CRETAP: FROM RECONSTRUCTION ERRORS TO RELIABLE ANOMALY PROBABILITIES ON IMBALANCED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstruction-based anomaly detectors, such as autoencoders, rank samples by an error whose scale depends on the dataset, preprocessing, and aggregation, so thresholds do not transfer and scores are not probabilities. Fixing this with post-hoc calibration withholds scarce labeled anomalies from training, and switching to supervised classifiers discards the element-wise residuals that explain each decision. We introduce CRETAP, which fits Platt scaling of the log-error—a two-parameter Hill function—jointly with the reconstruction network under binary cross-entropy (BCE) instead of afterwards, so the anomalies that train the detector also calibrate it. Because the probability depends on the input only through the error, we prove that the excess BCE upper-bounds twice the squared calibration error and that stationary points of training are mean-calibrated with unit calibration slope in-sample. The forward pass also yields feature-level attributions—each element's error contribution divided by the learned decision boundary—that sum to more than one exactly when the probability exceeds one half. On the 29 ADBench datasets with at most 5% anomalies and without any post-hoc calibration, CRETAP is the best-calibrated of eight natively probabilistic detectors on 9 datasets, more than any other, and second in mean expected calibration error (0.0073 vs. 0.0067 for XGBOD), while its mean area under the precision–recall curve, 71.19%, trails CatBoost's by 2.0 points. In this regime, a reconstruction detector can thus serve as its own calibrator; above 5% contamination, gradient-boosted trees are preferable.

1 INTRODUCTION

Reconstruction-based anomaly detection (AD) learns the regularities of nominal data and interprets the discrepancy between a sample and its reconstruction as evidence of abnormality (Ruff et al., 2021; Pang et al., 2021). The paradigm is attractive because nominal data are abundant whereas anomalies are rare and heterogeneous. Rare, however, does not mean absent: deployed systems often accumulate a handful of verified anomalies from past incidents, and even such limited supervision can substantially improve detection over purely unsupervised training (Pang et al., 2019; Ruff et al., 2020; Zhou et al., 2022; Han et al., 2022). We study this label-informed regime across tabular, image, and text data, training a reconstruction model on abundant, mostly nominal samples together with a small set of labeled anomalies.

Deploying such a detector in an anomaly detection pipeline requires more than discriminative power: it needs a *ranking* to prioritize samples for review and a *calibrated probability* to express uncertainty on a scale that is comparable across systems. Raw reconstruction errors provide only the former: their magnitude depends on the dataset, preprocessing, dimensionality, and aggregation, so an operating threshold tuned in one setting does not transfer to another. The latter is usually obtained in one of two ways, neither satisfactory when labeled anomalies are scarce. The first maps errors to probabilities after training with post-hoc calibrators such as Platt scaling and isotonic regression (Platt, 1999; Zadrozny & Elkan, 2002), which must be fit on held-out labeled data because training-set scores are optimistically biased. Every anomaly reserved for calibration is thus one withheld from training, and the calibration set contains few positives for a stable fit, particularly for isotonic regression (Niculescu-Mizil & Caruana, 2005). The second abandons reconstruction in favor of *natively probabilistic* detectors, i.e., supervised classifiers such as gradient-boosted trees (Zhao & Hryniewicki, 2018; Prokhorenkova et al., 2018), random forests, or multi-layer perceptrons (MLPs)

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.

that output $p(y{=}1 \mid x)$ directly. These models, however, can be miscalibrated without further correction (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017). They also discard the element-wise residuals that give reconstruction models a built-in explanation, so feature-level attribution requires perturbation- or gradient-based explainers, typically at the cost of many additional model evaluations (Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017).

We propose CRETAP, which removes this trade-off by making the reconstruction model itself natively probabilistic. CRETAP places a learnable two-parameter Hill function (Hill, 1910) directly after error aggregation, mapping the non-negative reconstruction error $e(x) \in [0, \infty)$ to an anomaly probability

$$p_{k,n}(x) \;=\; \frac{e(x)^n}{k^n + e(x)^n} \;=\; \sigma\big(n\,[\log e(x) - \log k]\big), \qquad k, n > 0, \tag{1}$$

where σ is the logistic sigmoid. As the second form makes explicit, the Hill function belongs to the same family as Platt scaling applied to the log-error; what differs is *where* it is fit. Rather than calibrating a frozen detector on a separate split, CRETAP trains (k, n) jointly with the reconstruction network under binary cross-entropy (BCE), a strictly proper scoring rule (Gneiting & Raftery, 2007). A single objective thus uses every labeled anomaly for detection and calibration: it contracts nominal errors below k, pushes anomalous errors above it, and fits the probability scale at once.

Because end-to-end BCE training can produce overconfident neural classifiers (Guo et al., 2017), it is not obvious that joint training suffices for calibration. In CRETAP, the probability depends on the input only through a scalar error and a two-parameter monotone function, so miscalibration can arise only from a mismatch between that function and the one-dimensional posterior given the error. We prove that the excess BCE of the function upper-bounds twice the squared mismatch (Proposition 2(ii)), so the training objective directly controls calibration. We also prove that at any stationary point of training the fitted probabilities are mean-calibrated with unit calibration slope (Proposition 2(iii)), so temperature scaling has nothing left to correct in-sample.

