Supplement

In this supplementary material we (1) provide additional theorems and proofs for Section 3, and (2) further describe the experimental results.

A Theoretical Results

Firstly, we provide the proof for Theorem 3.1.

Theorem 3.1 (Analysis of EXCEED) Let s be an anomaly score, and $\psi_n \in [0,1]$ the proportion of training scores $\leq s$. For $T \geq 4$, there exist $t_1 = t_1(n,\gamma,T) \in [0,1]$, $t_2 = t_2(n,\gamma,T) \in [0,1]$ such that

$$\psi_n \in [t_1, t_2] \implies \mathscr{M}_s \leq 1 - 2e^{-T}.$$

Proof. We split this proof into two parts: we show that the reverse inequalities, i.e. that **(a)** if $\psi_n \leq t_1$, then $\mathscr{M}_s \geq 1 - 2e^{-T}$, and **(b)** if $\psi_n \geq t_2$, then $\mathscr{M}_s \geq 1 - 2e^{-T}$, hold and prove the final statement because $\mathbb{P}(\hat{Y} = 1|s)$ is monotonic increasing on s.

(a) The probability $\mathbb{P}(\hat{Y} = 1|s)$ (as in Eq. 1) can be seen as the cumulative distribution F of a binomial random variable $\mathcal{B}(q_s, n)$ with at most $n\gamma - 1$ successes out of n trials, with $q_s = \frac{n(1-\psi_n)+1}{2+n}$ as the success probability. By applying Hoeffding's inequality, we obtain the upper bound

$$\mathbb{P}(\hat{Y} = 1|s) \leq \exp\left(-2n\left(\frac{n(1-\psi_n)+1}{2+n} - \frac{n\gamma - 1}{n}\right)^2\right)$$

that holds for the constraint $\psi_n \leq \frac{2+n}{n^2} + \frac{1-2\gamma}{n} + (1-\gamma)$. Because $\mathbb{P}(\hat{Y} = 1|s) \leq e^{-T}$ implies that $\mathscr{M}_s \geq 1 - 2e^{-T}$, we search for the values of ψ_n such that the upper bound is $\leq e^{-T}$. Forcing the upper bound $\leq e^{-T}$ results in

$$2n\left(\frac{n(1-\psi_n)+1}{2+n} - \frac{n\gamma - 1}{n}\right)^2 - T \geq 0,$$

which is satisfied for (I_1) $0 \leq \psi_n \leq A_1 - \sqrt{B_1}$ and (I_2) $A_1 + \sqrt{B_1} \leq \psi_n \leq 1$, where

$$A_1 = \frac{2 + n(n+1)(1-\gamma)}{n^2} \qquad B_1 = \frac{2n\left(-3\gamma^2 - 2n(1-\gamma)^2 + 4\gamma - 3\right) + T(n+2)^2 - 8}{2n^3}.$$

However, for $T \geq 4$, no values of n, γ, and T that satisfy the constraint on ψ_n also satisfy I_2. Moving to I_1, we find out that if ψ_n satisfies I_1, then it also satisfies the constraint on ψ_n for any n, γ, and T. Therefore, we we set $t_1(n,\gamma,T) = A_1 - \sqrt{B_1}$. As a result,

$$\psi_n \leq t_1 \implies \mathbb{P}(\hat{Y} = 1|s) \leq e^{-T} \implies \mathscr{M}_s \geq 1 - 2e^{-T}.$$

(b) Similarly, $\mathbb{P}(\hat{Y} = 0|s)$ can be seen as the cumulative distribution F of $\mathcal{B}(p_s, n)$, with $n(1-\gamma)$ successes and $p_s = \frac{1+n\psi_n(s)}{2+n}$. By seeing the binomial as a sum of Bernoulli random variables, and using the property of its cumulative distribution $F(n(1-\gamma), n, p_s) + F(n\gamma - 1, n, 1-p_s) = 1$, we apply the Hoeffding's inequality and compare such upper bound to the e^{-T}. We obtain

$$2n\left(\frac{1+\psi_n n}{2+n} - (1-\gamma)\right)^2 - T \geq 0$$

that holds with the constraint $\psi_n \geq \frac{(2+n)(1-\gamma)-1}{n}$. The quadratic inequality in ψ_n has solutions for (I_1) $0 \leq \psi_n \leq A_2 - \sqrt{B_2}$ and (I_2) $A_2 + \sqrt{B_2} \leq \psi_n \leq 1$, where $A_2 = \frac{(2+n)(1-\gamma)-1}{n}$, and $B_2 = \frac{T(n+2)^2}{2n^3}$. However, the constraint limits the solutions to I_2, i.e. for $\psi_n \geq A_2 + \sqrt{B_2}$. Thus, we set $t_2(n,\gamma,T) = A_2 + \sqrt{B_2}$ and conclude that

$$\psi_n \geq t_2 \implies \mathbb{P}(\hat{Y} = 1|s) \geq 1 - e^{-T} \implies \mathscr{M}_s \geq 1 - 2e^{-T}.$$

$\square$

Secondly, Theorem 3.6 relies on two important results: given S the anomaly score random variable, (1) if ψ_n was the *theoretical* cumulative of S, it would have a uniform distribution (Theorem A.1), but because in practice (2) ψ_n is the *empirical* cumulative of S, its distribution is close to uniform with high probability (Theorem A.2). We prove these results in the following theorems.

