55.74	91.57	85.48	88.52
MDS [40]	99.30	84.54	91.92	89.23	64.17	18.97	57.46	94.61	84.54	89.58
KNN [61]	100.00	98.59	99.30	85.95	86.65	95.78	89.46	94.15	97.89	96.02
SHE [76]	99.77	98.13	98.95	90.40	92.74	99.06	94.07	93.21	99.06	96.14
RMDS [52]	97.89	91.10	94.50	82.90	75.18	92.97	83.68	95.55	92.97	94.26
Relation [35]	99.53	97.42	98.48	83.14	76.35	64.64	74.71	96.02	100.00	98.01
fDBD [44]	95.78	90.16	92.97	84.07	82.44	83.84	83.45	96.96	99.53	98.24
SCALE [71]	100.00	99.06	99.53	90.40	97.42	100.00	95.94	77.99	86.89	82.44
ReAct [60]	97.19	94.38	95.78	88.76	83.37	89.70	87.28	98.36	100.00	99.18
ASH [15]	99.77	93.68	96.72	85.48	94.85	100.00	93.44	52.46	73.54	63.00
RankFeat [58]	96.02	95.32	95.67	92.51	89.70	94.61	92.27	97.66	91.80	94.73
OpenMax [8]	100.00	100.00	100.00	87.59	93.44	99.30	93.44	94.85	99.30	97.07
ODIN [42]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	50.12	76.81	63.47
GEN [46]	100.00	100.00	100.00	91.10	95.55	99.77	95.47	95.08	98.59	96.84
MSP [23]	100.00	100.00	100.00	90.87	96.25	100.00	95.71	94.38	98.59	96.49
Dropout [17]	100.00	100.00	100.00	90.63	96.96	100.00	95.86	94.38	98.36	96.37
TempScale [19]	100.00	100.00	100.00	90.63	96.02	100.00	95.55	94.38	98.59	96.49
NNGuide [49]	100.00	100.00	100.00	91.80	96.49	100.00	96.10	88.06	91.80	89.93
KLM [24]	86.89	88.76	87.82	96.96	76.35	61.83	78.38	93.91	95.55	94.73
EBO [45]	99.77	99.06	99.41	91.80	98.13	100.00	96.64	93.21	94.85	94.03
MLS [26]	99.77	99.06	99.41	91.80	98.13	100.00	96.64	93.21	94.85	94.03
DICE [59]	100.00	100.00	100.00	93.91	95.78	99.77	96.49	100.00	100.00	100.00

Table 14. Results from the PhaKIR benchmark for the $AUPR_{IN}$ and $AUPR_{OUT}$ metrics. M. Smoke and H. Smoke stands for Medium and Heavy Smoke. CAT. stands for CATARACTS.

	cs-ID			near-OOD				far-OOD		
	M. Smoke	H. Smoke	Avg	Cholec80	EndoSeg15	EndoSeg18	Avg	Kvasir	CAT.	Avg
$AUPR_{IN}$ ↑										
MDSEns [40]	29.60	69.21	49.41	73.26	98.59	99.80	90.55	99.96	87.55	93.75
ViM [70]	40.23	61.00	50.61	2.46	94.66	89.08	62.07	31.28	0.82	16.05
Residual [70]	27.14	52.52	39.83	1.16	92.18	94.58	62.64	32.41	1.11	16.76
MDS [40]	26.85	51.74	39.29	1.10	92.31	93.55	62.32	27.96	1.02	14.49
KNN [61]	22.77	27.39	25.08	1.41	84.54	32.37	39.44	26.53	0.33	13.43
SHE [76]	23.14	29.06	26.10	0.99	80.58	27.50	36.35	36.09	0.36	18.22
RMDS [52]	25.57	34.20	29.89	1.48	88.00	54.44	47.97	24.84	0.86	12.85
Relation [35]	22.78	25.66	24.22	1.41	87.45	59.25	49.37	23.63	0.27	11.95
fDBD [44]	25.35	29.13	27.24	1.57	84.08	41.98	42.54	20.71	0.28	10.50
SCALE [71]	23.45	27.33	25.39	0.90	73.50	22.14	32.18	45.45	0.84	23.14
ReAct [60]	23.66	28.07	25.86	1.01	82.82	37.00	40.28	19.93	0.27	10.10
ASH [15]	26.14	35.08	30.61	1.05	75.74	23.02	33.27	71.80	1.31	36.55