This parameterization has three practical consequences. First, the probability scale does not depend on the units of the error: rescaling the error or raising it to a power, as when switching from mean squared error (MSE) to root mean squared error (RMSE), is absorbed entirely by (k, n), so operating points are set on a probability scale shared across datasets rather than on the error scale, which is not. Second, the parameters are interpretable: the scale k is the error level at which a sample is equally likely to be nominal or anomalous, i.e., the learned decision boundary, and the shape n controls how sharply confidence changes around it. Third, the Hill function is strictly increasing and invertible in closed form, so ranking metrics such as the area under the precision–recall curve (AUPR) are identical for the error and the probability, and every probability threshold corresponds to an explicit error threshold.

Placing the function after aggregation brings a further benefit: the element-wise residuals that any reconstruction model exposes gain a common unit, the learned boundary k. Dividing each element's contribution to the error by k expresses a feature's attribution as a fraction of the error budget that separates nominal from anomalous, and for additive aggregations these attributions sum to more than one exactly when the sample is declared anomalous. CRETAP thus yields feature-level attributions that are complete with respect to the emitted probability and comparable across samples and datasets, in the same forward pass as the detection and without backpropagation or repeated model evaluations.

Our contributions are as follows:

- We introduce CRETAP, which makes reconstruction models natively probabilistic by learning a Hill function end to end under BCE, so that scarce labeled anomalies serve both detection and calibration.
- We prove that the excess BCE upper-bounds twice the squared calibration error and that the Hill function is ranking-preserving and unit-invariant; residual contributions divided by k give complete feature-level attributions in the same forward pass.
- We evaluate CRETAP on 29 ADBench datasets with at most 5% contamination against 28 baselines: it ranks third overall in mean AUPR (71.19%), has the second-lowest mean expected calibration error (ECE) among natively probabilistic detectors (0.0073; XGBOD: 0.0067), and lies on the AUPR–ECE Pareto frontier with CatBoost and XGBOD.

2 Related Work

Reconstruction-based and label-informed detection. Reconstruction methods keep anomalies from being reconstructed as well as nominal data through memory prototypes, simulated anomalies, pretrained feature targets, or architectures that avoid identical shortcuts (Gong et al., 2019; Zavrtanik et al., 2021; Deng & Li, 2022; You et al., 2022; He et al., 2024), or adapt the pretrained encoder to the target domain (Guo et al., 2023). When a few anomalies are labeled, a purely unsupervised objective wastes them. DeepSAD (Ruff et al., 2020) pushes labeled anomalies away from the center of the DeepSVDD hypersphere (Ruff et al., 2018), DevNet drives their scores into the upper tail of a Gaussian reference distribution (Pang et al., 2019), PReNet regresses ordinal targets over sample pairs (Pang et al., 2023), and FEAWAD scores an autoencoder's hidden representation, residual, and error with a separate network (Zhou et al., 2022). In CRETAP, by contrast, the Hill function is the only consumer of the error, whereas FEAWAD routes its margin-trained error through a separate scorer; the decision thus remains a threshold on the error however many labels are available, and the output is a probability rather than an unbounded, dataset-dependent score.

Calibration and probabilistic anomaly scores. Post-hoc calibrators such as Platt scaling, isotonic regression, and temperature scaling (Platt, 1999; Zadrozny & Elkan, 2002; Guo et al., 2017) depend on a held-out calibration set: isotonic regression overfits small sets (Niculescu-Mizil & Caruana, 2005), and current techniques cannot estimate how miscalibrated scaling methods are (Kumar et al., 2019). In anomaly detection, label-free alternatives rescale outlier scores under distributional assumptions (Kriegel et al., 2011), fit a Platt-type sigmoid by EM with latent labels (Gao & Tan, 2006), or estimate example-wise confidence (Perini et al., 2020), all post-hoc on a fixed detector. Closest to CRETAP, ABC trains an autoencoder end to end under a binary objective that treats $\exp(-e(x))$ as the probability of being nominal (Yamanaka et al., 2019); this map has no free parameters, its decay is tied to the raw error scale, and its outputs are not evaluated as probabilities. CRETAP instead fits (k, n) jointly with the reconstruction network, so the map adapts to the data and is invariant to the units of the error, and we measure its calibration.

Explaining and attributing an anomaly decision. Post-hoc attribution is costly: gradient-based methods need up to hundreds of backward passes (Selvaraju et al., 2017; Sundararajan et al., 2017), and perturbation-based methods such as Occlusion, RISE, LIME, and Kernel SHAP need hundreds to thousands of forward evaluations (Zeiler & Fergus, 2014; Petsiuk et al., 2018; Ribeiro et al., 2016; Lundberg & Lee, 2017). A reconstruction model already computes the element-wise residual map, and thresholding it is the prevailing way to localize defects with autoencoders, although Bergmann et al. (2019b) replace per-pixel ℓ^p residuals with a structural-similarity residual. Other detectors derive their map from a segmentation network or from patch distances to a memory bank or fitted distribution, and score the image by its (possibly reweighted) maximum (Zavrtanik et al., 2021; Defard et al., 2021; Roth et al., 2022). Residual maps are typically rescaled—often by per-sample min–max normalization—before they are shown or evaluated, and when that rescaling is done per sample the intensities are no longer comparable between samples;

CRETAP instead divides the residuals by the learned boundary k, so the attributions sum to $e(x)/k$, which determines the emitted probability, and are comparable across samples scored by the same model; the map is thus consistent with the decision by construction, and we examine its agreement with human annotations only qualitatively.