Theorem A.1. *Let S be the anomaly score random variable, and $\psi = F_S(S)$ be the cumulative distribution of S applied to S itself. Then $\psi \sim Unif(0,1)$.*

Proof. We prove that, if $\psi \sim Unif(0,1)$, then $F_\psi(t) = t$ for any $t \in [0,1]$:

$$F_\psi(t) = \mathbb{P}(\psi \leq t) = \mathbb{P}(F_S(S) \leq t) = \mathbb{P}(S \leq F_S^{-1}(t)) = F_S(F_S^{-1}(t)) = t \implies \psi \sim Unif(0,1).$$

□

Theorem A.2. *Let ψ be as in Theorem A.1, and F_{ψ_n} be its empirical distribution obtained from a sample of size n. For any small $\delta > 0$ and $t \in [0,1]$, with probability $> 1-\delta$*

$$F_{\psi_n}(t) \in \left[F_\psi(t) - \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}, F_\psi(t) + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}} \right].$$

Proof. For any $\varepsilon > 0$, the DKW inequality implies

$$\mathbb{P}\left(\sup_{t \in [0,1]} |F_{\psi_n}(t) - F_\psi(t)| > \varepsilon \right) \leq 2 \exp\left(-2n\varepsilon^2\right).$$

By setting $\delta = 2\exp\left(-2n\varepsilon^2\right)$, i.e. $\varepsilon = \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$, and using the complementary probability we conclude that

$$\mathbb{P}\left(\sup_{t \in [0,1]} |F_{\psi_n}(t) - F_\psi(t)| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2n}} \right) > 1 - \delta.$$

□

Finally, we give the proof sketch for Corollary 3.2, as most of the steps follow from simple algebra:

Corollay 3.2 Given t_1 and t_2 as in Theorem 3.1, the following properties hold for any $s, n, \gamma, T \geq 4$:

- P1. $\lim_{n \to +\infty} t_1 = \lim_{n \to +\infty} t_2 = 1 - \gamma$;
- P2. t_1 and t_2 are, respectively, monotonic decreasing and increasing as functions of T;
- P3. the interval always contains $1-\gamma$, i.e. $t_1 \leq 1-\gamma \leq t_2$;
- P4. for $n \to \infty$, there exists s^* with $\psi_n = t^* \in [t_1, t_2]$ such that $t^* \to 1-\gamma$ and $\mathscr{M}_s \to 0$.
- P5. $\psi_n \in [t_1, t_2]$ **iff** $s \in [\lambda - u_1, \lambda + u_2]$, where $u_1(n,\gamma,T), u_2(n,\gamma,T)$ are positive functions.

Proof sketch. For P1, it is enough to observe that $A_1, A_2 \to 1-\gamma$, while $B_1, B_2 \to 0$ for $n \to +\infty$. For P2 and P3, the result comes from simple algebraic steps. P4 follows from the surjectivity of $\mathscr{M}_s$ when $n \to +\infty$, the monotonicity of $\mathbb{P}(\hat{Y} = 1|s)$, from P1 with the squeeze theorem. Finally, P5 follows from $\psi_n \in [t_1, t_2] \implies s \in \left[\psi_n^{-1}(t_1), \psi_n^{-1}(t_2)\right]$, as ψ_n is monotonic increasing, where ψ_n^{-1} is the inverse-image of ψ_n. Because for P3 $1-\gamma \in [t_1, t_2]$, it holds that $\psi_n^{-1}(t_1) \leq \psi_n^{-1}(1-\gamma) = \lambda \leq \psi_n^{-1}(t_2)$. This implies that $s \in [\lambda - u_1, \lambda + u_2]$, where $u_1 = \lambda - \psi_n^{-1}(t_1)$, $u_2 = \lambda - \psi_n^{-1}(t_2)$. □

Table 2: Properties (number of examples, features, and contamination factor) of the 34 benchmark datasets used for the experiments.

Dataset	#Examples	#Features	γ
Aloi	20000	27	0.0315
Annthyroid	7062	6	0.0756
Campaign	20000	62	0.1127
Cardio	1822	21	0.0960
Cardiotocography	2110	21	0.2204
Census	20000	500	0.0854
Donors	20000	10	0.2146
Fault	1941	27	0.3467
Fraud	20000	29	0.0021
Glass	213	7	0.0423
Http	20000	3	0.0004
InternetAds	1966	1555	0.1872
Landsat	6435	36	0.2071
Letter	1598	32	0.0626
Lymphography	148	18	0.0405
Mammography	7848	6	0.0322
Musk	3062	166	0.0317
Optdigits	5198	64	0.0254
PageBlocks	5393	10	0.0946
Pendigits	6870	16	0.0227
Pima	768	8	0.3490
Satellite	6435	36	0.3164
Satimage	5801	36	0.0119
Shuttle	20000	9	0.0725
Thyroid	3656	6	0.0254
Vertebral	240	6	0.1250
Vowels	1452	12	0.0317
Waveform	3443	21	0.0290
WBC	223	9	0.0448
WDBC	367	30	0.0272
Wilt	4819	5	0.0533
Wine	129	13	0.0775
WPBC	198	33	0.2374
Yeast	1453	8	0.3310

