LUNAR: Unifying Local Outlier Detection Methods via Graph Neural Networks (Supplementary Material)

1 Local Outlier Methods

Method	$\mathbf{e}_{j,i}$	$\phi^{(1)}$	$\square^{(1)}$	$\gamma^{(1)}$	$\phi^{(2)}$	$\square^{(2)}$	$\gamma^{(2)}$
KNN (Anguilli,2002)	$\mathsf{dist}(i,j)$	$\mathbf{e}_{j,i}$	max	$\mathbf{h}_{N_i}$	-	-	-
AGGR-KNN (Anguilli,2002)	$\mathsf{dist}(i,j)$	$\mathbf{e}_{j,i}$	sum	$\mathbf{h}_{N_i}$	-	-	-
LOF (Breunig,2000)	r-$\mathsf{dist}(i,j)$*	$\mathbf{e}_{j,i}$	sum	$\mathbf{h}_{N_i}^{-1}$	$\frac{\mathbf{h}_j}{\mathbf{h}_i}$	mean	$\mathbf{h}_{N_i}$
SIMPLE-LOF (Schubert,2014)	$\mathsf{dist}(i,j)$	$\mathbf{e}_{j,i}$	sum	$\mathbf{h}_{N_i}^{-1}$	$\frac{\mathbf{h}_j}{\mathbf{h}_i}$	mean	$\mathbf{h}_{N_i}$
LOOP (Kriegel, 2009a)	$\mathsf{dist}(i,j)$	$\mathbf{e}_{j,i}^2$	mean	$\lambda\mathbf{h}_{N_i}$	$\frac{\mathbf{h}_j}{\mathbf{h}_i}-1$	sum	$\max\{0, \mathrm{erf}(\frac{1}{\sqrt{2}}\mathbf{x}_{N_i})\}$
INFLO (Jin,2006)	$\mathsf{dist}(i,j)$	$\mathbf{e}_{j,i}^2$	k^{th}-max	$\mathbf{h}_{N_i}^{-1}$	$\frac{\mathbf{h}_j}{\mathbf{h}_i}$	mean	$\mathbf{h}_{N_i}$
DBSCAN (Ester,1996)	$\mathsf{dist}(i,j)$	$H(\epsilon-\mathbf{e}_{j,i})$	sum	$H(\mathbf{h}_{N_i}-\text{minPts})$	$\mathbf{h}_j$	max	$1-\mathbf{h}_{N_i}$
ROS (Pei,2006)	1	$\mathbf{h}_j-\mathbf{h}_i$	sum	$1-\min_{1\le r\le n}\mathbf{h}_{N_j}$	-	-	-
SOD (Kriegel,2009b)	1	$\mathbf{h}_j$	mean	$\mathbf{h}_i-\mathbf{h}_{N_i}$	$\mathbf{h}_j$	mean	$\frac{\mathbf{h}_{N_i}\cdot v_c^{R(i)}}{\|v_c^{R(i)}\|_1}$***

* r-$\mathsf{dist}(i,j)$ is the reachability distance.
** erf is the Gauss error function
*** $v_c^{R(i)}=1$ for $c=1,..,d$.

2 Experiments

2.1 Datasets

Dataset	Link
HRSS	https://www.kaggle.com/init-owl/high-storage-system-data-for-energy-optimization
MI-F	https://www.kaggle.com/shasun/tool-wear-detection-in-cnc-mill
MI-V	https://www.kaggle.com/shasun/tool-wear-detection-in-cnc-mill
OPTDIGITS	http://odds.cs.stonybrook.edu (Rayana2016)
PENDIGITS	http://odds.cs.stonybrook.edu (Rayana2016)
SATELLITE	http://odds.cs.stonybrook.edu (Rayana2016)
SHUTTLE	http://odds.cs.stonybrook.edu (Rayana2016)
THYROID	http://odds.cs.stonybrook.edu (Rayana2016)

2.2 Statistical Significance

Table 1 shows the p values for which the performance of LUNAR is statistically significant over the local outlier methods (LOF, KNN and DN2) according to the one-sided Wilcoxon test, where all of the trials over all tested values of k are considered.

Table 2 shows the p values for the which the performance of LUNAR is statistically significant over the second best performing baseline for each dataset respectively (with $k = 100$ for local outlier methods specifically as in the main experiments). We perform the one-sided Wilcoxon test as well as T test. We see that overall, LUNAR significantly performs better than the other methods (p ¡ 0.01) for most datasets. The lowest p-value possible from Wilcoxon tests is limited by the number of trials, which explains so many values are at this lower limit.