3 Preliminaries

Setup. We train on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where each sample $x_i \in \mathbb{R}^d$ has d elements and its target $y_i \in \{0, 1\}$ is 1 for the few labeled anomalies and 0 for the remaining, mostly nominal samples. A reconstruction model scores a sample x by its reconstruction error $e(x) \geq 0$ (Section 4.1); e also denotes a generic error value, as in $h_{k,n}(e)$ below. The BCE of a probability $p \in [0, 1]$ for a target y is $\ell(y, p) = -y \log p - (1 - y) \log(1 - p)$.

Calibration. A detector $f : \mathbb{R}^d \to [0, 1]$ is *perfectly calibrated* if $\Pr\big(y = 1 \mid f(x) = q\big) = q$ for all q in its range. We measure deviations from this ideal by the L_1 calibration error and, for a partition $\mathcal{B} = \{B_1, \ldots, B_M\}$ of $[0, 1]$ into bins, the binned ECE (Pakdaman Naeini et al., 2015; Guo

et al., 2017),

$$\mathrm{CE}(f) = \mathbb{E}\,\Big|\,\Pr\big(y = 1 \mid f(x)\big) - f(x)\Big|, \tag{2}$$

$$\mathrm{ECE}_{\mathcal{B}}(f) = \sum_{m=1}^{M} \Pr\big(f(x) \in B_m\big)\,\Big|\,\mathbb{E}\big[\,y - f(x) \mid f(x) \in B_m\big]\Big|. \tag{3}$$

Both are population quantities over (x, y); equation 3 is estimated on test data by plug-in averages.

The Hill function. For $k, n > 0$, the two-parameter Hill function (Hill, 1910; 1913; Goutelle et al., 2008) and the anomaly probability of equation 1 are

$$h_{k,n} : [0,\infty) \to [0,1), \quad h_{k,n}(e) = \frac{e^n}{k^n + e^n}, \qquad p_{k,n}(x) = h_{k,n}\big(e(x)\big). \tag{4}$$

Proposition 1 (Properties of the Hill function)**.** *For all $k, n > 0$ (proof in Appendix C.1):*

(i) ***Logistic form.*** *For $e > 0$, with σ the logistic sigmoid and σ^{-1} its inverse, the logit,*

$$h_{k,n}(e) = \sigma\big(n\,[\log e - \log k]\big), \qquad \textit{equivalently} \qquad \frac{h_{k,n}(e)}{1 - h_{k,n}(e)} = \Big(\frac{e}{k}\Big)^n, \tag{5}$$

so $h_{k,n}$ is Platt scaling (Platt, 1999) on $\log e$: its logit is affine in $\log e$, with slope n and intercept $-n \log k$.

(ii) ***Monotonicity and inversion.*** *$h_{k,n}$ is a strictly increasing bijection from $[0,\infty)$ onto $[0,1)$, and for every $\tau \in (0,1)$,*

$$p_{k,n}(x) \geq \tau \iff e(x) \geq k\,\big(\tau/(1-\tau)\big)^{1/n}. \tag{6}$$

Hence $p_{k,n}$ and e order samples identically, including ties, so every ranking metric (e.g., AUPR) is identical for both.

(iii) ***Symmetry about k.*** *$h_{k,n}(k) = \frac{1}{2}$, and $h_{k,n}(kt) + h_{k,n}(k/t) = 1$ for all $t > 0$.*

(iv) ***Invariance to the units.*** *For any $c, a > 0$, let $T(e) = c\,e^a$. Then*

$$h_{k,n}(e) = h_{T(k),\,n/a}\big(T(e)\big) \qquad \textit{for all } e \geq 0, \tag{7}$$

so the errors e and $T \circ e$ yield the same family of detectors $x \mapsto h_{k,n}(e(x))$, $k, n > 0$, with $(k, n) \leftrightarrow (T(k), n/a)$.

(v) ***Bounded sensitivity under BCE.*** *For $e > 0$ and $p = h_{k,n}(e)$,*

$$\partial\ell/\partial \log e = -\,\partial\ell/\partial \log k = n\,(p - y), \qquad \partial\ell/\partial n = (p - y)\log(e/k), \tag{8}$$

so that $|\partial\ell/\partial \log e| \leq n$. Moreover, for fixed errors, ℓ is convex in the natural parameters $(n, -n \log k)$.

Why the Hill function. Unlike the Weibull CDF $1 - \exp(-(e/k)^n)$, which violates (iii) and (v), and a logistic function of the raw error, which violates (iii)–(v), the Hill function satisfies (iii)–(v); by (v), a $y = 0$ sample with a very large error, such as an unlabeled anomaly, has bounded influence (Appendix B).