B Experiments

Data. Table 2 shows the properties of the 34 datasets used for the experimental comparison, in terms of number of examples, features, and contamination factor γ. For the datasets with $> 20,000$ examples, we randomly sub-sample them to $20,000$ examples to limit the computational time. The datasets can be downloaded in the following link: `https://github.com/Minqi824/ADBench/tree/main/datasets/Classical`.

Q1. RejEx against the baselines. Table 3 and Table 4 show the results (mean $\pm$ std) aggregated by detectors in terms of, respectively, cost per example and ranking position. Results confirm that RejEx obtains an average cost per example lower than all the baselines for 9 out of 12 detectors, which is similar to the runner-up SS-Repen for the remaining 3 detectors. Moreover, RejEx has always the best (lowest) average ranking position.

Q2. Varying the costs c_{fp}, c_{fn}, c_r. Table 5 and Table 6 show the average cost per example and the ranking position (mean $\pm$ std) aggregated by detectors for three representative cost functions, as discussed in the paper. Results are similar in all three cases. For high false positives cost ($c_{fp} = 10$), RejEx obtains an average cost per example lower than all the baselines for 11 out of 12 detectors and always the best average ranking position. For high false negative cost ($c_{fn} = 10$) as well as for low rejection cost ($c_{fp} = 5, c_{fn} = 5, c_r = \gamma$), it has the lowest average cost for all detectors and always the best average ranking. Moreover, when rejection is highly valuable (low cost), RejEx's cost has a large gap with respect to the baselines, which means that it is particularly useful when rejection is less expensive.

Q5. Impact of training labels on RejEx. In this experiment we include an additional baseline Oracle, which uses ExCeeD's confidence metric and simulates having access to the training labels to set the optimal rejection threshold. Table 7 shows the average cost and rejection rates at test time obtained by the two methods. Overall, the two methods obtain similar costs, with Oracle only achieving a lower average cost by 0.01. In terms of rejection rate, Oracle rejects fewer examples: by finding an optimal threshold, it is able to reject fewer examples whenever the detector is accurate which reduces the number of rejected but otherwise correctly predicted examples.

Table 3: Cost per example (mean ± std) per detector aggregated over the datasets. Results show that RejEx obtains a lower average cost for 9 out of 12 detectors and similar average cost as the runner-up SS-Repen for the remaining 3 detectors. Moreover, RejEx has the best overall average (last row).

	Cost per example (mean ± std.)							
Det.	RejEx	SS-Repen	Mv	Ens	Udr	Em	Stability	NoReject
Ae	**0.122 ± 0.139**	0.124 ± 0.133	0.136 ± 0.143	0.134 ± 0.151	0.138 ± 0.148	0.143 ± 0.150	0.143 ± 0.149	0.148 ± 0.152
Copod	**0.123 ± 0.138**	0.125 ± 0.134	0.135 ± 0.142	0.133 ± 0.140	0.140 ± 0.144	0.143 ± 0.148	0.145 ± 0.147	0.146 ± 0.148
Ecod	**0.120 ± 0.136**	0.124 ± 0.135	0.129 ± 0.136	0.133 ± 0.143	0.139 ± 0.142	0.140 ± 0.145	0.143 ± 0.144	0.145 ± 0.145
Gmm	**0.122 ± 0.135**	0.124 ± 0.137	0.144 ± 0.141	0.136 ± 0.146	0.142 ± 0.145	0.156 ± 0.148	0.151 ± 0.147	0.156 ± 0.149
Hbos	**0.116 ± 0.129**	0.123 ± 0.138	0.131 ± 0.132	0.134 ± 0.136	0.135 ± 0.137	0.139 ± 0.141	0.140 ± 0.139	0.142 ± 0.142
IFor	**0.115 ± 0.128**	0.123 ± 0.135	0.130 ± 0.136	0.129 ± 0.136	0.134 ± 0.139	0.140 ± 0.143	0.139 ± 0.141	0.143 ± 0.144
Inne	**0.113 ± 0.129**	0.122 ± 0.133	0.145 ± 0.134	0.146 ± 0.140	0.145 ± 0.138	0.147 ± 0.140	0.146 ± 0.139	0.145 ± 0.140
Kde	**0.127 ± 0.140**	**0.127 ± 0.134**	0.143 ± 0.145	0.138 ± 0.145	0.144 ± 0.145	0.150 ± 0.148	0.145 ± 0.143	0.152 ± 0.148
Knn	**0.119 ± 0.123**	0.125 ± 0.135	0.140 ± 0.131	0.135 ± 0.131	0.135 ± 0.130	0.144 ± 0.132	0.141 ± 0.131	0.146 ± 0.133
Loda	**0.125 ± 0.133**	**0.125 ± 0.134**	0.131 ± 0.130	0.139 ± 0.137	0.140 ± 0.136	0.146 ± 0.141	0.141 ± 0.131	0.151 ± 0.142
Lof	**0.126 ± 0.131**	**0.126 ± 0.136**	0.155 ± 0.140	0.140 ± 0.139	0.142 ± 0.138	0.157 ± 0.140	0.151 ± 0.139	0.158 ± 0.140
Ocsvm	**0.120 ± 0.131**	0.125 ± 0.133	0.138 ± 0.138	0.132 ± 0.140	0.138 ± 0.140	0.141 ± 0.140	0.137 ± 0.136	0.147 ± 0.143
Avg.	**0.121 ± 0.133**	0.125 ± 0.135	0.138 ± 0.137	0.136 ± 0.140	0.139 ± 0.140	0.146 ± 0.143	0.144 ± 0.140	0.148 ± 0.144

