# SUBSELECT: UNSUPERVISED MODEL SELECTION FOR ANOMALY DETECTION USING SUBSET CONTRAST

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Unsupervised anomaly detection involves selecting an algorithm and its hyperparameters, but the absence of labels makes it difficult to validate this choice. We propose SUBSELECT, an unsupervised model selection approach for anomaly detection that uses only the models' anomaly scores and unlabeled data. It compares the highest-scoring observations, meaning the most anomalous ones according to the respective model, with the remaining data. If a model separates anomalies and normal observations well, its highest-scoring observations should differ from the remaining data in feature space. We quantify the *contrast* between the highest-scoring and the remaining observations with a combination of complementary measures. To avoid choosing a single top subset fraction, we aggregate the results across multiple fractions. We evaluate our approach against 34 unsupervised model selection methods and two baselines on 39 datasets, using 297 candidate anomaly detection models. On average precision (AP), SUBSELECT attains the best mean rank and outperforms 34 of the 36 compared approaches in pairwise Bayesian signed-rank tests, with no comparison decided against it. Our selections are not significantly worse than always choosing the 54th-best model out of the 297 models, compared with the 80th-best for the next selection method. SUBSELECT is the only method in our comparison that significantly exceeds the expected AP of an Isolation Forest baseline with randomly chosen hyperparameters. For ROC-AUC, SUBSELECT also significantly outperforms the Isolation Forest baseline, attains the best mean rank, and outperforms the majority of other methods in pairwise Bayesian signed-rank tests, with no comparison decided against it.

## 1 INTRODUCTION

Anomaly detection is often applied when labeled observations are unavailable (Chandola et al., 2009). To detect anomalies, researchers have developed a wide variety of algorithms based on distance (Ramaswamy et al., 2000), density (Breunig et al., 2000), and other principles. From these algorithms, practitioners must choose one with suitable hyperparameters, a choice that can lead to large performance differences (Campos et al., 2016; Han et al., 2022; Bouman et al., 2024).

We distinguish three groups of approaches that select models without labels on the dataset at hand: *stand-alone* scores that evaluate each model independently, *consensus methods* that use agreement among models, and *meta-learning* methods that transfer performance from labeled reference datasets (Goix, 2016; Marques et al., 2020; Zhao et al., 2021; 2022).

Ma et al. (2023) evaluated stand-alone and consensus model selection methods on 39 datasets and 297 anomaly detection models. On average precision (AP), the stand-alone scores show no significant advantage over random model selection, while consensus strategies generally improve on random selection but do not significantly outperform the IFOREST-R baseline. This baseline averages the performance of Isolation Forest (Liu et al., 2008) over 81 hyperparameter configurations. Although it is a strong baseline, other candidate models often outperform it, leaving room for better model selection.

We propose SUBSELECT, an unsupervised model selection approach that compares each candidate model's highest-scoring observations with the remaining data in feature space. If the highest-scoring observations of a model are anomalies, they should differ from the remaining data. We call the



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

proportion of observations included in the highest-scoring subset the *top subset fraction*. Rather than choosing a single fraction and contrast measures, we evaluate the contrast using multiple measures and aggregate their results across several top subset fractions. Each contrast uses only the individual model's output and the input data, requiring neither agreement among models nor labeled reference datasets.

None of the selection methods compared in this paper combines the three components of SUBSELECT: using the input features to compare a model's highest-scoring observations with the remaining data, complementary contrast measures, and aggregation over multiple top subset fractions. If a model concentrates anomalies near the top, the remaining observations provide a reference with a lower proportion of anomalies. Different measures assess different aspects of this contrast, while aggregation across fractions avoids choosing a single cutoff without labels to validate it. Section 2 discusses the three groups of model selection methods and explains how SUBSELECT differs from them in more detail.

Our contributions:

- **Method.** We propose SUBSELECT, an unsupervised selection approach for anomaly detection models based on contrasts between a model's highest-scoring observations and the remaining data in feature space. It combines multiple contrast measures to account for different ways in which anomalies can differ from normal observations.
- **Aggregation across subset fractions.** We aggregate candidate-model ranks across multiple top subset fractions rather than choosing one fraction for each dataset. Our empirical analysis supports this aggregation, and our recommended configuration requires no dataset-specific adjustment of selection hyperparameters.
- **Experiments.** We compare SUBSELECT with 34 selection methods and two baselines on 39 datasets and 297 candidate models from eight algorithms. SUBSELECT significantly outperforms IFOREST-R on both AP and ROC-AUC, and is the only compared method to do so on AP. It attains the best mean rank on both metrics, and Bayesian signed-rank tests decide 34 of the 36 comparisons in its favor on AP, with no comparison decided against it on either metric.
- **Software and research resources.** We provide the `subselect` Python package, which supports custom contrast measures and aggregation rules, together with precomputed experimental results (see the Reproducibility statement).

## 2 RELATED WORK

We group model selection methods into stand-alone, consensus and meta-learning methods (see Table 5 in Appendix A.3.3).

**Stand-alone.** Cluster-quality indices compare the high and low anomaly scores separated by a threshold, rather than observations in feature space (Ma et al., 2023). In feature space, Excess-Mass (EM) and Mass-Volume (MV) criteria assess whether regions that a model regards as normal contain many observations within a small volume (Goix et al., 2015; Goix, 2016). Focusing on individual observations, Internal, Relative Evaluation of Outlier Solutions (IREOS) favors models that assign high anomaly scores to observations a classifier can easily separate from the others (Marques et al., 2020). Fitting a separate classifier for each observation makes IREOS computationally expensive. Therefore, Similarity-based Internal, Relative Evaluation of Outlier Solutions (SIREOS) replaces classifier-based separability with average similarities between observations, favoring models that score the least similar observations highest (Marques et al., 2022). IREOS and SIREOS are similar to SUBSELECT because they assess a model's ranking of observations in the feature space. SUBSELECT, however, forms top and reference subsets from each candidate model's scores and compares them using multiple contrast measures fitted on the reference subset.

**Consensus.** Ma et al. (2023) adapt Unsupervised Disentanglement Ranking (UDR) (Duan et al., 2020), ModelCentrality (MC) (Lin et al., 2020), its sampled variant MCS (Ma et al., 2023), and HITS (Kleinberg, 1999) to select models whose rankings of the observations agree with those of other candidate models. UDR, MC, and MCS measure agreement using Spearman's $\rho$, Kendall's

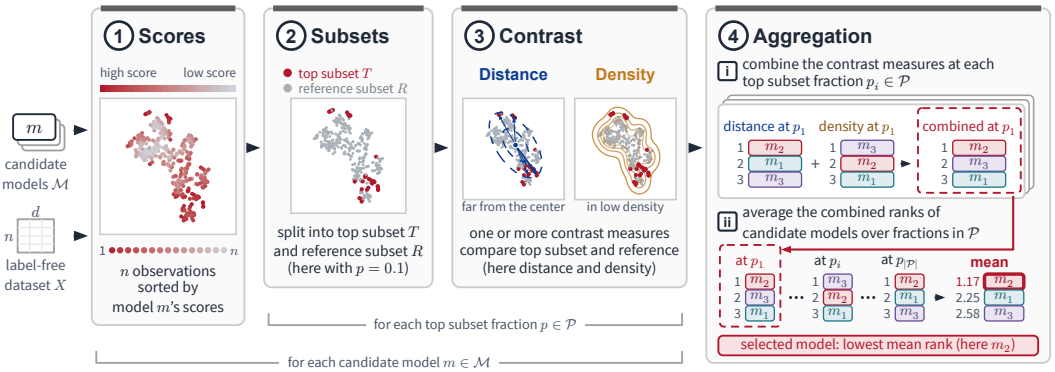


Figure 1: **Overview of SUBSELECT.** **(1)** For each candidate model, we sort the observations by their anomaly score. **(2)** For each top subset fraction $p \in \mathcal{P}$, the highest-scoring observations form the top subset $T$ and the remainder the reference subset $R$. **(3)** We measure the contrast between the subsets $R$ and $T$. **(4.i)** We combine the candidate models' ranks across measures and rank the combined values at each fraction $p$. **(4.ii)** The model with the lowest mean rank across all fractions $p \in \mathcal{P}$ is selected.

$\tau$, or normalized discounted cumulative gain (NDCG). The ensemble method SELECT (Rayana & Akoglu, 2016) is adapted as Ens, which selects the model whose ranking agrees most with an ensemble built from a subset of candidate models (Ma et al., 2023). HITS and Ens can also combine candidate-model scores into a single ensemble ranking of the observations, reported as HITS-auth and Ens-pseudo (Ma et al., 2023). Unlike consensus methods, SUBSELECT does not use agreement between model outputs as an indicator of quality. Its rank-based combination and aggregation depend on the candidate models, but each model's raw contrasts are computed independently.

**Meta-learning.** MetaOD learns from model performance on labeled reference datasets to recommend a model for a new dataset (Zhao et al., 2021). Similarly, ELECT predicts performance differences between models from internal evaluation measures and uses these predictions to identify the reference datasets used for selection (Zhao et al., 2022). Learning to Learn with Optimal Transport for Unsupervised Scenarios (LOTUS) instead identifies the most similar labeled reference dataset for a given dataset using optimal transport and recommends suitable algorithms, preprocessing steps, and hyperparameters (Singh et al., 2026). These approaches rely on labeled reference datasets, which SUBSELECT does not need.

## 3 METHOD

**Notation and Problem.** We are given a matrix of observations $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, where $\boldsymbol{x}_i \in \mathbb{R}^d$, and a set of trained anomaly detection models $\mathcal{M} = \{m_1, \ldots, m_M\}$. Each model $m \in \mathcal{M}$ computes *anomaly scores* $\boldsymbol{s}^{(m)} \in \mathbb{R}^n$ for the observations $\boldsymbol{X}$, where its $i$-th coordinate $[\boldsymbol{s}^{(m)}]_i \in \mathbb{R}$ indicates how anomalous the observation $\boldsymbol{x}_i$ is. In the following, higher anomaly scores indicate more anomalous observations. Using only the observations $\boldsymbol{X}$ and the anomaly scores $\boldsymbol{s}^{(m)}$ of all candidate models in $\mathcal{M}$, the task is to select from $\mathcal{M}$ a model with good detection performance, without access to anomaly labels.

**Assumptions.** The following assumptions motivate SUBSELECT. **(A1)** Normal observations concentrate in regions with high probability density in the feature space, whereas anomalies are comparatively rare and fall into regions with low probability density (Ruff et al., 2021; Chandola et al., 2009). **(A2)** Normal observations follow a common distribution that may contain several separate clusters (Ruff et al., 2021; Chandola et al., 2009). Because anomalies are rare (A1), we can estimate this normal behavior from a mostly normal sample (Rousseeuw & Hubert, 2018). **(A3)** Anomalies may be scattered or clustered and may deviate globally or locally from normal data (Han et al., 2022). Characterizing these diverse anomalies using a single criterion is challenging (Bouman et al., 2024). Their proportions differ between datasets (Han et al., 2022), and we assume that their pro-

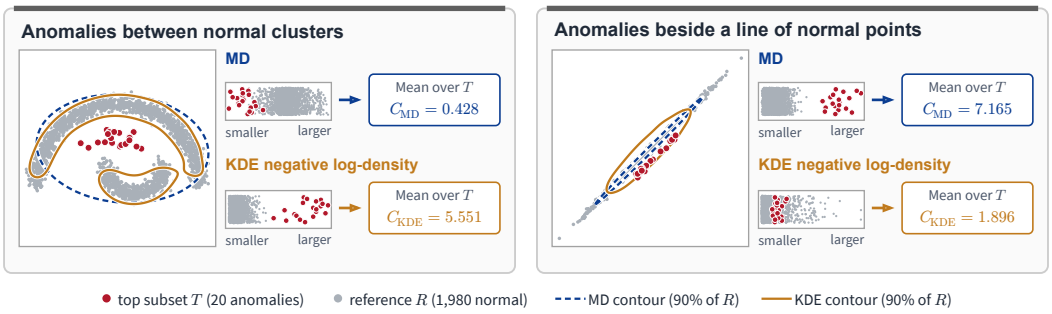


Figure 2: **Contrasts for two synthetic examples.** At $p = 0.01$, an ideal model places all 20 anomalies in the top subset $T$, leaving 1,980 normal observations in reference subset $R$. The left dataset is adapted from the Big Moon, Small Moon example of Ruff et al. (2021). Each strip plots pointwise contrasts $\phi_R(\boldsymbol{x})$ for $T$ (red) and $R$ (gray) on its own scale, with arrows indicating their mean over $T$, $C_\phi(m, p)$. Dashed MD and solid KDE contours enclose 90% of $R$.

portion and type are unknown. **(A4)** A model that ranks anomalies well should concentrate them among its highest-scoring observations, leaving fewer anomalies among the remaining observations.