4 CRETAP: Converting Reconstruction Error To Anomaly Probabilities

CRETAP has four components (Algorithm 1, Appendix A): a reconstruction network whose element-wise residuals are averaged into the error $e(x)$ (Section 4.1); the Hill function (Section 3), which maps $e(x)$ to $p_{k,n}(x)$; a BCE objective that trains both jointly and controls the calibration error (Sections 4.2 and 4.3); and an attribution map, from the same forward pass, that expresses each element's contribution to the error in units of k (Section 4.4).

4.1 Element-wise residuals and reconstruction error

Let $\psi_\theta : \mathbb{R}^d \to \mathbb{R}^d$ be a reconstruction network with parameters θ whose output $\hat{x} = \psi_\theta(x)$ is aligned with the input coordinates. For each element j of x (a feature, pixel–channel entry, or embedding coordinate), the element-wise residual and the reconstruction error are

$$r_j(x) = \big(x_j - \hat{x}_j\big)^2, \qquad e(x) = \frac{1}{d}\sum_{j=1}^{d} r_j(x), \tag{9}$$

and $r(x) = (r_1(x), \ldots, r_d(x)) \in [0,\infty)^d$ is the residual vector, which CRETAP retains for attribution. The squared error is not essential: Propositions 1 and 2 require only $e(x) \geq 0$, and Section 4.4 also requires the additive form of equation 9 with $r_j \geq 0$. Averaged and compared across elements, residuals must be commensurate: Proposition 1(iv) absorbs a common rescaling of all residuals but not an element-specific one, so we min–max scale each element with training-set statistics. Any input-aligned architecture can serve as ψ_θ; we use fully connected autoencoders for tabular and embedding data and convolutional ones for images (C channels on an $H \times W$ pixel grid, flattened to $d = CHW$; Appendix H).

4.2 End-to-end objective

To use every labeled anomaly for both detection and calibration, CRETAP trains the network and the Hill function jointly, with $k = \exp(\kappa)$ and $n = \exp(\nu)$ for $\kappa, \nu \in \mathbb{R}$, by minimizing the BCE of $\sigma(u(x))$, where u is the logit:

$$u(x) = n\,\big[\log\big(e(x) + \epsilon\big) - \log k\big], \qquad \sigma\big(u(x)\big) = h_{k,n}\big(e(x) + \epsilon\big), \tag{10}$$

$$\mathcal{L}_{\mathrm{BCE}}(\theta, \kappa, \nu) = \frac{1}{N}\sum_{i=1}^{N} \ell\big(y_i,\, \sigma(u(x_i))\big). \tag{11}$$

The offset $\epsilon > 0$ keeps the logarithm finite; the computed anomaly probability $\sigma(u(x))$ is $p_{k,n}(x)$ with $e + \epsilon$ in place of e, so Propositions 1 and 2 hold for this offset error, and Propositions 1(iv) and 2(iv) hold for e only up to the offset. The objective is computed stably from u, and its gradient reaches θ through e; Appendix A gives the two training protocols. Because equation 11 has no explicit reconstruction term, we measure, rather than assume, whether e is a faithful reconstruction error after training. At inference, one gradient-free forward pass returns $\sigma(u(x))$, the decision $\sigma(u(x)) \geq \tau$ at an operating point $\tau \in (0,1)$, and the attributions. If all targets are 0, the objective can be lowered by raising k as well as by lowering the errors, so calibration is not guaranteed (Appendix A); all our results use the fully labeled regime.

4.3 Calibration guarantees

Because $p_{k,n}$ depends on x only through $e(x)$, its calibration is governed by the posterior given the error alone, $\eta(e) = \Pr\big(y = 1 \mid e(x) = e\big)$. Let $R(k,n) = \mathbb{E}\big[\ell\big(y, p_{k,n}(x)\big)\big]$ be the population BCE, $R^\star_e = \inf_g \mathbb{E}\big[\ell\big(y, g(e(x))\big)\big]$ the smallest BCE of any measurable $g : [0,\infty) \to [0,1]$, and $\Delta_{\mathrm{BCE}}(k,n) = R(k,n) - R^\star_e$ the *excess BCE* of the Hill function; KL is the Kullback–Leibler divergence and $\mathrm{Ber}(q)$ the Bernoulli distribution with mean q.

Proposition 2 (Calibration of the Hill function)**.** *For all $k, n > 0$ (proof in Appendix C.2):*

(i) ***Calibration error is Hill function mismatch.***

$$\mathrm{CE}(p_{k,n}) = \mathbb{E}\,\big|\eta\big(e(x)\big) - h_{k,n}\big(e(x)\big)\big|. \tag{12}$$

(ii) ***BCE controls ECE.*** *For every binning $\mathcal{B}$,*

$$\Delta_{\mathrm{BCE}}(k,n) = \mathbb{E}\Big[\mathrm{KL}\big(\mathrm{Ber}(\eta(e(x))) \,\big\|\, \mathrm{Ber}(h_{k,n}(e(x)))\big)\Big], \tag{13}$$

$$\mathrm{ECE}_{\mathcal{B}}(p_{k,n}) \leq \mathrm{CE}(p_{k,n}) \leq \sqrt{\tfrac{1}{2}\,\Delta_{\mathrm{BCE}}(k,n)}, \tag{14}$$

so the excess BCE upper-bounds $2\,\mathrm{CE}(p_{k,n})^2$. As $R^\star_e$ does not depend on (k,n), lowering the BCE over (k,n) lowers Δ_{BCE} equally and tightens equation 14.