Table 4: Ranking positions (mean ± std) per detector aggregated over the datasets. Results show that RejEx obtains always the lowest average rank, despite being close to the runner-up SS-Repen when the detector is Loda.

	Ranking position (mean ± std.)							
Det.	RejEx	SS-Repen	Mv	Ens	Udr	Em	Stability	NoReject
Ae	**2.63 ± 1.63**	3.96 ± 2.94	4.78 ± 2.10	3.81 ± 2.11	4.98 ± 1.88	5.15 ± 1.80	4.92 ± 1.78	5.77 ± 1.84
Copod	**2.49 ± 1.65**	3.44 ± 2.75	3.97 ± 1.92	4.25 ± 2.17	5.24 ± 1.70	5.13 ± 1.67	5.59 ± 1.48	5.89 ± 2.13
Ecod	**2.43 ± 1.44**	3.62 ± 2.86	3.73 ± 1.96	4.13 ± 2.29	5.53 ± 1.57	4.75 ± 1.62	5.74 ± 1.39	6.07 ± 1.93
Gmm	**2.12 ± 1.08**	3.05 ± 2.49	5.26 ± 1.92	3.49 ± 2.00	4.43 ± 1.66	6.26 ± 1.24	5.20 ± 1.43	6.20 ± 1.45
Hbos	**2.29 ± 1.57**	3.64 ± 2.98	4.54 ± 2.11	4.39 ± 2.06	4.95 ± 1.79	5.04 ± 1.80	5.61 ± 1.40	5.52 ± 1.88
IFor	**2.23 ± 1.48**	3.78 ± 2.78	4.12 ± 1.90	4.26 ± 2.08	5.10 ± 1.88	5.27 ± 1.66	5.34 ± 1.38	5.91 ± 2.22
Inne	**1.73 ± 1.14**	3.18 ± 2.74	5.86 ± 2.42	5.57 ± 1.40	4.94 ± 1.62	5.57 ± 1.60	5.32 ± 1.37	3.83 ± 1.63
Kde	**2.33 ± 1.42**	3.99 ± 2.86	4.74 ± 2.06	3.79 ± 2.03	5.01 ± 1.90	5.43 ± 1.59	4.87 ± 1.92	5.84 ± 1.80
Knn	**2.02 ± 1.29**	3.58 ± 2.87	4.87 ± 1.81	3.94 ± 1.94	4.41 ± 1.83	5.75 ± 1.49	5.22 ± 1.62	6.21 ± 1.48
Loda	**2.89 ± 1.77**	3.17 ± 2.30	4.15 ± 2.26	4.36 ± 2.14	4.99 ± 2.00	5.50 ± 2.04	4.98 ± 2.11	5.95 ± 1.73
Lof	**2.04 ± 1.01**	3.16 ± 2.73	5.68 ± 1.40	3.32 ± 1.71	3.96 ± 1.63	6.15 ± 1.19	5.47 ± 1.49	6.22 ± 1.31
Ocsvm	**2.33 ± 1.29**	3.92 ± 2.84	4.89 ± 1.98	3.85 ± 2.17	4.86 ± 1.89	5.31 ± 1.80	5.06 ± 1.89	5.78 ± 1.66
Avg.	**2.29 ± 1.40**	3.54 ± 2.76	4.72 ± 1.99	4.10 ± 2.01	4.87 ± 1.78	5.44 ± 1.63	5.28 ± 1.60	5.77 ± 1.76

Table 5: Cost per example (mean ± std) per detector aggregated over the datasets. The table is divided into three parts, where each part has different costs (false positives, false negatives, rejection). Results show that REJEX obtains a lower average cost in all cases but one (KDE).

