MESA: SCORING THE VOLUME AND STABILITY OF COMPARATIVE CLAIMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid pace of AI/ML research produces a steady stream of methods claiming state-of-the-art performance, each backed by a leading benchmark score. However, reported scores offer practitioners limited guidance on adoption, as they are obtained under one configuration of the many ad-hoc choices an experiment requires. This leaves two questions unresolved: (1) the amount of the plausible configuration space in which a method outperforms others, and (2) whether that lead survives a small change to the configuration. Existing evaluation tactics fall short of answering these questions, as they either evaluate methods under fixed configurations or analyze each method in isolation. We introduce Meta-Experiment Stability Analysis (MESA), which answers both. MESA evaluates a set of candidate methods across plausible experimental configurations to determine the regions where each method is in the top-k after accounting for the noise between repeated runs. MESA scores each method's top-k region by (1) its volume, the plausibility-weighted fraction of the space it covers, and (2) its stability, the typical distance from a point in the region to its boundary. The same meta-experiment can then be reused to evaluate how well benchmark datasets separate the methods. We demonstrate MESA through two case studies from different areas of AI/ML. In both, no method leads on a majority of the configuration space on any dataset. These results suggest that a leading benchmark score is weak evidence that a method will lead under a practitioner's own choices, and that MESA's volume and stability scores give evaluation a more informative target[1].

1 INTRODUCTION

The pace of research in machine learning (ML) and artificial intelligence (AI) is rapidly accelerating. As the literature swells, many papers propose methods for the same task and make overlapping or contradictory claims about state-of-the-art (SOTA) performance. Autonomous research systems may exacerbate this problem by producing minor methodological variations that claim modest performance increases (Xu & Peng, 2025; Lu et al., 2026). The growing amount of AI/ML research is compounded by the evolving reproducibility crisis within the field (Semmelrock et al., 2025), which has lowered trust in published performance claims. Even when claims reproduce, they may break under equally reasonable choices; method performance has been shown to change with implementation (Henderson et al., 2018), hyperparameter optimization procedure (Cooper et al., 2021), or even random seed (Picard, 2021). This introduces a challenge for practitioners deciding which method to deploy and researchers choosing prior work to build on. It is not enough for a method to win under its authors' configuration, as the stated gains should also hold under plausible alternative settings.

Determining whether a comparative performance claim holds more generally requires two distinct measurements: how much of the plausible configuration space supports it, and the robustness of that support to small perturbations. The experiments used to obtain these measurements also reveal the efficacy of the benchmark datasets in separating the candidate methods. Existing evaluation frameworks fall short of achieving these goals. Traditional benchmarking relies on static environments and measures performance with respect to fixed experimental choices (Hardt, 2026), while more transparent reporting practices record the variability of a result but average away the configuration. Multiverse analysis describes sensitivity to preprocessing and modeling choices (Steegen et al.,

[1]Code is available at `https://anonymous.4open.science/r/mesa-81D7`.

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.

2016), and has been extended to comparative claims (Bell et al., 2022), but does not quantitatively measure the performance surface or account for run-to-run noise. Consequently, a practitioner has no way to discern whether a reported advantage holds on the settings they care about, or how robust that advantage is to plausible changes.

To address this gap, we introduce **M**eta-**E**xperiment **S**tability **A**nalysis (MESA), a principled approach to evaluating the stability of SOTA claims through regimented meta-experimentation. MESA evaluates a set of candidate methods using repeated runs across a grid of plausible experimental configurations, leading to the following contributions.

1. **A noise-aware definition of the SOTA region.** We say a method is top-k at a configuration if its margin over its k-th best competitor exceeds the noise between repeated runs. The subset of configurations where a method is top-k forms its top-k SOTA region (Section 3.1).
2. **Two measures of the SOTA region.** Volume represents the plausibility-weighted proportion of configuration space it covers, and stability represents the average distance from a point in the region to the closest boundary (Sections 3.2 and 3.3).
3. **Two scores of benchmark efficacy.** Resolution volume is the share of configuration space where the benchmark separates the top-k methods, and effective frontier count is the number of equally-sized top-k regions which would yield the same fragmentation (Section 4).
4. **Two case studies,** one in conformal prediction and the other in anomaly detection. We find that although no method leads on more than 47% of the configuration space on any dataset, every method leads somewhere, and volume and stability often disagree. Furthermore, the datasets differ greatly in their ability to separate the methods (Section 5).

2 Related Work

Fragility of comparative performance. Advantages held by AI/ML methods can diminish or disappear due to ancillary decisions. Comparative performance shifts with choice of implementation (Henderson et al., 2018), hyperparameter optimization procedure (Cooper et al., 2021), benchmark (Dehghani et al., 2021), and even random seed (Picard, 2021). The proposed solutions typically consist of reporting practices such as score distributions (Reimers & Gurevych, 2017), expected validation performance given tuning budget (Dodge et al., 2019), and probability of outperforming (Bouthillier et al., 2021). These approaches marginalize over variation, whereas MESA retains the configuration to identify the region where a method leads and measure its typical distance to failure.

Statistical method comparison. Statistical tests have been proposed to determine if one method outperforms another, using datasets, runs, or tasks as the unit of replication. This body of work includes rank-based tests across datasets (Demšar, 2006), Bayesian testing with a region of practical equivalence (Benavoli et al., 2017), distribution-free tests for almost stochastic dominance (Dror et al., 2019), and interval estimates using a handful of runs (Agarwal et al., 2021). Performance profiles (Dolan & Moré, 2002) evaluate methods using the cumulative distribution function of their performance ratio relative to the problem's best method, and are closest to our volume score. However, this approach aggregates over a fixed set of problems rather than the space of experimental choices, treating each problem as an observation instead of testing against the noise between runs.

Perturbation-based analysis. Empirical results depend on ad-hoc choices with defensible alternatives, motivating work which measures how conclusions change across these decisions. Multiverse analysis (Steegen et al., 2016) and specification curve analysis (Simonsohn et al., 2020) report results across all defensible choices, while the PCS framework (Yu & Kumbier, 2020; Yu & Barter, 2024; Rewolinski & Yu, 2026) unifies these ideas with classical sensitivity analysis (Frey & Patil, 2002) to establish that stability under perturbation is a prerequisite for trustworthy conclusions. Closest to our work is Bell et al. (2022), who apply multiverse analysis to ML, including comparative performance of optimizers. Specifically, they model the configuration space with a Gaussian process and use Bayesian experimental design for exploration, locating each optimizer's winning region visually rather than scoring it, and leaving run-to-run noise for future work. MESA complements this by scoring each method's winning region, after testing the performance differences against noise.

Multicriteria decision analysis. Our proposed scores build on predecessors from multicriteria decision analysis. In particular, SMAA (Lahdelma et al., 1998) explores the space of criteria weights

and computes an 'acceptability index', measuring the share of the space in which an alternative is preferable. This is then extended to rank-based settings in SMAA-2 (Lahdelma & Salminen, 2001). Our volume differs by integrating over experimental configurations instead of criteria weights, and separating the configurations from the baselines which set the noise levels. Wendell (1985) advocates for reporting the largest changes to a linear program's coefficients under which the solution remains optimal, a sentiment which our stability score builds on. Specifically, MESA's stability score averages boundary distance across the region rather than reporting it for a specific point.

Benchmark evaluation. Researchers have begun to evaluate benchmarks in addition to methods. Item-response theory focuses on the ability of individual test points to discriminate between models (Martínez-Plumed et al., 2019; Rodriguez et al., 2021). Zhang & Hardt (2024) evaluate multi-task benchmarks through the lens of social choice theory, demonstrating a trade-off between the diversity of tasks and sensitivity to irrelevant changes in the tasks. Our dataset scores answer a different question by measuring the proportion of configuration space on which a fixed dataset identifies a winning method after accounting for noise, as well as the effective number of distinct winning sets it supports.

3 The MESA Framework

3.1 Preliminaries

Consider some set of *candidate methods* $\mathbb{C} = \{h_1, \ldots, h_n\}$ designed to perform the same task. The performance of each candidate depends on unoptimized ad hoc decisions, which we refer to as *experimental parameters* and collect in the vector $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)$. These parameters encompass hyperparameter settings, architectural choices, experimental protocols, and more.

In addition, performance relies on *baseline* decisions that control the irreducible noise in the system. Each choice for a baseline decision is equally plausible; for example, any nonnegative integer is an equally valid random seed. We perform $r = 1, \ldots, R$ runs at each configuration, with each candidate method evaluated under the baselines $\mathbf{b}^{(r)} = \left(b_1^{(r)}, \ldots, b_\ell^{(r)}\right)$. Since baselines are shared across methods for each run, the comparison is paired, meaning that the dependence between method results should be accounted for when comparing performance.