RankFeat [58]	36.56	33.82	35.19	1.22	80.96	34.21	38.80	21.17	0.90	11.03
OpenMax [8]	22.75	25.59	24.17	1.11	84.98	40.47	42.19	26.16	0.28	13.22
ODIN [42]	34.63	34.94	34.78	12.38	78.09	33.44	41.30	77.69	12.47	45.08
GEN [46]	22.75	26.04	24.40	0.97	79.95	28.26	36.39	26.05	0.28	13.17
MSP [23]	22.67	25.79	24.23	0.96	79.29	27.31	35.86	25.80	0.29	13.04
Dropout [17]	22.66	25.76	24.21	0.96	79.27	27.33	35.85	26.00	0.29	13.14
TempScale [19]	22.62	25.66	24.14	0.95	78.64	26.30	35.30	25.72	0.30	13.01
NNGuide [49]	22.34	24.49	23.41	0.89	73.06	21.80	31.92	33.96	1.05	17.50
KLM [24]	42.03	42.02	42.02	0.76	84.61	54.99	46.79	26.08	0.35	13.21
EBO [45]	22.60	26.00	24.30	0.90	74.94	22.31	32.72	27.38	0.50	13.94
MLS [26]	22.60	26.00	24.30	0.90	74.94	22.31	32.72	27.38	0.50	13.94
DICE [59]	24.35	26.03	25.19	0.85	80.38	29.65	36.96	20.85	0.23	10.54
$AUPR_{OUT}$ ↑										
MDSEns [40]	63.55	92.10	77.83	99.97	83.02	99.92	94.31	99.99	99.98	99.99
ViM [70]	79.28	90.54	84.91	99.67	56.84	93.99	83.50	72.87	99.77	86.32
Residual [70]	60.54	87.27	73.91	99.52	44.71	96.78	80.34	67.50	99.79	83.65
MDS [40]	59.45	86.08	72.76	99.49	44.72	96.38	80.20	61.44	99.76	80.60
KNN [61]	53.15	66.78	59.96	99.57	26.56	55.74	60.62	55.42	99.61	77.51
SHE [76]	54.59	68.58	61.59	99.56	23.59	52.13	58.43	70.69	99.53	85.11
RMDS [52]	54.15	60.83	57.49	99.43	32.88	78.20	70.17	52.34	99.62	75.98
Relation [35]	52.51	61.55	57.03	99.56	26.44	61.97	62.66	53.14	99.49	76.31
fDBD [44]	51.58	57.02	54.30	99.50	22.32	53.05	58.29	50.25	99.53	74.89
SCALE [71]	53.78	66.55	60.17	99.56	19.02	44.81	54.46	61.78	99.59	80.68
ReAct [60]	51.03	59.10	55.07	99.49	21.66	51.43	57.52	49.90	99.52	74.71
ASH [15]	61.16	75.83	68.50	99.54	18.66	46.01	54.74	76.77	99.66	88.21
RankFeat [58]	68.24	65.31	66.77	99.07	21.02	55.80	58.63	49.21	99.48	74.34
OpenMax [8]	52.83	62.91	57.87	99.58	37.71	82.03	73.11	58.21	99.44	78.82
ODIN [42]	55.84	66.78	61.31	99.62	19.55	42.94	54.04	88.09	99.70	93.90
GEN [46]	53.06	64.94	59.00	99.53	26.20	57.69	61.14	54.70	99.47	77.09
MSP [23]	52.19	62.79	57.49	99.55	22.94	53.90	58.80	53.92	99.49	76.71
Dropout [17]	52.14	62.57	57.35	99.55	22.96	53.77	58.76	53.93	99.49	76.71
TempScale [19]	51.98	62.57	57.28	99.56	22.32	52.01	57.96	53.68	99.50	76.59
NNGuide [49]	50.90	60.83	55.87	99.53	18.50	44.13	54.05	56.01	99.62	77.82
KLM [24]	62.48	70.24	66.36	99.49	23.76	60.34	61.19	54.04	99.54	76.79
EBO [45]	51.59	62.26	56.93	99.55	19.68	45.10	54.77	53.35	99.56	76.45
MLS [26]	51.60	62.29	56.94	99.55	19.64	45.10	54.76	53.34	99.56	76.45
DICE [59]	59.34	68.38	63.86	99.56	27.24	59.65	62.15	52.39	99.28	75.84

Table 15. Results from the OASIS-3 benchmark for the AUROC and FPR@95 metrics.