DET.	COST PER EXAMPLE FOR THREE COST FUNCTIONS (MEAN ± STD)							
	REJEX	SS-REPEN	MV	ENS	UDR	EM	STABILITY	NOREJECT
FALSE POSITIVE COST = 10, FALSE NEGATIVE COST = 1, REJECTION COST = $\min\{10(1-\gamma), \gamma\}$								
AE	**0.504 ± 0.626**	0.584 ± 0.723	0.697 ± 0.763	0.661 ± 0.830	0.703 ± 0.829	0.766 ± 0.841	0.768 ± 0.826	0.825 ± 0.873
COPOD	**0.491 ± 0.637**	0.593 ± 0.706	0.686 ± 0.746	0.618 ± 0.726	0.707 ± 0.788	0.778 ± 0.825	0.785 ± 0.801	0.781 ± 0.833
ECOD	**0.479 ± 0.628**	0.584 ± 0.727	0.625 ± 0.705	0.642 ± 0.755	0.711 ± 0.774	0.748 ± 0.803	0.770 ± 0.783	0.771 ± 0.817
GMM	**0.568 ± 0.713**	0.589 ± 0.752	0.823 ± 0.878	0.715 ± 0.929	0.790 ± 0.925	0.941 ± 0.948	0.889 ± 0.929	0.950 ± 0.967
HBOS	**0.475 ± 0.595**	0.569 ± 0.758	0.666 ± 0.693	0.697 ± 0.732	0.709 ± 0.764	0.776 ± 0.803	0.771 ± 0.770	0.809 ± 0.816
IFOR	**0.477 ± 0.602**	0.575 ± 0.712	0.665 ± 0.718	0.634 ± 0.731	0.683 ± 0.786	0.776 ± 0.818	0.763 ± 0.788	0.808 ± 0.831
INNE	**0.479 ± 0.592**	0.567 ± 0.698	0.752 ± 0.724	0.820 ± 0.795	0.815 ± 0.787	0.819 ± 0.793	0.818 ± 0.792	0.823 ± 0.799
KDE	0.602 ± 0.827	**0.589 ± 0.704**	0.819 ± 0.947	0.740 ± 0.913	0.793 ± 0.939	0.897 ± 0.945	0.774 ± 0.906	0.914 ± 0.945
KNN	**0.498 ± 0.577**	0.596 ± 0.726	0.741 ± 0.734	0.669 ± 0.720	0.669 ± 0.736	0.777 ± 0.747	0.739 ± 0.735	0.800 ± 0.749
LODA	**0.518 ± 0.619**	0.574 ± 0.709	0.574 ± 0.647	0.689 ± 0.729	0.701 ± 0.748	0.762 ± 0.774	0.697 ± 0.682	0.827 ± 0.797
LOF	**0.539 ± 0.623**	0.603 ± 0.742	0.898 ± 0.840	0.685 ± 0.773	0.715 ± 0.790	0.891 ± 0.813	0.831 ± 0.821	0.887 ± 0.808
OCSVM	**0.479 ± 0.599**	0.589 ± 0.705	0.745 ± 0.790	0.632 ± 0.752	0.694 ± 0.782	0.760 ± 0.775	0.695 ± 0.737	0.818 ± 0.806
FALSE POSITIVE COST = 1, FALSE NEGATIVE COST = 10, REJECTION COST = $\min\{1-\gamma, 10\gamma\}$								
AE	**0.730 ± 0.747**	0.761 ± 0.756	0.909 ± 0.882	0.784 ± 0.825	0.780 ± 0.805	0.819 ± 0.843	0.789 ± 0.825	0.797 ± 0.821
COPOD	**0.761 ± 0.767**	0.765 ± 0.770	0.930 ± 0.888	0.794 ± 0.805	0.800 ± 0.801	0.844 ± 0.842	0.802 ± 0.815	0.827 ± 0.832
ECOD	**0.739 ± 0.759**	0.767 ± 0.766	0.900 ± 0.858	0.789 ± 0.811	0.788 ± 0.787	0.840 ± 0.839	0.791 ± 0.803	0.821 ± 0.819
GMM	**0.670 ± 0.676**	0.765 ± 0.767	0.845 ± 0.782	0.754 ± 0.755	0.739 ± 0.736	0.785 ± 0.757	0.760 ± 0.753	0.766 ± 0.750
HBOS	**0.687 ± 0.684**	0.776 ± 0.782	0.824 ± 0.808	0.750 ± 0.768	0.744 ± 0.747	0.785 ± 0.787	0.749 ± 0.765	0.753 ± 0.766
IFOR	**0.679 ± 0.680**	0.775 ± 0.776	0.847 ± 0.824	0.755 ± 0.771	0.743 ± 0.745	0.761 ± 0.772	0.757 ± 0.774	0.763 ± 0.770
INNE	**0.660 ± 0.685**	0.772 ± 0.779	0.695 ± 0.620	0.774 ± 0.742	0.748 ± 0.722	0.758 ± 0.737	0.744 ± 0.716	0.773 ± 0.754
KDE	**0.691 ± 0.692**	0.791 ± 0.773	0.887 ± 0.836	0.754 ± 0.760	0.755 ± 0.744	0.785 ± 0.807	0.758 ± 0.754	0.759 ± 0.760
KNN	**0.706 ± 0.657**	0.767 ± 0.764	0.839 ± 0.779	0.791 ± 0.736	0.769 ± 0.710	0.778 ± 0.736	0.799 ± 0.736	0.803 ± 0.729
LODA	**0.750 ± 0.714**	0.781 ± 0.775	0.880 ± 0.850	0.811 ± 0.768	0.806 ± 0.761	0.804 ± 0.783	0.827 ± 0.780	0.838 ± 0.784
LOF	**0.738 ± 0.679**	0.764 ± 0.764	0.999 ± 0.833	0.826 ± 0.757	0.810 ± 0.739	0.871 ± 0.770	0.867 ± 0.799	0.846 ± 0.747
OCSVM	**0.730 ± 0.711**	0.780 ± 0.774	0.953 ± 0.878	0.791 ± 0.786	0.795 ± 0.773	0.845 ± 0.833	0.787 ± 0.772	0.796 ± 0.783
FALSE POSITIVE COST = 5, FALSE NEGATIVE COST = 5, REJECTION COST = γ								
AE	**0.534 ± 0.611**	0.618 ± 0.666	0.671 ± 0.716	0.644 ± 0.741	0.655 ± 0.736	0.707 ± 0.748	0.705 ± 0.740	0.738 ± 0.762
COPOD	**0.545 ± 0.619**	0.627 ± 0.673	0.676 ± 0.724	0.629 ± 0.674	0.666 ± 0.716	0.719 ± 0.747	0.719 ± 0.730	0.731 ± 0.739
ECOD	**0.529 ± 0.609**	0.625 ± 0.675	0.629 ± 0.687	0.638 ± 0.702	0.662 ± 0.705	0.701 ± 0.736	0.708 ± 0.716	0.724 ± 0.727
GMM	**0.534 ± 0.599**	0.626 ± 0.687	0.719 ± 0.709	0.656 ± 0.716	0.675 ± 0.720	0.776 ± 0.736	0.746 ± 0.731	0.780 ± 0.743
HBOS	**0.499 ± 0.572**	0.622 ± 0.694	0.632 ± 0.661	0.650 ± 0.669	0.641 ± 0.681	0.695 ± 0.706	0.688 ± 0.686	0.710 ± 0.709
IFOR	**0.497 ± 0.569**	0.623 ± 0.677	0.643 ± 0.680	0.620 ± 0.667	0.629 ± 0.692	0.696 ± 0.712	0.688 ± 0.701	0.714 ± 0.719
INNE	**0.491 ± 0.569**	0.617 ± 0.668	0.622 ± 0.572	0.718 ± 0.691	0.691 ± 0.674	0.709 ± 0.685	0.705 ± 0.675	0.726 ± 0.698
KDE	**0.562 ± 0.639**	0.642 ± 0.673	0.709 ± 0.739	0.666 ± 0.711	0.684 ± 0.726	0.748 ± 0.742	0.679 ± 0.689	0.761 ± 0.742
KNN	**0.521 ± 0.544**	0.628 ± 0.677	0.687 ± 0.667	0.657 ± 0.646	0.634 ± 0.651	0.701 ± 0.664	0.693 ± 0.657	0.728 ± 0.664
LODA	**0.550 ± 0.595**	0.627 ± 0.677	0.601 ± 0.649	0.670 ± 0.668	0.665 ± 0.680	0.708 ± 0.698	0.683 ± 0.645	0.757 ± 0.711
LOF	**0.554 ± 0.580**	0.628 ± 0.681	0.809 ± 0.737	0.678 ± 0.682	0.674 ± 0.688	0.792 ± 0.703	0.759 ± 0.718	0.788 ± 0.698
OCSVM	**0.523 ± 0.582**	0.631 ± 0.671	0.704 ± 0.728	0.634 ± 0.685	0.657 ± 0.695	0.709 ± 0.704	0.660 ± 0.674	0.733 ± 0.716