We conduct a *meta-experiment* across the domain of valid experimental parameters $\boldsymbol{\Psi}$. At each $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, we evaluate whether some method of interest $h^* \in \mathbb{C}$ ranks among the top k candidates. To do so, we rank all competing methods $c_i \in \mathbb{C} \setminus \{h^*\}$ by their empirical mean performance $\bar{f}\left(c_i(\boldsymbol{\psi})\right)$ from highest to lowest. Let $c_{(k)}$ represent the k-th ranked competitor at configuration $\boldsymbol{\psi}$. The top-k margin at run r measures h^*'s performance relative to this k-th best competitor:

$$g_k\left(\boldsymbol{\psi}, \mathbf{b}^{(r)}\right) = f\left(h^*(\boldsymbol{\psi}), \mathbf{b}^{(r)}\right) - f\left(c_{(k)}(\boldsymbol{\psi}), \mathbf{b}^{(r)}\right), \tag{1}$$

where the sample mean is $\bar{g}_k(\boldsymbol{\psi}) = \frac{1}{R}\sum_{r=1}^{R} g_k\left(\boldsymbol{\psi}, \mathbf{b}^{(r)}\right)$ and $\hat{\varepsilon}_k^2(\boldsymbol{\psi})$ denotes the sample variance. We say that h^* is top-k state-of-the-art (SOTA) at $\boldsymbol{\psi}$ if the mean margin exceeds a noise threshold:

$$\mathcal{S}_{h^*}^{(k)} = \left\{ \boldsymbol{\psi} \in \boldsymbol{\Psi} : \bar{g}_k(\boldsymbol{\psi}) > t_{\gamma,R-1} \cdot \sqrt{\frac{\hat{\varepsilon}_k^2(\boldsymbol{\psi})}{R}} \right\}, \tag{2}$$

where $t_{\gamma,R-1}$ is the one-sided t-value at the γ-quantile with $R-1$ degrees of freedom. We note that since a method justifies adoption only when it clearly leads, a configuration is excluded from $\mathcal{S}_{h^*}^{(k)}$ if either h^* is outperformed or its margin cannot be distinguished from noise. In Appendix A.1, we give a diagnostic for distinguishing weak method performance from an underpowered test and show that the low volumes in our case studies are not driven by the number of runs.

Choosing a test. The threshold in Equation 2 requires a decision rule for determining if the margin exceeds the run-to-run noise. The appropriate rule should follow the experimental design and setup; for example, we use a paired t-test to determine the value of $t_{\gamma,R-1}$, as each method shares the same baselines for each run. Using a paired test allows us to make the weaker assumption of approximate normality of the differences $g_k^{(r)}$, as opposed to approximate normality of each method's scores. If methods are instead evaluated under independently drawn baselines, an unpaired test (e.g., Welch's) is more appropriate. However, any valid decision rule is able to be applied here, such as tests incorporating a region of practical equivalence (Benavoli et al., 2017) or relations of almost stochastic dominance (Dror et al., 2019).

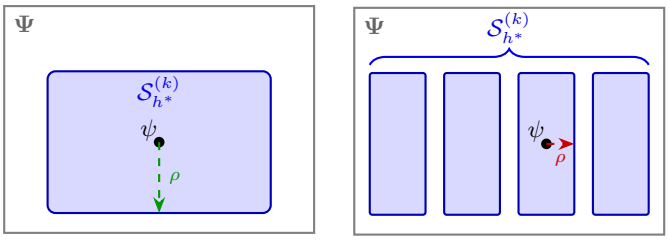


Figure 1: Both methods have top-k SOTA regions with equivalent volumes. However, local instabilities fragment the region on the right. This shrinks the distance to failure $\rho(\psi)$, yielding a lower stability score.

3.2 Volume of the SOTA Region

We assign the joint probability mass function $\pi(\boldsymbol{\psi})$ over the space of experimental parameters $\boldsymbol{\Psi}$ to reflect that some configurations may be more plausible than others. To measure the scope of a top-k claim, we calculate the *volume* given by:

$$\text{Volume}\left(\mathcal{S}_{h^*}^{(k)}\right) = \sum_{\boldsymbol{\psi} \in \mathcal{S}_{h^*}^{(k)}} \pi(\boldsymbol{\psi}). \tag{3}$$

The volume represents the total probability mass for which h^* finishes among the top k candidates. Thus, a small volume indicates limited coverage, while a high volume implies that h^* consistently finishes in the top k candidates.

3.3 Local Stability of the SOTA Region

While volume captures the total size of the top-k region, it does not tell us whether performance is robust to minor perturbations to the experimental parameters—a SOTA region may have high volume but contain narrow branches or gaps, as seen in Figure 1. To measure stability, we partition the experimental parameters into ordered parameters $\mathbf{o}$ (e.g., learning rate) and unordered categorical parameters $\mathbf{u}$ (e.g., choice of optimizer) such that $\boldsymbol{\psi} = (\mathbf{o}, \mathbf{u})$. For the ordered parameters, we define the standardized L_∞ distance

$$d_\pi(\mathbf{o}, \mathbf{o}') = \max_j \frac{|o_j - o_j'|}{\sigma_j}, \tag{4}$$

where σ_j is the π-weighted standard deviation of the jth ordered parameter. It follows that for any top-k configuration $(\mathbf{o}, \mathbf{u}) \in \mathcal{S}_{h^*}^{(k)}$, its distance to failure is the minimum distance to a non-top-k configuration while holding categorical choices fixed:

$$\rho(\mathbf{o}, \mathbf{u}) = \inf_{\mathbf{o}' : (\mathbf{o}', \mathbf{u}) \in \boldsymbol{\Psi} \setminus \mathcal{S}_{h^*}^{(k)}} d_\pi(\mathbf{o}, \mathbf{o}'). \tag{5}$$

Holding categorical $\mathbf{u}$ fixed ensures stability is not impacted by arbitrary categorical jumps. If the categories have a natural distance metric (e.g., Jaccard distance on feature sets), $d_\pi(\mathbf{o}, \mathbf{o}')$ can be straightforwardly adapted to include $\mathbf{u}$. We aggregate these distances into the local stability curve

$$\text{LocalStability}(r) = \frac{\sum_{(\mathbf{o},\mathbf{u}) \in \mathcal{S}_{h^*}^{(k)}} \pi(\mathbf{o}, \mathbf{u}) \mathbb{1}\{\rho(\mathbf{o}, \mathbf{u}) \geq r\}}{\text{Volume}\left(\mathcal{S}_{h^*}^{(k)}\right)}, \tag{6}$$

which tells us the probability that a configuration drawn from $\mathcal{S}_{h^*}^{(k)}$ is at least distance r from the region's boundary. Integrating over all possible distances up to the maximum parameter distance $r_{\max}$ provides the overall stability score, measuring the weighted mean distance of a top-k SOTA configuration to the nearest non-top-k configuration after controlling for the size of the region:

$$\text{Stability}\left(\mathcal{S}_{h^*}^{(k)}\right) = \frac{1}{r_{\max}} \int_0^{r_{\max}} \text{LocalStability}(r)\, dr. \tag{7}$$

The score lies in the unit interval $[0, 1]$. A small score indicates brittle regions prone to sudden performance collapses, whereas a large score implies stable neighborhoods.

4 EVALUATING BENCHMARK DATASETS

The MESA framework in Section 3 asks whether a method's comparative performance advantage is stable under perturbations to experimental parameters. The same meta-experiment outputs used to evaluate the candidate methods can answer a second question at no additional cost: whether the datasets are capable of discriminating method performance. In particular, we evaluate the efficacy of a dataset by observing the behavior of the cohort on it through two distinct quantities.

4.1 RESOLUTION VOLUME

Let $\mathbf{\Psi}_d$ denote the plausible experimental configurations for some dataset d. For each $\boldsymbol{\psi} \in \mathbf{\Psi}_d$, we order the cohort by empirical mean performance, yielding $h_{(1)}(\boldsymbol{\psi}), \ldots, h_{(n)}(\boldsymbol{\psi})$. Holding this order fixed, we define the frontier margin at run r as the gap between the kth and $(k+1)$th methods:

$$g_k^d(\boldsymbol{\psi}, \mathbf{b}^{(r)}) = f\left(h_{(k)}(\boldsymbol{\psi}), \mathbf{b}^{(r)}\right) - f\left(h_{(k+1)}(\boldsymbol{\psi}), \mathbf{b}^{(r)}\right), \tag{8}$$

with mean $\bar{g}_k^d(\boldsymbol{\psi})$ and sample variance $\hat{\varepsilon}_k^2(\boldsymbol{\psi})$. We say that d resolves the top-k frontier at $\boldsymbol{\psi}$ when the frontier margin exceeds the same noise threshold used in Equation 2, giving the resolved region:

$$\mathcal{R}_k(d) = \left\{ \boldsymbol{\psi} \in \mathbf{\Psi}_d : \bar{g}_k^d(\boldsymbol{\psi}) > t_{\gamma,R-1} \cdot \sqrt{\frac{\hat{\varepsilon}_k^2(\boldsymbol{\psi})}{R}} \right\}. \tag{9}$$

We then define the resolution volume as the plausibility-weighted size of this region,

$$\text{Volume}_k(d) = \sum_{\boldsymbol{\psi} \in \mathcal{R}_k(d)} \pi(\boldsymbol{\psi}). \tag{10}$$

The resolution volume thus represents the probability that a configuration drawn from π yields a statistically significant top-k set. Note that since this gap is defined with respect to the between-run variation, no separate effect-size threshold is required. A low resolution volume implies that d is either underpowered or saturated, which both negatively impact its capability as a benchmark.