	cs-ID			near-OOD				far-OOD		
	Modality	Scanner	Avg	ATLAS	BraTS	CT	Avg	MSD-H	CHAOS	Avg
AUROC↑										
MDSEns [40]	100.00	100.00	100.00	98.47	99.98	99.94	99.46	100.00	100.00	100.00
ViM [70]	100.00	97.63	98.82	98.19	97.01	100.00	98.40	100.00	100.00	100.00
Residual [70]	100.00	93.94	96.97	95.38	94.72	100.00	96.70	100.00	100.00	100.00
MDS [40]	100.00	92.62	96.31	94.65	93.80	100.00	96.15	100.00	100.00	100.00
KNN [61]	100.00	97.56	98.78	96.25	96.71	100.00	97.66	100.00	100.00	100.00
SHE [76]	97.20	86.51	91.85	85.03	87.37	100.00	90.80	99.28	100.00	99.64
RMDS [52]	63.91	52.09	58.00	58.88	56.72	99.46	71.69	99.78	100.00	99.89
Relation [35]	31.28	59.68	45.48	67.20	53.23	78.16	66.20	86.55	99.26	92.91
fDBD [44]	47.96	69.27	58.62	74.38	64.59	86.96	75.31	91.02	94.84	92.93
SCALE [71]	99.96	69.11	84.53	63.25	89.70	100.00	84.32	96.46	100.00	98.23
ReAct [60]	31.99	57.49	44.74	54.15	67.86	88.69	70.23	64.61	90.45	77.53
ASH [15]	92.53	48.88	70.71	57.64	75.81	96.12	76.52	86.82	96.13	91.48
RankFeat [58]	81.54	61.03	71.29	60.10	66.14	97.72	74.65	93.98	100.00	96.99
OpenMax [8]	37.81	49.56	43.69	58.97	54.48	44.35	52.60	68.56	72.38	70.47
ODIN [42]	55.01	52.00	53.51	34.86	51.09	92.63	59.53	42.76	74.49	58.63
GEN [46]	19.36	54.54	36.95	63.28	49.29	47.94	53.50	77.18	79.71	78.45
MSP [23]	19.36	54.54	36.95	63.28	49.29	47.94	53.50	77.18	79.71	78.45
Dropout [17]	19.69	54.52	37.10	62.32	49.31	48.02	53.22	74.70	79.56	77.13
TempScale [19]	19.36	54.54	36.95	63.28	49.29	47.94	53.50	77.18	79.71	78.45
NNGuide [49]	24.16	45.36	34.76	48.43	56.79	65.63	56.95	57.32	95.43	76.37
KLM [24]	61.10	47.92	54.51	54.38	46.71	32.38	44.49	64.89	63.08	63.98
EBO [45]	19.80	47.92	33.86	55.77	51.37	41.02	49.39	63.70	75.60	69.65
MLS [26]	19.80	48.07	33.94	56.01	51.34	41.05	49.47	63.92	75.54	69.73
DICE [59]	32.67	36.66	34.66	18.32	27.82	32.27	26.14	20.88	26.52	23.70
FPR@95↓										
MDSEns [40]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ViM [70]	0.00	11.05	5.52	8.29	13.81	0.00	7.37	0.00	0.00	0.00
Residual [70]	0.00	19.34	9.67	12.15	20.44	0.00	10.87	0.00	0.00	0.00
MDS [40]	0.00	22.10	11.05	15.47	22.10	0.00	12.52	0.00	0.00	0.00
KNN [61]	0.00	13.81	6.91	14.92	17.68	0.00	10.87	0.00	0.00	0.00
SHE [76]	4.97	35.91	20.44	32.60	30.94	0.00	21.18	2.21	0.00	1.10
RMDS [52]	95.58	98.90	97.24	89.50	91.71	0.00	60.41	2.21	0.00	1.10
Relation [35]	82.32	82.32	82.32	82.32	91.16	50.28	74.59	42.54	11.60	27.07
fDBD [44]	67.40	75.69	71.55	75.69	75.69	21.55	57.64	21.55	11.05	16.30
SCALE [71]	0.00	83.98	41.99	82.87	57.46	0.00	46.78	13.26	0.00	6.63
ReAct [60]	86.74	83.98	85.36	75.69	81.22	37.02	64.64	55.80	37.57	46.69
ASH [15]	18.78	86.19	52.49	91.71	71.82	3.87	55.80	23.76	3.87	13.81
RankFeat [58]	50.83	74.59	62.71	65.75	81.22	13.81	53.59	18.23	0.00	9.12
OpenMax [8]	86.19	87.29	86.74	87.29	81.77	86.19	85.08	71.82	59.67	65.75
ODIN [42]	64.09	83.43	73.76	86.74	76.80	47.51	70.35	79.56	65.19	72.38
GEN [46]	91.16	86.74	88.95	85.64	89.50	85.64	86.92	63.54	43.65	53.59
MSP [23]	91.16	86.74	88.95	85.64	89.50	85.64	86.92	63.54	43.65	53.59
Dropout [17]	90.06	87.29	88.67	85.64	90.06	85.64	87.11	69.06	41.44	55.25
TempScale [19]	91.16	86.74	88.95	85.64	89.50	85.64	86.92	63.54	43.65	53.59
NNGuide [49]	87.85	87.85	87.85	88.40	87.85	79.56	85.27	70.72	27.62	49.17
KLM [24]	82.87	97.24	90.06	93.37	96.13	97.24	95.58	74.59	79.01	76.80
EBO [45]	88.40	87.85	88.12	88.40	87.85	87.85	88.03	75.69	47.51	61.60
MLS [26]	88.40	87.85	88.12	88.40	87.85	87.85	88.03	75.69	47.51	61.60
DICE [59]	77.35	91.71	84.53	92.27	85.64	83.43	87.11	85.64	80.11	82.87

Table 16. Results from the OASIS-3 benchmark for the AUPR_{IN} and AUPR_{OUT} metrics.