Table 6: Rankings (mean ± std) per detector aggregated over the datasets, where lower positions mean lower costs (better). The table is divided into three parts, where each part has different costs for false positives, false negatives, and rejection. RejEx obtains the lowest (best) average ranking position for all the detectors and all cost functions.

	Rankings for the Three Cost Functions (mean ± std)							
Det.	RejEx	SS-Repen	Mv	Ens	Udr	Em	Stability	NoReject
False Positive Cost = 10, False Negative Cost = 1, Rejection Cost = $\min\{10(1-\gamma),\gamma\}$								
Ae	**2.35 ± 1.37**	3.84 ± 2.73	5.87 ± 2.40	3.68 ± 2.05	4.85 ± 1.96	5.33 ± 1.67	4.72 ± 1.68	5.36 ± 1.97
Copod	**2.25 ± 1.45**	3.63 ± 2.66	4.79 ± 2.27	3.89 ± 2.27	4.94 ± 1.76	5.46 ± 1.71	5.51 ± 1.45	5.54 ± 2.17
Ecod	**2.28 ± 1.30**	3.51 ± 2.73	4.63 ± 2.36	3.85 ± 2.21	5.34 ± 1.74	5.11 ± 1.66	5.38 ± 1.39	5.90 ± 2.09
Gmm	**2.13 ± 0.99**	3.10 ± 2.43	6.36 ± 2.23	3.31 ± 1.94	4.21 ± 1.69	6.28 ± 1.27	5.18 ± 1.53	5.44 ± 1.50
Hbos	**2.12 ± 1.42**	3.46 ± 2.85	5.41 ± 2.42	4.25 ± 2.06	4.78 ± 1.78	5.35 ± 1.66	5.42 ± 1.47	5.20 ± 1.95
IFor	**2.11 ± 1.49**	3.69 ± 2.61	4.73 ± 2.26	4.02 ± 2.11	5.07 ± 1.87	5.39 ± 1.61	5.36 ± 1.41	5.63 ± 2.24
Inne	**1.72 ± 1.24**	3.09 ± 2.68	5.42 ± 2.45	6.16 ± 1.44	5.47 ± 1.59	4.60 ± 1.51	5.32 ± 1.31	4.21 ± 1.80
Kde	**2.14 ± 1.25**	3.82 ± 2.68	5.73 ± 2.36	3.54 ± 1.92	4.75 ± 1.85	5.84 ± 1.58	4.83 ± 1.91	5.36 ± 1.75
Knn	**1.99 ± 1.28**	3.50 ± 2.74	5.55 ± 2.21	3.92 ± 2.02	4.37 ± 1.86	5.73 ± 1.46	5.23 ± 1.68	5.71 ± 1.71
Loda	**2.56 ± 1.53**	3.31 ± 2.31	4.29 ± 2.48	4.34 ± 2.14	5.03 ± 1.95	5.42 ± 1.88	4.96 ± 2.04	6.08 ± 1.75
Lof	**1.96 ± 1.03**	3.04 ± 2.46	7.12 ± 1.28	3.14 ± 1.60	3.73 ± 1.31	6.27 ± 1.14	5.59 ± 1.66	5.15 ± 1.39
Ocsvm	**2.15 ± 1.20**	3.93 ± 2.70	5.92 ± 2.30	3.58 ± 2.13	4.70 ± 1.92	5.40 ± 1.63	5.03 ± 1.89	5.29 ± 1.67
False Positive Cost = 1, False Negative Cost = 10, Rejection Cost = $\min\{1-\gamma,10\gamma\}$								
Ae	**2.98 ± 1.93**	3.82 ± 2.72	7.03 ± 1.95	4.30 ± 2.07	4.49 ± 1.82	4.96 ± 1.83	4.28 ± 1.62	4.14 ± 1.93
Copod	**2.91 ± 2.04**	3.56 ± 2.69	7.13 ± 1.56	4.15 ± 1.97	4.43 ± 1.87	5.30 ± 1.86	4.50 ± 1.60	4.03 ± 1.79
Ecod	**2.70 ± 1.96**	3.88 ± 2.82	6.87 ± 1.72	4.15 ± 2.02	4.74 ± 1.79	5.01 ± 2.03	4.23 ± 1.50	4.42 ± 1.90