**Overview.** SUBSELECT assesses a candidate model by comparing its highest-scoring observations with the remaining data (Figure 1). For each model $m$ and top subset fraction $p$, these groups form the top subset $T$ and reference subset $R$. If the ranking concentrates more anomalies in $T$, then $R$ contains fewer anomalies and provides a better reference for normal behavior (A1, A4). To capture different aspects of the contrast between the top subset $T$ and the reference subset $R$ (A2, A3), we combine complementary measures (Section 3.2). For each top subset fraction, we combine the ranks of candidate models from both measures. We then average the resulting ranks across fractions and select the model with the lowest mean rank (Sections 3.3 and 3.4). The recommended configuration requires no dataset-specific selection hyperparameters. The computational cost and opportunities for parallel execution are discussed in Appendix A.2.

### 3.1 FORMING SUBSETS FROM MODEL SCORES

For a model $m \in \mathcal{M}$ and a *top subset fraction* $p \in (0, 1/2]$, the *top subset* $T_p^{(m)}$ contains the indices of the $\max\{1, \lfloor pn \rfloor\}$ observations with the highest anomaly scores from $m$. The *reference subset* contains the remaining indices,

$$R_p^{(m)} = \{1, \ldots, n\} \setminus T_p^{(m)}, \tag{1}$$

with $|R_p^{(m)}| = n - |T_p^{(m)}|$. If there is no ambiguity, we write $T = T_p^{(m)}$ and $R = R_p^{(m)}$.

### 3.2 MEASURING CONTRAST AGAINST THE REFERENCE SUBSET

A *contrast measure* $\phi_R : \mathbb{R}^d \to \mathbb{R}$ is fitted using only the reference observations $R$ and assigns a contrast value to each observation in the top subset $T$. Larger values indicate a stronger deviation from the reference under that measure. For a candidate model $m$ and fraction $p$, we average these values over its top subset $T$ to obtain

$$C_\phi(m, p) = \frac{1}{|T|} \sum_{i \in T} \phi_R(\boldsymbol{x}_i). \tag{2}$$

As contrast measures, we suggest using Mahalanobis distance (MD) and kernel density estimation (KDE), two established distance- and density-based methods for anomaly detection (Ruff et al., 2021). MD measures the distance from the center of the reference subset $R$ to an observation in the top subset $T$, while accounting for variation and correlations among the reference features, and gives $C_{\text{MD}}$ (see Appendix A.1.1).

KDE measures whether observations of the top subset $T$ lie in regions of low density according to the reference subset $R$. Because our Gaussian KDE uses a single bandwidth across features,

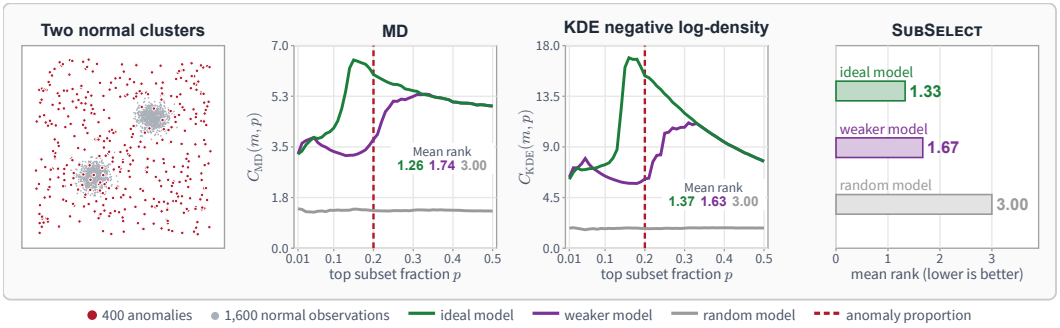


Figure 3: **Contrasts across top subset fractions. Left.** The 2,000 observations include two normal clusters of equal spread and 400 uniformly distributed anomalies. **Middle.** The *ideal model* ranks all anomalies first, the *weaker model* starts with 100 anomalies and then alternates anomalies and normal observations until all anomalies are included, and the *random model* uses a random ordering. The ideal and weaker models' top subsets coincide from $p = 0.35$. Dashed red lines mark the anomaly proportion 0.20, and colored numbers give mean ranks across all 50 fractions for each measure, with lower ranks preferred. At $p = 0.05$, selection based on MD, KDE, and their combination selects the weaker model, whereas aggregation across all 50 fractions selects the ideal model for all three. **Right.** Combining measures and aggregating across fractions gives the ideal model the lowest mean rank.

we first standardize both subsets using the reference mean and standard deviation of each feature (see Appendix A.1.2). Using the negative log-density as $\phi_R$ gives $C_{\mathrm{KDE}}$.

### 3.3 Combining contrast measures

We do not expect a single contrast measure to capture all anomaly patterns (A3), and the more informative measure may vary across datasets. Choosing one measure for each dataset would introduce another selection decision without labels for validation. Figure 2 shows two example datasets, where KDE highlights anomalies in the low-density region between normal clusters, while MD captures departures from a narrow band of normal observations. Therefore, we combine contrast measures to complement each other rather than choosing one, motivated by prior work on unsupervised model selection (Jaskowiak et al., 2016; Marques et al., 2022). Because the raw contrasts have different scales, we rank the candidate models separately under each measure at a fixed fraction $p$. Throughout, the largest value receives rank one and tied values receive their average rank. We convert these ranks into percentile ranks and average them across the set of contrast measures $\Phi$,

$$\rho_\phi(m,p) = 1 - \frac{\mathrm{rank}[C_\phi(\cdot,p)]_m - 1}{M}, \qquad \overline{\rho}(m,p) = \frac{1}{|\Phi|}\sum_{\phi\in\Phi}\rho_\phi(m,p). \tag{3}$$

Here, $\mathrm{rank}[C_\phi(\cdot,p)]_m$ is the rank of candidate model $m$ among all $M$ models at fraction $p$. Larger values of $\rho_\phi$ indicate larger contrasts relative to the other candidate models.

### 3.4 Aggregating across top subset fractions

The top subset fraction $p$ affects the contrasts $C_\phi(m,p)$, as illustrated in Figure 3 (middle). Choosing an appropriate top subset fraction $p$ without labels is challenging, especially because anomaly ratios are typically unknown (A3) and vary widely across datasets (Han et al., 2022). We therefore suggest aggregating the combined contrast scores $\overline{\rho}(m,p)$ over multiple top subset fractions $p \in \mathcal{P}$. Because the spread of the combined contrast scores $\overline{\rho}(m,p)$ across candidate models varies with $p$, we rank the candidate models at each fraction and average these ranks across the grid $\mathcal{P}$,

$$r(m,p) = \mathrm{rank}[\overline{\rho}(\cdot,p)]_m, \qquad \overline{r}(m) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} r(m,p). \tag{4}$$

The same rank convention as in Equation (3) applies, so smaller $\overline{r}(m)$ indicates a more favorable average position across fractions. We suggest using the grid $\mathcal{P} = \{0.01, 0.02, \ldots, 0.50\}$, with

$|\mathcal{P}| = 50$, for every dataset, as we do in our experiments. The upper limit $p = 0.50$ reflects the minority assumption (A1), keeping the reference subset at least as large as the top subset. The selected model is then

$$m^\star \in \arg\min_{m \in \mathcal{M}} \overline{r}(m). \tag{5}$$

# 4 EXPERIMENTS

## 4.1 DATASETS, CANDIDATE MODELS, AND SELECTION METHODS

Our experiments follow Ma et al. (2023), with 39 real-world datasets (Appendix A.3.1) and 297 candidate models from eight anomaly detection algorithms (Appendix A.3.2). We use the authors' anomaly scores, obtained by training each candidate model without labels on the entire dataset and scoring the same observations (Ma et al., 2023). The labels of the datasets are used for evaluation. For the 30 selection variants, we use the model-selection results supplied by Ma et al. (2023) and the corresponding AP and ROC-AUC values for each dataset. We then recompute all evaluation statistics described in Section 4.2. Additionally, we evaluated SIREOS (Marques et al., 2022), MetaOD (Zhao et al., 2021), ELECT (Zhao et al., 2022), and LOTUS (Singh et al., 2026), as described in Appendix A.3.3, giving 34 comparison methods in total.

## 4.2 EVALUATION

We measure model performance by AP and ROC-AUC. Appendix A.4.1 contains the ROC-AUC results, with the main findings summarized in Section 5.

**Comparison with baselines.** Following Ma et al. (2023), we compare every selection method with the baselines RANDOM, the expected performance of choosing uniformly from all 297 candidate models, and IFOREST-R, the expected performance of choosing uniformly from the 81 Isolation Forest models. We compare each method's performance with the baselines using the one-sided paired Wilcoxon signed-rank test (Wilcoxon, 1945) across the 39 datasets, without correction.

**Comparison within the candidate models.** For each selection method, the smallest-$q$ statistic of Ma et al. (2023) counts how many of the best candidate models of each dataset are significantly better than the models it selects, so lower values are better. It compares the selections with always choosing the $k$-th best candidate model for each dataset using a one-sided Wilcoxon test across the 39 datasets, for $k = 1, 2, \ldots$ until the difference is no longer significant (Ma et al., 2023). We also count the number of datasets in which the selected model lies among the best 10, 20, or 30 percent of the 297 candidate models, where 39 is the maximal achievable value.

**Comparison among selection methods.** We rank the 37 approaches, including SUBSELECT and the two baselines, by the AP and ROC-AUC of their selected models on each dataset. We then report their mean ranks and mean AP and ROC-AUC values across the 39 datasets. To compare performance differences, we apply the Bayesian signed-rank test to all $\binom{37}{2} = 666$ pairs of approaches using their paired AP or ROC-AUC differences across datasets (Benavoli et al., 2017). The test assigns posterior probabilities to either method being better or the two being *practically equivalent*, with differences within 0.01 treated as equivalent. This threshold has been used for accuracy (Benavoli et al., 2017), AP in multilabel classification (Tan et al., 2023), and ROC-AUC in model selection for anomaly detection (Singh et al., 2026). A comparison is *decided* when a posterior probability reaches 0.95 and remains *unresolved* otherwise (Benavoli et al., 2017). Following Benavoli et al. (2017), we do not apply a multiple-testing correction across pairs. Appendix A.4.2 provides a supplementary mean-rank analysis using the Friedman and Nemenyi tests (Demsar, 2006).

# 5 RESULTS

## 5.1 COMPARISON WITH BASELINES

On AP, SUBSELECT significantly outperforms both baselines under the one-sided Wilcoxon test, with $p < 0.001$ against RANDOM and $p = 0.021$ against IFOREST-R (Table 1). Among the 34 other

Table 1: **AP results across 39 datasets.** One-sided Wilcoxon $p$-values follow Ma et al. (2023), comparing with RANDOM (R) and IFOREST-R (IF). Lower mean rank and $q$ are better. Bold marks $p < 0.05$, posterior probabilities $\geq 0.95$, and column-best values among selection methods. SD denotes standard deviation and $q$ is zero-indexed smallest-$q$. Top counts give the numbers of datasets with performance among the best $10/20/30\%$ of the 297 candidate models. $P_>$ and $P_=$ are Bayesian posterior probabilities that SUBSELECT is better than or practically equivalent to the row approach. $W$ counts the row approach's decided wins (posterior probability $\geq 0.95$, maximum 36). (A)/(L) denote SELECT-ALL/SELECT-LAST. (U) denotes unified scores for HITS and HITS-auth.

