Unsupervised anomaly detection algorithms on real-world data: **how many do we need?**

Roel Bouman
(roel.bouman@ru.nl)
Zaharah Bukhsh
Tom Heskes

NEURAL INFORMATION PROCESSING SYSTEMS
Radboud Universiteit
SINCE 1923

SCAN ME

Introduction

- Several comparison studies have been conducted in the past,
    - many of them are **outdated**,
    - compare **few algorithms, or**
    - **do not evaluate on real-world data**.

We present t**he largest comparison to date**, comparing **33 algorithms** on **52 datasets**. All data and code is made publicly available.

Anomalies can have many properties. These include:

- Global – Local
- Enclosed – Peripheral
- Isolated – Clustered
- Univariate – Multivariate

Few of the existing studies take these properties into account in their comparison. **Our research enables researchers to incorporate this knowledge when choosing algorithms.**

In this study, we aim to answer the following questions:

- How many algorithms do we need for tackling real-world tabular data?
- Are the current benchmark datasets enough to cover the breadth of anomaly properties?

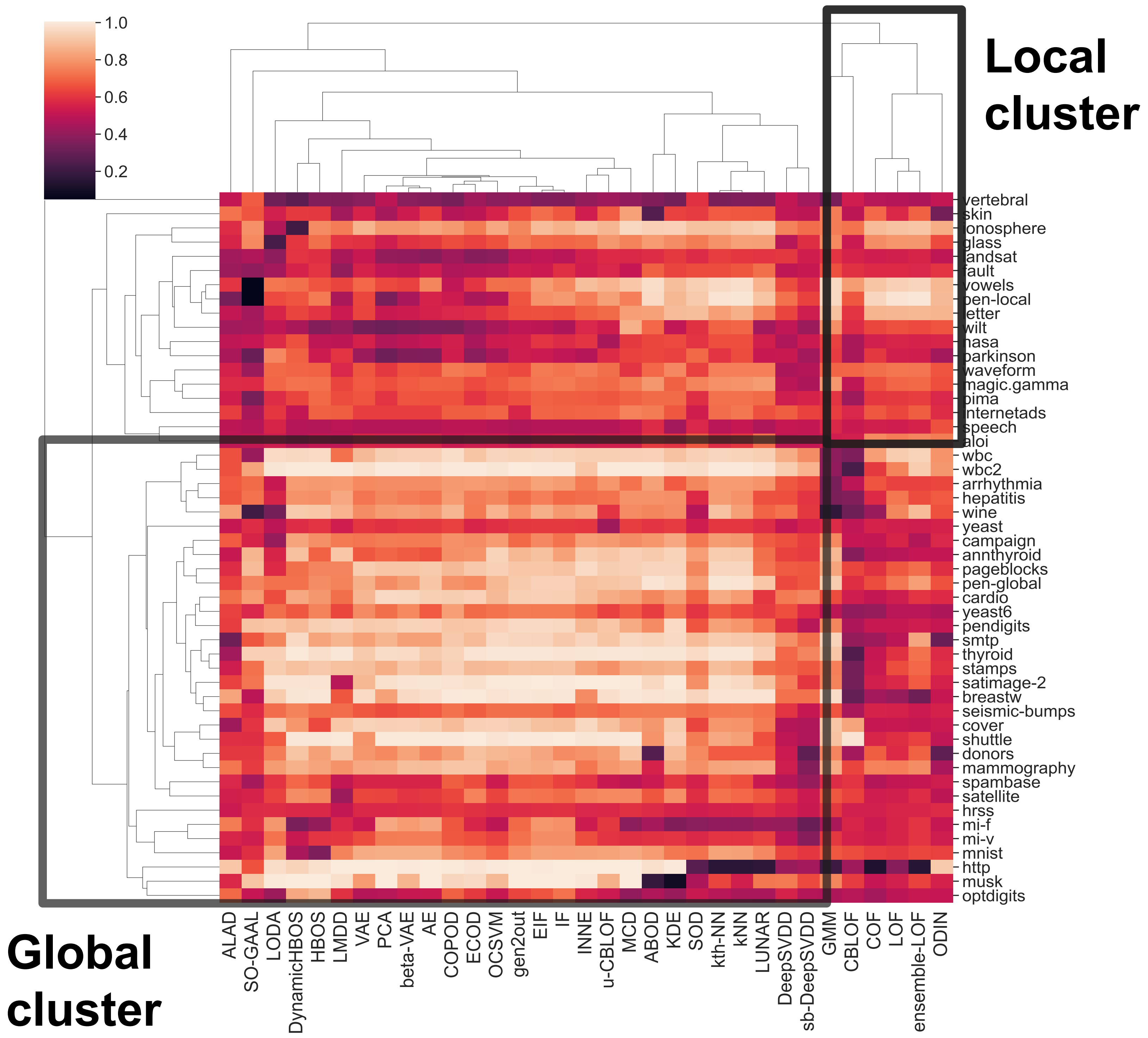

Figure 1: Clustered heatmap of the ROC/AUC performance of each algorithm. The algorithms and datasets are each clustered using hierarchical clustering with average linkage and the Pearson correlation as metric. A lighter color indicates a better performance.

