XStream-Static Temporary

1 Static Dataset Evaluation

Dataset	Number of samples	Dimensionality	Number of Anomalies
	High-dimensional Datasets		
gisette	3850	4970	351
isolet	4886	617	389
letter	4586	617	389
madelon	1430	500	130
	Low/medium-dimensional Datasets		
cancer	385	30	28
ionosphere	242	33	17
telescope	13283	10	951
indians	538	8	38

Table 1: Datasets used for the static evaluation.

Dataset	IF	HST	RSH	LODA	XS
		Original Datasets			
cancer	0.617 ± 0.021	0.646 ± 0.033	0.619 ± 0.03	0.826 ± 0.013	0.8459 ± 0.0082
ionosphere	0.705 ± 0.006	0.706 ± 0.007	0.764 ± 0.032	0.642 ± 0.067	0.8484 ± 0.018
telescope	0.367 ± 0.008	0.392 ± 0.012	0.391 ± 0.012	0.322 ± 0.007	0.344 ± 0.0094
indians	0.142 ± 0.003	0.146 ± 0.002	0.156 ± 0.007	0.177 ± 0.008	0.2165 ± 0.01
gisette	0.078 ± 0.002	0.08 ± 0.002	0.084 ± 0.007	0.087 ± 0.003	0.0907 ± 0.0035
isolet	0.099 ± 0.003	0.097 ± 0.005	0.108 ± 0.004	0.089 ± 0.004	0.1125 ± 0.0065
letter-recognition	0.093 ± 0.001	0.092 ± 0.002	0.104 ± 0.004	0.094 ± 0.006	0.1224 ± 0.0059
madelon	0.11 ± 0.003	0.101 ± 0.013	0.092 ± 0.005	0.101 ± 0.01	0.0978 ± 0.0043

Table 2: Average precision of static methods on original, unperturbed datasets. Mean and standard deviation are reported over 10 runs.

We conduct our experiments on datasets mentioned in Table 6. A Friedman test for differences in the best-performing method across all datasets, showed that we cannot reject the null hypothesis that the difference rankings between methods is statistically significant with $p = 0.1107$.

We did perform a posthoc-Friedman test, Nemenyi test to compare all methods to each other. We first compute average ranks of all the methods over the 8 original datasets, which is shown in Table 4. Setting significance level to be $\alpha = 0.05$, we get the $q_{0.05}$ for $k = 5$ pairs as 2.728. Difference in average ranks of any two methods will be significant if

$$(R_i - R_j) > q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{1}$$

Setting $N = 8$ and $k = 5$, for difference between a pair of methods to be significant, the difference

Dataset	IF	HST	RSH	LODA	XS
cancer(100.0,0.1)	0.599 ± 0.031	0.605 ± 0.031	0.646 ± 0.032	0.811 ± 0.012	0.8251 ± 0.012
cancer(1000.0,0.1)	0.406 ± 0.088	0.201 ± 0.024	0.425 ± 0.112	0.722 ± 0.056	0.813 ± 0.0226
cancer(2000.0,0.1)	0.306 ± 0.044	0.229 ± 0.029	0.337 ± 0.077	0.633 ± 0.092	0.8229 ± 0.0213
cancer(5000.0,0.1)	0.12 ± 0.04	0.158 ± 0.018	0.153 ± 0.07	0.336 ± 0.141	0.7962 ± 0.028
ionosphere (100.0,0.1)	0.651 ± 0.049	0.568 ± 0.026	0.622 ± 0.038	0.56 ± 0.056	0.8485 ± 0.0111
ionosphere (1000.0,0.1)	0.302 ± 0.072	0.231 ± 0.006	0.258 ± 0.07	0.589 ± 0.073	0.8199 ± 0.0193
ionosphere (2000.0,0.1)	0.211 ± 0.105	0.085 ± 0.007	0.233 ± 0.1	0.561 ± 0.092	0.7911 ± 0.0261
ionosphere (5000.0,0.1)	0.112 ± 0.035	0.15 ± 0.017	0.135 ± 0.062	0.494 ± 0.072	0.6855 ± 0.0654
telescope (100.0,0.1)	0.311 ± 0.012	0.26 ± 0.006	0.326 ± 0.015	0.322 ± 0.006	0.3408 ± 0.0088
telescope (1000.0,0.1)	0.156 ± 0.011	0.102 ± 0.004	0.164 ± 0.019	0.303 ± 0.01	0.3115 ± 0.0068
telescope (2000.0,0.1)	0.108 ± 0.01	0.098 ± 0.014	0.112 ± 0.019	0.296 ± 0.016	0.284 ± 0.0053
telescope (5000.0,0.1)	0.084 ± 0.005	0.079 ± 0.001	0.087 ± 0.011	0.248 ± 0.017	0.2717 ± 0.0053
indians (100.0,0.1)	0.123 ± 0.007	0.093 ± 0.003	0.128 ± 0.009	0.171 ± 0.008	0.1961 ± 0.0154
indians (1000.0,0.1)	0.086 ± 0.014	0.096 ± 0.009	0.087 ± 0.011	0.153 ± 0.028	0.1788 ± 0.006
indians (2000.0,0.1)	0.087 ± 0.013	0.076 ± 0.003	0.085 ± 0.008	0.139 ± 0.028	0.1513 ± 0.0138
indians (5000.0,0.1)	0.073 ± 0.007	0.075 ± 0.009	0.083 ± 0.018	0.126 ± 0.028	0.1529 ± 0.0203

