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	0.845 ± 0.008
ionosphere	0.705 ± 0.006	0.706 ± 0.007	0.764 ± 0.032	0.642 ± 0.067	0.848 ± 0.018
telescope	0.367 ± 0.008	0.392 ± 0.012	0.391 ± 0.012	0.322 ± 0.007	0.344 ± 0.009
indians	0.142 ± 0.003	0.146 ± 0.002	0.156 ± 0.007	0.177 ± 0.008	0.216 ± 0.01
gisette	0.078 ± 0.002	0.08 ± 0.002	0.084 ± 0.007	0.087 ± 0.003	0.09 ± 0.003
isolet	0.099 ± 0.003	0.097 ± 0.005	0.108 ± 0.004	0.089 ± 0.004	0.112 ± 0.006
letter-recognition	0.093 ± 0.001	0.092 ± 0.002	0.104 ± 0.004	0.094 ± 0.006	0.122 ± 0.005
madelon	0.11 ± 0.003	0.101 ± 0.013	0.092 ± 0.005	0.101 ± 0.01	0.097 ± 0.004

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 $k = 5$ methods as 2.728. Difference in average ranks of any two methods will be significant if

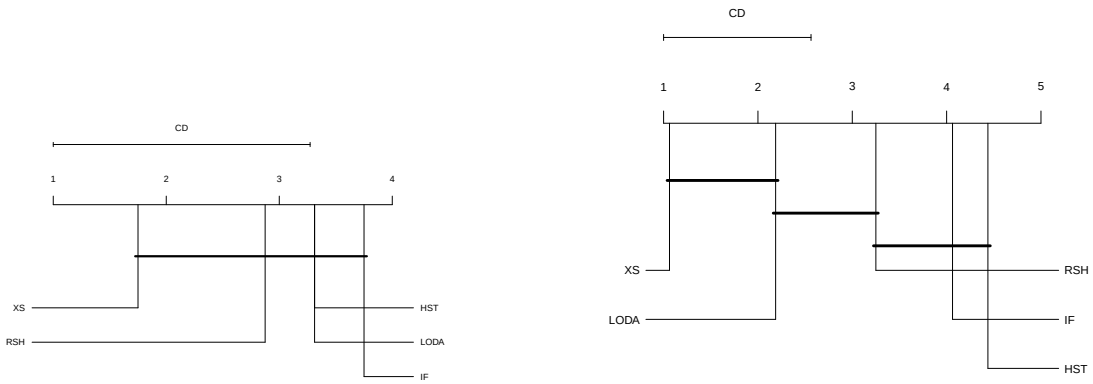
$$(R_i - R_j) > q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{1}$$

Setting $N = 8$ and $k = 5$, for difference between a pair of methods to be significant, the difference between the rank should be greater than 2.1567. Looking at the average ranks shown in Table 4,

we notice that X-Stream has the best ranking on average although no pairs that are significantly different from each other.

To show the effect of increasing 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 in 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

2 Alternative Dataset from ODDS

For evaluating our static approach, we compared the proposed method with four other baseline methods - iForest, LODA, RS-Hash and HS-Trees over 13 datasets we obtained from ODDS repository **ODDS** The list of datasets used and their properties is shown 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 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 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 1.692. We can see that difference between none of the pairs is significant. Therefore, we can safely say that on standard datasets, all the methods are similar.

