# Random Similarity Isolation Forests: Supplementary Results

## S1 Datasets

The used datasets were taken from outlier detection benchmarks and generators for different types of data:

- numerical and categorical data: `https://github.com/Minqi824/ADBench`
- time series: `https://outlier-detection.github.io/utsd`
- sequences of sets: `https://gingerbread.shinyapps.io/SequencesOfSetsGenerator/`
- images, text, and graphs: `https://github.com/GuansongPang/ADRepository-Anomaly-detection-datasets`

The exact versions of the datasets can be found in the code repository accompanying this paper. When looking for benchmarks, we favored those in which the examples marked as outliers constituted 5% or fewer examples in the dataset.

**Sensitivity analysis.** For the analysis of the hyperparameters of RSIF, we used datasets that were independent of those used for the experimental comparison with other methods. More precisely, we used 10 datasets: 4 numerical (`cardio`, `lymphography`, `optdigits`, `speech`), 1 categorical (`ad`), 3 graph (`cox2`, `bzr`, `dhfr`), 1 text embedding (`agnews`), and 1 time-series (`twoleadecg`).

**Scalar data.** For tests with scalar data, we used 10 popular benchmark datasets, including 5 numerical (`glass`, `musk`, `satimage`, `vowels`, `wbc`) and 5 categorical (`aid`, `apascal`, `cmc`, `reuters`, `solarflare`). The datasets were chosen for their variety in terms of the number of examples and features (Table S1).

**Complex data.** For experiments with complex objects, we used 5 graph (`aids`, `dd`, `enzymes`, `nci1`, `proteins`), 2 time series (`earthquakes`, `aibo`), 3 text (`amazon`, `imdb`, `yelp`), and 3 image (`cifar`, `fashionmnist`, `svhn`) datasets. For the text and image datasets we used embeddings of pretrained RoBERTa[1] and ViT[2] models, respectively.

**Mixed data.** For mixed-type data, we used 3 sequences of sets (`item`, `length`, `order`) and 3 multiomics (`ovarian`, `her2`, `rosmap`) datasets. Sequences of sets, by nature, can be treated as sets, as sequences, or as combinations of the two representations. The multiomics datasets consist of the results of different genetic measurements, represented as distributions (lengths of variants) or numbers (gene expression).

[1] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. **CoRR abs/1907.11692**, 1–13 (2019)

[2] Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. **CoRR abs/2010.11929**, 1–22 (2020)

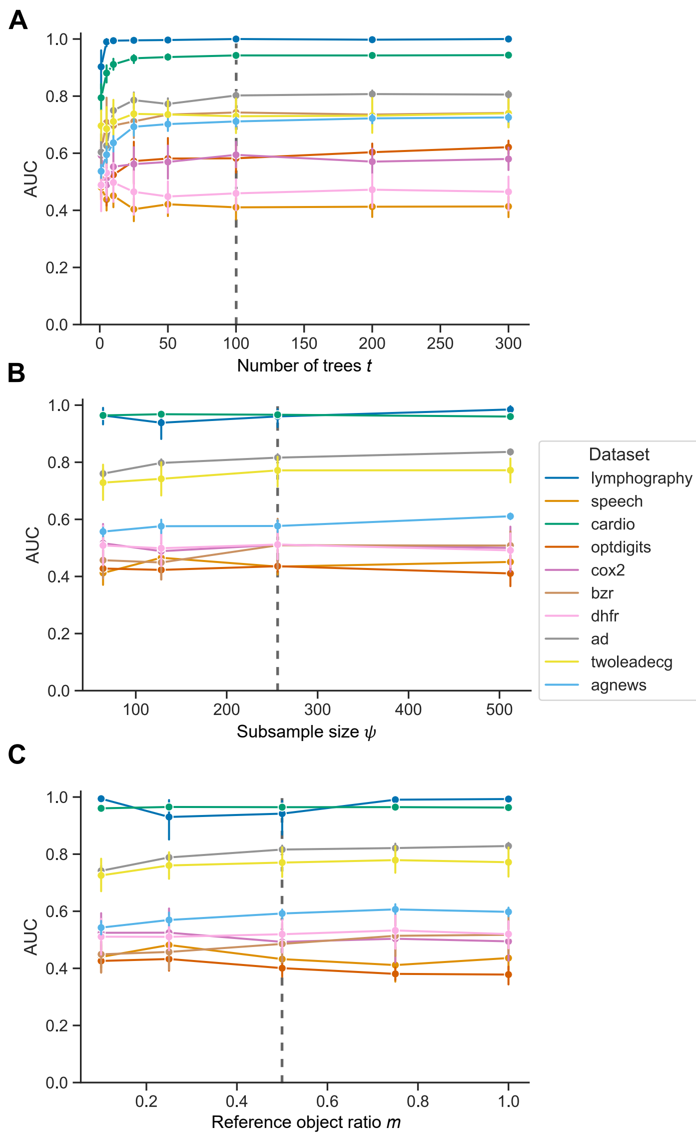


**Fig. S1.** Hyperparameter sensitivity analysis. Effect of (**A**) the number of trees $t$, (**B**) subsample size $\psi$, and (**C**) pool of reference objects on RSIF's predictive performance (ROC AUC). Bars represent 95% confidence intervals. The dashed gray lines show the selected defaults.