Table 3: Average precision of static methods on high-dimensional (top) and low/medium-dimensional (bottom) datasets. Mean and standard deviation reported over 10 runs. Numbers in the brackets indicate: (top) the percentage of features, fraction of samples to which noise is added, signal-to-noise ratio, and (bottom) noise column amount (as % of original dimensionality), relative noise factor.

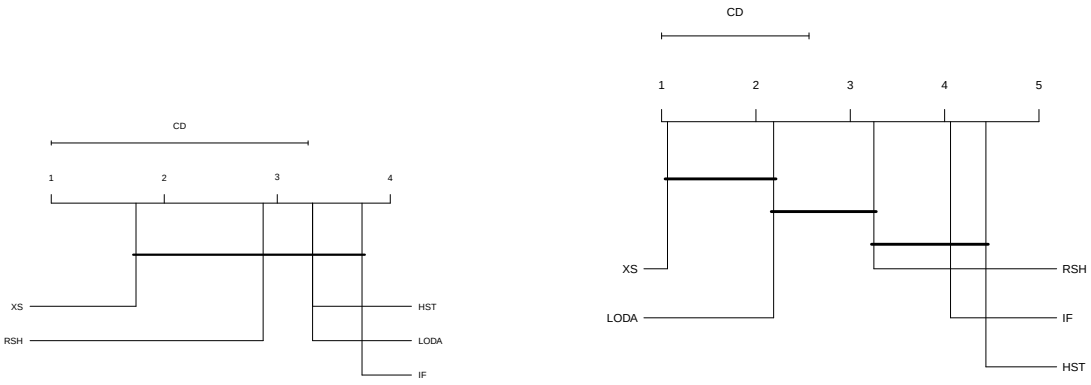
	IF	HST	RSH	LODA	XS
breast-cancer-wisconsin	0.617(5.0)	0.646(3.0)	0.619(4.0)	0.826(2.0)	0.8459(1.0)
ionosphere	0.705(4.0)	0.706(3.0)	0.764(2.0)	0.642(5.0)	0.8484(1.0)
magic-telescope	0.367(3.0)	0.392(1.0)	0.391(2.0)	0.322(5.0)	0.344(4.0)
pima-indians	0.142(5.0)	0.146(4.0)	0.156(3.0)	0.177(2.0)	0.2165(1.0)
gisette	0.078(5.0)	0.08(4.0)	0.084(3.0)	0.087(2.0)	0.0907(1.0)
isolet	0.099(3.0)	0.097(4.0)	0.108(2.0)	0.089(5.0)	0.1125(1.0)
letter-recognition	0.093(4.0)	0.092(5.0)	0.104(2.0)	0.094(3.0)	0.1224(1.0)
madelon	0.11(1.0)	0.101(2.5)	0.092(5.0)	0.101(2.5)	0.0978(4.0)
Avg Rank	3.75	3.3125	2.875	3.3125	1.75

Table 4: Average rank of method over 8 original datasets.

between the rank should be greater than 2.1567. Looking at the average ranks shown in Table 4, we can say there are no pairs that are significantly different from each other.