Dataset	LOF	KNN	DN2
HRSS	0.00000012**	0.00000012**	0.00000012**
MI-F	0.00037028**	0.00000012**	0.03093528*
MI-V	0.00000012**	0.00000012**	0.00000012**
OPTDIGITS	0.00020157**	0.00000053**	0.00000012**
PENDIGITS	0.00000029**	0.00000085**	0.00000012**
SATELLITE	0.00034885**	0.64070398	0.00000094**
SHUTTLE	0.00000012**	0.00000014**	0.00000012**
THYROID	0.00000086**	0.00000012**	0.00000012**

Table 1: p-values of the significance of the improvement of LUNAR over the selected local outlier methods, calculated with the Wilcoxon one-sided test, over all values of k tested in experiments. $p < 0.01$ is marked by ** and $p < 0.05$ is marked by *.

Dataset	Wilcoxon Test	T-test
HRSS	0.02155722*	0.00000000**
MI-F	0.44636920	0.84219156
MI-V	0.02155722*	0.00000086**
OPTDIGITS	0.02155722*	0.25849456
PENDIGITS	0.02155722*	0.00947708**
SATELLITE	0.97844278	0.33650830
SHUTTLE	0.02155722*	0.00002765**
THYROID	0.02155722*	0.00004539**

Table 2: p-values of the significance of the improvement of LUNAR over the second best-performing baseline method, calculated with the Wilcoxon one-sided test and T-test, for $k = 100$. $p < 0.01$ is marked by ** and $p < 0.05$ is marked by *.

2.3 Standard Deviations

Tables 3 and 4 show the standard deviations of the scores associated with each model and dataset, over the five trials performed for each.

Dataset	IFOREST	OC-SVM	LOF	KNN	AE	VAE	DAGMM	SO-GAAL	DN2	LUNAR
HRSS	0.54	0.20	0.20	0.19	1.47	0.6	0.28	4.58	4.81	0.26
MI-F	1.22	0.37	2.58	0.50	9.06	0.6	0.66	1.85	3.86	0.44
MI-V	1.83	0.82	0.83	0.67	3.22	2.11	0.55	14.62	1.26	0.20
OPTDIGITS	2.80	2.59	0.29	0.98	1.65	2.22	2.76	8.25	8.90	0.25
PENDIGITS	1.13	1.49	0.41	0.78	2.29	2.31	1.12	2.17	8.44	0.25
SATELLITE	2.22	1.53	0.88	0.75	2.25	3.36	1.08	1.30	9.33	1.18
SHUTTLE	0.05	0.01	0.03	0.01	0.48	0.02	0.20	0.03	1.31	0.02
THYROID	1.75	0.97	1.29	0.91	3.46	1.18	2.52	5.00	1.92	1.49

Table 3: Standard deviations of AUC scores for each method on each dataset.

k	LOF	KNN	DN2	LUNAR	LOF	KNN	DN2	LUNAR	LOF	KNN	DN2	LUNAR
			HRSS				MI-F				MI-V	
2	0.15	0.26	1.11	0.27	0.21	0.28	2.87	0.31	0.44	0.34	3.73	0.21
10	0.54	0.14	4.91	0.29	0.94	0.63	3.71	0.20	0.78	0.81	2.53	0.24
30	0.35	0.18	4.93	0.21	3.46	1.17	4.25	0.67	1.53	0.85	0.94	0.21
50	0.24	0.16	4.89	0.23	1.18	0.84	4.69	0.60	1.53	0.73	1.31	0.21
100	0.20	0.19	4.81	0.26	2.58	0.50	3.86	0.44	0.83	0.67	1.26	0.20
150	0.19	0.23	4.85	0.36	1.38	0.66	3.59	0.35	1.25	0.49	1.31	0.18
200	0.18	0.25	4.93	0.95	0.79	0.49	3.54	0.66	0.47	0.63	1.35	0.20
			OPTDIGITS				PENDIGITS				SATELLITE	
2	0.23	0.17	13.94	0.18	0.42	0.26	12.83	0.24	1.19	0.85	3.88	0.77
10	0.12	0.21	12.65	0.18	0.21	0.25	13.62	0.25	1.28	0.87	6.78	4.08
30	0.21	0.26	11.52	0.24	0.30	0.58	13.74	0.24	1.01	0.81	8.84	1.32
50	0.23	0.44	10.35	0.18	0.38	0.67	12.97	0.25	0.94	0.82	9.22	1.50
100	0.29	0.98	8.90	0.25	0.41	0.78	8.44	0.25	0.88	0.75	9.33	1.18
150	0.39	1.43	9.27	0.20	0.57	0.90	4.70	0.25	0.81	0.72	9.00	1.45
200	0.43	1.66	9.22	0.21	0.72	1.00	4.47	0.26	0.84	0.75	8.74	1.56

k	LOF	KNN	DN2	LUNAR	LOF	KNN	DN2	LUNAR
		SHUTTLE				THYROID		
2	0.02	0.01	1.57	0.01	2.18	1.33	2.81	1.43
10	0.01	0.01	1.69	0.01	1.69	1.41	1.63	2.34
30	0.04	0.01	1.57	0.02	1.51	1.08	1.14	3.08
50	0.07	0.01	1.35	0.01	1.53	1.04	1.38	1.31
100	0.03	0.01	1.31	0.02	1.29	0.91	1.92	1.49
150	0.03	0.01	1.40	0.03	0.89	1.03	1.98	0.87
200	0.03	0.01	1.55	0.02	0.75	1.02	2.09	1.13