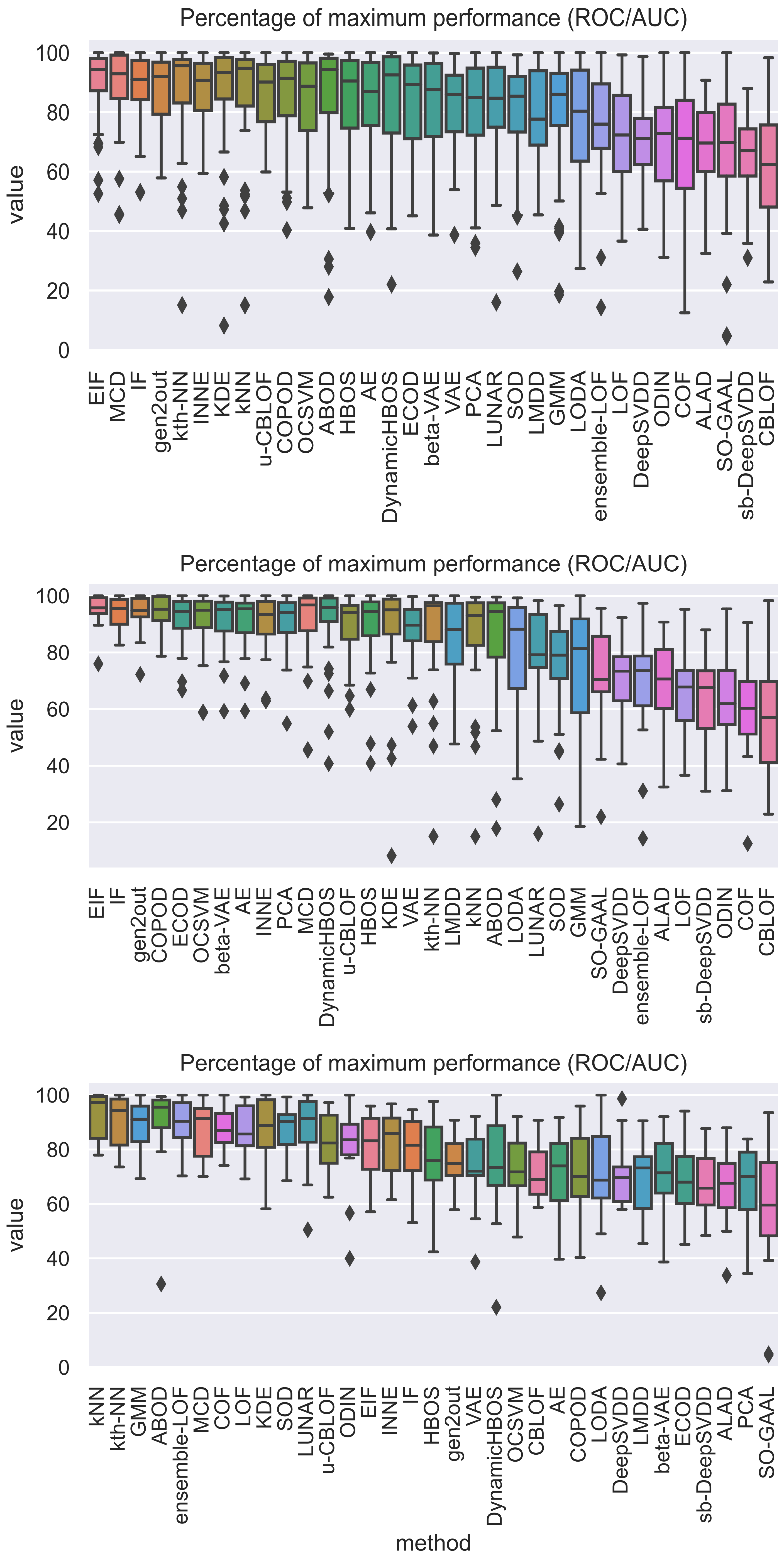

Figure 2: Boxplots of the performance of each algorithm on (a): all, (b): local, and (c): global datasets in terms of percentage of maximum AUC. The maximum AUC is the highest AUC value obtained by the best performing algorithm on that particular dataset.

Results

- The **Extended Isolation Forest (EIF)** outperforms 14 out of 33 algorithms(p=0.05) with an average AUC of 0.770 when considering all datasets. (Figure 2: top)
- By two-way clustering of the AUC scores (Fig 1) we found that two clear clusters of datasets emerge: local and global. This can be seen from the strong performance of the "local" algorithms on datasets known to contain local anomalies.
- These clusters seem to show anti-correlated patterns of performance, indicative of Simpson's paradox.
- Therefore we divided our dataset in two groups (global vs. local), and repeated the statistical analysis.
    - The top "local" algorithm is **kNN** (17 algorithms, avg. AUC 0.737), see Figure 2: middle.
    - The top "global" algorithm is **EIF** (13 algorithms, avg. AUC 0.849), see Figure 2: bottom.

Discussion

While we have conducted the largest comparison of anomaly detection algorithms to date, there are still many things left to be done.

- We only cover real-valued tabular datasets. specific types of data (text, images, etc.) have not been examined.
- The distinction global/local highly corresponds to unimodal/multimodal. Full disentanglement is only possible by targeted simulation studies.
- We only found a separation between local and global datasets. We likely need more datasets to identify the different properties of anomalies.
- We found neural networks to work subpar on many datasets. It is likely that they need much more data or more optimization than can be done in the unsupervised setting on this scale.

Conclusion

- **We need only 2** of the 33 analysed algorithms to achieve strong performance on most datasets.
- When a researcher does not know what properties the anomalies in their dataset have, or when dealing with global anomalies, the best choice is the **EIF** algorithm.
- When dealing with local anomalies the **kNN** method performs best.
- Many of the known properties of anomalies are not identifiably present in this large benchmark.
    - We therefore need more open-sourced datasets.