| Method | $p$ vs. baseline $p_{\mathrm{R}}$ | $p_{\mathrm{IF}}$ | Mean AP rank | Mean AP value ± SD | $q$ | Top counts 10/20/30% | $P_>$ | $P_=$ | $W$ |
|---|---|---|---|---|---|---|---|---|---|
| RANDOM | – | 1.000 | 20.69 | 0.342 ± 0.234 | 144 | 0/0/7 | **> 0.99** | 0.00 | 3 |
| IFOREST-R | **< 0.001** | – | 16.92 | 0.399 ± 0.300 | 84 | 0/5/16 | **0.98** | 0.00 | 17 |
| **SUBSELECT** | **< 0.001** | **0.021** | **13.46** | **0.424 ± 0.293** | **53** | **9/16/21** | – | – | **34** |
| MC-NDCG | **< 0.001** | 0.182 | 14.18 | 0.404 ± 0.291 | 82 | 5/9/18 | 0.81 | 0.13 | 19 |
| HITS-auth (U) | **< 0.001** | 0.364 | 15.12 | 0.397 ± 0.299 | 87 | 2/8/15 | 0.94 | 0.01 | 17 |
| HITS-auth | **< 0.001** | 0.577 | 15.13 | 0.401 ± 0.286 | 94 | 1/3/11 | **0.99** | 0.00 | 20 |
| MC-$\rho$ | **< 0.001** | 0.217 | 15.42 | 0.395 ± 0.289 | 89 | 5/8/16 | **0.97** | 0.01 | 17 |
| UDR-NDCG | **0.004** | 0.825 | 15.64 | 0.384 ± 0.270 | 93 | 8/11/14 | **0.99** | 0.00 | 16 |
| MC-$\tau$ | **0.002** | 0.062 | 15.72 | 0.396 ± 0.297 | 81 | 6/7/19 | **0.99** | 0.00 | 15 |
| ELECT | **0.001** | 0.186 | 15.99 | 0.395 ± 0.296 | 83 | 8/12/16 | **0.99** | 0.00 | 16 |
| Ens-pseudo (A) | **0.001** | 0.423 | 16.29 | 0.373 ± 0.282 | 79 | 4/10/18 | **0.98** | 0.00 | 15 |
| Ens (A) | **0.002** | 0.730 | 16.33 | 0.371 ± 0.282 | 81 | 6/8/18 | **0.98** | 0.00 | 15 |
| MCS-NDCG | **0.001** | 0.205 | 16.35 | 0.391 ± 0.285 | 83 | 4/13/17 | **0.97** | 0.02 | 14 |
| Ens-pseudo (L) | **0.001** | 0.489 | 16.55 | 0.374 ± 0.283 | 84 | 4/8/15 | **0.97** | 0.00 | 16 |
| HITS | **< 0.001** | 0.494 | 16.65 | 0.397 ± 0.299 | 95 | 3/5/11 | **0.98** | 0.01 | 16 |
| Ens (L) | **0.001** | 0.667 | 16.82 | 0.373 ± 0.284 | 87 | 5/7/15 | **0.97** | 0.00 | 16 |
| MCS-$\tau$ | **0.001** | 0.599 | 17.22 | 0.397 ± 0.305 | 90 | 5/7/15 | **0.99** | 0.00 | 16 |
| UDR-$\rho$ | **0.012** | 0.905 | 17.33 | 0.383 ± 0.283 | 104 | 7/9/12 | **> 0.99** | 0.00 | 14 |
| UDR-$\tau$ | **0.019** | 0.952 | 17.69 | 0.379 ± 0.282 | 109 | 5/9/11 | **> 0.99** | 0.00 | 13 |
| IREOS | 0.303 | 0.908 | 18.60 | 0.335 ± 0.261 | 99 | 6/11/14 | **> 0.99** | 0.00 | 3 |
| MCS-$\rho$ | **0.007** | 0.707 | 18.73 | 0.385 ± 0.289 | 108 | 5/6/10 | **0.99** | 0.01 | 11 |
| SIREOS | 0.097 | 0.910 | 19.46 | 0.346 ± 0.261 | 103 | 6/9/12 | **0.99** | 0.00 | 5 |
| LOTUS | 0.270 | 0.981 | 19.74 | 0.350 ± 0.282 | 116 | 6/8/10 | **> 0.99** | 0.00 | 5 |
| d | 0.572 | 0.990 | 20.58 | 0.350 ± 0.296 | 120 | 3/5/10 | **> 0.99** | 0.00 | 2 |
| EM | 0.539 | 0.979 | 20.68 | 0.322 ± 0.265 | 115 | 3/11/13 | **> 0.99** | 0.00 | 3 |
| rs | 0.500 | 0.981 | 21.22 | 0.354 ± 0.298 | 127 | 4/7/13 | **> 0.99** | 0.00 | 2 |
| ch | 0.500 | 0.981 | 21.22 | 0.354 ± 0.298 | 127 | 4/7/13 | **> 0.99** | 0.00 | 2 |
| xb | 0.500 | 0.981 | 21.22 | 0.354 ± 0.298 | 127 | 4/7/13 | **> 0.99** | 0.00 | 2 |
| MV | 0.687 | 0.994 | 21.44 | 0.311 ± 0.262 | 120 | 2/8/11 | **> 0.99** | 0.00 | 3 |
| db | 0.522 | 0.982 | 21.55 | 0.343 ± 0.296 | 123 | 6/10/16 | **> 0.99** | 0.00 | 2 |
| sd | 0.910 | 0.995 | 22.78 | 0.295 ± 0.254 | 156 | 7/10/13 | **> 0.99** | 0.00 | 1 |
| s | 0.953 | 0.999 | 22.79 | 0.289 ± 0.268 | 163 | 3/7/13 | **> 0.99** | 0.00 | 1 |
| h | 0.849 | 0.996 | 22.83 | 0.299 ± 0.250 | 131 | 2/6/9 | **> 0.99** | 0.00 | 1 |
| i | 0.882 | 0.997 | 23.09 | 0.299 ± 0.248 | 139 | 2/6/9 | **> 0.99** | 0.00 | 1 |
| HITS (U) | 0.999 | 1.000 | 24.17 | 0.219 ± 0.205 | 203 | 6/8/11 | **> 0.99** | 0.00 | 0 |
| MetaOD | 0.839 | 1.000 | 24.74 | 0.324 ± 0.273 | 146 | 2/4/7 | **> 0.99** | 0.00 | 2 |
| std | 1.000 | 1.000 | 28.64 | 0.223 ± 0.228 | 231 | 4/5/6 | **> 0.99** | 0.00 | 0 |

selection methods, 17 significantly outperform RANDOM on AP, but none of the 14 stand-alone comparison methods does. SUBSELECT has a higher AP than IFOREST-R on 25 of the 39 datasets, with a median paired gain across all 39 datasets of 0.012, compared with at most 0.006 for any other selection method. Thus, among the compared selection methods, only SUBSELECT provides a significant AP improvement over using a single Isolation Forest with random hyperparameters.

On ROC-AUC, SUBSELECT also significantly outperforms both baselines, with $p < 0.001$ against RANDOM and $p = 0.012$ against IFOREST-R (Table 6 in Appendix A.4.1). IREOS is the only stand-alone comparison method that significantly outperforms RANDOM, and four other methods significantly outperform IFOREST-R. SUBSELECT has higher ROC-AUC than IFOREST-R on 30 datasets, with a median paired gain of 0.008, compared with at most 0.006 for the other selection methods.

### 5.2 COMPARISON AMONG THE MODEL SELECTION APPROACHES

The Bayesian signed-rank test compares SUBSELECT with the other approaches using their paired AP differences (Table 1). It decides 34 of the 36 comparisons in favor of SUBSELECT, including both baselines and 32 of the 34 selection methods, with none decided against it. The unresolved comparisons are MC-NDCG and HITS-auth (unified), with posterior probabilities of 0.81 and 0.94, respectively, that SUBSELECT is better.

SUBSELECT also has the best mean rank on AP, 13.46 against 14.18 for MC-NDCG, and the highest mean AP value. On the smallest-$q$, the models selected by SUBSELECT are not significantly worse than always choosing the 54th-best candidate model, against the 80th-best for the next method, Enspseudo (SELECT-ALL). It leads the top-10%, top-20% and top-30% counts within the 297-model pool, with 9, 16 and 21 of the 39 datasets, against at most 8, 13 and 19 for any other method. UDR-NDCG and ELECT are closest at 10% but fall further behind at 30%, whereas MC-$\tau$ is closest at 30% but has fewer top-10% selections. On ROC-AUC (Table 6 in Appendix A.4.1), SUBSELECT again has the best mean rank, the lowest smallest-$q$, and the highest top-10, top-20, and top-30 percent counts. HITS-auth, however, has the highest mean ROC-AUC value. The Bayesian signed-rank test decides 24 of the 36 comparisons in favor of SUBSELECT, with none decided against it. On both metrics, the test decides more comparisons for SUBSELECT than for any other approach.

### 5.3 ABLATION STUDY

We examine the contrast measures and top subset fractions, with additional analyses in Appendix A.4.4. SUBSELECT retains similar mean AP values across pool sizes and compositions shown in Table 9 in Appendix A.4.4. Separating the results by dataset collection gives positive mean AP differences from both baselines in each collection (Table 10 in Appendix A.4.4).

Table 2: **Contrast-measure ablation on AP.** All variants aggregate over the same 50 top subset fractions and replace SUBSELECT in the 37-approach mean-rank comparison. Bold marks $p < 0.05$ and the best mean rank and $q$. $p_{\mathrm{R}}$ and $p_{\mathrm{IF}}$ compare with RANDOM and IFOREST-R. Both means MD and KDE combined, and $q$ denotes smallest-$q$.

| | $p$ vs. baseline | | Mean AP | |
|---|---|---|---|---|
| Measures | $p_{\mathrm{R}}$ | $p_{\mathrm{IF}}$ | rank | $q$ |
| **SUBSELECT (both)** | **< 0.001** | **0.021** | **13.46** | **53** |
| MD-only | **< 0.001** | 0.079 | 13.90 | 68 |
| KDE-only | **< 0.001** | 0.086 | 14.33 | 65 |

**Ablation regarding contrast measures.** Table 2 compares SUBSELECT with MD-only and KDE-only variants, all aggregating over the same 50 suggested top subset fractions. Both single-measure variants significantly outperform RANDOM on AP and ROC-AUC under the one-sided Wilcoxon test, but only their combination significantly outperforms IFOREST-R on AP. The combination also has the best mean AP rank, smallest-$q$ and mean AP value, although the Bayesian signed-rank test does not decide it better than either single-measure variant (Table 7 in Appendix A.4.4). The better single measure varies by dataset, and combining them can select a model with higher AP than either single-measure selection (Figure 7 in Appendix A.4.3). On both metrics, SUBSELECT performs better than each single-measure variant on more datasets than it performs worse (Figures 8 and 9 in Appendix A.4.4). Fitting both measures on all observations instead of the reference subset reduces the mean AP value from 0.424 to 0.360 without changing the combination or aggregation (Appendix A.4.4).

**Ablation regarding top subset fractions.** We next compare selecting at a single top subset fraction with aggregation over all 50 suggested fractions (Figure 4(a)). No single fraction of

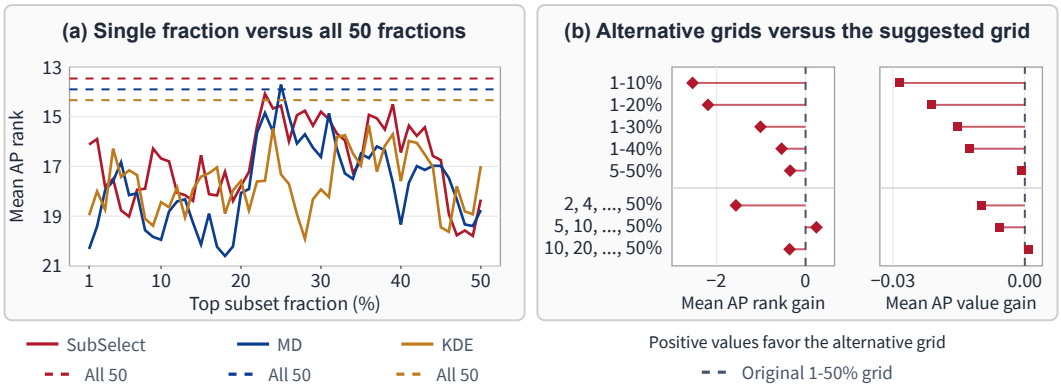


Figure 4: **Effect of the top subset fractions on AP. (a)** Mean AP rank for each single fraction, using the setup of Table 2. Ranks range from 1 (best) to 37 (worst) and lower ranks are plotted higher, with only ranks 13 to 21 shown. Dashed lines show aggregation over all 50 fractions. **(b)** Mean AP rank gain (diamonds, original minus alternative) and mean AP value gain (squares, alternative minus original) relative to SUBSELECT's original grid. Positive values favor the alternative, and dashed vertical lines mark zero. The first five rows use one-percentage-point steps.

SUBSELECT, MD-only or KDE-only achieves a better mean AP rank than SUBSELECT using all 50 fractions, and the same holds for mean ROC-AUC rank. The MD-only variant at 25% does improve on its own mean AP rank over all 50 fractions. Figure 4(b) varies the range and step size of the aggregation grid for SUBSELECT. Among the grids shown, restricting the range to 1%–10% or 1%–20% has the largest adverse effect on mean AP rank and mean AP value, while omitting fractions below 5% has a smaller effect. Keeping every second, fifth or tenth fraction produces mixed changes, with none improving both mean AP rank and mean AP value. These results support aggregation across a broad range without choosing a single fraction. On ROC-AUC, the original grid has a better mean rank and mean value than all tested changes to its range or step size (Figure 10 in Appendix A.4.4).

## 6 LIMITATIONS AND FUTURE WORK

We evaluated SUBSELECT using tabular datasets and eight anomaly detection algorithms. The method applies to deep models on tabular data as well, although we have not evaluated that setting. Extending it to other data types may require suitable feature representations or contrast measures, which we leave to future work.

SUBSELECT depends on how anomalies differ from normal observations in the available features. For example, anomalies near the reference center in a dense region can receive little contrast under MD and KDE. Further contrast measures could address such patterns.

## 7 CONCLUSION

SUBSELECT uses contrasts between a model's highest-scoring observations and the remaining data for unsupervised model selection. It combines MD and KDE across 50 top subset fractions using candidate-model ranks, without validation labels, labeled reference datasets, or agreement among model outputs. In pairwise Bayesian signed-rank tests, SUBSELECT outperforms 34 of the 36 compared approaches on AP and 24 on ROC-AUC, with no comparison decided against it, while attaining the best mean rank on both metrics. It is also the only compared selection method that significantly outperforms IFOREST-R baseline on AP. The ablations support combining contrast measures and aggregating across a broad range of top subset fractions without dataset-specific adjustment of selection hyperparameters.

