Unsupervised Surrogate Anomaly Detection - Supplementary Material

No Author Given

No Institute Given

1 Further Experiments

1.1 Runtime

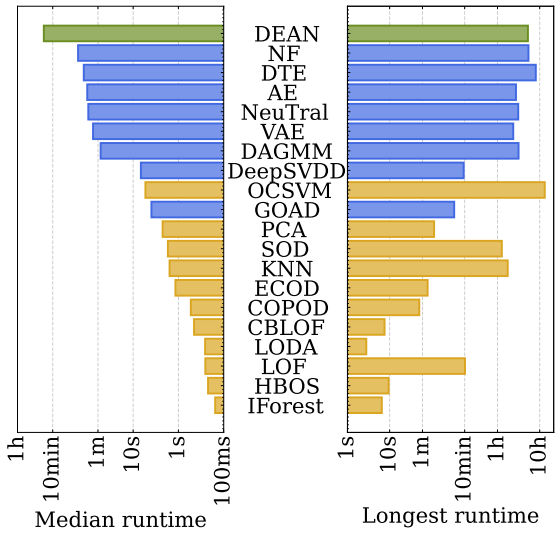

Fig. 1: Left: median runtime per dataset of our algorithm compared to our competitors, Right: Maximum runtime of each algorithm on the datasets. We show DEAN in green, the remaining deep learning algorithms in blue, and shallow algorithms in yellow.

We compare the median runtime of the algorithms considered in Figure 1. Since some datasets are significantly larger than others, the runtimes vary significantly between datasets. Because of this, we show the median runtime over each dataset as well as the maximum over all datasets. To keep the comparison fair, we only use a single CPU core per algorithm and dataset.

Generally, we see deep learning algorithms perform slowest, with DEAN being the slowest algorithm with a median runtime of about $15min$ per dataset. Still,

these algorithms are also the most affected when using GPUs. Additionally, DEAN is an ensemble method and thus can be parallelized almost perfectly. Also because of DEAN use of feature bagging, the comparison is significantly more even when considering the worst longest runtime instead. Here, most deep learning approaches require comparable amounts of time, and three even require more.

1.2 Ensemble Structure

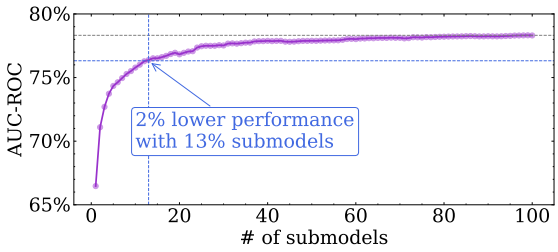

Fig. 2: Average DEAN AUC-ROC performance over all datasets as a function of the number of submodels used. We reach 2% lower performance with the first 13% of submodels. The same plot for AUC-PR can be found in Section 5

The hyperparameter affecting the runtime the most is the number of submodels used. And while the performance also depends on how many submodels are used, the relationship is not linear as Figure 2 shows. While even with the 100 submodels we employed for the benchmark experiments, we do still see an improvement from using more submodels, the change in average performance slows down significantly compared to the first submodels used.

This means we could train a version of DEAN in 87% less time, which would only have a 2% lower performance.

The continuous growth of DEAN with more submodels is a direct consequence of the simple submodels we use, and the high variance between submodels they allow. We similarly compare the performance of an Autoencoder or DeepSVDD ensemble in Section 2 and show that these ensembles stay almost constant.

2 Existing Surrogates as Ensembles

Since we compare our ensemble-based method mostly against non-ensemble methods, we also show in Fig. 3 the performance of the other deep learning-based surrogate methods when used in an ensemble. While DEAN's performance improves with more submodels, the performance of the other methods stays almost constant. We believe this to be a result of the simple submodel we use.

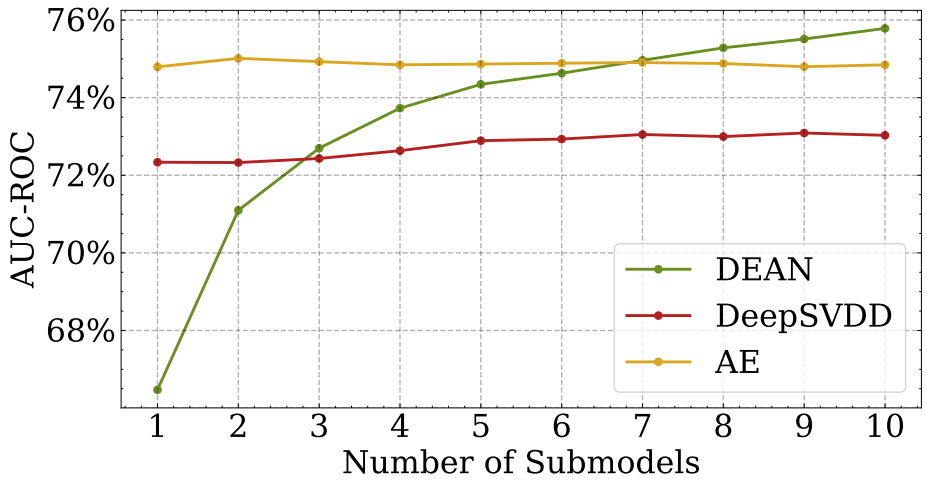

Fig. 3: Comparing other deep learning surrogate models as ensembles. The average performance over all datasets is shown here. While DEAN improves significantly when more submodels are used, both DeepSVDD and Autoencoder stay almost constant.

3 Importance of Learnable Shifts

Trivial solutions are a common problem also for DeepSVDD [14]. Namely when the last layer learns a zero multiplicative weight, but the learnable shift is equal to the desired $\boldsymbol{c}$. To combat this, Ruff et. al. propose to remove the learnable shifts entirely. And while this certainly helps in making this shift impossible, it also limits how complicated a function can be learned by the neural network [11].

We show this in Figure 4, where we task neural networks to approximate a simple sinus curve. Here, we use neural networks with three layers of 100 nodes and relu activation in each hidden layer. The three networks differ only by the learnable shifts they use. While the network with learnable shifts (green) is clearly able to approximate the sinus curve, the version without learnable shifts (blue) is not able to do so. And since real anomaly representations can be much more complicated than such a simple sinus curve, we do not think that limiting the neural network complexity is a reasonable choice.

Instead, we use other methods to remove the trivial solution of a constant network. This also includes using learnable shifts in each hidden layer but not in the output layer. This setup is still able to approximate complicated functions, as is shown in orange in Figure 4.

4 Anomaly Detection Beyond Accuracy

While our comparison in the main paper shows the competitive performance of DEAN, it also highlights the problem of creating a strictly better-performing

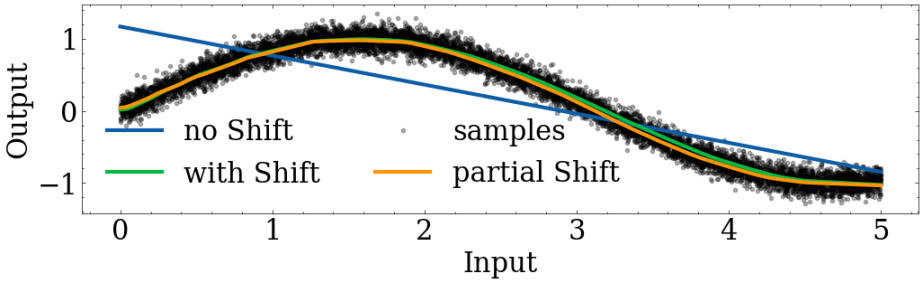

Fig. 4: Given complicated alinear data, the functions learned by three neural networks with relu activations are shown. The network without learnable shifts cannot capture the structure of the underlying data, while both a network with learnable shifts in each layer and a network with learnable shifts in all layers except the last can describe the alinearity.

anomaly detection algorithm. Even when using 121 datasets, we are not able to find many significant differences between the performances of various algorithms.

Thus, we argue that while accuracy is vital up to a certain level, usability might be more beneficial than marginally improving an algorithm's performance beyond competitive performance. Given our axiomatic algorithm design, DEAN is a highly practical algorithm. Specifically, we aim to demonstrate three key benefits: reliability, hyperparameter invariance, and modifiability.

4.1 Modifiability

Beyond achieving competitive performance reliably and with easy-to-configure parameterization, it is important to note that not every task has the same exact requirements. Other surrogate algorithms like DeepSVDD and Autoencoder are arguably prominent because they are easy to modify and extend. For instance, they allow the incorporation of fairness criteria [21] and sparse ground truth through a few labeled instances [15], or the integration of preprocessing [6] and ensemble techniques [3] to boost robustness and explainability. We argue that DEAN is even more versatile than these algorithms, thanks to its simple submodel design and inherent ensemble structure.

The ensemble structure, for example, can be used to explain why specific samples are considered normal or abnormal by leveraging shapley values [7]. Here, feature bagging provides a natural solution to the high computational complexity of shapley values. Similarly, the ensemble structure should make it possible to create a version of DEAN with federated learning [20]. Additionally, techniques such as subsampling [23] could be used to create ensembles that change what data they are trained on to be GDPR compliant [12]. A similar idea could be used to create a version of DEAN that uses active learning to either improve or personalize an anomaly detection algorithm [10]. Feature bagging [9] instead could likely be used to allow the model to handle missing values [4].