Gmm	**2.59 ± 1.70**	3.99 ± 2.85	6.84 ± 2.22	4.04 ± 2.12	4.08 ± 1.77	5.73 ± 1.37	4.62 ± 1.55	4.12 ± 1.58
Hbos	**2.96 ± 2.14**	4.32 ± 2.93	6.41 ± 2.20	4.49 ± 1.92	4.37 ± 1.82	5.15 ± 1.94	4.48 ± 1.65	3.81 ± 1.81
IFor	**2.71 ± 2.06**	4.51 ± 2.92	6.80 ± 2.00	4.47 ± 2.09	4.62 ± 1.72	4.33 ± 1.66	4.55 ± 1.47	4.01 ± 1.93
Inne	**2.64 ± 1.94**	4.71 ± 2.93	5.06 ± 2.95	5.85 ± 1.52	5.06 ± 1.80	4.03 ± 1.64	4.72 ± 1.50	3.94 ± 1.87
Kde	**3.00 ± 2.01**	4.49 ± 2.93	6.51 ± 2.27	4.00 ± 1.84	4.40 ± 1.68	5.01 ± 1.97	4.40 ± 1.78	4.18 ± 1.96
Knn	**2.64 ± 2.01**	4.11 ± 3.01	6.67 ± 2.23	4.17 ± 1.89	4.13 ± 1.87	4.99 ± 1.60	4.88 ± 1.54	4.41 ± 1.64
Loda	**3.44 ± 1.96**	3.66 ± 2.71	6.32 ± 2.30	4.22 ± 1.94	4.36 ± 1.95	4.47 ± 2.17	4.53 ± 2.09	4.99 ± 1.87
Lof	**2.22 ± 1.38**	3.43 ± 2.67	7.74 ± 0.67	3.47 ± 1.73	3.63 ± 1.40	5.95 ± 1.17	5.35 ± 1.57	4.22 ± 1.38
Ocsvm	**2.82 ± 1.71**	3.83 ± 2.63	7.30 ± 1.50	4.23 ± 2.13	4.35 ± 1.78	5.34 ± 1.65	4.32 ± 1.95	3.80 ± 1.72
False Positive Cost = 5, False Negative Cost = 5, Rejection Cost = γ								
Ae	**2.31 ± 1.38**	4.05 ± 2.78	5.85 ± 2.41	3.66 ± 2.12	4.69 ± 1.86	5.27 ± 1.68	4.84 ± 1.74	5.34 ± 1.92
Copod	**2.24 ± 1.49**	3.72 ± 2.70	4.62 ± 2.17	3.98 ± 2.34	4.89 ± 1.87	5.26 ± 1.58	5.57 ± 1.50	5.72 ± 2.10
Ecod	**2.18 ± 1.31**	3.92 ± 2.75	4.22 ± 2.22	3.93 ± 2.33	5.30 ± 1.82	4.91 ± 1.69	5.48 ± 1.38	6.06 ± 1.94
Gmm	**1.96 ± 0.97**	3.31 ± 2.44	6.39 ± 2.21	3.36 ± 1.89	4.09 ± 1.68	6.29 ± 1.25	5.11 ± 1.51	5.48 ± 1.50
Hbos	**1.98 ± 1.37**	3.95 ± 2.88	5.30 ± 2.46	4.18 ± 2.01	4.63 ± 1.83	5.36 ± 1.68	5.41 ± 1.46	5.18 ± 1.93
IFor	**2.01 ± 1.46**	4.10 ± 2.67	4.71 ± 2.26	3.96 ± 2.05	4.98 ± 1.96	5.35 ± 1.60	5.29 ± 1.42	5.60 ± 2.24
Inne	**1.70 ± 1.25**	3.75 ± 2.82	4.17 ± 2.42	6.02 ± 1.44	5.57 ± 1.78	4.56 ± 1.36	5.36 ± 1.38	4.87 ± 2.08
Kde	**2.22 ± 1.35**	4.24 ± 2.73	5.49 ± 2.55	3.62 ± 2.00	4.71 ± 1.95	5.60 ± 1.50	4.79 ± 1.86	5.34 ± 1.87
Knn	**1.98 ± 1.23**	3.88 ± 2.82	5.49 ± 2.39	3.91 ± 1.86	4.29 ± 1.86	5.56 ± 1.71	5.19 ± 1.63	5.69 ± 1.70
Loda	**2.58 ± 1.60**	3.59 ± 2.36	4.34 ± 2.58	4.26 ± 2.17	4.93 ± 1.98	5.33 ± 1.98	5.02 ± 1.98	5.94 ± 1.75
Lof	**1.88 ± 0.96**	3.26 ± 2.51	7.18 ± 1.29	3.16 ± 1.60	3.65 ± 1.32	6.24 ± 1.12	5.53 ± 1.69	5.10 ± 1.37
Ocsvm	**2.15 ± 1.19**	4.18 ± 2.77	5.73 ± 2.36	3.62 ± 2.20	4.59 ± 1.88	5.39 ± 1.59	5.05 ± 1.92	5.31 ± 1.67