(iii) ***Weak calibration at stationarity.*** *Let $\hat{R}(k,n) = \frac{1}{N}\sum_{i=1}^{N} \ell(y_i, p_i)$ be the unweighted empirical BCE, with $p_i = p_{k,n}(x_i)$ and $e(x_i) > 0$. At any stationary point of $\hat{R}$ with respect to (k,n), in particular at any stationary point of training,*

$$\frac{1}{N}\sum_{i=1}^{N}(p_i - y_i) = 0, \qquad \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)\log e(x_i) = 0. \tag{15}$$

Consequently, the single-bin ECE vanishes on the training set (mean calibration), and the Platt-type recalibration $\sigma(\gamma\,\sigma^{-1}(p_i) + \delta)$, which includes temperature scaling ($\delta = 0$), has its lowest in-sample BCE at $(\gamma, \delta) = (1, 0)$ (unit slope, zero intercept).

(iv) ***Invariance.*** *Under the reparameterization of Proposition 1(iv), the errors e and $T \circ e$ yield the same anomaly probabilities, hence the same* CE *and* $\mathrm{ECE}_{\mathcal{B}}$.

What the objective optimizes. As $R_e^\star$ is attained at $g = \eta$, $R = R_e^\star + \Delta_{\mathrm{BCE}}$ splits the BCE into a refinement term $R_e^\star$, which depends only on how well e separates the classes, as AUPR reflects, and a calibration term Δ_{BCE}, which ECE reflects (Bröcker, 2009; Kull & Flach, 2015); we report both. Post-hoc calibration reduces Δ_{BCE} only through (k,n) after the network is trained; CRETAP also passes the gradient of Δ_{BCE} to θ but minimizes $R_e^\star + \Delta_{\mathrm{BCE}}$ rather than Δ_{BCE} alone, so equation 14 does not by itself imply a lower ECE than post-hoc calibration. Appendix D explains why Proposition 2(iii) needs the logistic form and treats prior shift.

Limitations. Proposition 2(iii) presupposes a stationary point, which requires the training errors of the two classes to overlap: under complete separation, the BCE has no minimizer (Albert & Anderson, 1984), and when the anomalous errors all exceed the nominal ones, the anomaly probabilities become overconfident unless training stops early or n is regularized (Appendix D). It is also in-sample; its transfer to test data depends on how well the error distribution generalizes, since the network, unlike the Hill function, has high capacity.

4.4 Attributions from the forward pass

Since $e(x)$ averages the residuals, which exist before the Hill function is applied, element j contributes $r_j(x)/d$ to it; its attribution is this contribution in units of the learned boundary k,

$$b_j(x) = \frac{r_j(x)}{d\,k}, \qquad \sum_{j=1}^{d} b_j(x) = \frac{e(x)}{k}; \tag{16}$$

$b(x) = (b_1(x), \ldots, b_d(x))$ is the attribution map, and for images, identifying each pixel z with its C element indices gives $\tilde{b}_z(x) = \sum_{j \in z} b_j(x)$, with $\sum_z \tilde{b}_z(x) = e(x)/k$.

Corollary 1 (Completeness of the attributions)**.** *For every sample x (proof in Appendix C.3),*

$$\sum_{j=1}^{d} b_j(x) = \frac{e(x)}{k} = \left(\frac{p_{k,n}(x)}{1 - p_{k,n}(x)}\right)^{1/n}, \tag{17}$$

and for every $\tau \in (0,1)$, $p_{k,n}(x) \geq \tau$ if and only if $\sum_j b_j(x) \geq [\tau/(1-\tau)]^{1/n}$. In particular, the attributions sum to more than one exactly when $p_{k,n}(x) > \frac{1}{2}$. They are invariant to a common rescaling of the residuals, $r_j \mapsto c\,r_j$ with $k \mapsto ck$.

For $\sigma(u(x))$, the second equality and the threshold hold up to ϵ/k. The identity plays the role of the completeness axiom of Integrated Gradients (Sundararajan et al., 2017) and the local accuracy of SHAP (Lundberg & Lee, 2017), but comes from the forward pass alone; it fails for non-additive aggregations, e.g., the maximum residual.

CRETAP Score. Figures in Sections 6.2 and 6.3 show the *CRETAP Score* $s_j(x) = d\,b_j(x) = r_j(x)/k$, which averages to $e(x)/k$ and is summed over the channels of each pixel for images (Appendix E). Unlike b_j, s_j does not scale with d, so CRETAP Scores are comparable across datasets of different dimensionality; as k is one number per trained model and no per-sample normalization, smoothing, or clipping is applied, they are also comparable between samples scored by the same model.