4.2 EFFECTIVE FRONTIER COUNT

Consider the subspace $\mathcal{R}_k(d)$ where there is a definitive top-k set. Let $\pi_{\mathcal{R}}$ represent the probability distribution π restricted to $\mathcal{R}_k(d)$ and renormalized. Let $T_k(\boldsymbol{\psi})$ represent the top-k set for some $\boldsymbol{\psi} \in \mathcal{R}_k(d)$. For a top-$k$ set T, we then define the plausibility mass $p_d(T) = \sum_{\boldsymbol{\psi} \in \mathcal{R}_k(d) : T_k(\boldsymbol{\psi}) = T} \pi_{\mathcal{R}}(\boldsymbol{\psi})$. This allows us to borrow from the notion of entropy to define the effective frontier count

$$N_k(d) = 2^{-\sum_T p_d(T) \log_2 p_d(T)}, \tag{11}$$

which represents the number of equally plausible top-k sets that would produce the same entropy. For example, three top-k sets with equal-sized regions yields a $N_k(d)$ of 3, as would a larger number of top-k sets with an uneven distribution of region sizes, hence the framing as an effective count.

4.3 READING THE SCORES TOGETHER

Resolution volume gives information on the discriminatory abilities of a dataset while effective frontier count tells us how varied the sets of winning methods are. A dataset with high volume and low (close to 1) $N_k(d)$ is decisive: it supports a single top-k SOTA set which leads throughout the space of plausible experimental parameters. On the other hand, high volume with high (greater than 1) $N_k(d)$ is the richer case, since the dataset separates methods reproducibly but the leader depends on the regime of experimental parameters. Low volume indicates that a dataset is a poor benchmark for one of two reasons: either the dataset is saturated and methods perform equivalently, or real gaps are obscured by run-to-run noise. These reasons carry different implications for the dataset, and thus we describe a diagnostic for differentiating them in Appendix A.1. In either case, $N_k(d)$ is unreliable when computed over little resolved mass. We note that $N_k(d)$ is meant to be descriptive and is not a score for which higher or lower is uniformly better. Additionally, each measure tells us how well the dataset d specifically benchmarks the set of candidate methods $\mathbb{C}$ on the given experimental parameters $\mathbf{\Psi}$, and does not necessarily generalize to other settings or tasks.

5 CASE STUDIES

We now apply the MESA framework in two case studies: one in conformal inference and one in anomaly detection. These domains were chosen not because we suspected their results to be fragile, but rather because each contains a cluster of methods that solve the same task, compare against each other, and have publicly accessible implementations. We select candidate methods using two sources from their originating papers: the benchmark datasets used for evaluation and the methods included for comparison. Experimental parameters are selected by inspecting the method's code for ad-hoc choices (i.e., not chosen by optimization) which impact results. We describe the procedures for selecting candidate methods, experimental parameters, and baselines in more detail in Appendix A.2. Each case study is presented in two parts: the first describes the meta-experiment setup, and the second details the method and dataset results. Both case studies use the threshold in Equation 2 with a paired t-test, $\gamma = 0.9$, and $R = 10$ repeated runs per configuration over baselines including the data draw and any method-internal randomness. In Appendix A.1, we show that this choice of R is sufficiently large to support our conclusions. Equal mass is assigned across levels of each experimental parameter. Results instead using Welch's t-test, which is appropriate when methods do not share baseline draws, are included in Appendices A.4 and A.6. In Appendix A.7, we show that the two case studies have correlation structures of opposite signs, causing Welch's t-test to overestimate scores for the first and underestimate scores for the second. We focus on these case studies to show the practical use of the MESA framework rather than to detract from the methods being evaluated. Methods with advantages which do not survive perturbation may still be the right choice for use under the conditions their authors studied.

5.1 CASE STUDY #1: CLASS-CONDITIONAL CONFORMAL PREDICTION

Conformal prediction (CP) has become a popular framework for generating prediction sets with finite-sample coverage guarantees, largely due to its ability to process outputs from black-box models using only exchangeability assumptions (Vovk et al., 2005; Angelopoulos & Bates, 2023). When applied to multi-class classification, standard conformal prediction often achieves marginal coverage by over-covering the majority class and under-covering the minority classes (Löfström et al., 2015). Class-conditional CP seeks to fix this problem by ensuring that marginal coverage is not achieved at the expense of coverage for particular classes.

5.1.1 META-EXPERIMENT SETUP

Candidate methods. We examine the methods shown in Table A5. These methods were selected because they each attempt to improve class-wise prediction set coverage. Due to a lack of publicly available code, we excluded Kandinsky CP (Bairaktari et al., 2025), dynamically grouped CP (Papaioannou et al., 2026), tail-aware CP (Liu et al., 2026), and cluster frequency CP (Lavi et al., 2026).

Evaluation metrics. We would like the candidate methods to increase class-conditional coverage while keeping the average prediction set size small. To measure "success" in this task, we define two evaluation metrics. The first is under-coverage gap ($\mathrm{UnderCovGap}$), which measures the average amount by which class-conditional coverage falls below its target (Shi et al., 2024; Ding et al., 2026). Letting $\widehat{C}_y$ denote the empirical coverage for class y and $[z]_+ = \max\{0, z\}$, we define $\mathrm{UnderCovGap} = \frac{1}{K}\sum_{y=1}^{K}\left[(1-\alpha) - \widehat{C}_y\right]_+$. The second is average set size ($\mathrm{AvgSize}$), the average number of labels included in a prediction set (Angelopoulos & Bates, 2023; Ding et al., 2023). Since we would like to have small prediction sets subject to the undercoverage gap being small, we call a method feasible when $\mathrm{UnderCovGap} \leq \delta \cdot \alpha$, where α is the target undercoverage level and δ is an experimental parameter which controls the undercoverage tolerance. If at least one method is feasible, we say that a method is SOTA if it has the smallest $\mathrm{AvgSize}$ among the feasible methods. If no method is feasible, we say that the method with the smallest $\mathrm{UnderCovGap}$ is SOTA.

Experimental parameters. We evaluate the cohort methods over the experimental parameters shown in Table A6. These parameters were selected by identifying experimental quantities varied in one or more of the cohort papers and expressing them as shared axes wherever possible. For example, Ding et al. (2023) reports results over three different score functions and a range of calibration set sizes. The interpolation weight τ is specific to Interp-Q and directly controls its tradeoff between standard and classwise calibration. We additionally introduce the evaluation-stage toler-

	Top-1 SOTA Volume				Top-1 SOTA Stability			
Method	CIFAR	Places	Pl@nt	iNat	CIFAR	Places	Pl@nt	iNat
Classwise CP	**0.126**	0.048	0.010	0.004	0.172	0.130	0.058	0.028
Clustered CP	0.010	0.017	0.002	0.017	0.160	0.100	0.042	0.038
RC3P	0.089	**0.174**	0.002	<0.001	**0.242**	**0.257**	0.048	0.018
Interp-Q	0.112	0.121	0.157	**0.248**	0.175	0.124	0.146	0.152
Label-Weighted CP	0.002	0.001	**0.226**	0.133	0.073	0.036	**0.207**	**0.171**

Table 1: We observe that the volume leader on each dataset is different, and that RC3P is most stable on non-long-tailed datasets, whereas Label-Weighted CP is most stable on long-tailed datasets.

	$k=1$		$k=2$	
Dataset	Vol_1	N_1	Vol_2	N_2
CIFAR-100	0.339	**3.38**	0.173	6.23
Places365	0.361	3.14	0.164	**6.65**
Pl@ntNet	0.398	2.31	**0.313**	2.03
iNaturalist	**0.402**	2.32	0.182	3.08

Table 2: Each dataset resolves $< 41\%$ of the configuration space, yet has multiple regimes where distinct methods achieve superior performance.

	$k=1$		$k=2$	
Dataset	Vol_1	N_1	Vol_2	N_2
Cardio	0.674	**4.60**	0.570	**9.20**
Mammography	**0.747**	2.71	**0.702**	3.35
Thyroid	0.433	**4.60**	0.462	5.38
Vowels	0.489	4.00	0.549	2.65

Table 3: Mammography resolves a top-1 method on three quarters of the configuration space, while Thyroid and Vowels resolve less than half.

ance δ to specify how much undercoverage is acceptable when determining whether a method is feasible. We assign equal categorical weight to every listed value of each knob.

Datasets. We evaluate the cohort methods on the four datasets shown in Table A7. These datasets include both conventional image-classification benchmarks and naturally long-tailed species-classification datasets. All methods operate on the same released softmax predictions from fixed pretrained classifiers provided by the corresponding source paper.