Distorting ODDS datasets

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	0.825 ± 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 ± 0.022
cancer(2000.0,0.1)	0.306 ± 0.044	0.229 ± 0.029	0.337 ± 0.077	0.633 ± 0.092	0.822 ± 0.021
cancer(5000.0,0.1)	0.12 ± 0.04	0.158 ± 0.018	0.153 ± 0.07	0.336 ± 0.141	0.796 ± 0.028
ionosphere (100.0,0.1)	0.651 ± 0.049	0.568 ± 0.026	0.622 ± 0.038	0.56 ± 0.056	0.848 ± 0.011
ionosphere (1000.0,0.1)	0.302 ± 0.072	0.231 ± 0.006	0.258 ± 0.07	0.589 ± 0.073	0.819 ± 0.019
ionosphere (2000.0,0.1)	0.211 ± 0.105	0.085 ± 0.007	0.233 ± 0.1	0.561 ± 0.092	0.791 ± 0.026
ionosphere (5000.0,0.1)	0.112 ± 0.035	0.15 ± 0.017	0.135 ± 0.062	0.494 ± 0.072	0.685 ± 0.065
telescope (100.0,0.1)	0.311 ± 0.012	0.26 ± 0.006	0.326 ± 0.015	0.322 ± 0.006	0.34 ± 0.008
telescope (1000.0,0.1)	0.156 ± 0.011	0.102 ± 0.004	0.164 ± 0.019	0.303 ± 0.01	0.311 ± 0.006
telescope (2000.0,0.1)	0.108 ± 0.01	0.098 ± 0.014	0.112 ± 0.019	0.296 ± 0.016	0.284 ± 0.005
telescope (5000.0,0.1)	0.084 ± 0.005	0.079 ± 0.001	0.087 ± 0.011	0.248 ± 0.017	0.271 ± 0.005
indians (100.0,0.1)	0.123 ± 0.007	0.093 ± 0.003	0.128 ± 0.009	0.171 ± 0.008	0.196 ± 0.015
indians (1000.0,0.1)	0.086 ± 0.014	0.096 ± 0.009	0.087 ± 0.011	0.153 ± 0.028	0.178 ± 0.006
indians (2000.0,0.1)	0.087 ± 0.013	0.076 ± 0.003	0.085 ± 0.008	0.139 ± 0.028	0.151 ± 0.013
indians (5000.0,0.1)	0.073 ± 0.007	0.075 ± 0.009	0.083 ± 0.018	0.126 ± 0.028	0.152 ± 0.02

Table 3: Average precision of static methods on perturbed noisy datasets. Mean and standard deviation reported over 10 runs. Numbers in the brackets indicate: noise column amount (as % of original dimensionality), relative noise factor.

	IF	HST	RSH	LODA	XS
cancer	0.617(5.0)	0.646(3.0)	0.619(4.0)	0.826(2.0)	0.845(1.0)
ionosphere	0.705(4.0)	0.706(3.0)	0.764(2.0)	0.642(5.0)	0.848(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)	0.216(1.0)
gisette	0.078(5.0)	0.08(4.0)	0.084(3.0)	0.087(2.0)	0.09(1.0)
isolet	0.099(3.0)	0.097(4.0)	0.108(2.0)	0.089(5.0)	0.112(1.0)
letter-recognition	0.093(4.0)	0.092(5.0)	0.104(2.0)	0.094(3.0)	0.122(1.0)
madelon	0.11(1.0)	0.101(2.5)	0.092(5.0)	0.101(2.5)	0.097(4.0)
Avg Rank	3.75	3.3125	2.875	3.3125	1.75

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

	IF	HST	RSH	LODA	XS
cancer100	0.599(5.0)	0.605(4.0)	0.646(3.0)	0.811(2.0)	0.825(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)	0.822(1.0)
cancer5000	0.12(5.0)	0.158(3.0)	0.153(4.0)	0.336(2.0)	0.796(1.0)
ionos100	0.651(2.0)	0.568(4.0)	0.622(3.0)	0.56(5.0)	0.848(1.0)
ionos1000	0.302(3.0)	0.231(5.0)	0.258(4.0)	0.589(2.0)	0.819(1.0)
ionos2000	0.211(4.0)	0.085(5.0)	0.233(3.0)	0.561(2.0)	0.791(1.0)
ionos5000	0.112(5.0)	0.15(3.0)	0.135(4.0)	0.494(2.0)	0.685(1.0)
telescope100	0.311(4.0)	0.26(5.0)	0.326(2.0)	0.322(3.0)	0.34(1.0)
telescope1000	0.156(4.0)	0.102(5.0)	0.164(3.0)	0.303(2.0)	0.311(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)	0.271(1.0)
indians100	0.123(4.0)	0.093(5.0)	0.128(3.0)	0.171(2.0)	0.196(1.0)
indians1000	0.086(5.0)	0.096(3.0)	0.087(4.0)	0.153(2.0)	0.178(1.0)
indians2000	0.087(3.0)	0.076(5.0)	0.085(4.0)	0.139(2.0)	0.151(1.0)
indians5000	0.073(5.0)	0.075(4.0)	0.083(3.0)	0.126(2.0)	0.152(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.

To study the effect of adding noisy columns and high-dimensionality in datasets, we choose to distort by adding noisy columns only to the datasets for which all 5 methods are performing reasonably well i.e. breastw, cardio, ionosphere and lympho.