Cost. Being deterministic functions of the residuals, the attributions come from the same forward pass as the anomaly probability and the decision, without backpropagation, perturbations, a surrogate explainer, or a localization head (costs in Appendix E). A conventional residual map costs the same, so the difference is the unit, not the speed, and we claim no superior attribution quality.

5 EXPERIMENTAL SETTINGS

RQ1 (third contribution; Section 6.1): without post-hoc calibration, does CRETAP detect competitively with a low ECE? **RQ2** (Sections 4.4, 6.2, 6.3): where are its attributions, in units of k, largest? **RQ3** (Section 6.1): where does it stop being competitive?

Data. We use the 57 ADBench datasets (Han et al., 2022): 47 tabular, 5 image, and 5 text, the latter two as pretrained embeddings. As reconstruction presupposes that nominal samples dominate, the primary scope is the 29 datasets whose anomaly ratio (contamination) is at most the benchmark's median of 5% (20 tabular, 5 text, 4 image), a value not tuned by us; Appendix F covers the other 28. For RQ2, CRETAP is trained on the pixel images behind two embedding sets, for pixel-aligned residuals: 15 MVTec AD categories (Bergmann et al., 2019a), re-split as the release keeps all defective images in its test folder, and five MNIST-C corruptions (Mu & Gilmer, 2019); seven are shown (Appendices H.3, H.4).

Baselines. We compare with 28 of the 30 ADBench detectors at their published configurations, without tuning or post-hoc calibration (Appendix H.1); ResNet and FTTransformer, which ADBench trains on GPU, are omitted to compare CPU-trained models only, and with their published results CRETAP would fall from third to fourth in mean AUPR, with none of its significance verdicts on the 29 datasets changed (Appendix G.5). Calibration is compared among eight natively probabilistic detectors: CRETAP, NB, MLP, RF, and the *boosting group* (CatBoost, XGBoost, LightGBM, XGBOD); SVM, whose probabilities come from internal Platt calibration, enters detection only.

Protocol and metrics. All results except the label-noise runs of Section 6.1 use the fully labeled regime of Section 4.2 (every training anomaly labeled), with three seeds per dataset. CRETAP is tuned per dataset on a stratified 10% validation split of the training data, which also drives early stopping, so it is fit on the other 90% (Appendix H.2). Detection is measured by AUPR, the primary axis because an alarm for rare anomalies is judged by precision among its top-ranked samples and PR curves resolve differences that ROC curves compress under heavy class imbalance (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015), and by the area under the receiver operating characteristic curve (AUROC), both in % (↑; identical for e and the probability by Proposition 1(ii)); calibration by the ECE of equation 3 with 10 equal-width bins (↓). Among the eight, a model lies on a Pareto frontier if no other is at least as good on both axes and strictly better on one; the primary frontier pairs mean AUPR with mean ECE (secondary ones: Appendix G.1). *Significant* means that the Wilcoxon signed-rank test over datasets rejects the null of no difference at $\alpha = 0.05$ after Holm correction over all pairs (Appendix H.5); non-significance is not evidence of equality.

6 RESULTS

6.1 DETECTION, CALIBRATION, AND SCOPE (RQ1, RQ3)

Detection. CRETAP ranks third of 29 methods by mean AUPR (71.19; behind CatBoost and SVM) and fourth by mean AUROC (94.11); no tested baseline is significantly better on either metric, and CRETAP is significantly better than 16 baselines on AUROC and 17 on AUPR (Appendix G.1). CatBoost leads on both, by 2.0 AUPR and 0.58 AUROC points (absolute), and wins 14 datasets against CRETAP by AUROC (CRETAP wins 11; 4 ties); CRETAP differs significantly from no boosting model, but we claim no equality.

Calibration. Without post-hoc calibration, CRETAP has the second-lowest mean ECE (0.0073, against 0.0067 for XGBOD, a non-significant absolute gap of 0.0006) and the best average ECE rank (2.93); it is the best-calibrated model on 9 of the 29 datasets, more than any other, but ranks as low as seventh (median second; Appendix G.2). Its ECE is significantly lower than that of NB, MLP, LightGBM, and XGBoost and not significantly different from CatBoost's, XGBOD's, and RF's. Its train–test ECE gap is the smallest of the eight by the mean (0.0017), second to MLP's by the median (Appendix G.3), so its calibration largely carries over to test data, which Section 4.3 leaves open.

Table 1: Detection and calibration on the 29 datasets with anomaly ratio $\leq 5\%$ (fully labeled, no post-hoc calibration): means over datasets of means over three seeds, $\pm$ std over seeds. ECE rank: mean per-dataset rank; #best: datasets with lowest ECE; bold/underline: best/second of the eight; •: AUPR–ECE Pareto frontier. All methods: App. G.1.

