Supplement

This supplement contains the characteristics of the datasets used in our experimental analysis and addresses an additional research question to demonstrate SADAL's effectiveness in another common scenario for anomaly detection, where some annotated anomalies are available from the beginning.

Data

Table 1 illustrates the characteristics (number of instances, number of features, and contamination factor) of the 25 datasets used for the empirical evaluation of our method.

SADAL's performance when some annotated anomalies are available from the beginning.

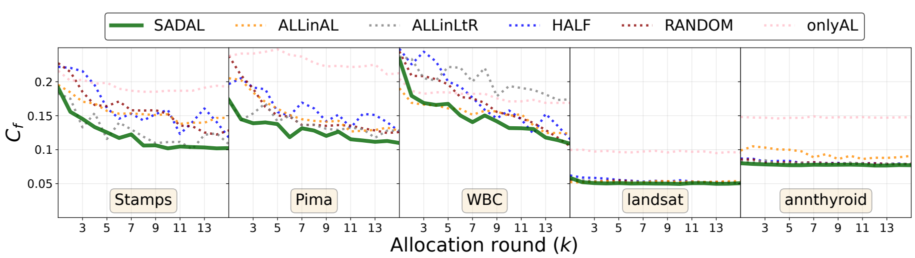

Figure 1: Average test cost per instance C_f for all the considered strategies on five representative datasets for 15 allocation rounds (k) starting from a training set with 10% of the anomalies labeled. Overall, SADAL outperforms the considered baselines and has for most of the allocation rounds the lowest cost per instance.

Anomaly detection spans from fully unsupervised to fully supervised tasks. In the paper, we assume no labels are available at the beginning. However, conceptually our approach also works if the training data contains a small number of already annotated anomalies, which is quite common in the anomaly detection literature. To show the effectiveness of our approach in this scenario,

Table 1: Characteristics of the 25 real-world and benchmark anomaly detection datasets used for the experiments.

Dataset	#(instances)	#(features)	contamination
annthyroid	7062	6	0.0756
Cardiotocography	1738	21	0.0535
celeba	10000	39	0.0255
cover	10000	10	0.0096
fraud	10000	29	0.0017
http	10000	3	0.0003
InternetAds	1672	1555	0.0443
landsat	5369	36	0.0497
letter	1598	32	0.0626
magic.gamma	10000	10	0.0964
mammography	7848	6	0.0322
PageBlocks	5393	10	0.0946
Pima	554	8	0.0974
satellite	6435	36	0.0634
skin	10000	3	0.0738
SpamBase	2864	57	0.1173
speech	3686	400	0.0165
Stamps	340	9	0.0912
vowels	1452	12	0.0317
Waveform	3443	21	0.0290
WBC	223	9	0.0448
Wilt	4819	5	0.0533
yeast	1453	8	0.0668
Turbine 15	42125	9	0.0668
Turbine 21	18529	9	0.0534

we perform experiments on five representative datasets (Stamps, Pima, WBC, landsat, annthyroid). We follow the same setup as in Q1 (paper), except that all methods start with a training set where 10% of the anomalies are already annotated. The anomalies to label are randomly selected.

Figure 1 provides a fine-grained view of the results by plotting the average test cost per instance C_f as a function of the allocation round k. On average, SADAL outperforms all baselines by reducing the test cost by approximately 5% vs ALLinLtR, 8% vs RANDOM, 10% vs HALF and ALLinAL, and 37% vs onlyAL. Moreover, SADAL achieves lower/similar (i.e., differences ≤ 0.001) test cost in around 68% and 73% of the experiments against the two runner-ups ALLinLtR and RANDOM.

For each experiment, we rank the methods from the best (rank 1) to the worst (rank 6) and report the average ranks in Table 2. Results show that

k	**Ranks** (avg. ± std.) SADAL	ALLinAL	ALLinLtR	HALF	RANDOM	onlyAL
1	**2.2 ± 1.3**	2.8 ± 1.7	2.4 ± 1.1	4.6 ± 1.1	4.4 ± 1.8	4.6 ± 1.6
2	**1.6 ± 0.6**	2.6 ± 1.7	3.0 ± 2.0	4.8 ± 0.8	4.0 ± 1.0	5.0 ± 1.4
3	**1.4 ± 0.6**	2.8 ± 1.5	3.0 ± 1.4	5.0 ± 1.2	3.6 ± 1.1	5.2 ± 1.3
4	**1.2 ± 0.5**	3.4 ± 1.5	2.6 ± 0.9	5.0 ± 0.7	3.4 ± 1.1	5.4 ± 1.3
5	**1.4 ± 0.6**	3.4 ± 1.7	2.6 ± 2.0	4.4 ± 0.6	3.8 ± 0.8	5.4 ± 1.3
6	**1.0 ± 0.0**	3.6 ± 1.5	3.2 ± 1.6	4.0 ± 1.0	3.6 ± 1.4	5.6 ± 0.9
7	**1.0 ± 0.0**	3.4 ± 1.4	3.2 ± 1.8	4.0 ± 1.0	3.6 ± 0.9	5.8 ± 0.5
8	**1.0 ± 0.0**	3.8 ± 1.1	3.0 ± 1.7	4.0 ± 1.0	3.4 ± 1.1	5.8 ± 0.5
9	**1.0 ± 0.0**	4.2 ± 0.8	2.8 ± 1.8	3.6 ± 1.1	3.6 ± 0.9	5.8 ± 0.5
10	**1.0 ± 0.0**	3.6 ± 1.1	3.4 ± 1.7	4.2 ± 1.3	3.0 ± 0.7	5.8 ± 0.5
11	**1.2 ± 0.5**	4.0 ± 1.2	3.0 ± 2.0	3.4 ± 1.1	3.6 ± 0.9	5.8 ± 0.5
12	**1.4 ± 0.6**	3.8 ± 1.3	3.0 ± 1.9	3.4 ± 1.8	3.6 ± 0.6	5.8 ± 0.5
13	**1.0 ± 0.0**	4.0 ± 1.2	3.2 ± 1.8	3.8 ± 1.3	3.2 ± 0.5	5.8 ± 0.5
14	**1.0 ± 0.0**	4.4 ± 0.9	3.4 ± 1.5	4.0 ± 0.7	2.4 ± 0.9	5.8 ± 0.5
15	**1.2 ± 0.5**	4.6 ± 0.6	3.6 ± 1.8	3.0 ± 0.7	2.8 ± 1.3	5.8 ± 0.5

Table 2: Average rank (± std.) for each method across all datasets for 15 allocation rounds. Overall, SADAL outperforms the competing baselines and always achieves the lowest (best) average rank, indicating that it consistently obtains good performances over the experiments.

SADAL consistently achieves the lowest (best) average rank when aggregating for each allocation round over all datasets. Additionally, also in this scenario all the baselines that include a reject option achieve similar average positions (around 3).