### Artificial intelligence (AI) use statement

We used OpenAI Codex and Anthropic Claude Code for assistance with implementing and testing the method and evaluation code. We also used these tools for feedback on methodology and interpretation of results. These tools assisted with literature search and review, proposing draft text and providing feedback on the manuscript. AI-assisted code was tested, and its outputs were checked. We made the research decisions and take responsibility for the final work.

### Reproducibility statement

Our anonymous repository provides the implementation of SubSelect and precomputed results per dataset, model, and method.[1] The implementation is distributed as the Python package `subselect`, with an interface compatible with anomaly detection models from scikit-learn (Pedregosa et al., 2011) and the Python Outlier Detection toolbox (PyOD) (Zhao et al., 2019). The repository includes an example that reproduces our model selections from the saved contrasts. Appendix A.3 gives the setup of the compared methods, the statistical procedures, the runtime measurement and the software versions.

## References

Alessio Benavoli, Giorgio Corani, Janez Demsar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *J. Mach. Learn. Res.*, 18:77:1–77:36, 2017.

Roel Bouman, Zaharah Bukhsh, and Tom Heskes. Unsupervised anomaly detection algorithms on real-world data: How many do we need? *J. Mach. Learn. Res.*, 25:105:1–105:34, 2024.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *SIGMOD Conference*, pp. 93–104. ACM, 2000.

Guilherme Oliveira Campos, Arthur Zimek, Jörg Sander, Ricardo J. G. B. Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E. Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Min. Knowl. Discov.*, 30(4): 891–927, 2016.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, 2009.

Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

Sunny Duan, Loic Matthey, Andre Saraiva, Nick Watters, Chris Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. In *ICLR*. OpenReview.net, 2020.

Nicolas Goix. How to evaluate the quality of unsupervised anomaly detection algorithms? *CoRR*, abs/1607.01152, 2016.

Nicolas Goix, Anne Sabourin, and Stéphan Clémençon. On anomaly ranking and excess-mass curves. In *AISTATS*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. In *KI-2012 Poster and Demo Track*, 2012.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *NeurIPS*, 2022.

[1] `https://anonymous.4open.science/r/subselect-review-iclr27`

Charles R. Harris, K. Jarrod Millman, Stéfan van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with numpy. *Nat.*, 585:357–362, 2020.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Pablo A. Jaskowiak, Davoud Moulavi, Antonio C. S. Furtado, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. On strategies for building effective ensembles of relative clustering validity criteria. *Knowl. Inf. Syst.*, 47(2):329–354, 2016.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

Zinan Lin, Kiran Koshy Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and model-centrality: Self-supervised model training and selection for disentangling gans. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6127–6139. PMLR, 2020.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *ICDM*, pp. 413–422. IEEE Computer Society, 2008.

Martin Q. Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. The need for unsupervised outlier model selection: A review and evaluation of internal evaluation strategies. *SIGKDD Explor.*, 25 (1):19–35, 2023.

Henrique O. Marques, Ricardo J. G. B. Campello, Jörg Sander, and Arthur Zimek. Internal evaluation of unsupervised outlier detection. *ACM Trans. Knowl. Discov. Data*, 14(4):47:1–47:42, 2020.

Henrique O. Marques, Arthur Zimek, Ricardo J. G. B. Campello, and Jörg Sander. Similarity-based unsupervised evaluation of outlier detection. In *SISAP*, volume 13590 of *Lecture Notes in Computer Science*, pp. 234–248. Springer, 2022.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

Tomáš Pevný. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102:275–304, 2016.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *SIGMOD Conference*, pp. 427–438. ACM, 2000.

Shebuti Rayana and Leman Akoglu. Less is more: Building selective anomaly ensembles. *ACM Trans. Knowl. Discov. Data*, 10(4):42:1–42:33, 2016.

Peter J. Rousseeuw and Mia Hubert. Anomaly detection by robust statistics. *WIREs Data Mining Knowl. Discov.*, 8(2), 2018.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proc. IEEE*, 109(5):756–795, 2021.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 1992.

Prabhant Singh, Pieter Gijsbers, Elif Ceren Gok Yildirim, Murat Onur Yildirim, and Joaquin Vanschoren. Automated machine learning for unsupervised tabular tasks. *Mach. Learn.*, 115(3):39, 2026.

Chao Tan, Sheng Chen, Xin Geng, and Genlin Ji. A label distribution manifold learning algorithm. *Pattern Recognit.*, 135:109112, 2023.

Jian Tang, Zhixiang Chen, Ada Wai-Chee Fu, and David W. Cheung. Enhancing effectiveness of outlier detections for low density patterns. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2002.

Maxim Terpilovski. scikit-posthocs: Pairwise multiple comparison tests in python. *J. Open Source Softw.*, 4(36):1169, 2019.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *J. Mach. Learn. Res.*, 20:96:1–96:7, 2019.

Yue Zhao, Ryan A. Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. In *NeurIPS*, pp. 4489–4502, 2021.

Yue Zhao, Sean Zhang, and Leman Akoglu. Toward unsupervised outlier model selection. In *ICDM*, pp. 773–782. IEEE, 2022.

# A Appendix

## A.1 Definitions of the contrast measures

### A.1.1 Mahalanobis distance (MD)

We first compute the reference mean $\boldsymbol{\mu}_R \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma}_R \in \mathbb{R}^{d\times d}$,

$$\boldsymbol{\mu}_R = \frac{1}{|R|}\sum_{i\in R} \boldsymbol{x}_i, \qquad \boldsymbol{\Sigma}_R = \frac{1}{|R|}\sum_{i\in R} (\boldsymbol{x}_i - \boldsymbol{\mu}_R)(\boldsymbol{x}_i - \boldsymbol{\mu}_R)^\top. \tag{6}$$

Using the Moore–Penrose pseudoinverse $\boldsymbol{\Sigma}_R^+$ to handle singular covariances, the mean MD contrast is

$$C_{\mathrm{MD}}(m,p) = \frac{1}{|T|}\sum_{i\in T} \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_R)^\top \boldsymbol{\Sigma}_R^+ (\boldsymbol{x}_i - \boldsymbol{\mu}_R)}. \tag{7}$$

Larger values of $C_{\mathrm{MD}}$ indicate that the top subset $T$ differs more from the reference subset $R$.

### A.1.2 KERNEL DENSITY ESTIMATION (KDE)

We want to measure whether observations of the top subset $T$ lie in regions of low density according to the reference subset $R$. Following the idea of KDE (Parzen, 1962), we estimate the reference density by placing a Gaussian kernel at every reference point and averaging their contributions.

Because the Gaussian kernel uses one bandwidth across features, we standardize each feature using its mean and standard deviation over the reference subset. We apply the same transformation to the top subset. For feature $\ell \in \{1, \dots, d\}$, the standardized $\ell$-th coordinate is

$$[\boldsymbol{z}_R(\boldsymbol{x})]_\ell = \frac{[\boldsymbol{x}]_\ell - [\boldsymbol{\mu}_R]_\ell}{[\boldsymbol{\sigma}_R]_\ell}, \tag{8}$$

where $[\boldsymbol{\sigma}_R]_\ell = \sqrt{[\boldsymbol{\Sigma}_R]_{\ell\ell}}$ is the reference standard deviation for feature $\ell$. For features constant on $R$, we use a denominator of one. For an observation $\boldsymbol{z} \in \mathbb{R}^d$ with standardized features, its estimated density is

$$\widehat{f}_R(\boldsymbol{z}) = \frac{1}{|R|(2\pi)^{d/2} h_p^d} \sum_{j \in R} \exp\left(-\frac{\|\boldsymbol{z} - \boldsymbol{z}_R(\boldsymbol{x}_j)\|_2^2}{2h_p^2}\right), \tag{9}$$

where the bandwidth $h_p$ controls the kernel width and follows the sample-size form of Scott's rule (Scott, 1992), $h_p = |R|^{-1/(d+4)}$. We evaluate this density at each standardized observation in $T$ and average its negative natural logarithm,

$$C_{\mathrm{KDE}}(m, p) = -\frac{1}{|T|} \sum_{i \in T} \log \widehat{f}_R\big(\boldsymbol{z}_R(\boldsymbol{x}_i)\big). \tag{10}$$

Larger values of $C_{\mathrm{KDE}}(m, p)$ indicate that the top subset $T$ tends to lie in regions of low density according to the reference estimate.

## A.2 COMPUTATIONAL COST

We analyze the computational cost of selecting a model, excluding the cost of training the candidate models. For each candidate model, we sort the scores once in $\mathcal{O}(n \log n)$ time and reuse that ordering for all $|\mathcal{P}|$ top subset fractions. For the MD and one top subset fraction $p$, computing the reference covariance, its pseudoinverse, and the distances of the top subset costs $\mathcal{O}(nd^2 + d^3)$. Direct KDE evaluation compares each of the $|T_p|$ points of the top subset $T_p$ with all $|R_p|$ reference points, requiring $\mathcal{O}(|T_p|\,|R_p|\,d)$ operations. Summing over all candidate models and top subset fractions, the total computational cost of SUBSELECT is

$$\mathcal{O}\left(\underbrace{Mn \log n}_{\text{score sorting}} + M \sum_{p \in \mathcal{P}} \left[\underbrace{nd^2 + d^3}_{\text{MD}} + \underbrace{|T_p|\,|R_p|\,d}_{\text{KDE}}\right] + \underbrace{|\mathcal{P}|\,M \log M}_{\text{ranking candidate models}}\right). \tag{11}$$

For a fixed grid of $|\mathcal{P}| = 50$ top subset fractions, the computational cost reduces to

$$\mathcal{O}\left(\underbrace{M(nd^2 + d^3)}_{\text{MD}} + \underbrace{Mn^2 d}_{\text{KDE}} + \underbrace{M \log M}_{\text{ranking}}\right). \tag{12}$$

For a fixed number of features $d$, the computational cost in Equation (12) grows quadratically with the number of observations $n$ because of KDE and, apart from the ranking term $M \log M$, linearly with the number of candidate models $M$. For KDE, the bounds count direct kernel evaluations, whereas our scikit-learn implementation uses a k-d tree to evaluate the density estimate. Runtime measurements are reported in Appendix A.4.5.

Contrast calculations for different candidate models, top subset fractions, and measures are independent and can run in parallel. Further reductions are possible, for example by evaluating the KDE on a subsample of the reference subset or by using fewer top subset fractions.

Table 3: **Datasets studied** (Section 4.1), with numbers of observations $n$, features $d$ and anomalies, and the anomaly ratio. Dataset numbering follows Ma et al. (2023).

| # | Dataset | $n$ | $d$ | Anomalies | Anomalies (%) |
|---|---|---|---|---|---|
| 1 | annthyroid (ODDS) | 7200 | 6 | 534 | 7.417 |
| 2 | arrhythmia (ODDS) | 452 | 274 | 66 | 14.602 |
| 3 | breastw (ODDS) | 683 | 9 | 239 | 34.993 |
| 4 | glass (ODDS) | 214 | 9 | 9 | 4.206 |
| 5 | ionosphere (ODDS) | 351 | 33 | 126 | 35.897 |
| 6 | letter (ODDS) | 1600 | 32 | 100 | 6.250 |
| 7 | lympho (ODDS) | 148 | 18 | 6 | 4.054 |
| 8 | mammography (ODDS) | 11183 | 6 | 260 | 2.325 |
| 9 | mnist (ODDS) | 7603 | 100 | 700 | 9.207 |
| 10 | musk (ODDS) | 3062 | 166 | 97 | 3.168 |
| 11 | optdigits (ODDS) | 5216 | 64 | 150 | 2.876 |
| 12 | pendigits (ODDS) | 6870 | 16 | 156 | 2.271 |
| 13 | pima (ODDS) | 768 | 8 | 268 | 34.896 |
| 14 | satellite (ODDS) | 6435 | 36 | 2036 | 31.639 |
| 15 | satimage-2 (ODDS) | 5803 | 36 | 71 | 1.224 |
| 16 | speech (ODDS) | 3686 | 400 | 61 | 1.655 |
| 17 | thyroid (ODDS) | 3772 | 6 | 93 | 2.466 |
| 18 | vertebral (ODDS) | 240 | 6 | 30 | 12.500 |
| 19 | vowels (ODDS) | 1456 | 12 | 50 | 3.434 |
| 20 | wbc (ODDS) | 378 | 30 | 21 | 5.556 |
| 21 | wine (ODDS) | 129 | 13 | 10 | 7.752 |
| 22 | Annthyroid (DAMI) | 7129 | 21 | 534 | 7.491 |
| 23 | Arrhythmia (DAMI) | 450 | 259 | 206 | 45.778 |
| 24 | Cardiotocography (DAMI) | 2114 | 21 | 466 | 22.044 |
| 25 | HeartDisease (DAMI) | 270 | 13 | 120 | 44.444 |
| 26 | InternetAds (DAMI) | 1966 | 1555 | 368 | 18.718 |
| 27 | PageBlocks (DAMI) | 5393 | 10 | 510 | 9.457 |
| 28 | Pima (DAMI) | 768 | 8 | 268 | 34.896 |
| 29 | SpamBase (DAMI) | 4207 | 57 | 1679 | 39.910 |
| 30 | Stamps (DAMI) | 340 | 9 | 31 | 9.118 |
| 31 | Wilt (DAMI) | 4819 | 5 | 257 | 5.333 |
| 32 | ALOI (DAMI) | 49534 | 27 | 1508 | 3.044 |
| 33 | Glass (DAMI) | 214 | 7 | 9 | 4.206 |
| 34 | PenDigits (DAMI) | 9868 | 16 | 20 | 0.203 |
| 35 | Shuttle (DAMI) | 1013 | 9 | 13 | 1.283 |
| 36 | Waveform (DAMI) | 3443 | 21 | 100 | 2.904 |
| 37 | WBC (DAMI) | 223 | 9 | 10 | 4.484 |
| 38 | WDBC (DAMI) | 367 | 30 | 10 | 2.725 |
| 39 | WPBC (DAMI) | 198 | 33 | 47 | 23.737 |

### A.3 Experimental details

#### A.3.1 Datasets

The datasets used in our experiments (Section 4.1) have between $129$ and $49{,}534$ observations, between $5$ and $1{,}555$ features and between $0.2$ and $45.8$ percent anomalies. They are described in Table 3. Dataset names carry a source label for the Outlier Detection DataSets (ODDS) library or the collection of Campos et al. (2016) (DAMI). The abbreviated dataset names refer to the Amsterdam Library of Object Images (ALOI) and Wisconsin breast cancer data, including the original (WBC), diagnostic (WDBC) and prognostic (WPBC) variants. The ODDS identifier wbc also refers to diagnostic breast cancer data, while mnist denotes a processed version of the Modified National Institute of Standards and Technology handwritten-digit database.