	cs-ID			near-OOD				far-OOD		
	Modality	Scanner	Avg	ATLAS	BraTS	CT	Avg	MSD-H	CHAOS	Avg
AUPR_{IN} ↑										
MDSEns [40]	100.00	100.00	100.00	85.77	99.89	99.22	94.96	100.00	100.00	100.00
ViM [70]	100.00	98.95	99.47	95.50	89.45	100.00	94.99	100.00	100.00	100.00
Residual [70]	100.00	97.54	98.77	93.47	87.17	100.00	93.55	100.00	100.00	100.00
MDS [40]	99.99	96.98	98.49	92.51	85.57	100.00	92.69	100.00	100.00	100.00
KNN [61]	100.00	98.93	99.46	92.93	88.62	100.00	93.85	100.00	100.00	100.00
SHE [76]	96.87	94.27	95.57	79.70	73.60	100.00	84.43	99.92	100.00	99.96
RMDS [52]	29.10	66.63	47.86	30.23	15.79	93.64	46.55	99.98	100.00	99.99
Relation [35]	31.03	78.53	54.78	38.63	17.06	61.19	38.96	98.31	99.92	99.12
fDBD [44]	47.17	84.04	65.60	53.32	32.83	82.61	56.25	98.99	99.50	99.25
SCALE [71]	99.87	80.84	90.35	38.98	53.18	100.00	64.05	99.56	100.00	99.78
ReAct [60]	27.46	78.03	52.74	42.07	32.32	72.90	49.09	95.20	98.96	97.08
ASH [15]	89.95	71.72	80.84	31.09	37.84	96.53	55.15	98.49	99.64	99.07
RankFeat [58]	70.72	81.47	76.09	51.59	27.79	91.86	57.08	99.32	100.00	99.66
OpenMax [8]	30.53	72.16	51.35	36.66	27.72	23.74	29.37	95.14	96.69	95.91
ODIN [42]	52.11	74.82	63.47	28.22	25.26	68.08	40.52	90.52	96.72	93.62
GEN [46]	22.64	75.14	48.89	39.94	21.96	22.42	28.11	96.51	97.78	97.15
MSP [23]	22.64	75.14	48.89	39.94	21.96	22.42	28.11	96.51	97.78	97.15
Dropout [17]	22.98	74.85	48.92	39.64	22.01	22.36	28.00	95.96	97.73	96.84
TempScale [19]	22.64	75.14	48.89	39.94	21.96	22.42	28.11	96.51	97.78	97.15
NNGuide [49]	25.77	70.64	48.20	32.35	25.67	30.65	29.55	93.63	99.51	96.57
KLM [24]	37.68	67.45	52.56	28.05	13.92	8.57	16.85	93.85	95.10	94.47
EBO [45]	24.40	71.31	47.85	34.94	23.90	20.58	26.47	94.46	97.21	95.84
MLS [26]	24.40	71.34	47.87	35.00	23.89	20.58	26.49	94.49	97.20	95.84
DICE [59]	35.94	64.33	50.13	21.46	19.18	22.00	20.88	85.76	88.78	87.27
AUPR_{OUT} ↑										
MDSEns [40]	100.00	100.00	100.00	99.62	100.00	99.99	99.87	100.00	100.00	100.00
ViM [70]	100.00	95.15	97.57	99.38	99.47	100.00	99.62	100.00	100.00	100.00
Residual [70]	100.00	84.90	92.45	97.84	98.98	100.00	98.94	100.00	100.00	100.00
MDS [40]	100.00	81.85	90.92	97.29	98.80	100.00	98.70	100.00	100.00	100.00
KNN [61]	100.00	95.02	97.51	98.49	99.42	100.00	99.30	100.00	100.00	100.00
SHE [76]	98.02	64.90	81.46	91.83	96.74	100.00	96.19	93.65	100.00	96.83
RMDS [52]	86.94	38.65	62.80	82.12	89.86	99.93	90.64	98.26	100.00	99.13
Relation [35]	63.39	39.49	51.44	85.29	87.29	94.44	89.01	51.59	96.11	73.85
fDBD [44]	69.37	46.84	58.11	88.23	90.16	95.26	91.22	37.96	46.06	42.01