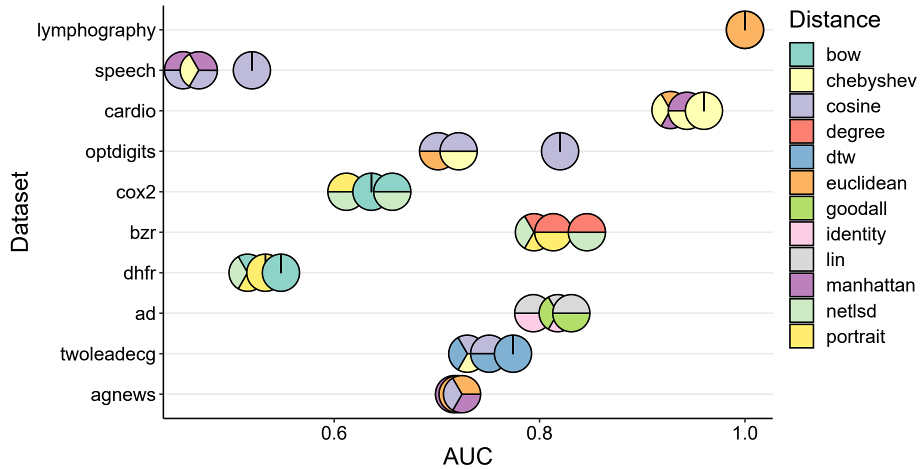


**Fig. S2.** Top 3 distance measure combinations for each of the sensitivity test datasets. Each set of measures used by RSIF is represented by a circle. If the set of distances consists of more than one measure, the circle is divided into multiple colored pieces, with colors defining the measures.

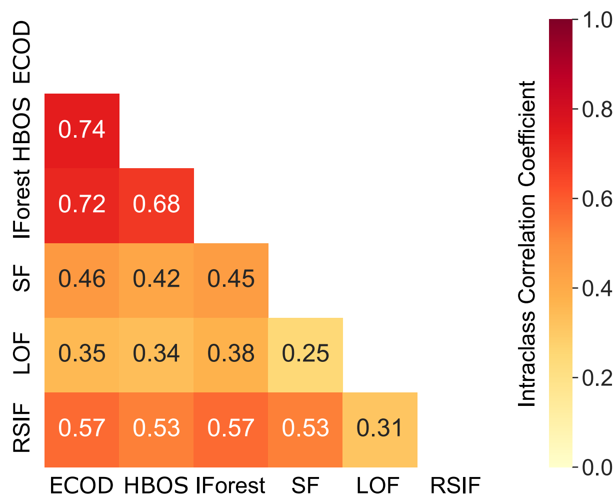


**Fig. S3.** Pairwise average intraclass correlation coefficients (ICC) based on outlier scores for the holdout test sets.

**Table S1.** AUC performance of RSIF and five competitive methods. The best results on each dataset are highlighted in bold, and the second best are underlined.