Table 4: **Candidate models.** Hyperparameter values of the eight anomaly detection algorithms, following the released code of Ma et al. (2023). Every combination of the listed hyperparameter values of an algorithm is one candidate model. The kernel labels denote polynomial (poly) and radial basis function (rbf) kernels.

| Algorithm | First hyperparameter | Second hyperparameter | Models |
|---|---|---|---|
| LOF | number of neighbors: 1, 5, 10, 15, 20, 25, 50, 60, 70, 80, 90, 100 | distance: manhattan, euclidean, minkowski | 36 |
| kNN | number of neighbors: 1, 5, 10, 15, 20, 25, 50, 60, 70, 80, 90, 100 | method: largest, mean, median | 36 |
| OCSVM | $\nu$: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 | kernel: linear, poly, rbf, sigmoid | 36 |
| COF | number of neighbors: 3, 5, 10, 15, 20, 25, 50 | – | 7 |
| ABOD | number of neighbors: 3, 5, 10, 15, 20, 25, 50 | – | 7 |
| Isolation Forest | number of trees: 10, 20, 30, 40, 50, 75, 100, 150, 200 | share of features: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 | 81 |
| HBOS | number of bins: 5, 10, 20, 30, 40, 50, 75, 100 | regularization parameter $\alpha$: 0.1, 0.2, 0.3, 0.4, 0.5 | 40 |
| LODA | number of bins: 5, 10, 15, 20, 25, 30 | number of random cuts: 10, 20, 30, 40, 50, 75, 100, 150, 200 | 54 |
| Total | – | – | 297 |

### A.3.2 Candidate models

We study eight anomaly detection algorithms, namely Angle-Based Outlier Detection (ABOD) (Kriegel et al., 2008), Connectivity-Based Outlier Factor (COF) (Tang et al., 2002), Histogram-based Outlier Score (HBOS) (Goldstein & Dengel, 2012), Isolation Forest (Liu et al., 2008), $k$-nearest neighbors (kNN) (Ramaswamy et al., 2000), Lightweight On-line Detector of Anomalies (LODA) (Pevný, 2016), Local Outlier Factor (LOF) (Breunig et al., 2000) and One-Class Support Vector Machine (OCSVM) (Schölkopf et al., 2001). Table 4 lists their hyperparameter configurations, which give the 297 candidate models, 81 of them for Isolation Forest.

### A.3.3 Selection methods and baselines

For the 30 variants, we use the per-dataset results supplied by Ma et al. to recompute the evaluation statistics, without rerunning the selection procedures. The 34 comparison methods comprise 14 stand-alone methods, 17 consensus methods and three meta-learning methods, following the categories in Table 5. HITS and SELECT provide both a selected model and an ensemble ranking that combines the anomaly scores of candidate models. We evaluate both outputs, as in Ma et al., and list the individual variants in Table 1. Following Ma et al. (2023), the cluster-quality indices evaluate two groups of anomaly scores, using the true anomaly count to set the cutoff between the highest-scores and the rest. This supplies the number of anomalies, but not their identities, to these methods.

The cluster-index rows are root-mean-square standard deviation (std), R-squared (rs), modified Hubert statistic (h), Calinski–Harabasz index (ch), silhouette index (s), I-index (i), Davies–Bouldin index (db), Xie–Beni index (xb) and Dunn index (d). The row sd uses the S_Dbw index, which measures within-cluster scattering and between-cluster density.

We run each of the four added methods with its authors' released code, adapted to our experimental setup. Where code and paper differ, we follow the code, so that we evaluate the authors' implementation instead of our own reading of the paper. When the released code leaves a step open, such as the preparation of the inputs, we follow the paper where we can. These runs adapt the methods to our experiments and are not exact reproductions of the experiments in their papers.

The three meta-learning methods learn from labeled reference datasets. We therefore train them separately for each test dataset using only the labeled results of the other 38 datasets, following the leave-one-dataset-out approach used by Zhao et al. (2021; 2022); Singh et al. (2026). With leave-one-dataset-out on the 39 datasets (Section 4.1), MetaOD learns from 38 meta-train tasks, fewer than the 61 or 80 in its original experiments (Zhao et al., 2021). We do not use the pretrained MetaOD model because it predicts over 302 models rather than the 297 candidate models of our experiments (Zhao et al., 2021). ELECT computes its internal performance measures on a small anchor set of representative candidate models, and because that set is not published, we compute them over all 297 candidate models. ELECT starts every fold with the same eight candidate models supplied in its released code (Zhao et al., 2022).

For LOTUS, we replace its model search with selection from the fixed pool of 297 candidate models (Singh et al., 2026). The reported results use ROC-AUC on the labeled reference dataset as the selection objective, following its paper (Singh et al., 2026). As a control for our AP evaluation, we also selected by AP on the same reference datasets, which lowered LOTUS's mean AP on the test datasets from $0.350$ to $0.264$.

Some of the 39 datasets derive from the same original data as another one, for example annthyroid (ODDS) and Annthyroid (DAMI) (Table 3). When one of them is the test dataset, the other remains among the labeled reference datasets. This overlap can favor the meta-learning methods, whereas SUBSELECT does not use labeled reference datasets for selection.

For SIREOS, the features are scaled to $[0, 1]$ and the anomaly scores of every candidate model are normalized with Gaussian scaling, as in the IREOS experiments (Marques et al., 2020; Campos et al., 2016). The bandwidth of the Gaussian kernel is the $0.01$ quantile of the pairwise distances, and the candidate model with the lowest index is selected (Marques et al., 2022).

The baselines RANDOM and IFOREST-R are the average AP or ROC-AUC of the corresponding candidate models per dataset, not the performance of a score ensemble (Ma et al., 2023).

### A.3.4 STATISTICAL EVALUATION

For the baseline comparisons and smallest-$q$ statistic, we use the normal approximation to calculate Wilcoxon $p$-values.

For the smallest-$q$ statistic (see Section 4.2), we use the one-sided Wilcoxon test without correction, following Ma et al. (2023). Because it tests whether candidate models at each rank outperform the selected models, its greater sensitivity in that direction can yield a larger, less favorable $q$ than a two-sided test. SUBSELECT also outperforms both baselines with a two-sided test and Holm's correction over the two baseline comparisons (Holm, 1979). The adjusted $p$-values are below $0.001$ against RANDOM on both metrics and are $0.042$ (AP) and $0.024$ (ROC-AUC) against IFOREST-R.

For the Bayesian signed-rank test (see Section 4.2), we use the prior strength of $0.5$ that Benavoli et al. (2017) choose, with the pseudo-observation at zero, and $150{,}000$ posterior samples as in their examples, with fixed random seeds. The test compares each pair on its own, so the result for SUBSELECT against a method does not change when other methods join or leave the comparison, which does not hold for post-hoc tests on mean ranks (Benavoli et al., 2017).

In the procedure of Demsar (2006), the Nemenyi test compares the mean ranks of two methods with a critical difference. At a fixed significance level, this critical difference depends only on the number of methods and datasets, grows with the number of methods, and shrinks only with the square root of the number of datasets (Demsar, 2006). The 39 datasets studied in our experiments carry limited information for 37 approaches and their 666 possible pairs. At the observed mean ranks on AP, the Nemenyi test would separate SUBSELECT from its three closest competitors only with more than $1{,}000$ datasets.

### A.3.5 RUNTIME MEASUREMENT

We measure contrast fitting and scoring for one representative Isolation Forest model on each dataset on a single core of an AMD EPYC 7713 processor, for all 50 top subset fractions. Each top subset fraction is computed twice and only the second timing is retained, so that one-time start-up costs are excluded. We report the mean over 30 repetitions.

Table 5: **Compared model selection approaches and the information they use.** The first three information columns concern the target dataset, while the last concerns labeled reference datasets. Types follow Ma et al. (2023) for stand-alone and consensus methods, with meta-learning shown separately. Rows group the variants compared in Section 5.

| | | Information used for the selection | | | |
|---|---|---|---|---|---|
| Method | Output | Scores | Features | Agreement | Labels |
| *Stand-alone* | | | | | |
| SUBSELECT (ours) [a] | selected model | yes | yes | no | no |
| Cluster-quality indices Ma et al. (2023) | selected model | yes | no | no | no |
| EM, MV (Goix et al., 2015; Goix, 2016) | selected model | yes | yes[b] | no | no |
| IREOS (Marques et al., 2020) | selected model | yes | yes | no | no |
| SIREOS (Marques et al., 2022) | selected model | yes | yes | no | no |
| *Consensus* | | | | | |
| UDR (Duan et al., 2020) | selected model | yes | no | yes | no |
| MC, MCS (Lin et al., 2020) | selected model | yes | no | yes | no |
| HITS (Kleinberg, 1999) | selected model | yes | no | yes | no |
| Ens (Rayana & Akoglu, 2016) | selected model | yes | no | yes | no |
| HITS-auth | ensemble ranking | yes | no | yes | no |
| Ens-pseudo | ensemble ranking | yes | no | yes | no |
| *Meta-learning* | | | | | |
| MetaOD (Zhao et al., 2021) | selected model | no[c] | yes[d] | no | yes |
| ELECT (Zhao et al., 2022) | selected model | yes | no | yes | yes |
| LOTUS (Singh et al., 2026) | selected model | no[c] | yes | no | yes |

**Information columns.** Scores indicates use of the candidate models' anomaly scores on the target dataset, and Features indicates use of that dataset's input features. Agreement indicates whether agreement between candidate-model outputs contributes to selection. Labels indicates use of labeled reference datasets for meta-learning.

**Method variants.** Cluster-quality indices comprise std, h, s, i, db, sd, d, rs, ch and xb. UDR, MC and MCS each use $\rho$, $\tau$ or NDCG. HITS and HITS-auth each use ranks or unified scores. Ens and Ens-pseudo each use SELECT-ALL or SELECT-LAST.

[a]For SUBSELECT, stand-alone refers to the raw contrasts. The rank-based aggregation depends on the candidate models, but does not measure agreement between model outputs.

[b]EM and MV use features to define a domain for uniform sampling.

[c]MetaOD and LOTUS do not require candidate-model scores on the target dataset, but use model performance measured on labeled reference datasets.

[d]MetaOD uses dataset-level meta-features rather than those of individual observations.

We implement SUBSELECT in Python 3.12 with NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020) and scikit-learn (Pedregosa et al., 2011). Scikit-learn provides empirical covariance with its pseudoinverse, standardization, and kernel density estimation. Scikit-learn evaluates the KDE in Equation (9) with a k-dimensional tree (k-d tree), using the default tree-evaluation settings and the Scott bandwidth defined in Appendix A.1.2. This numerical evaluation can give different values from direct summation of the kernel contributions. The released tables store mean KDE log-likelihoods, for which lower values are preferred. The released contrast values were computed with NumPy 2.4.6, SciPy 1.17.1, and scikit-learn 1.8.0. For the evaluation, SciPy 1.16.1 computes the Wilcoxon and Friedman tests and the Nemenyi critical difference, scikit-posthocs 0.17.0 (Terpilovski, 2019) computes the pairwise Nemenyi tests and draws the critical difference diagrams. Baycomp 1.0.3, the package of Benavoli et al. (2017), computes the posterior probabilities of the Bayesian signed-rank test. SIREOS, MetaOD, ELECT, and LOTUS run with the code released by their authors, at the commits listed in our repository.

## A.4 Additional results

### A.4.1 Comparison of model selection approaches on ROC-AUC

Table 6 shows the ROC-AUC results with the same columns as Table 1, as summarized in Sections 5.1 and 5.2.