SCALE [71]	99.99	57.82	78.90	83.51	98.29	100.00	93.94	87.52	100.00	93.76
ReAct [60]	64.56	33.10	48.83	72.11	92.78	98.00	87.63	12.48	62.71	37.60
ASH [15]	93.33	29.72	61.52	79.04	92.54	97.22	89.60	28.04	50.43	39.24
RankFeat [58]	92.48	33.99	63.24	75.50	92.52	99.65	89.22	65.40	100.00	82.70
OpenMax [8]	71.23	31.98	51.60	80.94	87.99	83.90	84.28	15.20	25.35	20.27
ODIN [42]	73.04	38.14	55.59	63.73	84.36	98.93	82.34	8.04	30.43	19.24
GEN [46]	59.71	35.28	47.50	82.87	85.71	84.52	84.37	22.30	19.47	20.88
MSP [23]	59.71	35.28	47.50	82.88	85.72	84.52	84.38	22.30	19.47	20.88
Dropout [17]	59.83	34.32	47.07	81.86	85.23	84.50	83.86	21.52	21.96	21.74
TempScale [19]	59.71	35.28	47.50	82.88	85.72	84.52	84.38	22.30	19.47	20.88
NNGuide [49]	61.07	27.28	44.17	70.65	89.81	93.02	84.49	10.64	79.77	45.20
KLM [24]	76.10	29.95	53.03	76.91	83.39	80.44	80.25	13.72	11.59	12.66
EBO [45]	59.79	29.39	44.59	76.63	87.03	81.74	81.80	12.59	27.42	20.00
MLS [26]	59.79	29.82	44.80	77.23	86.94	81.74	81.97	12.68	24.82	18.75
DICE [59]	63.38	30.60	46.99	58.38	74.31	78.08	70.26	6.25	6.07	6.16

Table 17. Results from the ImageNet1k benchmark for the AUROC and FPR@95 metrics.

	cs-ID				near-OOD			far-OOD			
	IN-V2 [51]	IN-C [22]	IN-R [25]	Avg	SSB-hard [69]	NINCO [9]	Avg	iNaturalist [28]	Textures [13]	OpenImage-O [70]	Avg
AUROC↑											
MDSEns [40]	51.15	76.74	74.87	67.58	48.30	60.66	54.48	56.42	93.30	73.82	74.51
ViM [70]	57.34	83.77	87.95	76.35	65.53	78.63	72.08	89.56	97.97	90.50	92.68
Residual [70]	49.83	67.78	65.52	61.04	42.14	54.59	48.37	52.13	87.81	61.02	66.99
MDS [40]	51.72	70.80	69.78	64.10	48.50	62.38	55.44	63.67	89.80	69.27	74.25
KNN [61]	56.44	83.94	87.64	76.01	62.57	79.64	71.10	86.41	97.09	87.04	90.18
SHE [76]	57.60	83.65	86.18	75.81	68.73	82.79	75.76	95.40	97.16	91.57	94.71
RMDS [52]	58.70	79.70	81.73	73.38	71.77	82.22	76.99	87.24	86.08	85.84	86.38
Relation [35]	56.92	79.05	84.38	73.45	65.90	79.82	72.86	91.26	91.32	88.34	90.31
fDBD [44]	58.56	81.77	87.64	75.99	70.65	82.60	76.63	93.70	93.44	91.17	92.77
SCALE [71]	58.36	84.07	83.81	75.41	77.34	85.37	81.36	98.02	97.63	93.95	96.53
ReAct [60]	58.49	80.99	85.98	75.15	73.02	81.73	77.38	96.34	92.79	91.87	93.67
ASH [15]	58.07	83.85	84.15	75.36	74.71	84.54	79.63	97.72	97.87	93.82	96.47
RankFeat [58]	53.05	66.11	64.42	61.19	58.87	54.27	56.57	58.64	74.79	60.20	64.54
OpenMax [8]	58.07	80.10	85.69	74.62	71.37	78.17	74.77	92.05	88.10	87.62	89.26
ODIN [42]	57.60	76.19	85.47	73.09	71.74	77.77	74.75	91.17	89.00	88.23	89.47
GEN [46]	58.89	80.60	86.44	75.31	72.00	81.69	76.85	92.44	87.59	89.26	89.76