| Dataset | Type | Features | #Ex. | #Feat. | %Outlier | AUC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | IF | LOF | HBOS | ECOD | SF | RSIF |
| glass | | | 214 | 7 | 4.21 | 0.72 | 0.74 | 0.76 | 0.60 | 0.73 | **0.80** |
| letter | | | 1600 | 32 | 6.25 | 0.61 | **0.92** | 0.59 | 0.55 | 0.75 | 0.77 |
| musk | | | 3062 | 166 | 3.17 | 1.00 | 0.60 | 1.00 | 0.96 | **1.00** | 1.00 |
| annthyroid | | | 7200 | 6 | 7.42 | 0.80 | 0.71 | 0.61 | 0.78 | 0.68 | **0.84** |
| satimage | | | 5803 | 36 | 1.22 | 0.99 | 0.34 | 0.97 | 0.96 | 0.98 | **0.99** |
| thyroid | | numeric | 3772 | 6 | 2.47 | 0.98 | 0.48 | 0.95 | 0.98 | 0.98 | **0.98** |
| vowels | | | 1456 | 12 | 3.43 | 0.69 | **0.93** | 0.66 | 0.59 | 0.59 | 0.91 |
| waveform | scalar | | 3443 | 21 | 2.90 | 0.73 | 0.74 | 0.69 | 0.59 | **0.83** | 0.76 |
| wbc | | | 223 | 9 | 4.48 | **1.00** | 0.92 | 0.99 | 1.00 | 0.99 | **1.00** |
| wdbc | | | 367 | 30 | 2.72 | 0.99 | 0.98 | 0.99 | 0.97 | **1.00** | 0.99 |
| wilt | | | 4819 | 5 | 5.33 | 0.46 | **0.69** | 0.41 | 0.39 | 0.34 | 0.53 |
| aid | | | 4278 | 114 | 1.40 | 0.65 | 0.58 | 0.66 | **0.66** | 0.61 | 0.64 |
| apascal | | | 12694 | 64 | 1.39 | 0.49 | 0.55 | 0.66 | **0.66** | 0.55 | 0.56 |
| cmc | | categorical | 1472 | 8 | 1.97 | 0.57 | 0.51 | **0.59** | **0.59** | 0.53 | 0.57 |
| reuters | | | 12896 | 100 | 1.84 | 0.98 | 0.95 | **0.99** | **0.99** | 0.98 | 0.98 |
| solarflare | | | 1065 | 11 | 4.04 | 0.80 | 0.55 | 0.84 | 0.84 | **0.85** | 0.81 |
| nci1 | | | 2160 | 1 | 4.77 | 0.48 | **0.56** | 0.46 | 0.49 | 0.47 | 0.53 |
| aids | | | 1680 | 1 | 4.76 | 0.92 | 0.83 | 0.96 | 0.92 | 0.99 | **0.99** |
| enzymes | | graph | 105 | 1 | 4.76 | **0.76** | 0.61 | 0.68 | 0.72 | 0.63 | 0.63 |
| proteins | | | 696 | 1 | 4.47 | 0.54 | 0.58 | 0.35 | 0.67 | 0.68 | **0.70** |
| dd | | | 726 | 1 | 4.82 | 0.66 | 0.46 | 0.31 | 0.75 | **0.79** | 0.78 |
| earthquakes | | | 387 | 512 | 4.91 | 0.61 | 0.57 | 0.49 | 0.56 | 0.43 | **0.64** |
| aibo | | time series | 367 | 70 | 4.90 | 0.50 | **0.63** | 0.50 | 0.46 | 0.55 | 0.55 |
| ECGFiveDays | complex | | 465 | 136 | 4.95 | 0.80 | **0.91** | 0.75 | 0.67 | 0.74 | 0.79 |
| MPOC | | | 583 | 80 | 4.97 | 0.68 | **0.75** | 0.62 | 0.53 | 0.66 | 0.61 |
| amazon | | | 10000 | 768 | 5.00 | 0.52 | **0.55** | 0.51 | 0.52 | 0.49 | 0.50 |
| imdb | | text | 10000 | 768 | 5.00 | 0.47 | **0.52** | 0.47 | 0.47 | 0.48 | 0.50 |
| yelp | | | 10000 | 768 | 5.00 | 0.54 | **0.59** | 0.55 | 0.56 | 0.50 | 0.54 |
| cifar | | | 5263 | 512 | 5.00 | 0.73 | **0.73** | 0.68 | 0.71 | 0.68 | 0.71 |
| fashionmnist | | image | 6315 | 512 | 5.00 | **0.84** | 0.74 | 0.76 | 0.83 | 0.81 | 0.83 |
| svhn | | | 5208 | 512 | 5.00 | 0.56 | **0.66** | 0.48 | 0.54 | 0.57 | 0.55 |
| items | | | 210 | 1 | 4.76 | 0.83 | 0.83 | **0.84** | **0.84** | 0.75 | 0.76 |
| length | | sequences of sets | 210 | 1 | 4.76 | 0.85 | 0.87 | **0.92** | **0.92** | 0.87 | 0.81 |
| order | mixed | | 210 | 1 | 4.76 | 0.53 | 0.53 | 0.55 | **0.59** | 0.51 | 0.54 |
| ovarian | | | 125 | 50 | 4.80 | 0.50 | 0.29 | 0.45 | 0.57 | 0.33 | **0.69** |
| breast | | multiomics | 770 | 50 | 3.64 | 0.62 | 0.83 | 0.49 | 0.63 | 0.83 | **0.83** |
| rosmap | | | 177 | 600 | 4.52 | 0.62 | 0.60 | 0.68 | 0.67 | **0.70** | 0.66 |
| | | | | | **Avg. rank** | 3.50 | 3.57 | 3.78 | 3.51 | 3.78 | **2.85** |