And since retraining singular submodels can be done in parallel to predictions, a continuously learning [18] version could be created. Ensembles also ensure specific properties of our anomaly detection algorithm, such as robustness against adversarial samples, by employing pruning or weighting techniques to remove the least robust ensemble submodels [2].

Similarly, the simplicity of our submodels allows for the easy modification of the training procedure. This allows the input of additional information [8] like, for example, in semi-supervised anomaly detection [15] or outlier exposure [5] and could even likely be used for privacy-preserving anomaly detection [1].

To summarize, we see three major ways DEAN can be adapted. Either we can use the ensemble structure to change the selection of submodels, we can weight the ensemble combination or we can change the training procedure of each submodel. As an example, we apply all three of these to the task of fair anomaly detection [21] in Section 6.

In addition to this, we also think that using different machine learning models could lead to interesting versions of DEAN. Neural symbolic computing [13] could be used to extract human understandable patterns and a more lightweight model could be directly applied on IoT devices[17]. Also, graph neural networks [22] and recurrent neural networks [16] could be used to create a version of DEAN finding anomalies on graphs or time series.

5 Additional Result Plots with AUC-PR

Since our results are very similar whether we use AUC-ROC or AUC-PR, we only state most of our results in AUC-ROC and add the redundant plots here.

Figure 5 shows the critical difference plot when we use AUC-PR instead of AUC-ROC to compare the performance of algorithms. Additionally, Figure 6 shows the AUC-PR score as a function of the submodels used.

6 DEAN-Fair

Since we believe DEAN to be adaptable to many various situations (See Section 4.1), we want to demonstrate this on a toy example. For this, we want to show how DEAN could be modified to increase its fairness [21].

For this, we use the COMPAS dataset [19] and consider re-offending citizens as abnormal. The COMPAS dataset contains criminal recidivism risk scores assigned by the COMPAS algorithm to defendants, along with various demographic and criminal history features. It is widely used in studies evaluating algorithmic fairness and bias.

As a fairness metric, we measure the deviation in AUC-ROC performance between two subgroups based on binarizing the dataset with respect to a protected attribute. We chose the AUC-ROC since it is a metric invariant to the fraction of anomalous samples and also understands non-binary anomaly scores. Since a model that observes no difference between distributions is optimal, we

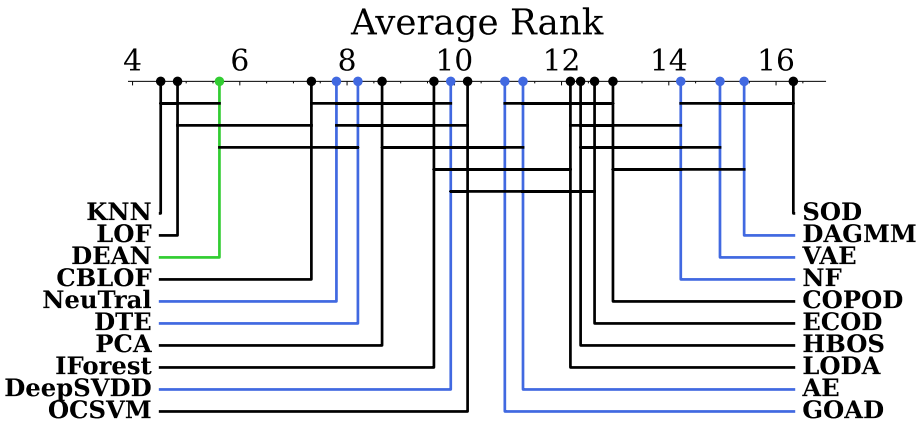

Fig. 5: Critical difference plot comparing DEAN to other algorithms. While our algorithm is not always the best, there is no algorithm that performs significantly better than DEAN. We use AUC-PR here to compare the performance of two algorithms. Pendant to Figure 2 in the main paper.

desire here a fairness AUC-ROC score of 0.5. We choose age (≤ 25 vs. > 25) as a protected attribute for this analysis to showcase how our algorithm can be guided towards equal treatment across different demographic groups in general.

Now we propose three different algorithms to improve the fairness. The first one uses a slightly changed loss function:

$$L = \sum_{\boldsymbol{x} \in X_{train}} \|f(\boldsymbol{x}) - 1\| + \theta \cdot L_{fair} \tag{1}$$

$$L_{fair} = \frac{\|L_1 - L_0\|}{\|L_1\| + \|L_0\|} \tag{2}$$

$$L_{1/0} = \frac{1}{\|X_{(un)protected}\|} \sum_{\boldsymbol{x} \in X_{(un)protected}} f(\boldsymbol{x}) \tag{3}$$

This loss function uses another term added to the loss that tries to make the mean for protected and unprotected samples equal. We choose here $\theta = 0.1$.

Our second model removes submodels. Here we choose the most unfair models to remove in a greedy fashion.

Our final model adds weights to each submodel in the ensemble to maximize fairness. Since the optimization is not continuous, we choose an evolutionary algorithm to choose these weights. More details can be found in our implementation.

The results of each of our adaptation methods are shown in Table 1. We repeat each experiment five times to generate uncertainties. Generally, the performance is relatively low, probably because reoffending citizens are not always clearly separable from those who don't. Still, our baseline model is unfair, with

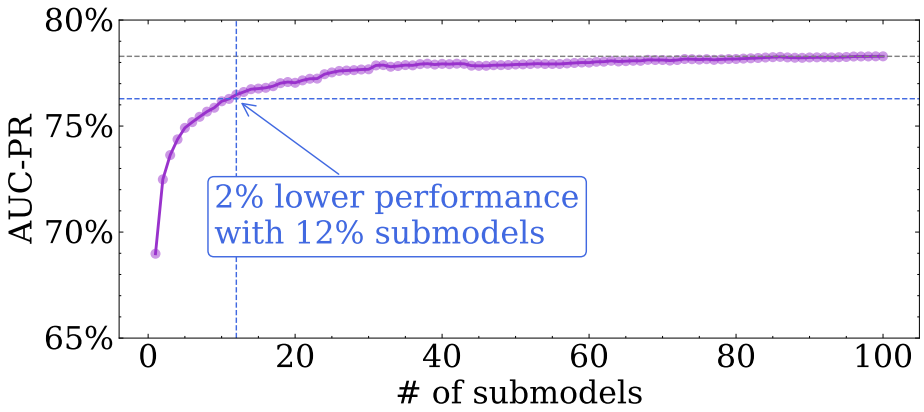

Fig. 6: DEAN AUC-PR performance as a function of the number of submodels used. We reach 2% less performance with the first 12% (instead of 13% for AUC-ROC) of submodels. Pendant to Figure 2.

Adjustment	AUC-ROC	Fairness
Baseline	0.583 ± 0.003	0.644 ± 0.020
Loss function	0.594 ± 0.012	0.453 ± 0.080
Pruning (1%)	0.583 ± 0.003	0.625 ± 0.019
Pruning (5%)	0.577 ± 0.003	0.555 ± 0.015
Pruning (10%)	0.574 ± 0.003	0.506 ± 0.014
Non-uniform weighting	0.566 ± 0.004	0.520 ± 0.011

Table 1: AUC-ROC performance and fairness AUC-ROC on the COMPAS dataset for various fairness improving variations of DEAN. Notice that the performance is better the higher the value is, while the fairness is optimal at 0.5.

a deviation of over 14% from a fair model. To fix this, we can prune our model. And even when just removing 1% of submodels, we already see an improvement in fairness. By removing 10% of submodels, we can almost remove the unfairness entirely (0.6% deviation, which is smaller than the uncertainty). Still, pruning comes at the cost of performance, with an about 1% drop in performance. Weighting has a more significant drop, with 1.7% lower performance and 2% unfairness. However, the most significant change is seen when optimizing with a different loss function. The fairness overshoots a bit, and there is still a 4.7% unfairness, and the performance increases by about 1.1%.

Overall all three of our adaptation methods could be used to increase the fairness of DEAN, confirming the adaptability of our method.

7 Individual Performance Scores

We state every performance in AUC-ROC in Tables 2, 3, 4, 5, 6 and 7. We also give the same performances in AUC-PR in Tables 8, 9, 10, 11, 12 and 13.