Model	AUPR ↑	AUROC ↑	ECE ↓	ECE rank ↓	#best ↑	Pareto
CatBoost	**73.20** $\pm$ 0.85	**94.69** $\pm$ 0.26	0.0087 $\pm$ 0.0002	3.45	2	•
CRETAP (ours)	<u>71.19</u> $\pm$ 1.73	94.11 $\pm$ 0.92	<u>0.0073</u> $\pm$ 0.0001	**2.93**	**9**	•
XGBoost	70.72 $\pm$ 1.86	94.52 $\pm$ 0.40	0.0139 $\pm$ 0.0003	5.07	0	
MLP	69.43 $\pm$ 2.01	93.27 $\pm$ 1.01	0.0115 $\pm$ 0.0005	4.93	2	
XGBOD	69.40 $\pm$ 1.15	<u>94.61</u> $\pm$ 0.27	**0.0067** $\pm$ 0.0002	<u>3.38</u>	<u>8</u>	•
LightGBM	69.33 $\pm$ 1.75	91.30 $\pm$ 1.68	0.0145 $\pm$ 0.0011	4.79	5	
RF	68.61 $\pm$ 0.96	92.54 $\pm$ 0.24	0.0087 $\pm$ 0.0002	4.17	2	
NB	44.76 $\pm$ 2.11	83.29 $\pm$ 0.84	0.2315 $\pm$ 0.0097	7.28	1	

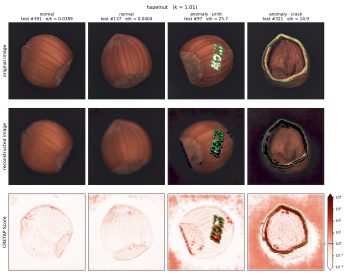

Figure 1: MVTec AD `hazelnut`, seed 1: input, reconstruction (clipped to $[0, 1]$), and CRETAP Score $\tilde{s}_z$ (channel-summed r_j/k, equation 22) on one log scale in units of k for all image figures (mark: $\tilde{s}_z = 1$); two nominal, two anomalous test images, headed by e/k; green: defect outline.

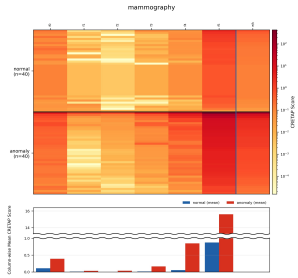

Figure 2: CRETAP Scores s_j, `mammography` ($d = 6$), seed 1: 40 random nominal (top) and 40 random anomalous test samples (log scale, e/k last); below, feature means (blue nominal, red anomalous; broken axis, App. H.4).

Both axes. CRETAP wins neither axis but is the only one of the eight in the top two on both, lying on the primary frontier with CatBoost and XGBOD (2.0 AUPR points below CatBoost at lower ECE; 1.8 above XGBOD at 0.0006 higher ECE). The answer depends on the scope and the detection axis: on mean AUROC and ECE, XGBOD dominates CRETAP (94.61 vs. 94.11; 0.0067 vs. 0.0073).

Scope and robustness (RQ3). Above a 5% anomaly ratio, CRETAP falls behind the boosting group and RF: on those 28 datasets it ranks sixth of 29 on both detection metrics, significantly behind CatBoost, LightGBM, and RF, and fourth of eight by mean ECE, significantly behind CatBoost, the only model on any frontier (Appendix F). The two-axis result is sensitive to the boundary: CRETAP is on the primary frontier at all 13 boundaries tested between 5% and 11.3%, but at none of the 15 below 5%, where the embedding datasets at 4.95–5.00% leave the scope. Under ADBench's noise on the 20 low-contamination tabular datasets (Appendix G.4; under label noise, CRETAP runs without early stopping, a post hoc choice, as noisy validation labels corrupt the stopping signal), its AUROC loss under duplicated anomalies or label noise does not differ significantly from any boosting model's, but it degrades significantly more than GANomaly at 25% label noise, and at 10% its ECE rises significantly more than XGBoost's and LightGBM's. Irrelevant features are a genuine weakness that re-tuning does not remove: at 50% added features CRETAP loses 3.6 AUROC points against 1.2–1.6 for the boosting group (unresolved on 20 datasets) and, on all 47 tabular datasets, 3.3 against 0.7–1.0 for the boosting group and 1.2 for RF, significantly more than each of them and NB (we hypothesize that reconstruction spends capacity on every element).

6.2 Attribution Maps on Images (RQ2)

On the anomalous images shown for five MVTec AD categories, the largest CRETAP Scores, $\tilde{s}_z \geq 10^3$, fall inside the annotated defects (Figure 1; Appendix I.1); for the two MNIST-C corruptions shown, which the release does not annotate, we can only say that the largest scores coincide with the

corrupted pixels. Over all nominal test images, the median fraction of pixels with $\tilde{s}_z \geq 1$ is 0.6–12% per dataset, except on `cable` (50%). Elevated scores also appear on backgrounds, borders, and in square blocks (we hypothesize: convolution window and stride). Section 4.2 leaves open whether e remains a faithful reconstruction error; on images, the reconstruction departs from the input at defects: without a reconstruction term, the BCE raises the error of anomalous samples ($\partial \ell / \partial \log e = n(p-1) < 0$ by equation 8), and on `cable`, the weakest of the seven detectors (AUROC 76.6), even nominal reconstructions carry a global color offset. The displayed images are the clearest cases (Appendix H.4); over whole test splits, the median anomalous e/k is 1.16 on `bottle`, 1.31 on `hazelnut`, and 0.877 on `cable`, below the boundary (Table 18 in Appendix I.1), so the figures show confident cases, do not show that the largest scores coincide with defects in general, and are not compared with post-hoc explainers.