5.1.2 Meta-Experiment Results

Methods. We report the top-1 SOTA volume and stability scores for each candidate method in Table 1. The largest volume is Interp-Q's 0.248 on iNaturalist-2018, showing that no method leads everywhere on any dataset. In fact, we observe that each dataset has a distinct volume leader. These results are consistent with the motivation and conditions each method was designed for. For example, Interp-Q and Label-Weighted CP were developed with long-tailed data distributions in mind; we find that they drastically outperform the other candidates on Pl@ntNet and iNaturalist-2018, which are both long-tailed datasets. On the other hand, RC3P shrinks prediction sets by discarding labels the classifier ranks poorly, and therefore helps mostly for classes whose true label is among the highest-scoring few (Shi et al., 2024). Thus, we observe that RC3P performs well on Places365 and CIFAR-100 where these conditions hold, but struggles on long-tailed datasets such as Pl@ntNet and iNaturalist-2018. We note that because stability is normalized by volume, the scores give us separate information; only on Places365 and Pl@ntNet do volume and stability agree. This allows stability to separate methods that have similar volume scores, as in the case of CIFAR-100. Notably, all 20 method-dataset pairs produce nonzero volume, meaning that there exists some configuration under which each method could be reported as SOTA on each dataset. For this reason, standard static benchmarking can fall short in providing meaningful advice for selecting methods in real-world settings. MESA addresses this by measuring where and how securely each method leads.

Datasets. Table 2 displays the benchmark dataset scores from Section 4 calculated on the conformal prediction meta-experiment results. All four datasets resolve a top-1 leader on similar amounts of the configuration space (between 34% and 40%), so none is markedly more decisive than the others. However, the resolved volume scores also mean that not one of these benchmarks is able to discern a top-performing method across even half of the experimental parameter space, after accounting for baseline noise. While the datasets are similar with respect to their resolution volumes, they differ in how many distinct regimes they contain. We observe that both N_1 and N_2 are greater than one for

each dataset, indicating that the leader(s) change not only with the choice of dataset but also with position in the configuration space. CIFAR-100 and Places365 have the most fragmented frontiers, with N_1 of 3.38 and 3.14 alongside N_2 of 6.23 and 6.65, respectively. The lower effective frontier counts for Pl@ntNet and iNaturalist-2018 are consistent with the dominance of the two long-tailed methods. For example, we see that Pl@ntNet has both the highest Vol_2 and the lowest N_2, indicating that it identifies which two methods lead far more reliably than the other benchmarks. We note that none of this structure is recoverable from a fixed benchmark leaderboard. While a benchmark reports performance of individual methods, it tells us nothing about whether the task's configuration space is contested or settled. The dataset scores MESA provides allow for direct evaluation of the benchmarks themselves, providing crucial information for evaluating performance.

5.2 Case Study #2: Isolation-Based Anomaly Detection

Anomaly detection methods aim to flag outliers in datasets in an unsupervised manner. This task has many high-importance real-world use cases, such as detection of credit card fraud or identifying possible cyberattacks. The most popular anomaly detection algorithm is isolation forest (iForest), which approaches this problem by recursively partitioning the data at random and giving larger anomaly scores to data points isolated in leaf nodes with short paths (Liu et al., 2008). The adoption of iForest encouraged a series of subsequent related methods, each of which claims to repair a specific weakness through alterations to the split rule, the tree structure, and other components of the algorithm. This creates a family of iForest-related methods and forms a natural test case for the MESA framework, as each proposed method makes a performance enhancement claim in a particular slice of the experimental space.

5.2.1 Meta-Experiment Setup

Candidate methods. We examine the seven methods shown in Table A10. Five of these candidates are direct tree-based descendants of iForest, while the other two are closely-related isolation kernel methods. Each candidate method has a publicly-available code implementation which takes as input an unlabeled data matrix and outputs continuous anomaly scores.

Evaluation metrics. The candidate methods yield continuous scores for the held-out data point, with higher scores indicating greater likelihood of anomaly status. We follow these works in using the mainstream evaluation protocols for anomaly detection (Liu et al., 2008; Hariri et al., 2021), evaluating performance using area under the receiver operating characteristic curve (AUROC). The AUROC describes the probability that an anomaly is scored above a randomly chosen normal observation. We also report results using area under the precision-recall curve (AUPRC) in Appendix A.6.

Experimental parameters. We evaluate the candidate methods over the experimental parameters shown in Table A11. We assign equal mass to each listed level of each experimental parameter. Five of the nine parameters apply to all candidate methods, and the remaining four apply to one method each. These parameters were detected by inspecting method implementation to determine which experimental values were varied in experimentation. Four of the five shared parameters specify the distribution and handling of the data, including anomaly presence, signal-to-noise feature ratio, preprocessing, and the difficulty of the anomalies. The remaining shared parameter controls the number of observations used to construct one tree or isolation partition. The four data-stage parameters are inspired by the benchmark dimensions of Emmott et al. (2016), who construct anomaly detection benchmarks by varying the feature relevance and anomaly difficulty, frequency, and clusteredness.

Datasets. We evaluate the candidate methods on the four datasets shown in Table A12. The anomalies included in these datasets are controlled semi-synthetic data points rather than the original outliers, as we are controlling the difficulty of the anomalies using an experimental parameter. The included datasets are used in the literature to evaluate anomaly detection performance and differ with respect to important characteristics such as number of features and proportion of anomalies.

5.2.2 Meta-Experiment Results

Methods. Table 4 displays the top-1 SOTA volume and stability of each candidate method. Unlike the class-conditional CP case study, we find that a single method—iNNE—leads in volume across all four datasets. Whereas the conformal prediction candidates split the configuration space between

	Top-1 SOTA Volume				Top-1 SOTA Stability			
Method	Cardio	Mamm.	Thyroid	Vowels	Cardio	Mamm.	Thyroid	Vowels
iForest	0.063	0.040	0.015	0.005	0.089	0.086	0.051	0.119
EIF	0.150	0.206	0.003	0.002	**0.245**	0.187	0.036	0.064
FCF	0.075	0.009	0.138	0.116	0.184	0.096	0.124	0.107
DIF	0.075	0.001	0.040	0.033	0.165	0.081	0.153	**0.145**
OptIForest	0.014	0.017	0.065	0.028	0.084	0.085	0.037	0.087
iNNE	**0.292**	**0.467**	**0.153**	**0.234**	0.193	**0.354**	0.109	0.127
IDK	0.005	0.006	0.019	0.071	0.222	0.088	**0.158**	0.051

Table 4: Unlike the conformal prediction case study, a single method, iNNE, holds the largest volume across datasets. In stability, iNNE is behind EIF on Cardio and multiple methods on Thyroid.

methods suited to different data regimes (i.e., short-tailed vs. long-tailed), the anomaly detection candidates provide a practitioner a defensible default choice. Nevertheless, we still observe that all method-dataset pairs produce nonzero volume, meaning that any method could be reported as SOTA under at least one defensible configuration. Volume alone does not settle the choice; stability is helpful here in yielding supplementary information. For example, in the case of Thyroid, we see that iNNE holds a narrow lead over FCF with respect to volume, but FCF has a larger stability score, potentially making it a more appealing choice to risk-averse practitioners.

Datasets. Table 3 displays the benchmark dataset scores calculated on the anomaly detection meta-experiment results. The four benchmark datasets differ drastically with respect to resolution volume. For example, we see that Mammography and Cardio resolve a top-1 method on 75% and 67% of the configuration space, whereas Thyroid and Vowels resolve less than half of the space. From the effective frontier counts, we observe that Cardio is the most contested dataset, since many methods win in different fragments of the configuration space, whereas Mammography is the most decisive, as its space is dominated by iNNE and EIF. Method performance is typically conveyed through tables with many benchmarks, where each benchmark is treated as equally informative. However, these results tell us that different benchmarks contain different amounts of information. MESA uncovers this asymmetry, allowing benchmarks to be weighted appropriately by their discriminatory abilities.

6 Discussion

The MESA framework measures whether a comparative performance claim survives reasonable perturbations to experimental parameters such as preprocessing choices, hyperparameter selection, and evaluation procedure. Specifically, MESA identifies the subspace of experimental parameters in which a method's lead over its kth best competitor exceeds the noise of the experiment. It then scores that subspace by both its size and stability. Furthermore, the MESA framework provides evaluations of benchmark datasets without requiring additional experimentation. We demonstrate the utility of the framework in two applied case studies. We find multiple recent methods achieve superior performance only in particular corners of the experimental parameter space, while several benchmark datasets are unable to meaningfully tease apart method performance.

Limitations. MESA scores are expectations over experimental parameters that the analyst selects. Therefore, the validity and utility of the scores depends on selections of parameters and their levels. Since top-k status is decided by a significance test to incorporate inter-run noise, a low volume score could either indicate that the test is underpowered and needs more runs or that the methods are similar in performance. The diagnostic check differentiating these cases (Appendix A.1) requires the analyst to select an effect size cutoff, which may require domain knowledge in the applied field.