Table 2: AUC-ROC Scores for each datasets and algorithm (1/3|low performing algorithms)

Dataset	DEAN	HBOS	GOAD	ECOD	COPOD	LODA	NF	DAGMM	VAE	SOD
$20news^{2}$	56%	44%	44%	44%	43%	46%	49%	44%	54%	46%
yeast	59%	42%	68%	45%	38%	55%	48%	49%	41%	44%
vertebral	68%	38%	67%	41%	34%	26%	50%	50%	50%	38%
$MNISTC^{identity}$	47%	49%	48%	48%	48%	48%	47%	49%	50%	46%
speech	59%	49%	50%	48%	50%	50%	47%	57%	49%	35%
imdb	53%	51%	54%	48%	53%	42%	53%	46%	42%	46%
$20news^{5}$	56%	49%	48%	48%	47%	55%	52%	48%	53%	44%
$WPBC$	54%	53%	45%	52%	54%	58%	49%	50%	48%	43%
$Wilt$	67%	36%	62%	38%	35%	37%	54%	70%	50%	32%
$20news^{4}$	59%	51%	53%	52%	50%	54%	42%	53%	48%	51%
$20news^{1}$	64%	50%	52%	47%	50%	54%	51%	46%	44%	52%
$agnews^{0}$	63%	51%	53%	49%	52%	49%	50%	49%	50%	49%
$20news^{3}$	50%	56%	55%	55%	56%	61%	57%	49%	41%	51%
$MVTecAD^{screw}$	60%	57%	52%	56%	56%	54%	47%	50%	57%	56%
$ALOI$	55%	54%	49%	54%	53%	52%	55%	53%	52%	54%
amazon	62%	57%	56%	55%	58%	61%	50%	50%	45%	54%
$SVHN^{6}$	65%	51%	64%	53%	52%	58%	54%	58%	52%	50%
$CIFAR10^{3}$	67%	48%	69%	52%	49%	59%	42%	54%	57%	51%
$CIFAR10^{5}$	67%	46%	69%	51%	47%	57%	52%	59%	53%	44%
$SVHN^{9}$	66%	51%	60%	54%	52%	54%	53%	55%	51%	56%
$SVHN^{3}$	65%	55%	60%	57%	56%	59%	46%	47%	50%	56%
$MNISTC^{rotate}$	67%	56%	48%	55%	55%	50%	56%	53%	55%	51%
$CIFAR10^{2}$	61%	55%	59%	56%	55%	58%	54%	55%	55%	52%
landsat	77%	71%	61%	36%	42%	43%	45%	53%	57%	49%
$SVHN^{8}$	70%	50%	62%	53%	51%	56%	53%	55%	47%	56%
yelp	67%	60%	61%	58%	60%	59%	47%	59%	42%	56%
$agnews^{3}$	64%	56%	54%	56%	56%	53%	50%	56%	50%	55%
$SVHN^{0}$	76%	50%	65%	53%	51%	59%	40%	59%	54%	50%
$CIFAR10^{1}$	75%	45%	73%	52%	47%	63%	47%	54%	53%	50%
$SVHN^{5}$	70%	57%	61%	59%	58%	64%	53%	57%	51%	57%
$MVTecAD^{pill}$	62%	64%	52%	61%	65%	66%	51%	50%	50%	65%
$agnews^{1}$	69%	58%	50%	58%	52%	58%	49%	52%	58%	50%
$SVHN^{4}$	66%	61%	54%	61%	61%	64%	52%	58%	62%	58%
$SVHN^{2}$	69%	58%	62%	60%	58%	64%	55%	50%	54%	61%
census	63%	66%	59%	66%	67%	55%	52%	61%	62%	58%
$fault$	75%	67%	69%	46%	45%	48%	57%	47%	52%	64%
$Hepatitis$	44%	82%	50%	73%	81%	74%	50%	50%	52%	52%
$SVHN^{7}$	66%	62%	62%	63%	62%	61%	51%	56%	62%	53%
$SVHN^{1}$	68%	62%	68%	63%	61%	55%	47%	62%	63%	51%
$Pima$	65%	70%	61%	59%	65%	62%	49%	50%	50%	52%
$20news^{0}$	75%	62%	61%	60%	61%	61%	56%	56%	51%	64%
$CIFAR10^{7}$	69%	56%	71%	61%	57%	65%	54%	60%	65%	54%
$MNISTC^{translate}$	85%	54%	54%	56%	56%	57%	51%	49%	50%	61%
$agnews^{2}$	74%	63%	61%	63%	63%	62%	54%	57%	48%	66%
$MVTecAD^{grid}$	65%	61%	66%	62%	62%	59%	71%	50%	34%	60%
$MVTecAD^{capsule}$	66%	67%	64%	66%	65%	65%	56%	50%	49%	70%
$MNISTC^{shear}$	74%	64%	59%	64%	64%	60%	50%	54%	50%	60%
letter	90%	57%	51%	53%	51%	52%	56%	53%	49%	58%
$MVTecAD^{metal\_nut}$	67%	63%	69%	64%	62%	67%	55%	50%	46%	66%
SpamBase	68%	79%	44%	66%	69%	71%	71%	58%	50%	55%

Table 3: AUC-ROC Scores for each datasets and algorithm (1/3|high performing algorithms)

Dataset	DEAN	LOF	KNN	CBLOF	NeuTral AE	IFor	PCA	D.SVDD	OCSVM	DTE
$20news^{2}$	56%	50%	48%	48%	57%	44%	46%	43%	44%	46%	49%
yeast	59%	47%	45%	47%	57%	42%	42%	43%	45%	45%	47%
vertebral	68%	53%	41%	46%	59%	66%	43%	41%	62%	41%	39%
$MNISTC^{identity}$	47%	49%	48%	50%	49%	50%	48%	48%	50%	47%	49%
speech	59%	53%	51%	49%	44%	49%	52%	49%	49%	48%	56%
imdb	53%	54%	52%	53%	48%	52%	49%	49%	48%	48%	56%
$20news^{5}$	56%	52%	50%	55%	57%	49%	46%	48%	51%	49%	55%
$WPBC$	54%	55%	55%	51%	43%	48%	55%	54%	49%	54%	47%
$Wilt$	67%	90%	82%	48%	80%	22%	45%	24%	44%	85%	35%
$20news^{4}$	59%	55%	50%	55%	62%	53%	53%	51%	51%	52%	46%
$20news^{1}$	64%	60%	61%	51%	62%	57%	47%	48%	52%	50%	47%
$agnews^{0}$	63%	67%	61%	56%	59%	61%	54%	51%	48%	50%	54%
$20news^{3}$	50%	55%	66%	55%	69%	53%	54%	55%	58%	53%	48%
$MVTecAD^{screw}$	60%	56%	59%	56%	58%	55%	58%	60%	60%	56%	47%
ALOI	55%	76%	70%	55%	54%	52%	56%	56%	56%	52%	54%
amazon	62%	59%	62%	57%	54%	58%	56%	56%	58%	55%	49%
$SVHN^{6}$	65%	61%	59%	55%	57%	55%	56%	56%	56%	59%	59%
$CIFAR10^{3}$	67%	66%	60%	62%	61%	59%	53%	56%	51%	60%	57%
$CIFAR10^{5}$	67%	63%	57%	60%	67%	57%	54%	57%	57%	60%	59%
$SVHN^{9}$	66%	64%	62%	58%	61%	60%	54%	57%	54%	57%	59%
$SVHN^{3}$	65%	66%	61%	58%	64%	56%	58%	59%	59%	56%	60%
$MNISTC^{rotate}$	67%	75%	67%	59%	58%	59%	57%	56%	55%	54%	63%
$CIFAR10^{2}$	61%	65%	60%	60%	56%	58%	58%	59%	55%	59%	61%
landsat	77%	75%	77%	67%	83%	60%	61%	40%	49%	47%	58%
$SVHN^{8}$	70%	67%	64%	59%	62%	63%	55%	57%	65%	55%	58%
yelp	67%	67%	67%	63%	62%	65%	60%	59%	42%	57%	52%
$agnews^{3}$	64%	75%	65%	60%	66%	63%	60%	58%	59%	57%	60%
$SVHN^{0}$	76%	74%	69%	62%	68%	68%	55%	59%	61%	61%	59%
$CIFAR10^{1}$	75%	76%	63%	63%	73%	58%	52%	62%	63%	62%	72%
$SVHN^{5}$	70%	66%	64%	63%	61%	65%	59%	62%	57%	58%	61%
$MVTecAD^{pill}$	62%	66%	67%	64%	67%	53%	65%	64%	61%	61%	52%
$agnews^{1}$	69%	83%	69%	61%	69%	65%	61%	60%	57%	57%	75%
$SVHN^{4}$	66%	65%	66%	63%	61%	64%	61%	60%	59%	61%	63%
$SVHN^{2}$	69%	69%	65%	62%	66%	65%	60%	62%	60%	60%	65%
census	63%	55%	67%	66%	54%	60%	66%	71%	69%	55%	61%
$fault$	75%	63%	80%	71%	73%	71%	66%	55%	54%	59%	72%
$Hepatitis$	44%	60%	53%	48%	55%	51%	82%	85%	70%	47%	83%
$SVHN^{7}$	66%	66%	64%	65%	61%	67%	64%	65%	65%	66%	67%
$SVHN^{1}$	68%	63%	67%	66%	66%	66%	63%	65%	65%	67%	66%
$Pima$	65%	67%	69%	68%	58%	70%	74%	72%	64%	62%	70%
$20news^{0}$	75%	78%	72%	64%	69%	64%	63%	63%	67%	61%	53%
$CIFAR10^{7}$	69%	71%	65%	65%	63%	61%	62%	65%	62%	68%	66%
$MNISTC^{translate}$	85%	91%	81%	66%	76%	69%	58%	61%	63%	55%	69%
$agnews^{2}$	74%	75%	74%	68%	64%	71%	65%	65%	61%	63%	53%
$MVTecAD^{grid}$	65%	68%	72%	65%	73%	70%	65%	64%	67%	65%	76%
$MVTecAD^{capsule}$	66%	67%	68%	71%	66%	64%	68%	66%	63%	65%	63%
$MNISTC^{shear}$	74%	79%	74%	70%	68%	70%	65%	66%	65%	59%	75%
letter	90%	88%	88%	78%	76%	81%	64%	54%	50%	90%	77%
$MVTecAD^{metal\_nut}$	67%	71%	73%	72%	72%	73%	68%	71%	71%	68%	75%
SpamBase	68%	64%	75%	70%	42%	70%	82%	80%	83%	76%	67%