We present the results on those 4 datasets in Table 10. Using Friedman test, we can reject the null hypothesis that the difference rankings between methods is not statistically significant with $p =$

Dataset	IF	HST	RSH	LODA	XS
annthyroid	0.312 ± 0.025	0.147 ± 0.004	0.19 ± 0.016	0.196 ± 0.023	0.161 ± 0.012
arrhythmia	0.462 ± 0.009	0.458 ± 0.006	0.467 ± 0.023	0.496 ± 0.029	0.494 ± 0.014
breastw	0.948 ± 0.008	0.98 ± 0.003	0.967 ± 0.003	0.966 ± 0.005	0.96 ± 0.005
cardio	0.538 ± 0.044	0.681 ± 0.006	0.522 ± 0.034	0.523 ± 0.041	0.504 ± 0.015
glass	0.082 ± 0.009	0.092 ± 0.004	0.089 ± 0.008	0.116 ± 0.03	0.129 ± 0.024
ionosphere	0.822 ± 0.006	0.822 ± 0.002	0.824 ± 0.013	0.78 ± 0.024	0.888 ± 0.006
lympho	0.948 ± 0.036	0.99 ± 0.013	0.967 ± 0.033	0.919 ± 0.067	0.581 ± 0.063
pendigits	0.275 ± 0.029	0.266 ± 0.03	0.23 ± 0.021	0.27 ± 0.074	0.147 ± 0.014
thyroid	0.554 ± 0.035	0.228 ± 0.007	0.299 ± 0.024	0.253 ± 0.042	0.183 ± 0.019
vertebral	0.093 ± 0.002	0.081 ± 0.001	0.086 ± 0.002	0.084 ± 0.002	0.087 ± 0.003
vowels	0.176 ± 0.056	0.162 ± 0.006	0.211 ± 0.044	0.115 ± 0.044	0.398 ± 0.034
wbc	0.594 ± 0.025	0.582 ± 0.01	0.597 ± 0.02	0.591 ± 0.039	0.437 ± 0.034
wine	0.174 ± 0.023	0.186 ± 0.014	0.209 ± 0.015	0.588 ± 0.134	0.252 ± 0.033

Table 7: Average precision of static methods on dataset from ODDS. Mean and standard deviation reported over 10 runs.

	IF	HST	RSH	LODA	XS
annthyroid	0.312(1.0)	0.147(5.0)	0.19(3.0)	0.196(2.0)	0.161(4.0)
arrhythmia	0.462(4.0)	0.458(5.0)	0.467(3.0)	0.496(1.0)	0.494(2.0)
breastw	0.948(5.0)	0.98(1.0)	0.967(2.0)	0.966(3.0)	0.96(4.0)
cardio	0.538(2.0)	0.681(1.0)	0.522(4.0)	0.523(3.0)	0.504(5.0)
glass	0.082(5.0)	0.092(3.0)	0.089(4.0)	0.116(2.0)	0.129(1.0)
ionosphere	0.822(3.5)	0.822(3.5)	0.824(2.0)	0.78(5.0)	0.888(1.0)
lympho	0.948(3.0)	0.99(1.0)	0.967(2.0)	0.919(4.0)	0.581(5.0)
pendigits	0.275(1.0)	0.266(3.0)	0.23(4.0)	0.27(2.0)	0.147(5.0)
thyroid	0.554(1.0)	0.228(4.0)	0.299(2.0)	0.253(3.0)	0.183(5.0)
vertebral	0.093(1.0)	0.081(5.0)	0.086(3.0)	0.084(4.0)	0.087(2.0)
vowels	0.176(3.0)	0.162(4.0)	0.211(2.0)	0.115(5.0)	0.398(1.0)
wbc	0.594(2.0)	0.582(4.0)	0.597(1.0)	0.591(3.0)	0.437(5.0)
wine	0.174(5.0)	0.186(4.0)	0.209(3.0)	0.588(1.0)	0.252(2.0)
Avg Rank	2.807	3.346	2.692	2.923	3.23