Future work. The MESA framework provides ample opportunities for follow-up work in evaluation and auto-research. The framework could be extended to the multi-objective setting by computing similar scores over Pareto frontiers. Auto-research systems can use MESA to both develop better benchmark datasets and more rigorously develop novel methods which don't simply optimize for an unimportant edge case. Furthermore, MESA dataset scoring can aid practitioners in running more efficient evaluations by narrowing their focus on the most informative benchmarks.

AI DISCLOSURE

In this work, we used generative AI tools to assist in implementing methods alongside substantial human-written code; specifically, AI was predominantly used for documentation and speeding up our existing implementation. We have not used generative AI tools for any other tasks with required disclosure. Additionally, we used generative AI tools for the following tasks with recommended disclosure: brainstorming, identifying relevant literature (particularly for the case studies in Section 5), creating and editing software code, and creating scientific figures. We have reviewed all AI-assisted work. All literature suggested by generative AI was read by the authors to confirm its relevance and applicability. Figure 1 was designed by the authors, and generative AI was used to turn the authors' hand-drawn sketch into TikZ code. Generative AI was used to assist in searching through candidate method's codebases in Section 5 for the identification of possible experimental parameters, as well as creating the scripts to run the candidate methods. These scripts and selected experimental parameters were manually checked by the authors. All writing, mathematical content, and other aspects of this work were performed by the authors without AI assistance. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f514cec81cb148559cf475e7426eed5e-Paper.pdf.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Konstantina Bairaktari, Jiayun Wu, and Steven Wu. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=IHAnkPkoiX.

Tharindu R. Bandaragoda, Kai Ming Ting, David Albrecht, Fei Tony Liu, Ye Zhu, and Jonathan R. Wells. Isolation-based anomaly detection using nearest-neighbor ensembles. *Computational Intelligence*, 34(4):968–998, 2018. doi: https://doi.org/10.1111/coin.12156. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12156.

Samuel J. Bell, Onno Kampman, Jesse Dodge, and Neil Lawrence. Modeling the machine learning multiverse. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18416–18429. Curran Associates, Inc., 2022. doi: 10.52202/068431-1338. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/750337e1301941f81ae31a90e0a1c181-Paper-Conference.pdf.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017. URL http://jmlr.org/papers/v18/16-305.html.

Aabesh Bhattacharyya, Tiffany Ding, and Rina Foygel Barber. Conformal prediction with macro-coverage guarantees. *arXiv preprint arXiv:2606.28598*, 2026.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In A. Smola, A. Dimakis, and I. Stoica (eds.), *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

A. Feder Cooper, Yucheng Lu, Jessica Forde, and Christopher De Sa. Hyperparameter optimization is deceiving us, and how to stop it. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3081–3095. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/17fafe5f6ce2f1904eb09d2e80a4cbf6-Paper.pdf.

David Cortes. Revisiting randomized choices in isolation forests, 2021. URL https://arxiv.org/abs/2110.13402.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery, 2021. URL https://arxiv.org/abs/2107.07002.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006. ISSN 1532-4435.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. In *International Conference on Learning Representations*, volume 2026, pp. 82431–82462, 2026.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL https://aclanthology.org/D19-1224/.

Elizabeth D Dolan and Jorge J Moré. Benchmarking optimization software with performance profiles. *Mathematical programming*, 91(2):201–213, 2002.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1266. URL https://aclanthology.org/P19-1266/.

Andrew Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. A meta-analysis of the anomaly detection problem, 2016. URL https://arxiv.org/abs/1503.01158.

H Christopher Frey and Sumeet R Patil. Identification and review of sensitivity analysis methods. *Risk analysis*, 22(3):553–578, 2002.

Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark, 2022. URL https://arxiv.org/abs/2206.09426.

Moritz Hardt. *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press, 2026.

Sahand Hariri, Matias Carrasco Kind, and Robert J. Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1479–1489, 2021. doi: 10.1109/TKDE.2019.2947676.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Risto Lahdelma and Pekka Salminen. Smaa-2: Stochastic multicriteria acceptability analysis for group decision making. *Operations research*, 49(3):444–454, 2001.

Risto Lahdelma, Joonas Hokkanen, and Pekka Salminen. Smaa-stochastic multiobjective acceptability analysis. *European journal of operational research*, 106(1):137–143, 1998.

Tomer Lavi, Bracha Shapira, and Nadav Rappoport. Cluster frequency conformal prediction for local coverage, 2026. URL https://arxiv.org/abs/2605.24872.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Shuqi Liu, Jianguo Huang, and Luke Ong. Conformal prediction meets long-tail classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23828–23836, 2026.

Tuve Löfström, Henrik Boström, Henrik Linusson, and Ulf Johansson. Bias reduction through conditional conformal prediction. *Intell. Data Anal.*, 19(6):1355–1375, November 2015. ISSN 1088-467X. doi: 10.3233/IDA-150786. URL https://doi.org/10.3233/IDA-150786.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271:18–42, 2019.

Jens-Michalis Papaioannou, Sebastian Jäger, Alexei Figueroa, David Stutz, Betty van Aken, Keno Bressem, Wolfgang Nejdl, Felix Gers, Alexander Löser, and Felix Biessmann. Robust conformal prediction for infrequent classes. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=nJ4p8rh3Ig.

David Picard. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1035. URL https://aclanthology.org/D17-1035/.

Zachary T. Rewolinski and Bin Yu. Predictability–computability–stability workflow for veridical data science in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 384(2321):20240605, 05 2026. ISSN 1364-503X. doi: 10.1098/rsta.2024.0605. URL https://doi.org/10.1098/rsta.2024.0605.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346/.

Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine*, 46(2):e70002, 2025.

Yuanjie Shi, Subhankar Ghosh, Taha Belkhouja, Janardhan Rao Doppa, and Yan Yan. Conformal prediction for class-wise coverage via augmented label rank calibration. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 132133–132178. Curran Associates, Inc., 2024. doi: 10.52202/079017-4200. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ee66188f019df7199c4c06320f698fa1-Paper-Conference.pdf.

Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. Specification curve analysis. *Nature human behaviour*, 4(11):1208–1214, 2020.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

Kai Ming Ting, Bi-Cun Xu, Takashi Washio, and Zhi-Hua Zhou. Isolation distributional kernel: A new tool for kernel based anomaly detection. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 198–206, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403062. URL https://doi.org/10.1145/3394486.3403062.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Richard E Wendell. The tolerance approach to sensitivity analysis in linear programming. *Management Science*, 31(5):564–578, 1985.

Haolong Xiang, Xuyun Zhang, Hongsheng Hu, Lianyong Qi, Wanchun Dou, Mark Dras, Amin Beheshti, and Xiaolong Xu. Optiforest: Optimal isolation forest for anomaly detection, 2023. URL https://arxiv.org/abs/2306.12703.

Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604, 2023. doi: 10.1109/TKDE.2023.3270293.

Renjun Xu and Jingwen Peng. A comprehensive survey of deep research: Systems, methodologies, and applications, 2025. URL https://arxiv.org/abs/2506.12594.

Bin Yu and Rebecca L Barter. *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press, 2024.

Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020. doi: 10.1073/pnas.1901326117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1901326117.

Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmarks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58984–59002. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhang24u.html.

A Appendix

A.1 A Diagnostic to Separate Saturated Performance & Underpowered Tests

Recall from Equation 2 that a configuration is excluded from the top-k SOTA region if

$$\bar{g}_k(\boldsymbol{\psi}) \leq t_{\gamma,R-1} \cdot \sqrt{\frac{\hat{\varepsilon}_k^2(\boldsymbol{\psi})}{R}}.$$

We can divide by $\hat{\varepsilon}_k(\boldsymbol{\psi})$ to collect the terms dependent on $\boldsymbol{\psi}$ on the same side, yielding

$$\frac{\bar{g}_k(\boldsymbol{\psi})}{\hat{\varepsilon}_k(\boldsymbol{\psi})} \leq \frac{t_{\gamma,R-1}}{\sqrt{R}}.$$

Let $D(\boldsymbol{\psi}) = \bar{g}_k(\boldsymbol{\psi})/\hat{\varepsilon}_k(\boldsymbol{\psi})$ represent the effect size and notice the right-hand side only depends on the confidence level and number of runs. We use $D(\boldsymbol{\psi})$ to reason about method and dataset scores.

Dataset diagnostics. Recall from Equation 8 that since the frontier margin compares the kth best method to the $(k+1)$th, we know that $\bar{g}_k^d(\boldsymbol{\psi}) \geq 0$, and thus we have $D(\boldsymbol{\psi}) \geq 0$. Given some effect size cutoff ξ (e.g., 0.2 for a small effect, as suggested by Cohen (2013)), we then have the following two regimes:

- *Saturated* ($D(\boldsymbol{\psi}) \leq \xi$). There is no detectable gap, the methods perform similarly.
- *Underpowered* ($\xi < D(\boldsymbol{\psi}) \leq t_{\gamma,R-1}/\sqrt{R}$). There is a meaningful gap, but it cannot be distinguished from run-to-run noise.