Table 4: AUC-ROC Scores for each datasets and algorithm (2/3|low performing algorithms)

Dataset	DEAN	HBOS	GOAD	ECOD	COPOD	LODA	NF	DAGMM	VAE	SOD
celeba	68%	77%	64%	76%	75%	58%	80%	62%	69%	44%
$CIFAR10^{9}$	77%	60%	76%	64%	61%	65%	48%	62%	62%	59%
$FashionMNIST^{6}$	82%	52%	68%	60%	55%	68%	55%	62%	66%	44%
Waveform	73%	69%	79%	58%	73%	69%	67%	55%	30%	49%
optdigits	99%	88%	52%	52%	60%	50%	55%	46%	44%	21%
$MNISTC^{scale}$	89%	59%	56%	59%	57%	80%	53%	48%	61%	17%
$MVTecAD^{cable}$	67%	72%	66%	71%	71%	71%	51%	50%	51%	69%
$CIFAR10^{8}$	74%	66%	78%	68%	66%	68%	58%	61%	65%	58%
Cardiotocography	84%	57%	76%	79%	66%	79%	50%	74%	60%	39%
$CIFAR10^{6}$	77%	70%	72%	71%	71%	70%	56%	56%	63%	65%
InternetAds	86%	55%	75%	69%	69%	57%	79%	61%	71%	41%
$CIFAR10^{0}$	76%	70%	76%	71%	69%	72%	74%	58%	65%	63%
campaign	73%	80%	70%	77%	78%	65%	73%	56%	69%	63%
$MNISTC^{brightness}$	93%	64%	60%	64%	63%	67%	51%	46%	61%	45%
$MVTecAD^{carpet}$	74%	75%	74%	71%	74%	74%	60%	50%	53%	64%
satellite	77%	87%	79%	59%	64%	71%	54%	72%	50%	54%
$MVTecAD^{hazelnut}$	68%	74%	70%	69%	72%	71%	58%	65%	66%	70%
annthyroid	77%	71%	46%	81%	79%	60%	94%	67%	68%	62%
$MNISTC^{canny\_edges}$	93%	73%	48%	69%	68%	72%	43%	59%	83%	39%
cover	50%	65%	98%	92%	88%	95%	50%	69%	50%	10%
magic.gamma	83%	75%	76%	64%	68%	67%	70%	70%	68%	73%
glass	89%	85%	93%	65%	75%	67%	64%	50%	59%	68%
$MVTecAD^{toothbrush}$	72%	81%	72%	77%	73%	59%	67%	50%	57%	83%
$MVTecAD^{wood}$	74%	76%	75%	76%	76%	72%	78%	50%	76%	75%
mnist	53%	73%	92%	75%	78%	80%	49%	72%	50%	47%
$CIFAR10^{4}$	77%	76%	78%	76%	76%	74%	78%	57%	71%	71%
$MNISTC^{shot\_noise}$	93%	71%	69%	71%	70%	74%	44%	58%	66%	55%
PageBlocks	85%	88%	72%	90%	87%	76%	53%	92%	50%	45%
$FashionMNIST^{8}$	93%	70%	79%	73%	71%	74%	50%	67%	68%	52%
$MVTecAD^{transistor}$	75%	80%	75%	78%	79%	80%	69%	50%	76%	73%
backdoor	94%	65%	78%	84%	79%	25%	89%	56%	90%	53%
vowels	94%	65%	90%	56%	45%	71%	91%	52%	58%	54%
$MVTecAD^{zipper}$	77%	80%	77%	77%	80%	78%	67%	50%	59%	84%
$MVTecAD^{tile}$	79%	82%	80%	79%	81%	81%	71%	50%	54%	75%
$FashionMNIST^{4}$	90%	70%	85%	77%	73%	80%	53%	62%	71%	54%
wine	99%	85%	92%	68%	84%	78%	50%	50%	99%	19%
$MNISTC^{zigzag}$	95%	79%	66%	79%	77%	76%	47%	57%	67%	62%
skin	97%	77%	89%	49%	47%	82%	93%	90%	50%	55%
$MNISTC^{dotted\_line}$	95%	75%	68%	76%	74%	69%	66%	67%	78%	64%
$FashionMNIST^{2}$	92%	66%	85%	74%	70%	79%	80%	74%	65%	53%
$MNISTC^{spatter}$	93%	81%	80%	79%	79%	82%	42%	76%	49%	69%
$MNISTC^{motion\_blur}$	98%	79%	78%	77%	77%	77%	44%	76%	45%	63%
musk	53%	100%	87%	97%	96%	99%	53%	89%	50%	4%
$FashionMNIST^{0}$	91%	77%	81%	81%	78%	84%	59%	72%	80%	62%
donors	100%	79%	50%	89%	82%	60%	67%	90%	84%	60%
smtp	92%	82%	84%	90%	92%	87%	96%	85%	21%	63%
$FashionMNIST^{3}$	93%	82%	83%	84%	82%	77%	67%	58%	82%	61%
$MNISTC^{fog}$	100%	79%	83%	79%	78%	89%	66%	71%	49%	37%
mammography	84%	84%	86%	90%	90%	90%	78%	87%	50%	64%
Ionosphere	86%	72%	82%	71%	78%	79%	96%	50%	75%	88%

Table 5: AUC-ROC Scores for each datasets and algorithm (2/3|high performing algorithms)

Dataset	DEAN	LOF	KNN	CBLOF	NeuTral	AE	IFor	PCA	D.SVDD	OCSVM	DTE
celeba	68%	46%	62%	59%	48%	67%	69%	80%	78%	72%	84%
$CIFAR10^{9}$	77%	78%	71%	71%	74%	73%	65%	70%	62%	69%	75%
$FashionMNIST^{6}$	82%	82%	81%	74%	75%	78%	63%	71%	72%	65%	77%
Waveform	73%	80%	81%	83%	67%	70%	68%	64%	68%	84%	66%
optdigits	99%	100%	100%	89%	64%	98%	86%	52%	47%	100%	50%
$MNISTC^{scale}$	89%	94%	91%	84%	80%	83%	66%	73%	82%	65%	68%
$MVTecAD^{cable}$	67%	78%	81%	75%	71%	72%	72%	71%	62%	74%	72%
$CIFAR10^{8}$	74%	76%	72%	71%	74%	73%	69%	72%	67%	72%	70%
Cardiotocography	84%	77%	76%	72%	64%	82%	79%	82%	73%	83%	51%
$CIFAR10^{6}$	77%	77%	79%	75%	76%	76%	74%	74%	62%	68%	76%
InternetAds	86%	86%	82%	73%	87%	71%	47%	79%	76%	72%	73%
$CIFAR10^{0}$	76%	76%	75%	71%	76%	68%	71%	73%	65%	71%	74%
campaign	73%	59%	74%	68%	78%	69%	75%	77%	70%	69%	74%
$MNISTC^{brightness}$	93%	98%	92%	80%	80%	87%	73%	72%	76%	66%	84%
$MVTecAD^{carpet}$	74%	77%	78%	77%	74%	74%	76%	76%	75%	75%	71%
satellite	77%	83%	87%	84%	80%	62%	79%	66%	68%	87%	70%
$MVTecAD^{hazelnut}$	68%	81%	80%	77%	74%	79%	73%	72%	71%	69%	73%
annthyroid	77%	78%	78%	68%	85%	63%	92%	84%	80%	57%	58%
$MNISTC^{canny\_edges}$	93%	98%	93%	84%	80%	83%	73%	76%	68%	70%	80%
cover	50%	100%	100%	69%	99%	50%	88%	94%	93%	52%	50%
magic.gamma	83%	83%	84%	76%	78%	76%	78%	71%	69%	73%	86%
glass	89%	80%	100%	100%	97%	63%	89%	65%	73%	46%	59%
$MVTecAD^{toothbrush}$	72%	64%	87%	85%	88%	90%	87%	73%	76%	65%	85%
$MVTecAD^{wood}$	74%	77%	80%	77%	80%	78%	79%	78%	77%	75%	72%
mnist	53%	96%	94%	87%	98%	96%	87%	91%	87%	50%	50%
$CIFAR10^{4}$	77%	76%	80%	79%	78%	73%	77%	77%	79%	76%	79%
$MNISTC^{shot\_noise}$	93%	95%	96%	90%	81%	86%	78%	79%	74%	76%	84%
PageBlocks	85%	91%	66%	64%	97%	52%	92%	93%	90%	61%	66%
$FashionMNIST^{8}$	93%	93%	92%	86%	72%	88%	77%	80%	72%	75%	90%
$MVTecAD^{transistor}$	75%	85%	79%	81%	81%	74%	82%	81%	74%	81%	72%
backdoor	94%	95%	95%	83%	90%	86%	76%	64%	57%	87%	89%
vowels	94%	97%	97%	90%	98%	90%	76%	61%	79%	81%	98%
$MVTecAD^{zipper}$	77%	88%	87%	84%	90%	79%	81%	81%	78%	79%	78%
$MVTecAD^{tile}$	79%	85%	86%	83%	79%	79%	84%	80%	79%	80%	84%
$FashionMNIST^{4}$	90%	88%	88%	85%	87%	86%	78%	84%	84%	82%	82%
wine	99%	99%	99%	99%	84%	100%	85%	90%	89%	90%	2%
$MNISTC^{zigzag}$	95%	96%	94%	85%	89%	89%	84%	85%	88%	78%	92%
skin	97%	93%	100%	91%	89%	89%	89%	60%	66%	90%	92%
$MNISTC^{dotted\_line}$	95%	97%	95%	84%	87%	87%	80%	82%	80%	80%	86%
$FashionMNIST^{2}$	92%	88%	91%	89%	90%	89%	79%	83%	81%	78%	90%
$MNISTC^{spatter}$	93%	96%	93%	86%	88%	90%	83%	85%	82%	77%	92%
$MNISTC^{motion\_blur}$	98%	98%	97%	89%	93%	92%	85%	86%	84%	75%	97%
musk	53%	100%	100%	100%	100%	100%	97%	100%	100%	50%	46%
$FashionMNIST^{0}$	91%	91%	92%	88%	90%	90%	82%	86%	81%	81%	88%
donors	100%	99%	100%	93%	40%	85%	92%	89%	92%	87%	99%
smtp	92%	93%	95%	86%	78%	80%	90%	84%	78%	84%	90%
$FashionMNIST^{3}$	93%	93%	92%	89%	87%	91%	83%	88%	86%	84%	93%
$MNISTC^{fog}$	100%	100%	100%	97%	98%	99%	89%	91%	87%	82%	99%
mammography	84%	84%	86%	87%	74%	91%	88%	90%	91%	88%	86%
Ionosphere	86%	91%	94%	93%	95%	88%	87%	87%	90%	80%	93%