Table 6: **ROC-AUC results across 39 datasets.** One-sided Wilcoxon $p$-values follow Ma et al., comparing with RANDOM (R) and IFOREST-R (IF). Lower mean rank and $q$ are better. Bold marks $p < 0.05$, posterior probabilities $\geq 0.95$, and column-best values among selection methods. SD denotes standard deviation and $q$ is zero-indexed smallest-$q$. Top counts give the numbers of datasets with performance among the best $10/20/30\%$ of the 297 candidate models. $P_>$ and $P_=$ are Bayesian posterior probabilities that SUBSELECT is better than or practically equivalent to the row approach. $W$ counts the row approach's decided wins (posterior probability $\geq 0.95$, maximum 36). (A)/(L) denote SELECT-ALL/SELECT-LAST. (U) denotes unified scores for HITS and HITS-auth.

| Method | $p$ vs. baseline $p_{\mathrm{R}}$ | $p_{\mathrm{IF}}$ | Mean ROC-AUC rank | Mean ROC-AUC value ± SD | $q$ | Top counts 10/20/30% | $P_>$ | $P_=$ | $W$ |
|---|---|---|---|---|---|---|---|---|---|
| RANDOM | – | 1.000 | 26.03 | $0.704 \pm 0.133$ | 183 | 0/0/2 | $>$ **0.99** | 0.00 | 2 |
| IFOREST-R | $<$ **0.001** | – | 17.82 | $0.763 \pm 0.166$ | 102 | 0/2/14 | **0.97** | 0.00 | 16 |
| **SUBSELECT** | $<$ **0.001** | **0.012** | **10.95** | $0.779 \pm 0.177$ | **46** | **10/21/26** | – | – | **24** |
| HITS-auth (U) | $<$ **0.001** | 0.057 | 13.47 | $0.771 \pm 0.167$ | 84 | 5/10/19 | 0.89 | 0.06 | 17 |
| HITS-auth | $<$ **0.001** | **0.018** | 13.74 | $\mathbf{0.785 \pm 0.163}$ | 77 | 3/6/19 | 0.77 | 0.02 | 18 |
| MC-NDCG | $<$ **0.001** | **0.034** | 14.17 | $0.771 \pm 0.170$ | 86 | 4/10/19 | 0.80 | 0.02 | 17 |
| UDR-$\tau$ | $<$ **0.001** | 0.186 | 14.67 | $0.768 \pm 0.180$ | 75 | 7/13/15 | 0.92 | 0.03 | 17 |
| MC-$\tau$ | $<$ **0.001** | **0.011** | 14.77 | $0.769 \pm 0.167$ | 91 | 3/8/18 | 0.949 | 0.01 | 17 |
| UDR-NDCG | $<$ **0.001** | 0.175 | 14.79 | $0.769 \pm 0.183$ | 75 | 9/11/16 | 0.92 | 0.01 | 17 |
| MC-$\rho$ | $<$ **0.001** | **0.036** | 14.92 | $0.767 \pm 0.168$ | 92 | 4/10/17 | 0.90 | 0.05 | 16 |
| UDR-$\rho$ | $<$ **0.001** | 0.279 | 15.24 | $0.763 \pm 0.180$ | 82 | 7/10/16 | **0.95** | 0.02 | 17 |
| ELECT | $<$ **0.001** | 0.303 | 15.59 | $0.760 \pm 0.168$ | 86 | 5/12/17 | **0.96** | 0.01 | 17 |
| MCS-$\tau$ | $<$ **0.001** | 0.121 | 16.01 | $0.761 \pm 0.167$ | 94 | 3/9/15 | **0.98** | 0.01 | 17 |
| Ens-pseudo (A) | $<$ **0.001** | 0.135 | 16.19 | $0.749 \pm 0.184$ | 87 | 3/8/17 | 0.91 | 0.01 | 16 |
| HITS | $<$ **0.001** | 0.148 | 16.35 | $0.762 \pm 0.169$ | 97 | 2/6/14 | **0.96** | 0.01 | 17 |
| Ens (L) | $<$ **0.001** | 0.265 | 16.38 | $0.749 \pm 0.185$ | 90 | 5/8/17 | 0.94 | 0.00 | 16 |
| Ens (A) | $<$ **0.001** | 0.230 | 16.50 | $0.749 \pm 0.183$ | 86 | 4/8/19 | 0.90 | 0.02 | 16 |
| Ens-pseudo (L) | $<$ **0.001** | 0.293 | 16.53 | $0.750 \pm 0.183$ | 91 | 5/8/15 | 0.93 | 0.01 | 16 |
| MCS-NDCG | $<$ **0.001** | 0.274 | 16.76 | $0.766 \pm 0.165$ | 94 | 2/8/18 | 0.94 | 0.02 | 16 |
| MCS-$\rho$ | $<$ **0.001** | 0.483 | 17.28 | $0.763 \pm 0.173$ | 100 | 5/5/15 | **0.95** | 0.00 | 16 |
| IREOS | **0.006** | 0.702 | 17.92 | $0.730 \pm 0.203$ | 83 | 3/12/16 | **0.99** | 0.00 | 10 |
| SIREOS | 0.243 | 0.943 | 19.12 | $0.704 \pm 0.207$ | 116 | 5/11/15 | **0.98** | 0.00 | 4 |
| rs | 0.358 | 0.980 | 20.36 | $0.690 \pm 0.206$ | 138 | 5/11/14 | $>$ **0.99** | 0.00 | 2 |
| ch | 0.358 | 0.980 | 20.36 | $0.690 \pm 0.206$ | 138 | 5/11/14 | $>$ **0.99** | 0.00 | 2 |
| xb | 0.358 | 0.980 | 20.36 | $0.690 \pm 0.206$ | 138 | 5/11/14 | $>$ **0.99** | 0.00 | 2 |
| LOTUS | 0.073 | 0.985 | 20.85 | $0.718 \pm 0.196$ | 128 | 5/6/10 | $>$ **0.99** | 0.00 | 8 |
| d | 0.110 | 0.998 | 21.13 | $0.700 \pm 0.197$ | 143 | 3/5/10 | $>$ **0.99** | 0.00 | 3 |
| EM | 0.358 | 0.992 | 21.72 | $0.682 \pm 0.216$ | 142 | 2/8/13 | $>$ **0.99** | 0.00 | 2 |
| db | 0.662 | 0.996 | 22.29 | $0.659 \pm 0.264$ | 166 | 6/12/14 | $>$ **0.99** | 0.00 | 1 |
| MV | 0.369 | 0.997 | 22.68 | $0.675 \pm 0.219$ | 147 | 2/6/10 | $>$ **0.99** | 0.00 | 2 |
| h | 0.818 | 1.000 | 23.38 | $0.670 \pm 0.181$ | 177 | 3/6/8 | $>$ **0.99** | 0.00 | 2 |
| i | 0.821 | 1.000 | 23.69 | $0.671 \pm 0.177$ | 178 | 3/6/7 | $>$ **0.99** | 0.00 | 2 |
| sd | 0.905 | 0.999 | 23.90 | $0.640 \pm 0.241$ | 187 | 4/6/9 | $>$ **0.99** | 0.00 | 1 |
| HITS (U) | 0.890 | 0.996 | 24.41 | $0.606 \pm 0.238$ | 185 | 6/8/10 | $>$ **0.99** | 0.00 | 1 |
| MetaOD | 0.859 | 1.000 | 25.83 | $0.673 \pm 0.199$ | 200 | 2/4/7 | $>$ **0.99** | 0.00 | 1 |
| s | 0.994 | 1.000 | 26.58 | $0.630 \pm 0.188$ | 213 | 2/5/7 | $>$ **0.99** | 0.00 | 1 |
| std | 1.000 | 1.000 | 30.26 | $0.528 \pm 0.187$ | 250 | 1/2/4 | $>$ **0.99** | 0.00 | 0 |

### A.4.2 STATISTICAL MEAN-RANK RESULTS

In the procedure of Demsar (2006), the Friedman test, corrected for ties as implemented in SciPy, rejects equal performance of the 37 approaches of the rank pool on AP ($F = 4.30$) and on ROC-AUC ($F = 7.83$), both with $p < 0.001$. Figures 5 and 6 show the mean ranks with the Nemenyi critical difference for all pairwise comparisons, which is $9.44$ ranks for 37 approaches on 39 datasets. It places 32 approaches within one critical difference of SUBSELECT on AP and 23 on ROC-AUC.

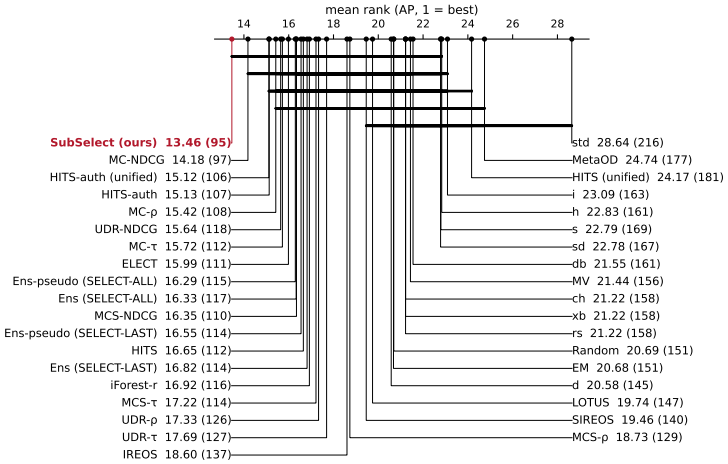


Figure 5: **Mean ranks on AP with the Nemenyi critical difference** of $9.44$ ranks at $\alpha = 0.05$ for the 37 approaches of the rank pool. The tie-corrected Friedman test rejects equal mean ranks ($F = 4.30$, $p < 0.001$). A horizontal bar joins every maximal group of methods whose mean ranks differ by less than the critical difference, so that the Nemenyi test does not reject equal performance at $\alpha = 0.05$ for any pair in the group (Demsar, 2006). Such a bar shows insufficient evidence of a difference and does not establish practical or statistical equivalence (Benavoli et al., 2017). The number in parentheses is the mean pool rank of the selected model among the 297 candidate models, for the two baselines of their expected performance.

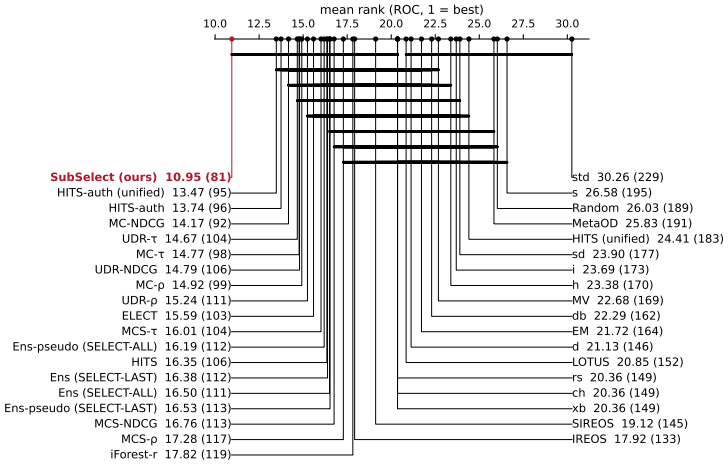


Figure 6: **Mean ranks on ROC-AUC with the Nemenyi critical difference** of $9.44$ ranks at $\alpha = 0.05$ for the 37 approaches of the rank pool. The tie-corrected Friedman test rejects equal mean ranks ($F = 7.83$, $p < 0.001$). Bars and the numbers in parentheses as in Figure 5.

### A.4.3 EXAMPLES OF COMPLEMENTING CONTRAST MEASURES

Figure 7 shows three datasets on which the two contrast measures complement each other (see Section 5.3). On Pima (DAMI), the combination selects the same model as MD, which outperforms KDE on this dataset. On Annthyroid (DAMI), where KDE selects the best model, the combination selects the second-best model by AP among the 297 candidate models. On Glass (DAMI), the combination selects a model with higher AP than both single-measure selections.