Method diagnostics. The top-k mean margin for a method can be negative, since we compare each method to its kth best competitor. Thus, the non-SOTA regimes are slightly different:

- *Saturated* ($|D(\boldsymbol{\psi})| \leq \xi$). There is no detectable gap, the methods perform similarly.
- *Underpowered* ($\xi < D(\boldsymbol{\psi}) \leq t_{\gamma,R-1}/\sqrt{R}$). There is a meaningful gap, but it cannot be distinguished from run-to-run noise.
- *Outperformed* ($D(\boldsymbol{\psi}) < -\xi$). There is a significant gap, showing the method is outperformed by the kth-best competitor.

Dataset results. Table A1 displays the diagnostic results computed on the datasets from each case study in Section 5. Specifically, we partition the configuration space into three categories: resolved, saturated, and underpowered. We observe that the long-tailed datasets from the conformal case study are the most saturated (since Interp-Q and Label-Weighted CP both perform well) and CIFAR-100 and Places365 are the most unresolved. In the anomaly detection case study, we see that the unresolved mass splits fairly evenly between saturated and underpowered.

Case Study	**Dataset**	Vol_1	Saturated	Underpowered
Conformal Prediction	CIFAR-100	0.339	0.395	0.266
	Places365	0.361	0.371	0.268
	Pl@ntNet	0.398	0.502	0.100
	iNaturalist	0.402	0.527	0.071
Anomaly Detection	Cardio	0.674	0.167	0.160
	Thyroid	0.433	0.324	0.243
	Mammography	0.747	0.121	0.133
	Vowels	0.489	0.264	0.247

Table A1: Dataset diagnostic results on case study datasets using a cutoff of $\xi = 0.2$. As in Section 5, these results use a paired t-test, and the anomaly detection performance is evaluated by AUROC.

Method results. Tables A2 and A3 display the diagnostic results computed on the methods from the conformal prediction and anomaly detection case studies, respectively. We again partition the configuration space into categories dictated by the diagnostic regimes described above. Across both

case studies, the 'underpowered' regime makes up a small proportion of the configuration space. We note that the top-1 SOTA volume and underpowered proportion together account for a rough ceiling on what more runs may be able to reveal. Under this ceiling, only iNNE on Mammography would reach over 50%, and thus the finding from Section 5 that no method leads on a majority of the space is shown to be not primarily due to run budget.

Dataset	**Method**	Top-1 SOTA Volume	Underpowered	Saturated	Outperformed
CIFAR-100	Classwise CP	0.126	0.065	0.211	**0.599**
	Clustered CP	0.010	0.036	0.214	**0.741**
	RC3P	0.089	0.086	0.219	**0.606**
	Interp-Q	0.112	0.064	0.207	**0.617**
	Label-Weighted CP	0.002	0.016	0.149	**0.832**
Places365	Classwise CP	0.048	0.047	0.179	**0.727**
	Clustered CP	0.017	0.050	0.238	**0.694**
	RC3P	0.174	0.071	0.160	**0.595**
	Interp-Q	0.121	0.090	0.217	**0.572**
	Label-Weighted CP	0.001	0.010	0.095	**0.894**
Pl@ntNet	Classwise CP	0.010	0.010	0.442	**0.538**
	Clustered CP	0.002	0.004	0.025	**0.969**
	RC3P	0.002	0.007	0.445	**0.545**
	Interp-Q	0.157	0.044	0.112	**0.686**
	Label-Weighted CP	0.226	0.035	0.165	**0.573**
iNaturalist	Classwise CP	0.004	0.002	0.474	**0.520**
	Clustered CP	0.017	0.011	0.116	**0.856**
	RC3P	$<$0.001	0.001	0.440	**0.558**
	Interp-Q	0.248	0.030	0.166	**0.555**
	Label-Weighted CP	0.133	0.026	0.390	**0.451**

Table A2: Method diagnostic results on the conformal prediction candidates, using a cutoff of ξ=0.2.

Sensitivity to change in ξ. In Table A4, we show how the proportion of the unresolved mass that is saturated/underpowered changes as we vary the cutoff ξ. When ξ is small (e.g., 0.05), we observe that the mass is more underpowered, since we have lowered the boundary for what we consider a sizeable gap. As expected, this reverses when ξ is large (e.g., 0.3), where we observe that the mass is more saturated. We see that this split moves a lot as we vary ξ, and that at each value of ξ the long-tailed conformal datasets are the most saturated.

Dataset	Method	Top-1 SOTA Volume	Underpowered	Saturated	Outperformed
Cardio	iForest	0.063	0.024	0.057	**0.856**
	EIF	0.150	0.020	0.060	**0.770**
	FCF	0.075	0.029	0.050	**0.847**
	DIF	0.075	0.027	0.056	**0.842**
	OptIForest	0.014	0.008	0.030	**0.948**
	iNNE	0.292	0.040	0.065	**0.604**
	IDK	0.005	0.012	0.038	**0.945**
Thyroid	iForest	0.015	0.013	0.082	**0.889**
	EIF	0.003	0.009	0.050	**0.937**
	FCF	0.138	0.068	0.163	**0.630**
	DIF	0.040	0.020	0.073	**0.867**
	OptIForest	0.065	0.015	0.052	**0.868**
	iNNE	0.153	0.073	0.196	**0.578**
	IDK	0.019	0.043	0.173	**0.765**
Mammography	iForest	0.040	0.016	0.045	**0.899**
	EIF	0.206	0.040	0.060	**0.695**
	FCF	0.009	0.008	0.028	**0.954**
	DIF	0.001	0.002	0.013	**0.983**
	OptIForest	0.017	0.009	0.021	**0.954**
	iNNE	**0.467**	0.040	0.054	0.439
	IDK	0.006	0.018	0.037	**0.938**
Vowels	iForest	0.005	0.006	0.032	**0.957**
	EIF	0.002	0.005	0.035	**0.958**
	FCF	0.116	0.028	0.081	**0.774**
	DIF	0.033	0.027	0.074	**0.867**
	OptIForest	0.029	0.023	0.063	**0.886**
	iNNE	0.234	0.112	0.179	**0.475**
	IDK	0.071	0.046	0.166	**0.716**

Table A3: Method diagnostics on the anomaly detection candidates, using ξ=0.2 and f=AUROC.

Case Study	Dataset	$D \leq 0.05$	$D \leq 0.1$	$D \leq 0.2$	$D \leq 0.3$
Conformal Prediction	CIFAR-100	21%	21%	60%	79%
	Places365	22%	22%	58%	77%
	Pl@ntNet	71%	71%	83%	88%
	iNaturalist	77%	77%	88%	91%
Anomaly Detection	Cardio	13%	27%	51%	72%
	Thyroid	16%	31%	57%	77%
	Mammography	11%	23%	48%	71%
	Vowels	15%	28%	52%	74%

Table A4: Percentage of each dataset's unresolved mass that is saturated (vs. underpowered) as we vary the cutoff ξ.

A.2 Selecting Candidate Methods & Experimental Parameters

The scores for both method performance and dataset efficacy are defined relative to a set of candidate methods and a distribution of experimental parameters. It follows that making these choices is an important step of meta-experimentation and should be done in a principled manner. While subject-matter expertise is helpful in the selection process, it is not strictly necessary for making informed decisions: neither of the two case studies included in Section 5 are within our areas of expertise. We found that deep research systems alongside agentic code reviewers were adept at finding "researcher degrees of freedom" that impact outputs.

Candidate methods. In Section 3, we say that candidate methods are designed to perform the same task. To discern whether a pair of methods is comparable, we use two sources from their originating papers: the benchmark datasets on which they are evaluated and the methods they compare against.

We rely on these sources instead of the paper's method description because they are concrete and easily verifiable. These conditions allow us to start with one or more seed papers and grow the candidate set outward by following the comparisons and references in the literature. Specifically, we add a paper's proposed method to our candidate set if it exceeds a user-set threshold for benchmark and method comparison overlap. Methods without publicly-available code are excluded.

Experimental parameters. After the set of candidate methods has been chosen, we must select the experimental parameters and the baselines. The experimental parameters are ad-hoc choices (i.e., not chosen by optimization) which are ancillary to the task but impact results, whereas baselines control the irreducible noise. Both are identified by an agent which inspects the publicly-available code of each method, alongside optional human input. A decision is treated as a baseline if its value cannot change the expectation of the result. For example, under standard exchangeability assumptions, the choice of train-test split changes the result but not its expectation. Every decision for which no such argument exists is treated as a plausible experimental parameter. These include hyperparameter settings, preprocessing choices, and evaluation protocols. We evaluate potential experimental parameters by assessing the author's justification and the importance of the decision, selecting those that exceed a user-defined threshold. We then inspect the values and probability distributions of each parameter used across codebases, pooling the supports of different candidate methods.