Table 6: AUC-ROC Scores for each datasets and algorithm (3/3|low performing algorithms)

Dataset	DEAN	HBOS	GOAD	ECOD	COPOD	LODA	NF	DAGMM	VAE	SOD
shuttle	100%	98%	82%	99%	99%	82%	9%	95%	50%	31%
pendigits	99%	94%	90%	92%	90%	88%	67%	64%	89%	21%
$MNISTC^{glass\_blur}$	100%	90%	92%	89%	88%	90%	74%	62%	52%	54%
cardio	89%	84%	95%	93%	91%	95%	90%	69%	95%	45%
http	100%	99%	1%	97%	99%	43%	99%	99%	100%	40%
$MVTecAD^{bottle}$	96%	96%	92%	92%	96%	95%	91%	50%	7%	97%
Stamps	89%	90%	88%	90%	91%	91%	89%	72%	91%	52%
*satimage*2	100%	98%	92%	97%	98%	99%	61%	99%	50%	46%
WDBC	100%	99%	93%	97%	100%	100%	50%	82%	50%	79%
Lymphography	100%	100%	100%	100%	100%	58%	69%	50%	94%	58%
WBC	99%	99%	99%	100%	100%	99%	85%	50%	99%	75%
$FashionMNIST^{5}$	96%	92%	96%	92%	91%	94%	73%	67%	79%	78%
$MNISTC^{stripe}$	100%	99%	90%	97%	97%	98%	40%	66%	87%	52%
fraud	94%	96%	91%	95%	95%	93%	92%	93%	95%	67%
thyroid	98%	99%	74%	98%	94%	93%	99%	83%	86%	66%
$FashionMNIST^{1}$	99%	92%	96%	94%	93%	95%	80%	73%	93%	76%
$FashionMNIST^{9}$	98%	94%	97%	95%	94%	94%	68%	83%	91%	83%
$MVTecAD^{leather}$	99%	99%	99%	97%	98%	98%	92%	50%	72%	98%
$MNISTC^{impulse\_noise}$	100%	99%	100%	98%	98%	100%	73%	98%	94%	41%
$FashionMNIST^{7}$	98%	95%	96%	96%	95%	95%	91%	89%	92%	89%
breastw	100%	99%	99%	99%	100%	99%	97%	50%	100%	91%
Average	78%	71%	71%	70%	70%	69%	61%	61%	61%	56%
Rank	5.64	12.12	11.08	12.83	12.86	12.05	15.26	15.81	15.37	16.47

Table 7: AUC-ROC Scores for each datasets and algorithm (3/3|high performing algorithms)

Dataset	DEAN	LOF	KNN	CBLOF	NeuTral	AE	IFor	PCA	D.SVDD	OCSVM	DTE
shuttle	100%	100%	100%	99%	100%	100%	100%	99%	99%	100%	50%
pendigits	99%	99%	100%	97%	62%	89%	98%	93%	94%	94%	98%
$MNISTC^{glass\_blur}$	100%	99%	100%	98%	96%	99%	95%	96%	97%	92%	99%
cardio	89%	93%	91%	92%	86%	91%	93%	95%	92%	94%	92%
http	100%	93%	100%	99%	100%	99%	99%	100%	100%	100%	99%
$MVTecAD^{bottle}$	96%	96%	96%	97%	96%	96%	97%	96%	96%	96%	95%
Stamps	89%	93%	95%	93%	99%	90%	92%	92%	93%	91%	92%
*satimage*2	100%	99%	100%	100%	100%	99%	100%	98%	97%	97%	50%
WDBC	100%	100%	100%	100%	96%	100%	100%	100%	100%	100%	40%
Lymphography	100%	97%	100%	100%	72%	100%	100%	100%	97%	100%	94%
WBC	99%	92%	99%	100%	72%	99%	99%	99%	93%	99%	40%
$FashionMNIST^{5}$	96%	93%	96%	96%	96%	95%	93%	94%	94%	94%	95%
$MNISTC^{stripe}$	100%	100%	100%	100%	100%	100%	99%	100%	100%	97%	100%
fraud	94%	74%	97%	96%	92%	95%	96%	96%	94%	95%	96%
thyroid	98%	98%	97%	94%	99%	95%	99%	98%	97%	88%	93%
$FashionMNIST^{1}$	99%	98%	99%	97%	97%	99%	95%	97%	95%	96%	99%
$FashionMNIST^{9}$	98%	98%	97%	96%	98%	97%	95%	96%	96%	96%	97%
$MVTecAD^{leather}$	99%	98%	99%	99%	99%	99%	99%	99%	99%	99%	99%
$MNISTC^{impulse\_noise}$	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
$FashionMNIST^{7}$	98%	97%	97%	96%	97%	97%	95%	96%	96%	96%	96%
breastw	100%	96%	100%	100%	86%	99%	100%	99%	99%	99%	91%
Average	78%	80%	80%	76%	75%	75%	74%	73%	72%	72%	71%
Rank	5.64	5.00	4.33	7.13	7.35	8.31	8.93	9.00	10.45	10.81	9.20

Table 8: AUC-PR Scores for each datasets and algorithm (1/3|low performing algorithms)