MSP [23]	58.54	77.06	80.51	72.04	72.09	79.95	76.02	88.41	82.43	84.86	85.23
Dropout [17]	58.51	76.97	80.41	71.96	71.99	79.81	75.90	88.21	82.26	84.64	85.04
TempScale [19]	58.89	78.78	83.16	73.61	72.87	81.41	77.14	90.50	84.95	87.22	87.56
NNGuide [49]	58.63	82.90	87.57	76.37	73.42	81.97	77.69	95.44	95.16	92.39	94.33
KLM [24]	57.57	76.77	81.64	71.99	71.40	81.91	76.65	90.79	84.71	87.30	87.60
EBO [45]	58.75	80.99	86.82	75.52	72.42	80.29	76.35	91.14	88.50	89.19	89.61
MLS [26]	58.77	80.93	86.70	75.46	72.51	80.41	76.46	91.17	88.39	89.17	89.57
DICE [59]	57.01	80.19	80.38	72.53	72.91	77.56	75.23	94.56	92.28	88.61	91.81
FPR@95↓											
MDSEns [40]	93.49	78.06	67.45	79.67	93.47	84.65	89.06	81.54	34.07	71.69	62.43
ViM [70]	92.14	65.75	46.55	68.15	80.41	62.28	71.35	30.69	10.49	32.82	24.67
Residual [70]	94.34	82.00	77.94	84.76	96.93	89.96	93.45	89.61	53.62	86.53	76.59
MDS [40]	93.40	75.47	71.00	79.96	92.10	78.80	85.45	73.81	42.79	72.15	62.92
KNN [61]	92.81	68.14	53.23	71.39	83.36	58.39	70.87	40.80	17.31	44.27	34.13
SHE [76]	92.31	65.04	51.32	69.56	79.56	54.20	66.88	20.78	15.57	33.29	23.21
RMDS [52]	91.66	72.44	59.37	74.49	77.88	52.20	65.04	33.67	48.80	40.27	40.91
Relation [35]	92.22	75.78	62.74	76.91	86.50	59.92	73.21	32.94	33.61	41.66	36.07
fDBD [44]	92.03	69.80	50.46	70.76	77.28	52.08	64.68	22.02	27.71	29.93	26.55
SCALE [71]	91.24	67.63	62.02	73.63	67.72	51.86	59.79	9.52	11.91	28.13	16.52
ReAct [60]	91.92	74.40	53.71	73.34	77.55	55.88	66.72	16.70	29.65	32.57	26.31
ASH [15]	91.29	66.91	62.12	73.44	70.81	53.11	61.96	10.99	11.01	28.61	16.87
RankFeat [58]	93.18	83.82	80.79	85.93	87.41	88.28	87.85	81.56	69.95	81.77	77.76
OpenMax [8]	91.59	73.88	51.32	72.27	77.33	60.81	69.07	25.29	40.26	37.39	34.31
ODIN [42]	92.51	79.99	59.46	77.32	76.83	68.16	72.50	35.98	49.24	46.66	43.96
GEN [46]	91.95	74.35	54.37	73.56	75.72	54.89	65.31	26.09	46.26	34.53	35.63
MSP [23]	91.85	77.54	66.24	78.54	74.48	56.85	65.66	43.35	60.86	50.16	51.46
Dropout [17]	91.67	77.52	65.92	78.37	74.58	57.31	65.94	43.72	61.72	50.57	52.00
TempScale [19]	91.76	77.24	63.31	77.44	73.90	55.13	64.52	37.56	56.94	45.43	46.65
NNGuide [49]	91.61	72.16	53.18	72.32	74.70	56.85	65.77	20.36	26.02	31.98	26.12
KLM [24]	92.85	83.54	72.56	82.98	84.73	59.61	72.17	38.49	52.29	48.80	46.52
EBO [45]	91.93	74.88	53.46	73.42	76.27	59.83	68.05	30.49	46.27	37.79	38.19
MLS [26]	91.95	74.65	53.61	73.40	76.20	59.44	67.82	30.61	46.17	37.88	38.22
DICE [59]	91.79	75.54	62.68	76.67	75.86	66.26	71.06	25.83	42.80	48.50	39.04

Table 18. Results from the ImageNet1k benchmark for the AUPR_{IN} and AUPR_{OUT} metrics.