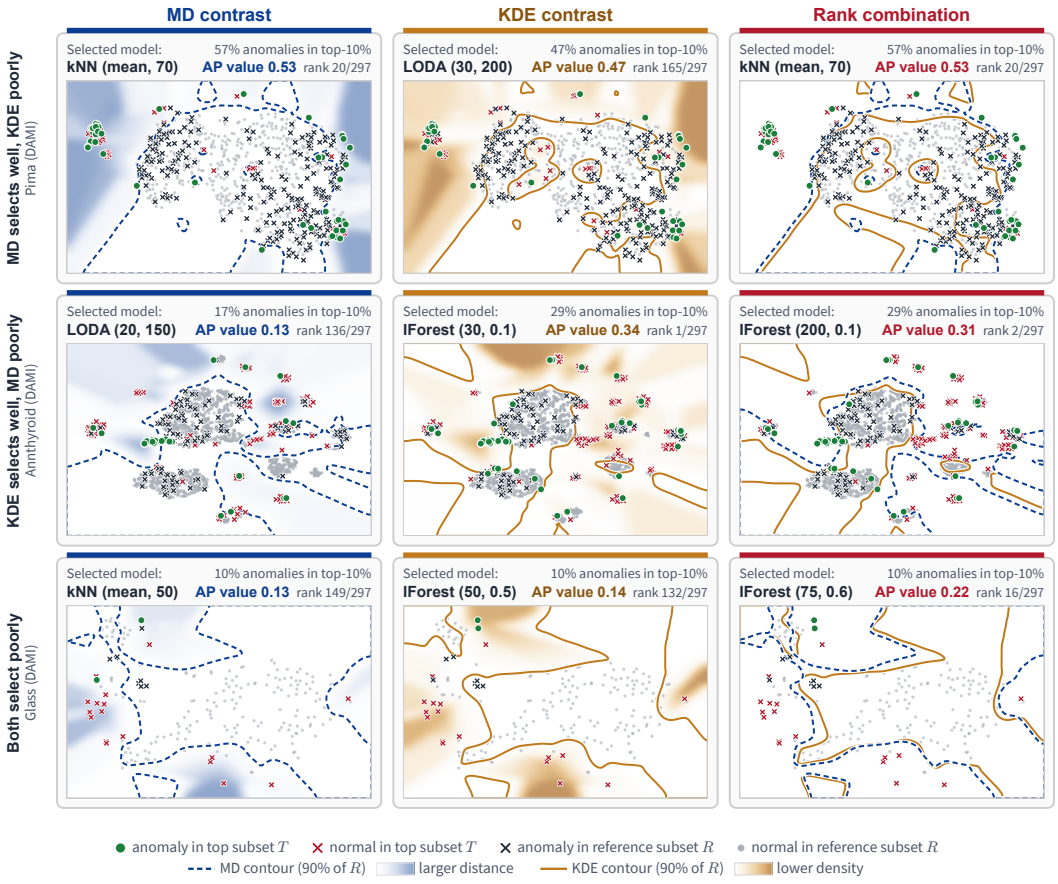


Figure 7: **Examples illustrating how MD and KDE complement each other.** Points show a t-distributed stochastic neighbor embedding (t-SNE) of at most 2,500 observations, for display only, and mark the top subset at $p = 0.1$. The contrasts are computed on all features. The model label IForest denotes Isolation Forest. Lines enclose 90% of the selected model's reference subset in all features, drawn approximately in the t-SNE view, by MD (dashed) and KDE density (solid). Shading darkens with larger distance in the first column and with lower density in the second. We chose the datasets after evaluation to illustrate the behavior of the contrast measures. Section 5.3 gives the results aggregated over all datasets. The rows show, from top to bottom, good selection by MD and poor selection by KDE, the reverse pattern, and poor selection by both measures. The columns show the model selected by the MD contrast, by the KDE contrast, and by their combination, each with all 50 top subset fractions, with its AP and rank among the 297 candidate models. In the last row, the combination of contrast measures selects a model with higher AP than either measure selects alone.

### A.4.4 ABLATION AND SENSITIVITY RESULTS

**Ablation regarding contrast measures.** To examine whether the combination of contrast measures improves on either measure alone at the dataset level, Figures 8 and 9 show the per-dataset paired differences between SUBSELECT and each single-measure variant on AP and ROC-AUC. Table 7 reports the corresponding aggregated statistics, including Bayesian posterior probabilities.

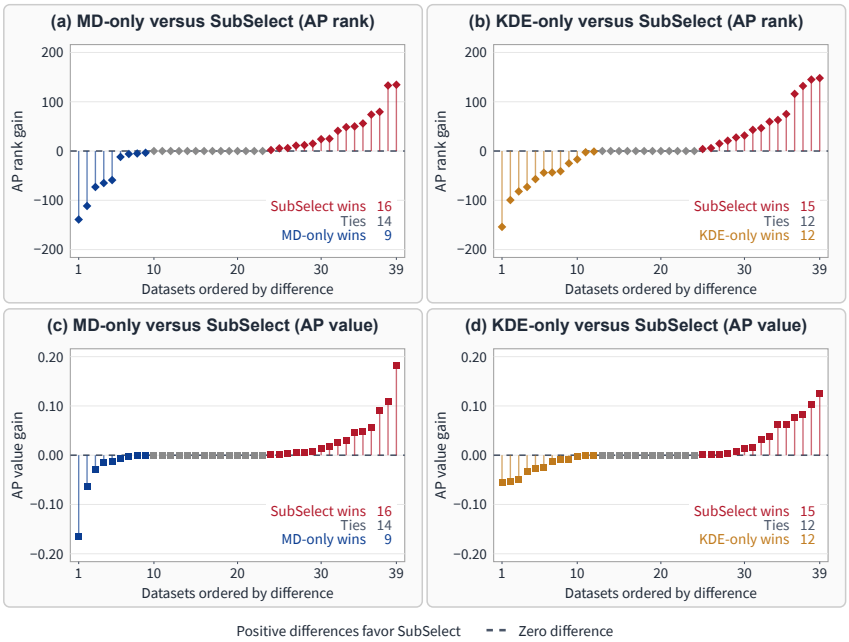


Figure 8: **Comparison of SUBSELECT with its single-measure variants on AP.** Each panel shows paired differences across all 39 datasets, using the same candidate-model pool and all 50 fractions. **(a, b)** Rank gains among the 297 candidate models, calculated as single-measure minus SUBSELECT, with average ranks for ties. **(c, d)** AP value gains, calculated as SUBSELECT minus single-measure. Positive differences favor SUBSELECT, with diamonds for ranks and squares for values. Datasets are ordered separately in each panel, so horizontal positions need not refer to the same dataset. Counts summarize strict wins and exact ties, not statistical-test decisions.

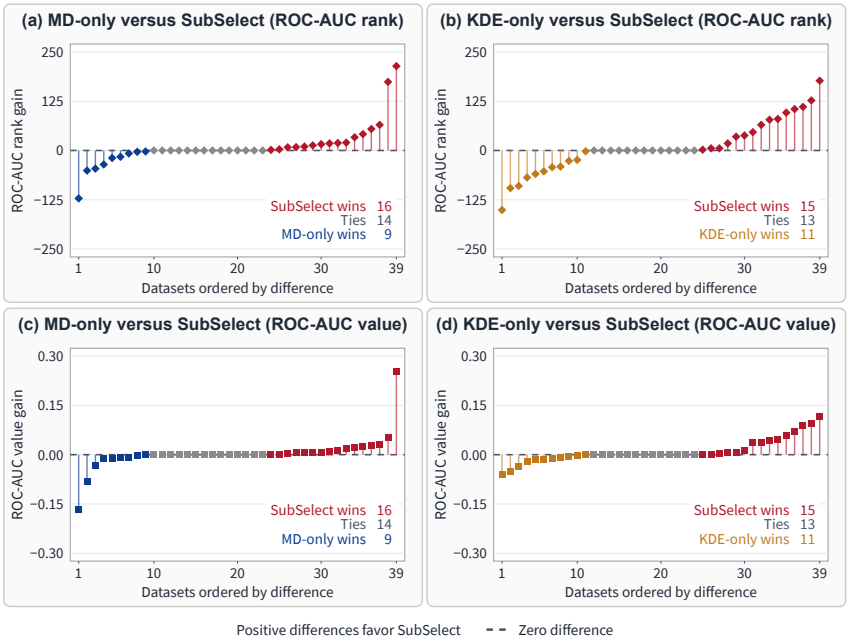


Figure 9: **Comparison of SUBSELECT with its single-measure variants on ROC-AUC.** Panels, rank definition, symbols and difference directions follow Figure 8, with ROC-AUC instead of AP. Each panel orders the 39 datasets separately, so horizontal positions need not match those in other panels or in Figure 8.

Table 7: **Extended contrast-measure ablation on AP and ROC-AUC.** The AP panel extends Table 2, and the ROC-AUC panel reports the corresponding secondary evaluation. Baseline tests, mean ranks and $q$ follow Table 2, with $p_{\mathrm{R}}$ and $p_{\mathrm{IF}}$ comparing against RANDOM and IFOREST-R. SD denotes standard deviation, and $P_{>}$ and $P_{=}$ are Bayesian posterior probabilities that SUBSELECT is better than or practically equivalent to the row variant. Bold marks $p < 0.05$, posterior probabilities $\geq 0.95$ and the best mean rank, mean value and $q$ within each panel.

**(a) AP**

| Measures | $p_{\mathrm{R}}$ | $p_{\mathrm{IF}}$ | Mean AP rank | Mean AP value ± SD | $q$ | $P_{>}$ | $P_{=}$ |
|---|---|---|---|---|---|---|---|
| **SUBSELECT (both)** | **< 0.001** | **0.021** | **13.46** | **0.424 ± 0.293** | **53** | – | – |
| MD-only | **< 0.001** | 0.079 | 13.90 | 0.415 ± 0.297 | 68 | 0.143 | 0.855 |
| KDE-only | **< 0.001** | 0.086 | 14.33 | 0.415 ± 0.291 | 65 | 0.261 | 0.709 |

**(b) ROC-AUC**

| Measures | $p_{\mathrm{R}}$ | $p_{\mathrm{IF}}$ | Mean ROC-AUC rank | Mean ROC-AUC value ± SD | $q$ | $P_{>}$ | $P_{=}$ |
|---|---|---|---|---|---|---|---|
| **SUBSELECT (both)** | **< 0.001** | **0.012** | **10.95** | **0.779 ± 0.177** | **46** | – | – |
| MD-only | **< 0.001** | **0.027** | 12.03 | 0.774 ± 0.182 | 59 | 0.027 | **0.971** |
| KDE-only | **< 0.001** | 0.051 | 12.24 | 0.769 ± 0.186 | 55 | 0.305 | 0.691 |

**Ablation regarding reference fitting.** To assess whether fitting the contrast measures on the reference subset rather than all observations matters, we fit each contrast measure either on its candidate-dependent reference subset $R$ or on all observations $X$, while retaining both measures, the 297 candidate models, and all 50 fractions. Fitting both measures on $R$ gives the best mean rank and smallest-$q$ among the four combinations on both metrics (Table 8).

Table 8: **Reference-fitting ablation on AP and ROC-AUC.** All rows use both measures and all 50 fractions, varying only whether each measure is fitted on the reference subset $R$ or all observations $X$. The $R, R$ row is the original SUBSELECT. Tests and mean-rank comparisons follow Table 2, with bold marking $p < 0.05$ and the best mean rank and $q$ within each metric.

**(a) AP**

| Fitted on MD | Fitted on KDE | $p$ vs. RANDOM | $p$ vs. IFOREST-R | Mean AP rank | $q$ |
|---|---|---|---|---|---|
| $R$ | $R$ | **< 0.001** | **0.021** | **13.46** | **53** |
| $X$ | $R$ | **0.001** | 0.338 | 14.64 | 67 |
| $R$ | $X$ | **0.002** | 0.795 | 17.12 | 78 |
| $X$ | $X$ | 0.379 | 0.969 | 20.08 | 119 |

**(b) ROC-AUC**

| Fitted on MD | Fitted on KDE | $p$ vs. RANDOM | $p$ vs. IFOREST-R | Mean ROC-AUC rank | $q$ |
|---|---|---|---|---|---|
| $R$ | $R$ | **< 0.001** | **0.012** | **10.95** | **46** |
| $X$ | $R$ | **< 0.001** | **0.027** | 12.79 | 65 |
| $R$ | $X$ | **< 0.001** | 0.086 | 14.33 | 61 |
| $X$ | $X$ | **< 0.001** | 0.712 | 18.69 | 85 |

**Ablation regarding top subset fractions and aggregation.** To test the robustness of the aggregation grid on ROC-AUC, Figure 10 repeats the single-fraction and grid comparisons of Figure 4 for ROC-AUC. The range checks retain every integer percentage within the stated interval, while the three spaced grids retain $2, 4, \ldots, 50\%$, $5, 10, \ldots, 50\%$ or $10, 20, \ldots, 50\%$. We apply SUBSELECT to the stored contrasts using each grid, with its other settings unchanged. These retrospective comparisons assess the original recommendation rather than choosing a replacement using the evaluation labels.

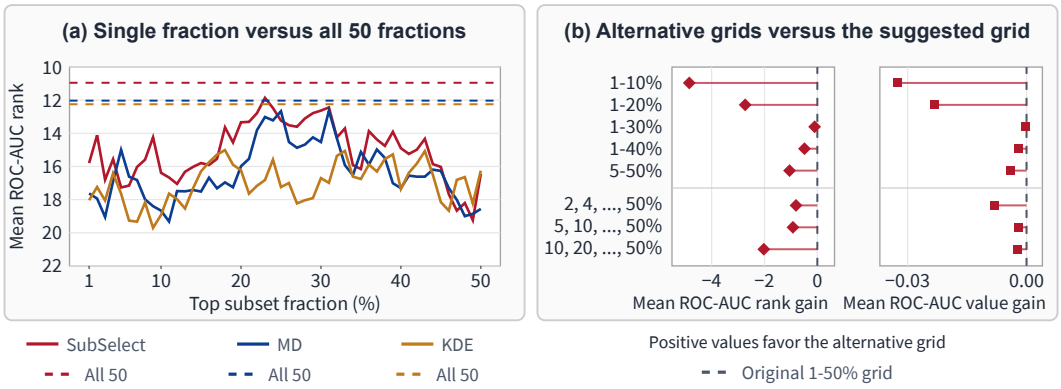


Figure 10: **Effect of the top subset fractions on ROC-AUC.** Panels, grid labels, and dashed lines follow Figure 4, with ROC-AUC instead of AP. Ranks range from 1 (best) to 37 (worst) and lower ranks are plotted higher, with only ranks 10 to 22 shown. Aggregation over all 50 fractions achieves the best mean ROC-AUC rank among the single-fraction choices for each selection approach.