Dataset	DEAN	GOAD	HBOS	ECOD	COPOD	LODA	NF	VAE	DAGMM	SOD
$20news^{2}$	56%	45%	44%	45%	44%	44%	44%	52%	45%	45%
imdb	53%	50%	48%	46%	49%	48%	54%	44%	46%	46%
$WPBC$	54%	46%	48%	47%	49%	50%	49%	49%	50%	45%
$MNISTC^{identity}$	47%	49%	49%	49%	49%	49%	48%	50%	50%	48%
vertebral	68%	58%	42%	43%	39%	37%	75%	50%	50%	41%
yeast	59%	61%	49%	50%	46%	54%	49%	45%	53%	45%
$20news^{1}$	64%	50%	49%	47%	49%	45%	47%	45%	47%	50%
speech	59%	50%	50%	50%	52%	48%	52%	50%	55%	39%
$20news^{5}$	56%	50%	50%	50%	48%	49%	51%	56%	47%	49%
$20news^{4}$	59%	52%	51%	51%	50%	52%	45%	52%	53%	48%
$20news^{3}$	50%	51%	54%	51%	56%	49%	59%	42%	48%	50%
$Wilt$	67%	57%	41%	43%	39%	42%	57%	50%	59%	38%
$agnews^{0}$	63%	51%	52%	50%	52%	46%	50%	49%	50%	50%
amazon	62%	54%	55%	53%	56%	49%	51%	47%	50%	52%
census	63%	52%	57%	58%	59%	39%	48%	54%	56%	53%
$CIFAR10^{2}$	61%	57%	52%	53%	53%	59%	53%	54%	58%	50%
$MVTecAD^{screw}$	60%	53%	59%	58%	58%	55%	47%	59%	50%	55%
$ALOI$	55%	50%	54%	54%	52%	57%	56%	55%	55%	54%
$MNISTC^{rotate}$	67%	49%	54%	53%	53%	55%	53%	56%	55%	50%
$CIFAR10^{5}$	67%	67%	47%	50%	47%	61%	55%	54%	62%	48%
$agnews^{3}$	64%	53%	54%	54%	54%	54%	52%	50%	53%	54%
landsat	77%	59%	68%	42%	45%	44%	45%	52%	53%	49%
$SVHN^{3}$	65%	59%	53%	55%	54%	56%	68%	49%	46%	54%
$CIFAR10^{3}$	67%	67%	50%	53%	51%	53%	47%	56%	51%	53%
$agnews^{1}$	69%	48%	53%	55%	49%	50%	48%	57%	51%	50%
$SVHN^{9}$	66%	58%	52%	54%	53%	51%	54%	51%	55%	58%
yelp	67%	58%	59%	56%	59%	54%	50%	45%	59%	54%
$SVHN^{8}$	70%	60%	50%	53%	52%	49%	56%	49%	54%	59%
$CIFAR10^{1}$	75%	68%	45%	49%	47%	58%	49%	52%	56%	50%
$SVHN^{6}$	65%	63%	53%	55%	54%	64%	57%	53%	57%	54%
$SVHN^{0}$	76%	62%	52%	53%	52%	56%	44%	55%	60%	54%
$SVHN^{5}$	70%	61%	56%	58%	57%	56%	57%	50%	58%	58%
$SVHN^{1}$	68%	66%	60%	60%	59%	59%	49%	63%	63%	48%
$20news^{0}$	75%	57%	59%	58%	59%	60%	55%	53%	55%	59%
$SVHN^{2}$	69%	62%	59%	61%	59%	57%	57%	55%	52%	60%
$Hepatitis$	44%	55%	74%	60%	67%	66%	75%	50%	50%	50%
$SVHN^{4}$	66%	54%	62%	62%	62%	63%	56%	63%	60%	56%
$Pima$	65%	59%	67%	61%	67%	60%	25%	50%	50%	55%
$fault$	75%	65%	66%	47%	46%	49%	60%	54%	49%	62%
$MNISTC^{translate}$	85%	55%	52%	54%	54%	59%	50%	51%	53%	57%
$SVHN^{7}$	66%	60%	64%	63%	63%	67%	55%	63%	60%	51%
$MVTecAD^{pill}$	62%	57%	65%	64%	66%	63%	60%	53%	50%	67%
$CIFAR10^{7}$	69%	70%	58%	61%	59%	64%	55%	64%	58%	57%
$agnews^{2}$	74%	62%	63%	61%	63%	64%	56%	50%	59%	64%
$MNISTC^{scale}$	89%	51%	53%	53%	52%	60%	53%	55%	46%	34%
$MNISTC^{shear}$	74%	61%	64%	64%	64%	67%	51%	52%	57%	60%
$FashionMNIST^{6}$	82%	68%	49%	55%	51%	61%	53%	61%	61%	47%
$letter$	90%	51%	54%	55%	53%	50%	58%	49%	58%	57%
$MVTecAD^{capsule}$	66%	69%	68%	69%	68%	63%	59%	54%	50%	72%
$MVTecAD^{grid}$	65%	68%	61%	64%	64%	70%	77%	44%	50%	64%

Table 9: AUC-PR Scores for each datasets and algorithm (1/3|high performing algorithms)

Dataset	DEAN	LOF	KNN	NeuTral	CBLOF	PCA	DTE	IFor	OCSVM	D.SVDD	AE
$20news^{2}$	56%	49%	47%	57%	44%	44%	51%	44%	45%	43%	47%
imdb	53%	52%	48%	50%	49%	47%	53%	49%	46%	45%	49%
WPBC	54%	50%	50%	46%	47%	51%	48%	52%	52%	51%	48%
$MNISTC^{identity}$	47%	50%	49%	50%	51%	49%	50%	50%	48%	49%	54%
vertebral	68%	51%	43%	60%	47%	42%	40%	42%	53%	44%	68%
yeast	59%	50%	49%	57%	48%	47%	49%	47%	48%	45%	48%
$20news^{1}$	64%	60%	61%	63%	50%	48%	48%	49%	48%	52%	53%
speech	59%	55%	53%	46%	51%	51%	58%	48%	51%	58%	51%
$20news^{5}$	56%	53%	51%	58%	53%	50%	54%	50%	51%	55%	47%
$20news^{4}$	59%	54%	49%	62%	52%	52%	48%	51%	52%	52%	51%
$20news^{3}$	50%	52%	67%	67%	52%	54%	50%	55%	50%	52%	51%
Wilt	67%	89%	70%	78%	47%	36%	39%	47%	85%	40%	37%
$agnews^{0}$	63%	65%	60%	60%	55%	52%	51%	54%	50%	51%	58%
amazon	62%	54%	56%	57%	55%	55%	51%	55%	54%	52%	56%
census	63%	50%	59%	55%	53%	65%	56%	55%	50%	65%	53%
$CIFAR10^{2}$	61%	62%	58%	56%	59%	56%	58%	53%	56%	55%	56%
$MVTecAD^{screw}$	60%	55%	59%	59%	54%	63%	52%	60%	57%	61%	50%
ALOI	55%	74%	71%	55%	56%	56%	55%	54%	55%	56%	55%
$MNISTC^{rotate}$	67%	74%	67%	58%	56%	56%	62%	54%	52%	55%	56%
$CIFAR10^{5}$	67%	66%	59%	68%	59%	58%	59%	52%	60%	53%	50%
$agnews^{3}$	64%	75%	64%	66%	58%	55%	58%	55%	54%	57%	56%
landsat	77%	80%	75%	82%	64%	45%	58%	62%	48%	42%	56%
$SVHN^{3}$	65%	66%	60%	64%	57%	58%	61%	57%	54%	58%	52%
$CIFAR10^{3}$	67%	68%	63%	64%	62%	56%	59%	56%	59%	54%	59%
$agnews^{1}$	69%	83%	66%	70%	57%	55%	71%	55%	54%	56%	57%
$SVHN^{9}$	66%	66%	65%	62%	59%	59%	63%	55%	55%	59%	60%
yelp	67%	63%	63%	62%	59%	59%	52%	61%	57%	58%	60%
$SVHN^{8}$	70%	68%	66%	64%	60%	60%	60%	56%	54%	61%	59%
$CIFAR10^{1}$	75%	76%	61%	73%	59%	60%	71%	51%	59%	54%	55%
$SVHN^{6}$	65%	62%	61%	59%	63%	59%	63%	56%	59%	63%	51%
$SVHN^{0}$	76%	72%	69%	67%	64%	60%	62%	57%	58%	59%	61%
$SVHN^{5}$	70%	66%	65%	63%	63%	62%	63%	60%	57%	62%	59%
$SVHN^{1}$	68%	58%	64%	66%	66%	61%	62%	62%	66%	56%	58%
$20news^{0}$	75%	75%	71%	69%	61%	59%	52%	61%	58%	58%	62%
$SVHN^{2}$	69%	66%	65%	67%	62%	62%	65%	61%	60%	63%	58%
Hepatitis	44%	64%	50%	57%	50%	76%	87%	63%	60%	67%	58%
$SVHN^{4}$	66%	61%	66%	61%	66%	60%	63%	61%	62%	62%	58%
Pima	65%	65%	69%	59%	68%	68%	67%	69%	71%	66%	68%
fault	75%	60%	76%	71%	70%	57%	71%	64%	60%	61%	72%
$MNISTC^{translate}$	85%	89%	77%	75%	61%	60%	70%	57%	55%	62%	60%
$SVHN^{7}$	66%	64%	64%	61%	64%	62%	67%	62%	67%	61%	62%
$MVTecAD^{pill}$	62%	69%	68%	68%	67%	65%	60%	66%	64%	65%	60%
$CIFAR10^{7}$	69%	73%	67%	63%	65%	66%	67%	61%	66%	61%	57%
$agnews^{2}$	74%	74%	73%	65%	67%	64%	53%	66%	63%	60%	62%
$MNISTC^{scale}$	89%	91%	89%	81%	77%	70%	66%	66%	60%	69%	72%
$MNISTC^{shear}$	74%	80%	75%	67%	69%	67%	77%	65%	62%	67%	62%
$FashionMNIST^{6}$	82%	86%	82%	74%	71%	68%	77%	59%	62%	73%	68%
letter	90%	89%	86%	75%	78%	56%	79%	58%	89%	54%	78%
$MVTecAD^{capsule}$	66%	70%	71%	67%	73%	69%	68%	69%	68%	68%	55%
$MVTecAD^{grid}$	65%	73%	77%	74%	69%	67%	78%	69%	67%	67%	57%

Table 10: AUC-PR Scores for each datasets and algorithm (2/3|low performing algorithms)