To show the effect of adding high-dimensionality, we add noisy columns to the 4 original datasets. The noise is generated as mentioned above in Approach 1. The average ranks for all the methods is shown in Table 5. Using Friedman test, we can reject the null hypothesis that the difference rankings between methods is not statistically significant with $p = 5.565 \times 10^{-10}$. We again perform posthoc-Friedman Nemenyi test; setting $N = 16$ and $k = 5$, we obtain that difference in average ranks should be greater than 1.524. Observing the table, we find difference to be significant between (LODA, I-Forest), (X-Stream, I-Forest), (LODA, HS-Trees), (X-Stream, HS-Trees) and (X-Stream, RS-Hash).

Another form of presenting this could be the visualization, which we present below in Fig 1.



(a) Nemenyi Test visualization for 8 original datasets.

(b) Nemenyi Test for 16 perturbed datasets.

Figure 1: Nemenyi Test Visualization

	IF	HST	RSH	LODA	XS
cancer100	0.599(5.0)	0.605(4.0)	0.646(3.0)	0.811(2.0)	0.8251(1.0)
cancer1000	0.406(4.0)	0.201(5.0)	0.425(3.0)	0.722(2.0)	0.813(1.0)
cancer2000	0.306(4.0)	0.229(5.0)	0.337(3.0)	0.633(2.0)	0.8229(1.0)
cancer5000	0.12(5.0)	0.158(3.0)	0.153(4.0)	0.336(2.0)	0.7962(1.0)
ionos100	0.651(2.0)	0.568(4.0)	0.622(3.0)	0.56(5.0)	0.8485(1.0)
ionos1000	0.302(3.0)	0.231(5.0)	0.258(4.0)	0.589(2.0)	0.8199(1.0)
ionos2000	0.211(4.0)	0.085(5.0)	0.233(3.0)	0.561(2.0)	0.7911(1.0)
ionos5000	0.112(5.0)	0.15(3.0)	0.135(4.0)	0.494(2.0)	0.6855(1.0)
telescope100	0.311(4.0)	0.26(5.0)	0.326(2.0)	0.322(3.0)	0.3408(1.0)
telescope1000	0.156(4.0)	0.102(5.0)	0.164(3.0)	0.303(2.0)	0.3115(1.0)
telescope2000	0.108(4.0)	0.098(5.0)	0.112(3.0)	0.296(1.0)	0.284(2.0)
telescope5000	0.084(4.0)	0.079(5.0)	0.087(3.0)	0.248(2.0)	0.2717(1.0)
indians100	0.123(4.0)	0.093(5.0)	0.128(3.0)	0.171(2.0)	0.1961(1.0)
indians1000	0.086(5.0)	0.096(3.0)	0.087(4.0)	0.153(2.0)	0.1788(1.0)
indians2000	0.087(3.0)	0.076(5.0)	0.085(4.0)	0.139(2.0)	0.1513(1.0)
indians5000	0.073(5.0)	0.075(4.0)	0.083(3.0)	0.126(2.0)	0.1529(1.0)
Avg Rank	4.0625	4.4375	3.25	2.1875	1.0625

Table 5: Average rank of methods over the 16 perturbed datasets.

Dataset	**Number of samples**	**Dimensionality**	**Number of Anomalies**	**Anomaly Rate**
annthyroid	7200	6	534	7.4%
arrhythmia	452	274	66	14.6%
breastw	683	9	239	34.99%
cardio	1831	21	176	9.6%
glass	214	9	9	4.2%
ionosphere	351	33	126	35.89%
lympho	148	18	6	4.05%
pendigits	6870	16	156	2.27%
thyroid	3772	6	93	2.46%
vertebral	240	6	30	12.5%
vowels	1456	12	50	3.43%
wbc	378	30	21	5.55%
wine	129	13	10	7.75%

Table 6: Datasets from ODDS.