Table 4: Standard deviations of AUC scores of LOF, KNN, DN2 and LUNAR for different values of k

2.4 Runtimes

Table 5 shows the average runtime (in seconds) over the five trials for each method and dataset, with $k = 100$ for the local outlier methods. Our method is faster than the other deep methods tested in all cases.

Dataset	IFOREST	OC-SVM	LOF	KNN	AE	VAE	DAGMM	SO-GAAL	LUNAR
HRSS	2.95	137.86	13.73	14.64	579.67	245.34	55.92	34.54	33.71
MI-F	1.37	26.23	18.53	18.91	170.70	149.18	18.67	68.26	12.21
MI-V	1.83	14.41	11.34	11.74	127.02	93.01	14.05	57.56	10.35
OPTDIGITS	0.45	1.26	1.69	1.71	40.27	58.17	5.31	35.13	4.00
PENDIGITS	0.38	0.86	0.59	0.59	68.44	76.55	6.19	38.20	4.00
SATELLITE	0.33	0.54	0.42	0.46	41.27	12.18	4.41	5.30	3.88
SHUTTLE	1.71	38.53	2.83	3.15	374.56	123.62	32.44	18.09	17.76
THYROID	0.35	35.18	0.54	0.58	61.44	28.16	5.90	35.18	5.35

Table 5: Runtime of each method on each dataset. DN2 is omitted as its runtime is virtually equivalent to AE as it uses the same model for feature extraction.

2.5 Ablation Study: Negative Sampling

Table 6 shows the performance of LUNAR when different formulations of negative samples are used. SP refers to only 'Subspace Perturbation' samples while U refers to only 'Uniform' samples. 'Mixed' refers to combining both types.

k	SP	U	Mixed	SP	U	Mixed	SP	U	Mixed	SP	U	Mixed
	HRSS			MI-F			MI-V			OPTDIGITS		
2	93.08	91.06	93.88	81.37	79.91	81.50	96.05	95.54	96.06	00.31	99.94	99.91
10	93.63	75.97	92.67	82.69	79.22	82.39	96.18	94.97	96.09	03.05	99.94	99.79
30	93.81	73.48	92.47	82.54	73.49	82.84	96.18	77.95	96.17	79.38	99.91	99.78
50	93.90	76.66	92.21	83.71	69.58	83.58	96.39	70.37	96.38	91.47	99.89	99.81
100	93.32	66.34	92.17	84.17	57.76	84.37	96.64	67.99	96.73	93.81	99.86	99.76
150	93.27	52.46	91.61	83.05	52.78	82.82	96.40	66.35	96.53	96.05	99.88	99.73
200	91.42	47.62	90.09	84.57	49.11	84.47	96.29	65.75	96.30	96.99	99.86	99.78
	PENDIGITS			SATELLITE			SHUTTLE			THYROID		
2	99.62	99.84	99.84	87.79	87.83	87.83	99.95	99.98	99.98	83.34	81.72	83.38
10	99.78	99.83	99.82	86.91	87.53	84.98	99.94	99.96	99.95	85.24	79.80	84.24
30	99.66	99.83	99.80	87.62	87.56	87.58	99.95	99.91	99.94	85.76	56.05	84.59
50	99.74	99.82	99.80	86.98	87.17	87.08	99.95	99.89	99.97	85.91	45.95	86.01
100	99.78	99.82	99.81	85.37	85.12	85.35	99.96	99.54	99.97	85.99	45.42	85.44
150	99.77	99.82	99.76	84.07	85.63	83.95	99.95	95.67	99.95	86.34	45.48	86.08
200	99.74	99.81	99.71	84.44	86.10	84.70	99.97	93.82	99.97	86.64	46.26	86.67

Table 6: Performance of LUNAR for each setting of k and for different formulation strategies for negative sampling

2.6 Ablation Study: Hidden Layer Size

Table 7 shows the performance when we vary the size of the hidden layers used in the aggregation network in $\{64, 128\}$. A value of 256 is used in the experiments shown in the main paper. The general pattern is that wider layers give better performance across the range of k values for most datasets.