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

Dataset	**IF**	**HST**	**RSH**	**LODA**	**XS**
breastw (1000,0.1)	0.789 ± 0.028	0.841 ± 0.024	0.782 ± 0.03	0.98 ± 0.006	0.967 ± 0.008
breastw (2000,0.1)	0.649 ± 0.05	0.703 ± 0.024	0.642 ± 0.07	0.966 ± 0.014	0.967 ± 0.004
breastw (3000,0.1)	0.551 ± 0.041	0.663 ± 0.014	0.566 ± 0.052	0.958 ± 0.013	0.97 ± 0.006
breastw (5000,0.1)	0.461 ± 0.029	0.626 ± 0.034	0.458 ± 0.054	0.91 ± 0.046	0.971 ± 0.003
cardio (1000.0,0.1)	0.245 ± 0.029	0.161 ± 0.006	0.226 ± 0.036	0.56 ± 0.043	0.493 ± 0.017
cardio (2000,0.1)	0.155 ± 0.023	0.133 ± 0.004	0.153 ± 0.027	0.51 ± 0.074	0.508 ± 0.017
cardio (3000,0.1)	0.136 ± 0.018	0.107 ± 0.006	0.148 ± 0.019	0.484 ± 0.071	0.537 ± 0.016
cardio (5000,0.1)	0.11 ± 0.007	0.115 ± 0.003	0.108 ± 0.01	0.458 ± 0.073	0.488 ± 0.017
ionosphere (1000,0.1)	0.58 ± 0.03	0.549 ± 0.022	0.541 ± 0.042	0.735 ± 0.035	0.861 ± 0.007
ionosphere (2000,0.1)	0.515 ± 0.039	0.421 ± 0.004	0.472 ± 0.028	0.707 ± 0.027	0.866 ± 0.009
ionosphere (3000,0.1)	0.439 ± 0.03	0.433 ± 0.015	0.43 ± 0.018	0.704 ± 0.035	0.822 ± 0.009
ionosphere (5000,0.1)	0.4 ± 0.031	0.428 ± 0.016	0.395 ± 0.024	0.7 ± 0.028	0.792 ± 0.008
lympho (1000,0.1)	0.362 ± 0.16	0.422 ± 0.064	0.287 ± 0.174	0.667 ± 0.115	0.484 ± 0.114
lympho (2000,0.1)	0.148 ± 0.093	0.057 ± 0.007	0.091 ± 0.073	0.5 ± 0.173	0.321 ± 0.183
lympho (3000,0.1)	0.189 ± 0.097	0.037 ± 0.003	0.099 ± 0.074	0.47 ± 0.191	0.414 ± 0.154
lympho (5000,0.1)	0.073 ± 0.056	0.103 ± 0.026	0.065 ± 0.061	0.274 ± 0.203	0.317 ± 0.091

Table 9: Average precision of static methods on perturbed dataset from ODDS. Mean and standard deviation reported over 10 runs. Numbers in the brackets indicate: noise column amount (as % of original dimensionality), relative noise factor.

	IF	HST	RSH	LODA	XS
breastw(1000,0.1)	0.789(4.0)	0.841(3.0)	0.782(5.0)	0.98(1.0)	0.967(2.0)
breastw(2000,0.1)	0.649(4.0)	0.703(3.0)	0.642(5.0)	0.966(2.0)	0.967(1.0)
breastw(3000,0.1)	0.551(5.0)	0.663(3.0)	0.566(4.0)	0.958(2.0)	0.970(1.0)
breastw(5000,0.1)	0.461(4.0)	0.626(3.0)	0.458(5.0)	0.91(2.0)	0.971(1.0)
cardio(1000,0.1)	0.245(3.0)	0.161(5.0)	0.226(4.0)	0.56(1.0)	0.493(2.0)
cardio(2000,0.1)	0.155(3.0)	0.133(5.0)	0.153(4.0)	0.51(1.0)	0.508(2.0)
cardio(3000,0.1)	0.136(4.0)	0.107(5.0)	0.148(3.0)	0.484(2.0)	0.537(1.0)
cardio(5000,0.1)	0.11(4.0)	0.115(3.0)	0.108(5.0)	0.458(2.0)	0.488(1.0)
ionosphere(1000,0.1)	0.58(3.0)	0.549(4.0)	0.541(5.0)	0.735(2.0)	0.861(1.0)
ionosphere(2000,0.1)	0.515(3.0)	0.421(5.0)	0.472(4.0)	0.707(2.0)	0.866(1.0)
ionosphere(3000,0.1)	0.439(3.0)	0.433(4.0)	0.43(5.0)	0.704(2.0)	0.822(1.0)
ionosphere(5000,0.1)	0.4(4.0)	0.428(3.0)	0.395(5.0)	0.7(2.0)	0.792(1.0)
lympho(1000,0.1)	0.362(4.0)	0.422(3.0)	0.287(5.0)	0.667(1.0)	0.484(2.0)
lympho(2000,0.1)	0.148(3.0)	0.057(5.0)	0.091(4.0)	0.5(1.0)	0.321(2.0)
lympho(3000,0.1)	0.189(3.0)	0.037(5.0)	0.099(4.0)	0.47(1.0)	0.414(2.0)
lympho(5000,0.1)	0.073(4.0)	0.103(3.0)	0.065(5.0)	0.274(2.0)	0.317(1.0)
Avg Rank	3.625	3.876	4.5	1.625	1.375

Table 10: Average rank of methods over the 16 perturbed datasets.