A.3 Case Study Setup: Class-Conditional Conformal Prediction

Table A5 summarizes the candidate methods. As mentioned in Section 5.1.1, four methods identified as potential candidates were excluded due to a lack of publicly-available code references in their corresponding papers. These excluded methods are Kandinsky CP (Bairaktari et al., 2025), dynamically grouped CP (Papaioannou et al., 2026), tail-aware CP (Liu et al., 2026), and cluster frequency CP (Lavi et al., 2026).

Table A6 displays the experimental parameters and their respective levels. The interpolation weight τ is only relevant for Interp-Q; other methods share the same runs for configurations differing only with respect to τ. We note that K_D represents the number of classes the specific benchmark dataset has. If the number of evaluated classes K is less than K_D, we retain a random subset of K classes, renormalize the softmax over the selected K classes, and apply the logit temperature before conformal scores are computed. Each combination of configurations is run $R = 10$ times with a paired seed controlling the subset of K classes, the calibration allocation, and the randomization of the conformity score.

Table A7 lists the benchmarks used in the class-conditional conformal prediction meta-experiment.

Method Name	**Method Description**	**Source Paper**
Classwise CP	Splits the calibration data by class and runs conformal prediction separately on each class.	Vovk (2012)
Clustered CP	Clusters classes with similar conformal scores and runs CP separately on each cluster.	Ding et al. (2023)
Rank-Calibrated Class-conditional CP (RC3P)	Improves Classwise CP by using label rank calibration to remove unlikely candidate classes.	Shi et al. (2024)
Interp-Q	Interpolates between standard and classwise CP to allow for tradeoffs between smaller prediction sets and better class-conditional coverage.	Ding et al. (2026)
Label-Weighted CP	Upweights minority classes to create wider prediction sets for their labels, satisfying a softened class-conditional "macro-coverage" constraint.	Bhattacharyya et al. (2026)

Table A5: Overview of the candidate class-conditional conformal prediction methods.

Experimental Parameter	Stage	Method(s)	Values
Target Miscoverage α	Method	All	$0.01, 0.05, 0.10, 0.20$
Conformal Score Function	Method	All	HPS/LAC, APS, RAPS
Interpolation Weight τ	Method	Interp-Q	$0.25, 0.5, 0.75, 0.9, 0.95, 0.975, .99, .999$
# of Evaluated Classes K	Experiment	All	$10, 25, 50, 100, 250, 500, K_D$
Calibration Fraction	Experiment	All	$0.25, 0.50, 1.00$
Logit Temperature	Experiment	All	$0.5, 1, 2$
Undercoverage Tolerance δ	Evaluation	All	$0, 0.05, 0.10, 0.20$

Table A6: Experimental parameters and levels in the class-conditional conformal meta-experiment. Here, K_D is the total number of classes in dataset D; listed values of K exceeding K_D are omitted.

Dataset	K_D	Calibration Pool	Test Set Size	Source
CIFAR-100	100	28,000	2,000	Ding et al. (2023)
Places365	365	176,696	7,300	Ding et al. (2023)
Pl@ntNet-300K-truncated	330	17,906	33,000	Ding et al. (2026)
iNaturalist-2018-truncated	857	12,361	85,700	Ding et al. (2026)

Table A7: Datasets used in the class-conditional CP meta-experiment. The reported test set sizes reflect the dataset under its full number of evaluated classes K_D. When using other values of K, the datasets only keep examples from selected classes.

A.4 Extended Case Study Results: Class-Conditional Conformal Prediction

	Top-1 SOTA Volume				Top-1 SOTA Stability			
Method	CIFAR	Places	Pl@nt	iNat	CIFAR	Places	Pl@nt	iNat
Classwise CP	**0.162**	0.066	0.009	0.001	0.185	0.141	0.070	0.018
Clustered CP	0.024	0.039	0.005	0.024	0.165	0.128	0.043	0.057
RC3P	0.132	**0.211**	0.003	0.000	**0.261**	**0.262**	0.043	—
Interp-Q	0.145	0.169	0.189	**0.272**	0.184	0.125	0.157	0.154
Label-Weighted CP	0.008	0.002	**0.250**	0.155	0.121	0.048	**0.206**	**0.181**

Table A8: Case study results under Welch's t-test. The volume leader on each dataset is different, and RC3P is most stable on non-long-tailed datasets, whereas Label-Weighted CP is most stable on long-tailed datasets. These conclusions are unchanged from the results under the paired t-test.

Tables A8 and A9 display the results of the conformal prediction case study using Welch's t-test instead of a paired t-test. The qualitative findings stay the same, as we observe that each dataset keeps the same volume and stability leaders as in Table 1. We note that resolution volumes are larger under Welch's test, rising from 0.34-0.40 to 0.45-0.49, with Places365 now resolving a top-1 the most frequently. This is consistent with the negative between-method correlation of this case study (Appendix A.7), which causes the paired variance to be larger than that used by Welch's test. Even with the inflated volumes under Welch's test, we do not see any method volume exceed 0.28, and each dataset resolves a top-1 method on less than half of the space. For dataset results, we observe that the ordering of the effective frontier counts N_1 and N_2 is nearly unchanged, with the leaders remaining constant.

	$k=1$		$k=2$	
Dataset	Vol_1	N_1	Vol_2	N_2
CIFAR-100	0.470	**3.68**	0.168	6.32
Places365	**0.488**	3.41	0.174	**6.85**
Pl@ntNet	0.457	2.35	**0.347**	1.90
iNaturalist	0.451	2.31	0.184	2.96

Table A9: Case study dataset results under Welch's t-test. Each dataset resolves $< 50\%$ of the configuration space, yet has multiple regimes where distinct methods achieve superior performance.

A.5 Case Study Setup: Isolation-Based Anomaly Detection

Tables A10, A11, and A12 summarize the candidate methods, experimental parameters, and datasets, respectively. We follow Han et al. (2022) in replacing each dataset's outliers with semi-synthetic local anomalies using a Gaussian mixture model. The train/test split is 70%/30%. Specifically, this mixture model is fit on the non-anomalous points in the training set, with anomalies in the test set then drawn from the model with covariance multiplied by the local anomaly covariance parameter. Thus, larger values of local anomaly covariance push the anomalies further into the tails, making them easier to detect. The "Native" level is a special case where we use the dataset's real anomalies instead of semi-synthetic ones. The "multiplier of anomaly frequency" scales the anomaly rates from Table A12 by removing either non-anomalous or anomalous points. The preprocessing levels either apply no scaling or `scikit-learn`'s StandardScaler or RobustScaler. Each method is unsupervised, and thus does not see the labels of points in the training set. Each combination of configurations is run $R = 10$ times with a paired seed controlling the train/test split, anomaly generation, noise feature draws, randomness of the Gaussian mixture model, and the internal randomness of the candidates.

Method Name	**Method Description**	**Source Paper**
iForest	Builds trees using random axis-aligned splits; shorter average paths indicate more anomalous observations.	Liu et al. (2008)
Extended Isolation Forest (EIF)	Replaces axis-aligned splits with random hyperplanes to reduce artifacts in the anomaly-score surface.	Hariri et al. (2021)
Fair-Cut Forest (FCF)	Selects splits using information gain to produce fairer partitions and retain randomized isolation trees.	Cortes (2021)
Deep Isolation Forest (DIF)	Uses neural networks to map original data into random representation ensembles, allowing for high freedom of partitions in the original data space.	Xu et al. (2023)
OptIForest	Establishes theory on isolation efficiency to design an optimal isolation forest.	Xiang et al. (2023)
iNNE	Proposes using nearest-neighbor ensembles to isolate outliers rather than trees.	Bandaragoda et al. (2018)
Isolation Distributional Kernel (IDK)	Introduces an isolation kernel feature map as a new way to measure similarity.	Ting et al. (2020)

Table A10: Overview of the candidate isolation-based anomaly-detection methods.

Experimental Parameter	Stage	Method(s)	Values
Multiplier of anomaly frequency	Data	All	$0.25, 0.5, 1, 2, 4$
Fraction of noise features	Data	All	$0, 0.2, 0.33, 0.5$
Scaling performed in preprocessing	Data	All	None, Standard, Robust
Local anomaly covariance	Data	All	$1, 5, 10$, Native
Isolation subsample size	Method	All	$64, 256, 1024$
Extension fraction	Method	EIF	$0, 0.5, 1$
Pooled-gain probability	Method	FCF	$0, 0.5, 1$
Representation count	Method	DIF	$10, 50, 100$
Branching factor	Method	OptIForest	$2, e, 3, 4$

Table A11: Experimental parameters and levels in the isolation-based anomaly detection meta-experiment. Here, 'Native' corresponds to the anomaly covariance in the original dataset. The values of 'scaling performed in preprocessing' correspond to the types of `Scaler` objects in `scikit-learn`.

Dataset	Rows	Features	Native Anomaly Rate	Source
Cardio	1,831	21	9.61%	Han et al. (2022)
Mammography	11,183	6	2.32%	Han et al. (2022)
Thyroid	3,772	6	2.47%	Han et al. (2022)
Vowels	1,456	12	3.43%	Han et al. (2022)

Table A12: Datasets used in the isolation-based anomaly detection meta-experiment.