2 Alternative Dataset from ODDS

A Friedman test for differences in the best-performing method across all datasets, showed that we cannot reject the null hypothesis that the difference rankings between methods is NOT statistically significant with $p = 0.8049$. We did perform a posthoc-Friedman test, Nemenyi test to compare all methods to each other. We first compute average ranks of all the methods over the

Dataset	IF	HST	RSH	LODA	XS
annthyroid	0.312 ± 0.025	0.147 ± 0.004	0.19 ± 0.016	0.196 ± 0.023	0.1617 ± 0.0129
arrhythmia	0.462 ± 0.009	0.458 ± 0.006	0.467 ± 0.023	0.496 ± 0.029	0.4945 ± 0.0146
breastw	0.948 ± 0.008	0.98 ± 0.003	0.967 ± 0.003	0.966 ± 0.005	0.9603 ± 0.0058
cardio	0.538 ± 0.044	0.681 ± 0.006	0.522 ± 0.034	0.523 ± 0.041	0.5041 ± 0.0151
glass	0.082 ± 0.009	0.092 ± 0.004	0.089 ± 0.008	0.116 ± 0.03	0.1294 ± 0.0248
ionosphere	0.822 ± 0.006	0.822 ± 0.002	0.824 ± 0.013	0.78 ± 0.024	0.8883 ± 0.0063
lympho	0.948 ± 0.036	0.99 ± 0.013	0.967 ± 0.033	0.919 ± 0.067	0.5819 ± 0.0633
pendigits	0.275 ± 0.029	0.266 ± 0.03	0.23 ± 0.021	0.27 ± 0.074	0.1476 ± 0.0147
thyroid	0.554 ± 0.035	0.228 ± 0.007	0.299 ± 0.024	0.253 ± 0.042	0.1831 ± 0.0198
vertebral	0.093 ± 0.002	0.081 ± 0.001	0.086 ± 0.002	0.084 ± 0.002	0.0875 ± 0.003
vowels	0.176 ± 0.056	0.162 ± 0.006	0.211 ± 0.044	0.115 ± 0.044	0.3985 ± 0.0341
wbc	0.594 ± 0.025	0.582 ± 0.01	0.597 ± 0.02	0.591 ± 0.039	0.4378 ± 0.0343
wine	0.174 ± 0.023	0.186 ± 0.014	0.209 ± 0.015	0.588 ± 0.134	0.2527 ± 0.0337

Table 7: Average precision of static methods on dataset from ODDS. Mean and standard deviation reported over 10 runs. Numbers in the brackets indicate: (top) the percentage of features, fraction of samples to which noise is added, signal-to-noise ratio, and (bottom) noise column amount (as % of original dimensionality), relative noise factor.

	IF	HST	RSH	LODA	XS
annthyroid	0.312(1.0)	0.147(5.0)	0.19(3.0)	0.196(2.0)	0.1617(4.0)
arrhythmia	0.462(4.0)	0.458(5.0)	0.467(3.0)	0.496(1.0)	0.4945(2.0)
breastw	0.948(5.0)	0.98(1.0)	0.967(2.0)	0.966(3.0)	0.9603(4.0)
cardio	0.538(2.0)	0.681(1.0)	0.522(4.0)	0.523(3.0)	0.5041(5.0)
glass	0.082(5.0)	0.092(3.0)	0.089(4.0)	0.116(2.0)	0.1294(1.0)
ionosphere	0.822(3.5)	0.822(3.5)	0.824(2.0)	0.78(5.0)	0.8883(1.0)
lympho	0.948(3.0)	0.99(1.0)	0.967(2.0)	0.919(4.0)	0.5819(5.0)
pendigits	0.275(1.0)	0.266(3.0)	0.23(4.0)	0.27(2.0)	0.1476(5.0)
thyroid	0.554(1.0)	0.228(4.0)	0.299(2.0)	0.253(3.0)	0.1831(5.0)
vertebral	0.093(1.0)	0.081(5.0)	0.086(3.0)	0.084(4.0)	0.0875(2.0)
vowels	0.176(3.0)	0.162(4.0)	0.211(2.0)	0.115(5.0)	0.3985(1.0)
wbc	0.594(2.0)	0.582(4.0)	0.597(1.0)	0.591(3.0)	0.4378(5.0)
wine	0.174(5.0)	0.186(4.0)	0.209(3.0)	0.588(1.0)	0.2527(2.0)
Avg Rank	2.80769231	3.34615385	2.69230769	2.92307692	3.23076923

Table 8: Average rank of methods over the 13 ODDS datasets.

13 datasets, which is shown in Table 8. Setting $N = 13$ and $K = 5$, we obtain that difference in average ranks should be greater than 1.629. We can see that difference between none of the pairs is significant.

Distorting ODDS datasets We choose to distort and add noisy columns only to those datasets - that were not very high dimensional already, all algorithms were performing fairly reasonably in terms of Avg. Precision (AP), and anomaly rate percentage was not too high.