Dataset	DEAN	GOAD	HBOS	ECOD	COPOD	LODA	NF	VAE	DAGMM	SOD
$MVTecAD^{metal\_nut}$	67%	74%	60%	62%	60%	71%	56%	48%	50%	68%
$SpamBase$	68%	54%	76%	61%	63%	72%	77%	50%	55%	53%
$celeba$	68%	66%	78%	77%	76%	58%	73%	69%	59%	42%
$optdigits$	99%	48%	84%	47%	52%	46%	64%	46%	45%	35%
$CIFAR10^{9}$	77%	75%	62%	65%	63%	71%	53%	64%	63%	59%
$CIFAR10^{8}$	74%	78%	63%	65%	64%	75%	56%	65%	62%	57%
$CIFAR10^{6}$	77%	72%	65%	66%	65%	62%	60%	61%	56%	61%
$Waveform$	73%	74%	64%	58%	68%	65%	77%	39%	59%	51%
$MNISTC^{brightness}$	93%	57%	60%	60%	59%	65%	51%	58%	51%	47%
$CIFAR10^{0}$	76%	75%	69%	69%	68%	66%	73%	64%	57%	61%
$MVTecAD^{cable}$	67%	70%	72%	72%	71%	80%	56%	54%	50%	74%
$MNISTC^{canny\_edges}$	93%	44%	67%	63%	62%	65%	47%	81%	60%	41%
$skin$	97%	80%	66%	45%	44%	64%	85%	50%	76%	50%
$campaign$	73%	72%	80%	77%	78%	66%	74%	70%	54%	59%
$Cardiotocography$	84%	74%	63%	76%	67%	76%	75%	59%	72%	46%
$annthyroid$	77%	49%	77%	79%	72%	57%	93%	69%	70%	61%
$cover$	50%	97%	65%	89%	85%	89%	25%	50%	70%	32%
$MVTecAD^{carpet}$	74%	77%	77%	73%	76%	75%	66%	59%	50%	71%
$MVTecAD^{hazelnut}$	68%	69%	78%	73%	76%	77%	58%	70%	64%	69%
$InternetAds$	86%	80%	60%	75%	75%	43%	86%	78%	64%	43%
$MNISTC^{shot\_noise}$	93%	63%	66%	67%	66%	82%	52%	64%	56%	52%
$CIFAR10^{4}$	77%	79%	75%	76%	75%	76%	78%	71%	56%	69%
$FashionMNIST^{8}$	93%	74%	67%	69%	68%	76%	74%	64%	67%	52%
$MVTecAD^{toothbrush}$	72%	75%	85%	80%	75%	54%	70%	63%	50%	87%
$backdoor$	94%	67%	58%	78%	73%	38%	94%	93%	61%	62%
$MNISTC^{zigzag}$	95%	59%	73%	73%	71%	73%	49%	65%	57%	60%
$vowels$	94%	90%	67%	62%	45%	65%	86%	58%	51%	50%
$donors$	100%	46%	71%	84%	78%	39%	67%	71%	85%	62%
$satellite$	77%	82%	89%	67%	71%	81%	55%	50%	82%	55%
$FashionMNIST^{4}$	90%	84%	65%	73%	68%	77%	57%	68%	60%	54%
$MNISTC^{dotted\_line}$	95%	59%	70%	71%	69%	72%	62%	78%	67%	61%
$FashionMNIST^{2}$	92%	83%	58%	66%	61%	77%	77%	60%	73%	53%
$mnist$	53%	91%	68%	68%	73%	81%	56%	50%	72%	50%
$PageBlocks$	85%	74%	84%	87%	83%	75%	66%	50%	91%	55%
$magic.gamma$	83%	81%	76%	67%	71%	73%	76%	73%	73%	75%
$MVTecAD^{zipper}$	77%	75%	79%	77%	79%	75%	74%	65%	50%	82%
$MVTecAD^{wood}$	74%	79%	79%	80%	80%	76%	82%	80%	50%	76%
$glass$	89%	90%	86%	71%	79%	64%	69%	68%	50%	82%
$MVTecAD^{transistor}$	75%	79%	83%	83%	83%	76%	71%	80%	50%	76%
$wine$	99%	89%	85%	59%	71%	61%	75%	99%	50%	35%
$MNISTC^{spatter}$	93%	78%	76%	75%	75%	79%	43%	48%	77%	66%
$MNISTC^{motion\_blur}$	98%	73%	76%	73%	73%	82%	53%	47%	72%	58%
$MVTecAD^{tile}$	79%	84%	86%	84%	86%	82%	79%	58%	50%	81%
$FashionMNIST^{0}$	91%	78%	73%	77%	74%	82%	58%	78%	71%	61%
$MNISTC^{fog}$	100%	81%	73%	74%	73%	78%	61%	45%	71%	41%
$FashionMNIST^{3}$	93%	83%	77%	79%	78%	82%	68%	79%	62%	62%
$http$	100%	26%	86%	75%	85%	34%	90%	100%	87%	32%
$Stamps$	89%	89%	76%	82%	77%	78%	82%	84%	71%	48%
$Ionosphere$	86%	80%	62%	74%	76%	69%	96%	79%	50%	92%
$mammography$	84%	89%	82%	92%	92%	89%	72%	50%	88%	58%

Table 11: AUC-PR Scores for each datasets and algorithm (2/3|high performing algorithms)

Dataset	DEAN	LOF	KNN	NeuTral	CBLOF	PCA	DTE	IFor	OCSVM	D.SVDD	AE
$MVTecAD^{metal\_nut}$	67%	77%	78%	72%	77%	68%	80%	68%	70%	71%	59%
SpamBase	68%	64%	75%	45%	70%	79%	66%	80%	76%	79%	72%
celeba	68%	45%	62%	49%	67%	81%	77%	74%	73%	76%	67%
optdigits	99%	100%	100%	66%	72%	46%	75%	79%	100%	43%	97%
$CIFAR10^{9}$	77%	78%	72%	75%	70%	70%	73%	67%	69%	68%	63%
$CIFAR10^{8}$	74%	76%	70%	74%	69%	70%	68%	66%	72%	66%	66%
$CIFAR10^{6}$	77%	76%	75%	75%	71%	70%	75%	67%	66%	72%	68%
Waveform	73%	84%	83%	68%	85%	62%	67%	73%	85%	51%	76%
$MNISTC^{brightness}$	93%	98%	90%	81%	77%	70%	84%	66%	62%	73%	72%
$CIFAR10^{0}$	76%	74%	73%	76%	70%	72%	74%	68%	70%	74%	64%
$MVTecAD^{cable}$	67%	81%	82%	71%	74%	70%	75%	76%	74%	65%	63%
$MNISTC^{canny\_edges}$	93%	97%	91%	80%	78%	72%	79%	69%	63%	84%	68%
skin	97%	83%	100%	90%	79%	52%	80%	76%	76%	60%	51%
campaign	73%	53%	74%	76%	70%	77%	75%	73%	72%	71%	68%
Cardiotocography	84%	75%	74%	64%	75%	81%	55%	77%	82%	78%	84%
annthyroid	77%	81%	78%	83%	70%	84%	59%	91%	57%	84%	61%
cover	50%	100%	100%	98%	59%	90%	75%	78%	76%	78%	51%
$MVTecAD^{carpet}$	74%	81%	81%	73%	80%	78%	74%	78%	78%	76%	67%
$MVTecAD^{hazelnut}$	68%	85%	82%	75%	80%	76%	79%	78%	73%	77%	60%
InternetAds	86%	89%	86%	86%	80%	82%	76%	45%	78%	82%	77%
$MNISTC^{shot\_noise}$	93%	94%	95%	79%	87%	77%	86%	72%	73%	80%	74%
$CIFAR10^{4}$	77%	77%	80%	78%	78%	78%	79%	76%	76%	76%	64%
$FashionMNIST^{8}$	93%	93%	89%	73%	81%	76%	88%	73%	70%	73%	78%
$MVTecAD^{toothbrush}$	72%	64%	88%	87%	86%	80%	88%	89%	72%	79%	56%
backdoor	94%	96%	96%	91%	72%	58%	92%	67%	78%	61%	77%
$MNISTC^{zigzag}$	95%	96%	93%	88%	81%	84%	93%	77%	72%	82%	73%
vowels	94%	96%	96%	98%	80%	63%	96%	78%	82%	69%	89%
donors	100%	99%	100%	42%	83%	82%	97%	84%	76%	68%	89%
satellite	77%	88%	90%	79%	89%	78%	70%	85%	89%	77%	71%
$FashionMNIST^{4}$	90%	90%	89%	86%	84%	84%	86%	72%	79%	84%	74%
$MNISTC^{dotted\_line}$	95%	96%	94%	86%	81%	81%	87%	73%	75%	78%	71%
$FashionMNIST^{2}$	92%	90%	90%	90%	86%	83%	92%	73%	73%	83%	78%
mnist	53%	96%	94%	96%	86%	90%	75%	84%	75%	88%	96%
PageBlocks	85%	92%	70%	96%	63%	91%	74%	90%	71%	88%	58%
magic.gamma	83%	85%	86%	79%	79%	74%	88%	78%	77%	72%	79%
$MVTecAD^{zipper}$	77%	88%	86%	89%	84%	81%	82%	80%	80%	78%	70%
$MVTecAD^{wood}$	74%	83%	84%	79%	83%	83%	80%	82%	80%	84%	60%
glass	89%	88%	100%	96%	100%	71%	61%	88%	61%	67%	72%
$MVTecAD^{transistor}$	75%	88%	83%	81%	83%	84%	78%	86%	84%	79%	66%
wine	99%	99%	99%	83%	99%	85%	31%	74%	92%	86%	100%
$MNISTC^{spatter}$	93%	97%	94%	87%	87%	85%	93%	84%	76%	85%	77%
$MNISTC^{motion\_blur}$	98%	98%	96%	91%	86%	85%	97%	80%	72%	85%	83%
$MVTecAD^{tile}$	79%	88%	88%	77%	88%	80%	87%	87%	83%	79%	58%
$FashionMNIST^{0}$	91%	91%	92%	91%	86%	84%	88%	78%	78%	80%	81%
$MNISTC^{fog}$	100%	100%	100%	97%	95%	90%	98%	81%	81%	85%	93%
$FashionMNIST^{3}$	93%	93%	92%	87%	87%	88%	93%	79%	81%	89%	78%
http	100%	59%	100%	99%	91%	99%	93%	79%	100%	99%	99%
Stamps	89%	89%	91%	98%	90%	85%	84%	79%	83%	89%	86%
Ionosphere	86%	91%	94%	94%	95%	89%	94%	86%	84%	85%	88%
mammography	84%	86%	88%	74%	85%	91%	86%	90%	89%	91%	92%