A.6 Extended Case Study Results: Isolation-Based Anomaly Detection

Tables A13-A18 display the results of the anomaly detection case study under different combinations of test (paired and Welch's) and metric (AUPRC and AUROC). Under the paired test with AUPRC (Tables A13 and A14), we again observe that iNNE leads in top-1 SOTA volume across all four datasets. Under this setup, iNNE reaches a top-1 SOTA volume of 0.620 on Mammography, which is the only method-dataset pair across all of our analyses where a single method leads on more than half of the configuration space.

We find that using Welch's test results in deflated volumes throughout both AUROC and AUPRC, which is consistent with the positive between-method correlation in this case study (Appendix A.7), causing the paired variance to be smaller than that used by Welch's test. When Welch's test is used along with AUPRC (Tables A15 and A16), iNNE still leads in top-1 SOTA volume across datasets. However, we see that under Welch's and AUROC (Tables A17 and A18), FCF leads in top-1 SOTA volume on both Thyroid and Vowels.

	Top-1 SOTA Volume				Top-1 SOTA Stability			
Method	Cardio	Mamm.	Thyroid	Vowels	Cardio	Mamm.	Thyroid	Vowels
iForest	0.043	0.027	0.010	0.004	0.093	0.082	0.052	0.086
EIF	0.115	0.061	0.004	0.001	**0.208**	0.088	0.034	0.108
FCF	0.073	0.005	0.084	0.070	0.159	0.082	0.048	0.104
DIF	0.065	0.001	0.086	0.107	0.191	0.066	**0.235**	**0.166**
OptIForest	0.019	0.010	0.055	0.027	0.108	0.087	0.035	0.096
iNNE	**0.333**	**0.620**	**0.348**	**0.183**	0.178	**0.331**	0.139	0.087
IDK	0.000	0.007	0.003	0.062	—	0.074	0.135	0.039

Table A13: Case study method results using a paired t-test with AUPRC. As in the results using AUROC, a single method, iNNE, holds the largest volume across datasets.

	$k=1$		$k=2$	
Dataset	Vol_1	N_1	Vol_2	N_2
Cardio	0.647	4.08	0.584	**7.88**
Mammography	**0.731**	1.84	**0.761**	1.93
Thyroid	0.591	3.40	0.602	4.62
Vowels	0.454	**4.44**	0.410	3.62

Table A14: Case study dataset results using a paired t-test with AUPRC. As in the results using AUROC, Mammography resolves the most volume for both k=1 and k=2.

	Top-1 SOTA Volume				**Top-1 SOTA Stability**			
Method	Cardio	Mamm.	Thyroid	Vowels	Cardio	Mamm.	Thyroid	Vowels
iForest	0.024	0.017	0.001	0.001	0.107	0.082	0.034	0.062
EIF	0.094	0.035	0.000	0.000	**0.230**	0.090	—	—
FCF	0.033	0.002	0.051	0.041	0.159	0.077	0.036	0.062
DIF	0.046	0.000	0.071	0.036	0.212	—	**0.217**	**0.096**
OptIForest	0.010	0.002	0.039	0.004	0.069	0.088	0.034	0.076
iNNE	**0.249**	**0.327**	**0.138**	**0.074**	0.139	**0.285**	0.114	0.087
IDK	0.000	0.000	0.000	0.037	—	—	—	0.034

Table A15: Case study method results using Welch's t-test with AUPRC. Across all datasets, iNNE leads in top-1 SOTA volume. Volumes are deflated in comparison to the paired t-test results.

	$k=1$		$k=2$	
Dataset	Vol_1	N_1	Vol_2	N_2
Cardio	**0.456**	3.73	0.459	**6.19**
Mammography	0.383	1.72	**0.701**	1.49
Thyroid	0.301	3.62	0.477	3.77
Vowels	0.193	**4.14**	0.287	1.85

Table A16: Case study dataset results using Welch's t-test with AUPRC. Resolution volumes fall compared to the paired t-test results from Table A14, and Cardio now finds a unique winner most frequently.

	Top-1 SOTA Volume				**Top-1 SOTA Stability**			
Method	Cardio	Mamm.	Thyroid	Vowels	Cardio	Mamm.	Thyroid	Vowels
iForest	0.044	0.013	0.003	0.001	0.092	0.086	0.060	0.061
EIF	0.135	0.153	0.000	<0.001	**0.255**	0.195	—	0.034
FCF	0.040	0.001	**0.068**	**0.075**	0.179	0.066	0.096	0.055
DIF	0.046	0.000	0.006	0.005	0.185	—	**0.129**	**0.116**
OptIForest	0.007	0.005	0.036	0.013	0.069	0.088	0.034	0.051
iNNE	**0.223**	**0.329**	0.021	0.069	0.176	**0.296**	0.089	0.111
IDK	0.000	0.000	0.000	0.040	—	—	—	0.035

Table A17: Case study method results using Welch's t-test with AUROC. We observe that iNNE and EIF account for most of the volume on Cardio and Mammography, while Thyroid and Vowels spread volume across multiple methods. FCF overtakes iNNE for volume on Thyroid and Vowels.

	$k=1$		$k=2$	
Dataset	Vol_1	N_1	Vol_2	N_2
Cardio	0.495	**4.10**	0.426	**7.27**
Mammography	**0.501**	2.21	**0.487**	2.31
Thyroid	0.134	3.40	0.276	3.33
Vowels	0.203	3.88	0.426	1.52

Table A18: Case study dataset results using Welch's t-test with AUROC. Thyroid only finds a unique winner in 13.4% of the space. We again see that each dataset has multiple regimes where distinct methods win.

A.7 Correlations Across Baselines and Their Impacts

In Equations 2 and 9, we are testing $\bar{g}_k$ and $\bar{g}_k^d$, respectively, which are both defined as differences between the performance of two methods. We henceforth use $\bar{g}_k$ for simplicity of notation, but the following also applies to $\bar{g}_k^d$. At some configuration $\boldsymbol{\psi}$, the two methods whose differences define $\bar{g}_k$ are each run under the same R baseline draws. Let σ_1 and σ_2 denote their standard deviations and ϕ denote the correlation between their per-run scores. We then have that the variance of $\bar{g}_k$ is

$$\mathrm{Var}(\bar{g}_k) = \frac{\sigma_1^2 + \sigma_2^2 - 2\phi\sigma_1\sigma_2}{R}.$$

Using an unpaired test (e.g., Welch's) assumes that the two methods are independent, i.e., that $\phi = 0$. Thus, it follows that if $\phi > 0$, it overestimates the variance (noise) and thus underestimates the resulting volume. Similarly, if $\phi < 0$, the reverse happens, and the variance is underestimated, causing the volume to be overestimated. To measure the size of this effect, we report the ratio of the paired test statistic to Welch's test statistic, which we define as the test-statistic ratio:

$$\mathrm{TestStatisticRatio} = \frac{\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)/R}}{\mathrm{sd}(\bar{g}_k)},$$

where $\mathrm{sd}(\bar{g}_k)$ represents the standard deviation of $\bar{g}_k$. We note that since $\mathrm{sd}(\bar{g}_k)$ also has a $\sqrt{R}$ term, these cancel and the ratio is independent of the number of runs. When the test-statistic ratio is larger than 1, it means that Welch's test is more conservative, whereas a ratio less than one implies the paired test is more conservative. Table A19 shows the median correlation and test-statistic ratio of all frontier pairs, defined as the top-ranked method against its second-ranked competitor(s).

	Median ϕ	% of $\phi > 0$	% of $\phi < 0$	Median Test-Statistic Ratio
Conformal Prediction	-0.655	18%	81%	0.778
Anomaly Detection	+0.813	96%	4%	2.102

Table A19: Correlation and test-statistic ratio information from each case study. Results are computed over all pairs of frontier methods, i.e., top-ranked and second-ranked methods. We additionally report the percentage of paired correlations that are positive and negative. Note the percentages of positive and negative do not add to 100% for conformal prediction due to zero-correlation pairs.

Conformal prediction. We observe that correlations at the frontier are mostly negative (81%). Since the median test-statistic ratio is less than one, we have that Welch's noise estimate is about 22% narrower than the paired estimate, explaining the larger volumes under Welch's t-test in Appendix A.4. The negative correlation occurs because f (as defined in Section 5.1.1) is an indicator reflecting smallest set size among feasible methods.

Anomaly detection. We observe that correlations at the frontier are overwhelmingly positive (96%). The median test-statistic ratio is 2.102, meaning that Welch's noise estimate is twice as wide as the paired estimate, explaining the smaller volumes under Welch's t-test in Appendix A.6. The positive correlations are likely due to the shared baseline controlling the train/test split, anomaly creation, and aspects such as noise features. If the conditions under a baseline make it more difficult to detect the outliers, all methods are more likely to struggle; and if the conditions make it easier, all methods are more likely to succeed.