	cs-ID				near-OOD			far-OOD			
	IN-V2 [51]	IN-C [22]	IN-R [25]	Avg	SSB-hard [69]	NINCO [9]	Avg	iNaturalist [28]	Textures [13]	OpenImage-O [70]	Avg
AUPR_{IN} ↑											
MDSEns [40]	82.96	91.56	82.45	85.66	48.23	92.32	70.28	87.13	98.47	88.79	91.46
ViM [70]	85.14	95.03	91.50	90.56	65.67	96.49	81.08	97.59	99.74	96.58	97.97
Residual [70]	82.20	90.00	75.91	82.70	42.63	90.66	66.65	84.35	98.17	81.52	88.01
MDS [40]	83.11	91.37	79.81	84.76	49.37	93.24	71.30	89.91	98.60	87.35	91.95
KNN [61]	84.66	95.01	90.80	90.16	62.33	96.74	79.53	96.66	99.61	95.04	97.10
SHE [76]	85.13	95.15	90.40	90.23	67.41	97.28	82.35	98.80	99.64	96.82	98.42
RMDS [52]	85.56	93.71	87.13	88.80	69.90	97.24	83.57	97.02	97.98	94.82	96.61
Relation [35]	84.78	93.24	87.87	88.63	62.79	96.69	79.74	97.85	98.85	95.55	97.42
fDBD [44]	85.44	94.32	91.02	90.26	69.62	97.31	83.46	98.56	99.14	96.90	98.20
SCALE [71]	85.61	94.98	87.51	89.36	76.95	97.64	87.29	99.54	99.69	97.67	98.97
ReAct [60]	85.46	93.74	89.58	89.59	70.70	97.06	83.88	99.14	99.03	96.96	98.37
ASH [15]	85.50	94.94	87.63	89.36	74.47	97.48	85.98	99.47	99.73	97.61	98.94
RankFeat [58]	83.55	89.31	74.24	82.37	57.98	90.77	74.38	87.67	96.08	82.64	88.79
OpenMax [8]	85.31	93.75	89.93	89.66	70.13	96.42	83.28	98.19	98.40	95.57	97.39
ODIN [42]	85.11	92.17	88.58	88.62	70.52	96.13	83.33	97.72	98.21	95.23	97.05
GEN [46]	85.61	93.73	89.71	89.68	71.25	97.11	84.18	98.23	98.13	96.11	97.49
MSP [23]	85.54	92.60	85.22	87.78	71.50	96.76	84.13	97.09	97.18	94.11	96.13
Dropout [17]	85.55	92.57	85.15	87.76	71.40	96.74	84.07	97.04	97.15	94.01	96.07
TempScale [19]	85.62	93.08	86.96	88.55	72.15	97.02	84.59	97.67	97.58	95.09	96.78
NNGuide [49]	85.60	94.39	90.37	90.12	72.37	97.06	84.72	98.89	99.28	97.09	98.42
KLM [24]	84.74	91.99	84.76	87.17	66.68	97.00	81.84	97.41	97.63	94.63	96.55
EBO [45]	85.56	93.77	89.92	89.75	71.27	96.79	84.03	97.87	98.25	95.97	97.36
MLS [26]	85.56	93.75	89.86	89.73	71.34	96.81	84.07	97.88	98.24	95.96	97.36
DICE [59]	85.14	93.57	85.75	88.15	71.56	96.15	83.85	98.61	98.74	95.26	97.54
AUPR_{OUT} ↑											
MDSEns [40]	18.13	51.08	61.55	43.59	49.72	14.67	32.19	19.39	74.53	47.45	47.12
ViM [70]	21.97	60.77	79.98	54.24	63.24	29.11	46.18	59.24	88.60	74.43	74.09
Residual [70]	17.71	30.25	49.77	32.57	45.90	12.15	29.02	18.30	56.39	34.43	36.37
MDS [40]	18.37	32.54	53.23	34.71	49.36	14.55	31.96	23.28	57.96	40.03	40.42
KNN [61]	21.68	66.29	82.29	56.75	61.97	32.07	47.02	56.41	87.84	70.98	71.74
SHE [76]	22.45	61.99	78.42	54.28	68.82	38.62	53.72	84.83	85.49	79.59	83.31
RMDS [52]	22.92	47.28	70.59	46.93	70.70	31.93	51.31	50.31	45.81	61.60	52.58
Relation [35]	22.87	56.09	78.43	52.46	66.04	31.31	48.68	69.45	57.85	70.18	65.83
fDBD [44]	23.64	59.90	81.67	55.07	69.13	34.62	51.87	75.23	68.31	75.48	73.01
SCALE [71]	23.11	65.10	77.30	55.17	76.51	46.66	61.59	92.35	87.69	86.05	88.70
ReAct [60]	23.54	60.14	78.41	54.03	73.13	38.68	55.91	86.41	66.02	80.52	77.65