Table 12: AUC-PR Scores for each datasets and algorithm (3/3|low performing algorithms)

Dataset	DEAN	GOAD	HBOS	ECOD	COPOD	LODA	NF	VAE	DAGMM	SOD
musk	53%	74%	100%	97%	95%	99%	76%	50%	92%	31%
pendigits	99%	84%	92%	90%	87%	96%	58%	88%	56%	37%
smtp	92%	89%	88%	91%	93%	92%	96%	36%	89%	65%
$MNISTC^{glass\_blur}$	100%	89%	86%	84%	83%	93%	79%	49%	63%	52%
cardio	89%	94%	85%	90%	89%	89%	91%	91%	69%	48%
shuttle	100%	87%	99%	99%	100%	90%	32%	50%	95%	42%
WBC	99%	99%	99%	100%	100%	94%	71%	99%	50%	70%
*satimage*2	100%	86%	98%	98%	98%	99%	53%	50%	99%	53%
$MVTecAD^{bottle}$	96%	95%	97%	93%	97%	96%	93%	31%	50%	97%
WDBC	100%	93%	99%	98%	100%	100%	75%	50%	77%	79%
thyroid	98%	73%	99%	98%	90%	91%	99%	86%	83%	58%
$FashionMNIST^{5}$	96%	97%	93%	94%	93%	96%	77%	82%	68%	76%
$MNISTC^{stripe}$	100%	86%	98%	96%	96%	99%	55%	86%	65%	51%
Lymphography	100%	100%	100%	100%	100%	96%	82%	94%	50%	49%
$FashionMNIST^{1}$	99%	93%	91%	92%	91%	95%	75%	92%	73%	74%
fraud	94%	94%	97%	97%	96%	97%	95%	97%	95%	67%
$FashionMNIST^{9}$	98%	97%	94%	95%	94%	95%	78%	93%	84%	81%
breastw	100%	99%	99%	99%	100%	99%	94%	100%	50%	88%
$MVTecAD^{leather}$	99%	99%	99%	97%	98%	91%	95%	82%	50%	98%
$MNISTC^{impulse\_noise}$	100%	100%	98%	97%	96%	100%	84%	95%	97%	44%
$FashionMNIST^{7}$	98%	97%	96%	96%	96%	97%	94%	93%	91%	90%
Average	78%	70%	69%	69%	68%	68%	64%	62%	61%	57%
Rank	5.72	10.90	12.34	12.62	12.95	12.14	14.19	14.95	15.38	16.28

Table 13: AUC-PR Scores for each datasets and algorithm (3/3|high performing algorithms)

Dataset	DEAN	LOF	KNN	NeuTral	CBLOF	PCA	DTE	IFor	OCSVM	D.SVDD	AE
musk	53%	100%	100%	99%	100%	100%	43%	96%	75%	100%	100%
pendigits	99%	98%	100%	61%	98%	90%	98%	96%	91%	78%	84%
smtp	92%	95%	95%	77%	90%	89%	92%	80%	88%	87%	65%
$MNISTC^{glass\_blur}$	100%	99%	100%	94%	97%	95%	99%	91%	89%	95%	93%
cardio	89%	91%	90%	85%	89%	94%	90%	93%	91%	97%	91%
shuttle	100%	100%	99%	98%	98%	99%	75%	100%	100%	99%	99%
WBC	99%	92%	99%	73%	99%	99%	45%	99%	99%	97%	98%
*satimage*2	100%	99%	100%	97%	100%	99%	75%	99%	97%	83%	99%
$MVTecAD^{bottle}$	96%	97%	97%	93%	97%	97%	97%	97%	97%	97%	87%
$WDBC$	100%	100%	100%	95%	100%	100%	40%	100%	100%	100%	100%
thyroid	98%	98%	97%	98%	92%	98%	93%	99%	88%	98%	85%
$FashionMNIST^{5}$	96%	95%	97%	96%	96%	96%	96%	95%	96%	95%	90%
$MNISTC^{stripe}$	100%	100%	100%	98%	100%	100%	100%	99%	97%	100%	99%
Lymphography	100%	97%	100%	71%	100%	100%	94%	100%	100%	100%	100%
$FashionMNIST^{1}$	99%	98%	98%	98%	94%	96%	98%	94%	94%	96%	96%
$fraud$	94%	71%	98%	92%	97%	97%	97%	96%	97%	96%	97%
$FashionMNIST^{9}$	98%	98%	98%	96%	97%	96%	98%	95%	96%	96%	92%
breastw	100%	92%	100%	83%	100%	99%	89%	100%	99%	98%	99%
$MVTecAD^{leather}$	99%	98%	99%	97%	99%	99%	99%	99%	99%	98%	97%
$MNISTC^{impulse\_noise}$	100%	100%	100%	98%	100%	100%	100%	99%	100%	100%	100%
$FashionMNIST^{7}$	98%	98%	98%	95%	97%	97%	97%	97%	97%	96%	90%
Average	78%	80%	80%	75%	75%	73%	73%	72%	72%	72%	70%
Rank	5.72	4.88	4.57	7.87	7.33	8.62	8.19	9.61	10.24	9.91	11.31

References

1. Alabdulatif, A., Kumarage, H., Khalil, I., Yi, X.: Privacy-preserving anomaly detection in cloud with lightweight homomorphic encryption.
2. Böing, B., Klüttermann, S., Müller, E.: Post-robustifying deep anomaly detection ensembles by model selection. p. In: 2022 IEEE International Conference on Data Mining (ICDM)
3. Chen, J., Sathe, S., Aggarwal, C.C., Turaga, D.S.: Outlier detection with autoencoder ensembles. pp. In: International Conference on Data Mining – ICDM
4. Dietterich, T.G., Zemicheal, T.: Anomaly detection in the presence of missing values
5. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure
6. Kieu, T., Yang, B., Guo, C., Jensen, C.S., Zhao, Y., Huang, F., Zheng, K.: Robust and explainable autoencoders for unsupervised time series outlier detection. p. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE)
7. Klüttermann, S., Balestra, C., Müller, E.: On the efficient explanation of outlier detection ensembles through shapley values. pp. In: Yang, D.N., Xie, X., Tseng, V.S., Pei, J., Huang, J.W., Lin, J.C.W. (eds.) Advances in Knowledge Discovery and Data Mining. Springer Nature Singapore, Singapore
8. Klüttermann, S., Müller, E.: About test-time training for outlier detection
9. Lazarevic, A., Kumar, V.: Feature bagging for outlier detection. p. vol. 21
10. Lochner, M., Bassett, B.: Astronomaly: Personalised active anomaly detection in astronomical data.
11. Lu, Z., Pu, H., Wang, F., Hu, Z., Wang, L.: The expressive power of neural networks: a view from the width. p. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, Curran Associates Inc., Red Hook, NY, USA
12. Menges, F., Latzo, T., Vielberth, M., Sobola, S., Pöhls, H.C., Taubmann, B., Köstler, J., Puchta, A., Freiling, F., Reiser, H.P., Pernul, G.: Towards gdpr-compliant data processing in modern siem systems.
13. Paisner, M., Cox, M.T., Perlis, D.: Symbolic anomaly detection and assessment using growing neural gas. p. In: 2013 IEEE 25th International Conference on Tools with Artificial Intelligence
14. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. p. In: ICML
15. Ruff, L., Vandermeulen, R.A., Görnitz, N., Binder, A., Müller, E., Müller, K., Kloft, M.: Deep semi-supervised anomaly detection.
16. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: p. Learning internal representations by error propagation
17. Sedjelmaci, H., Senouci, S.M., Al-Bahri, M.: A lightweight anomaly detection technique for low-resource iot devices: A game-theoretic methodology. p. In: 2016 IEEE International Conference on Communications (ICC)
18. Stocco, A., Tonella, P.: Towards anomaly detectors that learn continuously. p. In: 2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)
19. W Flores, A., Bechtel, K., Lowenkamp, C.: False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.".

20. Wang, X., Wang, Y., Javaheri, Z., Almutairi, L., Moghadamnejad, N., Younes, O.S.: Federated deep learning for anomaly detection in the internet of things.
21. Zhang, H., Davidson, I.: Towards fair deep anomaly detection. pp. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, Association for Computing Machinery, New York, NY, USA
22. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph Neural Networks: A Review of Methods and Applications. arXiv preprint: 1812.08434
23. Zimek, A., Gaudet, M., Campello, R.J., Sander, J.: Subsampling for efficient and effective unsupervised outlier detection ensembles. p. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '13, Association for Computing Machinery, New York, NY, USA