Table 7: Mean ± std. for the **cost per example** (on the left) and the **rejection rate** (on the right) at test time on a per detector basis and aggregated over the datasets.

	Cost per example (mean ± std.)		**Rejection Rate** (mean ± std.)	
Det.	RejEx	Oracle	RejEx	Oracle
Ae	0.126 ± 0.139	0.126 ± 0.139	0.131 ± 0.132	0.118 ± 0.125
Copod	0.123 ± 0.140	0.121 ± 0.140	0.123 ± 0.131	0.101 ± 0.114
Ecod	0.119 ± 0.138	0.118 ± 0.138	0.125 ± 0.130	0.107 ± 0.114
Gmm	0.123 ± 0.135	0.122 ± 0.134	0.139 ± 0.143	0.132 ± 0.136
Hbos	0.118 ± 0.129	0.118 ± 0.129	0.139 ± 0.148	0.114 ± 0.128
IFor	0.118 ± 0.129	0.118 ± 0.128	0.127 ± 0.131	0.118 ± 0.130
Inne	0.115 ± 0.129	0.115 ± 0.128	0.132 ± 0.132	0.122 ± 0.125
Kde	0.129 ± 0.140	0.129 ± 0.139	0.121 ± 0.129	0.105 ± 0.120
Knn	0.119 ± 0.123	0.118 ± 0.123	0.127 ± 0.129	0.112 ± 0.117
Loda	0.125 ± 0.133	0.122 ± 0.130	0.126 ± 0.124	0.110 ± 0.114
Lof	0.126 ± 0.131	0.125 ± 0.131	0.129 ± 0.126	0.118 ± 0.115
Ocsvm	0.120 ± 0.131	0.120 ± 0.131	0.126 ± 0.128	0.107 ± 0.115
Avg.	0.122 ± 0.133	0.121 ± 0.133	0.129 ± 0.132	0.114 ± 0.121