ASH [15]	22.88	64.69	78.18	55.25	73.88	44.73	59.31	91.31	88.56	85.68	88.52
RankFeat [58]	19.29	28.12	50.22	32.54	58.81	12.49	35.65	21.76	27.58	31.98	27.11
OpenMax [8]	22.68	45.87	73.70	47.42	69.52	26.64	48.08	63.57	37.49	61.82	54.29
ODIN [42]	22.90	51.31	81.18	51.80	71.05	32.73	51.89	72.11	61.00	73.72	68.95
GEN [46]	23.99	55.83	79.52	53.11	69.90	31.74	50.82	71.07	45.50	69.78	62.12
MSP [23]	23.61	51.40	73.00	49.34	70.34	31.17	50.76	63.88	40.27	64.60	56.25
Dropout [17]	23.49	51.04	72.81	49.12	70.22	31.04	50.63	63.51	39.75	64.18	55.81
TempScale [19]	23.95	54.76	76.33	51.68	70.96	32.42	51.69	67.17	44.86	68.06	60.03
NNGuide [49]	23.60	64.80	82.37	56.93	72.98	39.73	56.36	84.55	81.21	82.23	82.66
KLM [24]	23.15	46.91	73.33	47.80	72.54	35.77	54.16	70.16	36.45	68.81	58.47
EBO [45]	23.91	59.66	80.68	54.75	71.03	31.39	51.21	66.31	53.59	71.63	63.84
MLS [26]	23.94	59.20	80.41	54.52	71.10	31.60	51.35	66.61	52.74	71.46	63.60
DICE [59]	21.88	57.70	71.42	50.34	72.91	34.49	53.70	82.84	73.85	77.05	77.92

Table 19. Overview of the hyperparameter ranges for each OOD detection method with tunable hyperparameters employed in this work.

Method	Parameter 1	Parameter Range	Parameter 2	Parameter Range
MDSEns	noise	[0, 0.0025, 0.0014, 0.005, 0.01, 0.02, 0.04, 0.08]	–	
ViM	dim	[1, 16, 32, 64, 128, 256]	–	
Residual	dim	[32, 64, 128, 256, 512, 1024]	–	
KNN	k	[1, 2, 5, 10, 25, 50, 100, 200, 500, 750, 1000]	–	
SHE	metric	[inner product, euclidean, cosine]	–	
Relation	pow	[1, 2, 4, 6, 8]	–	
fDBD	normalized	[False,True]	–	
SCALE	percentile	[65, 70, 75, 80, 85, 90, 95]	–	
ReAct	percentile	[85, 90, 95, 99]	–	
ASH	percentile	[65, 70, 75, 80, 85, 90, 95]	–	
RankFeat	acc	[False, True]	temp	[0.1, 1, 10, 100, 1000]
OpenMax	sampling ratio	[0.01]	neighbors	[9]
ODIN	temperature	[1, 10, 100, 1000]	noise	[0.0014, 0.0028]
GEN	gamma	[0.01, 0.1, 0.5, 1, 2, 5, 10]	M	[1, 2, 3, 4, 5, 6, 7, 50, 100, 200]
Dropout	p	[0.5]	times	[15]
NNGuide	k	[1, 2, 5, 10, 50, 100, 200, 500, 750]	alpha	[0.01, 0.1, 0.25, 0.5, 0.75, 1.0]
EBO	temperature	[0.1, 0.5, 1, 1.5, 2.0]	–	
DICE	percentile	[60, 65, 70, 75, 80, 85, 90, 95]	–	

Table 20. Overview of the automatically selected hyperparameters for the three MIBs and ImageNet1k.

	MIDOG		PhaKIR		OASIS-3		ImageNet1k	
Method	Parameter 1	Parameter 2	Parameter 1	Parameter 2	Parameter 1	Parameter 2	Parameter 1	Parameter 2
MDSEns	0.0	–	0.0014	–	0.0	–	0.0	–
ViM	256	–	256	–	256	–	256	–
Residual	128	–	128	–	64	–	1024	–
KNN	5	–	50	–	5	–	200	–
SHE	cosine	–	cosine	–	cosine	–	cosine	–
Relation	8	–	8	–	8	–	1	–
fDBD	False	–	False	–	True	–	True	–
SCALE	65	–	65	–	95	–	85	–
ReAct	95	–	90	–	85	–	95	–
ASH	65	–	65	–	95	–	85	–
RankFeat	False	0.1	True	10	False	1	True	0.1
OpenMax	0.01	9	0.01	9	0.01	9	0.01	9
ODIN	1	0.0014	1	0.0014	10	0.0028	10	0.0014
GEN	0.01	2	0.01	2	0.01	1	0.01	100
Dropout	0.5	15	0.5	15	0.5	15	0.5	15
NNGuide	1	0.1	5	0.01	5	0.01	500	0.01
EBO	2.0	–	0.1	–	0.1	–	0.5	–
DICE	85	–	95	–	80	–	65	–