2.7 Ablation Study: Network Depth

In Table 8, we vary the depth of the network between 2 and 3 hidden layers. 4 layers are used in the experiments shown in the main paper. The general pattern is that a deeper network gives better performance across the range of k values for most datasets, with the difference between more significant for larger k.

Layer Size = 64							
k	2	10	30	50	100	150	200
HRSS	92.98	90.96	86.94	84.83	85.27	85.02	83.07
MI-F	80.68	80.40	80.55	81.14	82.49	82.28	83.53
MI-V	95.80	95.66	95.87	96.08	96.62	96.50	96.42
OPTDIGITS	99.88	99.86	99.77	99.67	99.74	99.75	99.75
PENDIGITS	99.84	99.83	99.81	99.80	99.81	99.55	99.53
SATELLITE	87.81	87.42	87.72	87.04	85.04	84.05	83.92
SHUTTLE	99.98	99.96	99.94	99.87	99.86	99.91	99.93
THYROID	82.17	84.05	84.37	85.88	84.23	85.50	85.83
Layer Size = 128							
k	2	10	30	50	100	150	200
HRSS	93.75	92.05	91.21	90.78	90.64	89.43	88.26
MI-F	81.33	81.82	82.27	82.59	83.40	82.69	84.23
MI-V	96.00	96.01	96.07	96.28	96.69	96.58	96.39
OPTDIGITS	99.88	99.76	99.72	99.73	99.68	99.72	99.73
PENDIGITS	99.84	99.83	99.80	99.81	99.81	99.69	99.62
SATELLITE	87.84	87.38	87.52	87.00	85.30	84.15	84.18
SHUTTLE	99.98	99.96	99.93	99.91	99.95	99.95	99.96
THYROID	83.00	84.26	85.34	85.30	85.05	85.77	86.52

Table 7: Average AUC scores over five trials for LUNAR with different layer sizes/widths in the neural network.

Depth = 2							
k	2	10	30	50	100	150	200
HRSS	92.37	82.64	76.41	76.29	72.26	69.95	67.89
MI-F	81.30	78.17	77.87	78.54	80.32	81.15	82.69
MI-V	95.87	94.77	93.85	94.52	94.86	95.18	95.19
OPTDIGITS	99.89	99.90	99.68	99.78	99.55	99.58	99.58
PENDIGITS	99.84	99.83	99.81	99.80	99.80	99.76	99.74
SATELLITE	87.59	87.60	87.54	87.23	85.35	83.95	82.88
SHUTTLE	99.94	99.97	99.92	99.88	99.77	99.74	99.77
THYROID	82.39	80.61	79.91	80.19	81.09	80.16	80.59
Depth = 3							
k	2	10	30	50	100	150	200
HRSS	93.39	91.58	88.67	86.57	86.87	85.55	84.80
MI-F	80.91	80.73	80.62	81.63	82.69	82.50	84.33
MI-V	95.88	95.81	95.85	96.03	96.55	96.39	96.21
OPTDIGITS	99.92	99.76	99.76	99.78	99.77	99.76	99.71
PENDIGITS	99.84	99.82	99.79	99.81	99.82	99.70	99.70
SATELLITE	87.85	86.47	87.55	87.03	85.35	84.19	83.81
SHUTTLE	99.98	99.96	99.91	99.87	99.88	99.92	99.93
THYROID	82.31	84.28	85.47	86.05	84.86	85.42	85.59

Table 8: Average AUC scores over five trials for LUNAR with different layer counts/depth in the neural network.

Angiulli, F.; and Pizzuti, C. 2002. Fast outlier detection in high dimensional spaces. In European conference on principles of data mining and knowledge discovery, 15–27. Springer.

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 93–104.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd, volume 96, 226–231. Jin, W.; Tung, A. K. H.; Han, J.; and Wang, W. 2006. Ranking outliers using

symmetric neighborhood relationship. In PAKDD, 577–593. Springer.

Kriegel, H.-P.; Kroger, P.; Schubert, E.; and Zimek, A. 2009a. LoOP: local outlier probabilities. In Proceedings of the 18th ACM conference on Information and knowledge management, 1649–1652.

Kriegel, H.-P.; Kroger, P.; Schubert, E.; and Zimek, A. 2009b. Outlier detection in axis-parallel subspaces of high dimensional data. In Pacific-asia conference on knowledge discovery and data mining, 831–838. Springer.

Pei, Y.; Zaiane, O. R.; and Gao, Y. 2006. An efficient reference-based approach to outlier detection in large datasets. In Sixth International Conference on Data Mining (ICDM'06), 478–487. IEEE.

Schubert, E.; Zimek, A.; and Kriegel, H.-P. 2014. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data mining and knowledge discovery, 28(1): 190–237

Shebuti Rayana (2016). ODDS Library [http://odds.cs.stonybrook.edu]. Stony Brook, NY: Stony Brook University, Department of Computer Science.