**Sensitivity to the candidate-model pool.** To assess whether SUBSELECT's performance depends on the size and composition of the candidate-model pool, we vary the pool in three ways. (1) For each dataset, we sample 200 pools at each size 10, 30, 75 and 150, uniformly without replacement from the 297 candidates. (2) A separate balanced design samples seven configurations without replacement from each of the eight algorithms, giving 56 candidates and 200 pools per dataset. (3) We also remove each algorithm in turn, without sampling, to assess whether performance depends on that algorithm being available. Pool construction uses no labels, and all candidate ranks are recomputed within each pool before combining measures and aggregating fractions. The stored contrast values and 1% through 50% grid remain unchanged.

Removing all Isolation Forest candidates gives a mean AP value of 0.426, compared with 0.424 for the original pool. For the balanced pools, the mean AP value is 0.415, exceeding RANDOM and IFOREST-R within those same pools by 0.097 and 0.016. For each pool, RANDOM averages the evaluation values of its candidate models, and IFOREST-R averages those of its Isolation Forest candidates. The latter comparison uses only sampled pools containing at least one Isolation Forest candidate and is not defined when that algorithm is removed entirely. Results are averaged over sampled pools within each dataset, then over the 39 datasets, so repeated pools do not count as additional datasets. These comparisons concern SUBSELECT and the baselines within the modified pools, not the other selection methods, whose results remain based on the full 297-model pool.

Table 9: **Sensitivity of SUBSELECT regarding the candidate models on (a) AP and (b) ROC-AUC.** Ranks place the selected model among the original 297 candidate models, regardless of the tested pool, with average ranks for ties. These ranks and metric values do not depend on other model selection methods. Rank gains are baseline minus SUBSELECT, and value gains are SUBSELECT minus baseline, so positive gains favor SUBSELECT. For each tested pool, RANDOM averages the ranks or values of all its candidate models, whereas IFOREST-R averages those of its Isolation Forest models. Results are averaged over sampled pools for each dataset and then over the 39 datasets. A dash denotes a pool without Isolation Forest candidates.

**(a) AP**

| | AP rank among 297 models | | | AP value | | |
|---|---|---|---|---|---|---|
| | | Rank gain over | | | Value gain over | |
| Pool | Mean | RANDOM | IFOREST-R | Mean | RANDOM | IFOREST-R |
| Original 297 | 95.44 | 53.56 | 19.14 | 0.424 | 0.082 | 0.025 |
| Sampled 10 | 111.93 | 37.39 | 3.19 | 0.404 | 0.063 | 0.006 |
| Sampled 30 | 107.05 | 41.69 | 7.52 | 0.412 | 0.070 | 0.013 |
| Sampled 75 | 102.46 | 46.61 | 12.17 | 0.417 | 0.075 | 0.018 |
| Sampled 150 | 98.95 | 49.99 | 15.56 | 0.420 | 0.078 | 0.021 |
| Balanced 56 | 106.82 | 49.42 | 7.55 | 0.415 | 0.097 | 0.016 |
| Without ABOD | 98.08 | 50.86 | 16.50 | 0.420 | 0.077 | 0.021 |
| Without COF | 97.03 | 51.31 | 17.55 | 0.423 | 0.078 | 0.024 |
| Without HBOS | 95.59 | 54.97 | 18.98 | 0.425 | 0.088 | 0.026 |
| Without Isolation Forest | 101.51 | 60.40 | – | 0.426 | 0.105 | – |
| Without LODA | 95.92 | 48.78 | 18.65 | 0.423 | 0.081 | 0.024 |
| Without LOF | 93.83 | 52.84 | 20.74 | 0.424 | 0.073 | 0.026 |
| Without OCSVM | 96.92 | 44.58 | 17.65 | 0.424 | 0.068 | 0.025 |
| Without kNN | 95.22 | 56.21 | 19.36 | 0.416 | 0.077 | 0.017 |

**(b) ROC-AUC**

| | ROC-AUC rank among 297 models | | | ROC-AUC value | | |
|---|---|---|---|---|---|---|
| | | Rank gain over | | | Value gain over | |
| Pool | Mean | RANDOM | IFOREST-R | Mean | RANDOM | IFOREST-R |
| Original 297 | 81.09 | 67.91 | 30.24 | 0.779 | 0.074 | 0.016 |
| Sampled 10 | 100.64 | 48.87 | 11.80 | 0.767 | 0.063 | 0.004 |
| Sampled 30 | 93.61 | 55.04 | 17.60 | 0.771 | 0.066 | 0.008 |
| Sampled 75 | 89.14 | 59.98 | 22.32 | 0.772 | 0.068 | 0.009 |
| Sampled 150 | 85.43 | 63.51 | 25.90 | 0.774 | 0.070 | 0.011 |
| Balanced 56 | 90.80 | 65.79 | 20.34 | 0.772 | 0.080 | 0.009 |
| Without ABOD | 83.12 | 65.92 | 28.21 | 0.777 | 0.074 | 0.014 |
| Without COF | 82.65 | 65.51 | 28.67 | 0.778 | 0.072 | 0.015 |
| Without HBOS | 80.87 | 69.27 | 30.45 | 0.780 | 0.080 | 0.017 |
| Without Isolation Forest | 84.76 | 78.37 | – | 0.773 | 0.091 | – |
| Without LODA | 83.19 | 59.83 | 28.13 | 0.783 | 0.075 | 0.020 |
| Without LOF | 86.41 | 61.29 | 24.92 | 0.778 | 0.072 | 0.015 |
| Without OCSVM | 83.71 | 55.78 | 27.62 | 0.776 | 0.048 | 0.013 |
| Without kNN | 84.49 | 68.97 | 26.84 | 0.774 | 0.079 | 0.011 |

**Sensitivity regarding dataset collection.** We summarize the existing results separately for the 21 ODDS and 18 DAMI datasets in Table 10, without rerunning the methods. The comparisons use both baselines, MC-NDCG as the overall runner-up by mean AP value, and HITS-auth as the approach with the highest overall mean ROC-AUC value. Mean differences, computed as SUBSELECT minus the comparison approach, are positive against both baselines in each collection on both metrics, with smaller gains on DAMI. The ROC-AUC mean differences from MC-NDCG and HITS-auth are positive on ODDS but negative on DAMI.

Table 10: **Performance differences by dataset collection on (a) AP and (b) ROC-AUC.** Mean differences are SUBSELECT minus the comparison approach, with equal weight per dataset in each collection. Positive differences favor SUBSELECT, and W/T/L counts datasets with higher/equal/lower metric values for SUBSELECT, not statistical decisions.

**(a) AP**

| | All (39) | | ODDS (21) | | DAMI (18) | |
|---|---|---|---|---|---|---|
| Compared with | Mean diff. | W/T/L | Mean diff. | W/T/L | Mean diff. | W/T/L |
| RANDOM | +0.082 | 29/0/10 | +0.111 | 17/0/4 | +0.048 | 12/0/6 |
| IFOREST-R | +0.025 | 25/0/14 | +0.037 | 14/0/7 | +0.012 | 11/0/7 |
| MC-NDCG | +0.020 | 18/2/19 | +0.035 | 10/2/9 | +0.002 | 8/0/10 |
| HITS-AUTH | +0.023 | 27/0/12 | +0.036 | 15/0/6 | +0.008 | 12/0/6 |

**(b) ROC-AUC**

| | All (39) | | ODDS (21) | | DAMI (18) | |
|---|---|---|---|---|---|---|
| Compared with | Mean diff. | W/T/L | Mean diff. | W/T/L | Mean diff. | W/T/L |
| RANDOM | +0.074 | 33/0/6 | +0.095 | 19/0/2 | +0.050 | 14/0/4 |
| IFOREST-R | +0.016 | 30/0/9 | +0.025 | 17/0/4 | +0.005 | 13/0/5 |
| MC-NDCG | +0.007 | 24/2/13 | +0.016 | 13/2/6 | −0.002 | 11/0/7 |
| HITS-AUTH | −0.007 | 27/0/12 | +0.005 | 17/0/4 | −0.020 | 10/0/8 |

### A.4.5 RUNTIME

We measure the time needed to fit the contrast measures and score the top subsets across all 50 top subset fractions for one representative candidate model per dataset. Appendix A.3.5 describes the runtime evaluation. Figure 11 shows the runtime per dataset. The array sizes depend on $n$, $d$ and the grid of top subset fractions, but measured runtimes can also depend on how the model partitions the observations. The measurements exclude model training, score sorting and rank aggregation across the model pool. Summed across all 39 datasets and all 50 fractions, the single-core measurements total 2.1 minutes for MD and 22.3 minutes for KDE for the representative candidate model. ALOI (DAMI), which has the most observations (49,534), accounts for 13.6 minutes of the KDE total, while InternetAds (DAMI), which has the most features (1,555), accounts for 1.9 minutes of the MD total.

Figure 12 shows the measured mean contrast-computation time per top subset fraction across the 39 datasets. Per top subset fraction, the KDE time rises with the fraction and the MD time only slightly, as expected from the operation counts in Appendix A.2. Different datasets, candidate models, top subset fractions, and contrast measures can be processed independently, including MD and KDE for the same candidate model (Appendix A.2). With enough workers to run all dataset and measure pairs simultaneously, the ideal elapsed time for the measured candidate model is the longest individual task, not the summed time reported above. Actual elapsed time also depends on resource contention and scheduling overhead, and is not measured here.

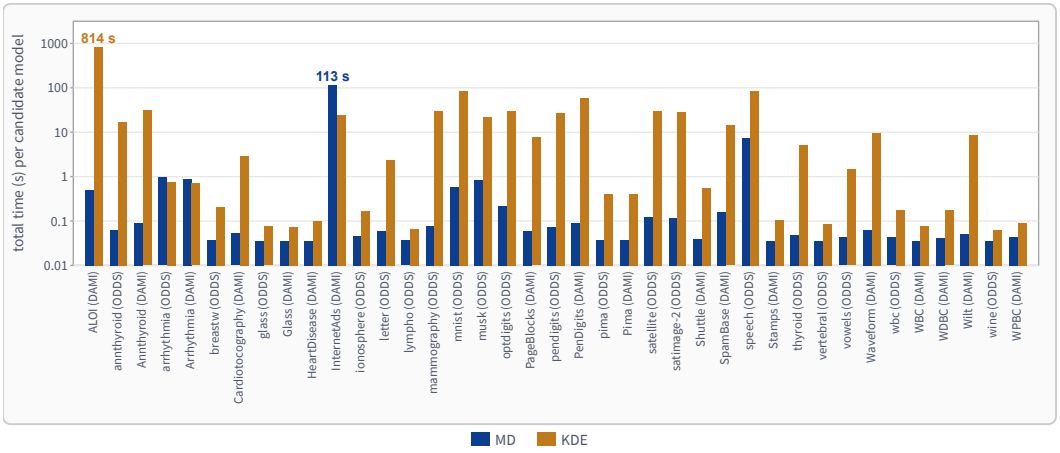


Figure 11: **Measured contrast-computation time for one representative candidate model** per dataset and contrast measure on a log scale. Each bar is the mean of 30 repetitions of the measurement. The measurement variation is too small to show on this scale (the interquartile range is below 2 percent of the median for every dataset and measure). The computation is measured on one core (see Appendix A.3.5).

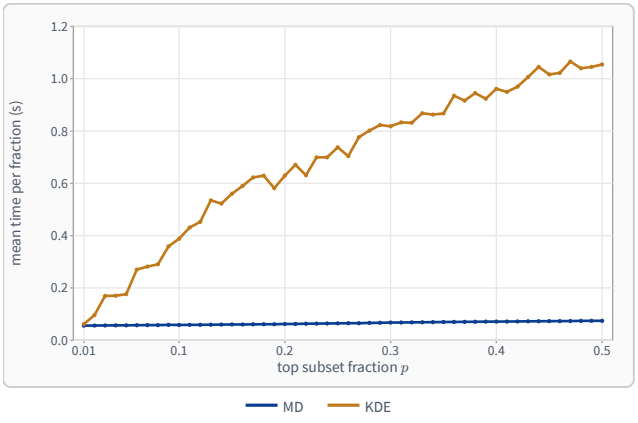


Figure 12: **Mean contrast-computation time per top subset fraction** over the 39 datasets, measured as for Figure 11. ALOI (DAMI) accounts for 61 percent of the KDE time and InternetAds (DAMI) for 90 percent of the MD time. The KDE time rises with the top subset fraction and the MD time only slightly, as the operation counts of Appendix A.2 lead us to expect.