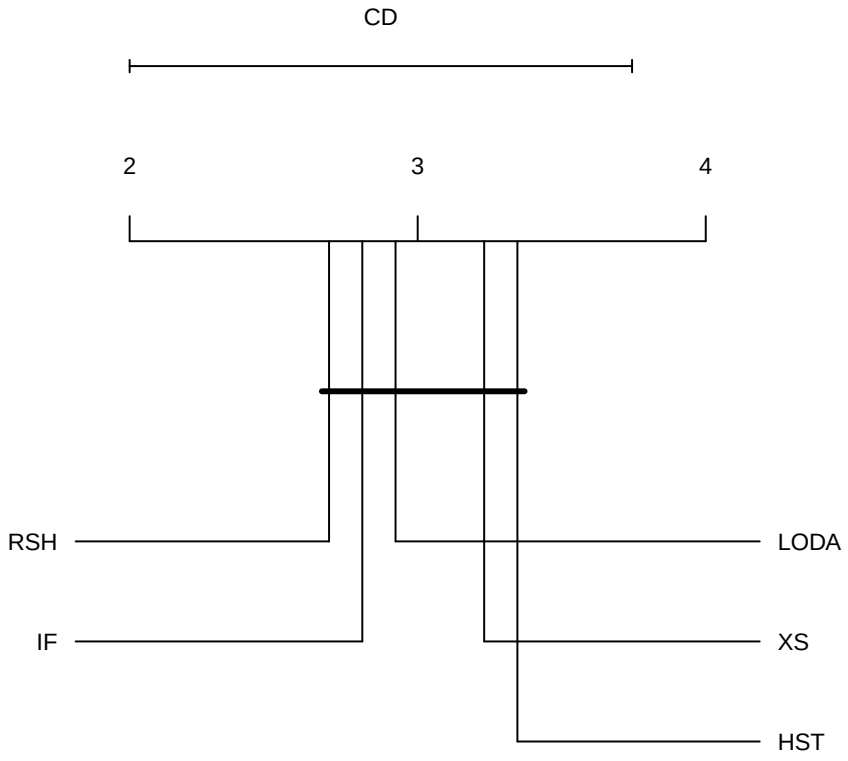


Figure 2: Nemenyi Test visualization for 13 original datasets from ODDS.

2.458×10^{-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 10, we find difference to be significant between (LODA, I-Forest), (X-Stream, I-Forest), (LODA, HS-Trees), (X-Stream, HS-Trees), (LODA, RS-Hash), and (X-Stream, RS-Hash). Though X-Stream is not significantly different from LODA, but the average rank of X-Stream is better than that of LODA. We visualize the same in Fig: 3.

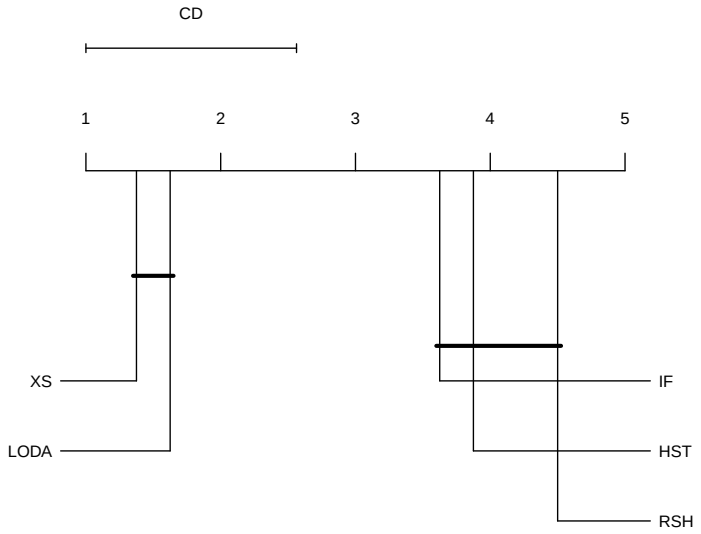


Figure 3: Nemenyi Test visualization for 16 perturbed ODDS datasets.