6.3 Attribution maps on tabular data (RQ2)

On `mammography`, feature `f5` carries the largest mean CRETAP Score in both classes, 0.88 for nominal and 15.6 for anomalous samples (Figure 2). Across the 20 low-contamination tabular datasets, anomalous samples' top-3 attributions recur on the same features more often than nominal samples' on 16 of the 19 datasets where this is defined (not on `ALOI` and `speech`; `http` tied; Appendix I.2), which locates where reconstruction fails, not what causes the anomaly.

7 Limitations

Section 6.1 bounds the use of CRETAP: nominal samples must dominate, and many irrelevant features hurt it more than tree ensembles. Baselines ran untuned and uncalibrated, whereas CRETAP was tuned per dataset on validation splits that overlap other seeds' test splits (Appendix H.2), which may favor it; calibrating the baselines post hoc would narrow the gap wherever enough held-out anomalies exist, though the validation split, also used by CRETAP for early stopping, holds fewer than ten on 15 of the 29 datasets on seed average. No ablation separates joint from post-hoc fitting of (k, n) or compares with ABC's parameter-free map (Yamanaka et al., 2019), and whether e stays a faithful reconstruction error is examined only qualitatively, on images (Section 6.2). ECE uses 10 equal-width bins, a number not fixed in advance, on test splits with as few as one anomaly. The number of labeled anomalies is never varied, and apart from label noise, unlabeled training anomalies (Proposition 1(v)) are not evaluated. The attributions are claimed to be cheap and in units of k, not better explanations (Sections 6.2–6.3), and all defective MVTec AD images come from the release's test folder, so acquisition differences could be confounded with defects (unexamined). Tabular and text evaluation is confined to ADBench (no equipment-monitoring data), without deployment data.

8 Conclusion

CRETAP makes reconstruction models natively probabilistic by learning a Hill function of the error jointly with the network under BCE, so scarce labeled anomalies serve detection and calibration; the Hill function preserves rankings and absorbs error units, the excess BCE controls the calibration error, and the attributions sum to $e(x)/k$ in one forward pass. On 29 ADBench datasets with $\leq 5\%$ contamination, CRETAP ranks third of 29 methods in mean AUPR, has the second-lowest mean ECE of eight natively probabilistic detectors without post-hoc calibration, and lies on the AUPR–ECE Pareto frontier with CatBoost and XGBOD, a position that depends on detection axis and scope. Next steps (Section 7): higher anomaly ratios, irrelevant features (feature weighting), post-hoc-calibrated baselines, scarcer and unlabeled training anomalies, and quantitative attribution evaluation.

Reproducibility statement

Appendix A gives the training and inference procedure; Appendix H the datasets, baselines, architectures, search spaces, splits, and statistical procedure; Appendix H.3 the construction of the image data, checked against published file hashes; and Appendices C.1–C.3 the proofs. Per-dataset calibration results are in Appendix G.2 and per-dataset attribution statistics in Appendix I.2.

AI USE STATEMENT

The authors conceived the method and its conceptual framework, designed the study and the experiments, formed the hypotheses, and interpreted the results; every design decision is theirs. Within that design, we used generative AI tools to provide feedback on research methodology or experiments, to implement methods, to clean and reformat dataset, to formulate mathematical claims, to provide critical ingredients for proving mathematical claims, to assist in the writing of proofs, and to assist with translation, and, among the recommended tasks, for code, figures, artifacts, drafting and language editing, references, literature search and summarization, and brainstorming. We have not used them to design research methodology or experiments, to help develop theoretical models or conceptual frameworks, to propose or refine hypotheses, or to interpret results; generating synthetic data sets and supporting qualitative and thematic data analysis do not apply to this work.

We reviewed all AI-assisted work, including every mathematical statement and proof, and take responsibility for the final content, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Adelin Albert and John A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019a. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Bergmann_MVTec_AD_--_A_Comprehensive_Real-World_Dataset_for_Unsupervised_Anomaly_CVPR_2019_paper.html.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 372–380, 2019b. doi: 10.5220/0007364503720380.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.

David R. Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4): 562–565, 1958.

Jesse Davis and Mark H. Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, 2006. doi: 10.1145/1143844.1143874.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, volume 12664 of *Lecture Notes in Computer Science*, pp. 475–489. Springer, 2021. doi: 10.1007/978-3-030-68799-1_35.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9737–9746, 2022.

Jing Gao and Pang-Ning Tan. Converting output scores from outlier detection algorithms into probability estimates. In *Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM)*, pp. 212–221, 2006.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019.

Sylvain Goutelle, Michaël Maurin, Florent Rougier, Xavier Barbaut, Laurent Bourguignon, Michel Ducher, and Pascal Maire. The Hill equation: A review of its capabilities in pharmacological modelling. *Fundamental & Clinical Pharmacology*, 22(6):633–648, 2008. doi: 10.1111/j.1472-8206.2008.00633.x.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. ReContrast: Domain-specific anomaly detection via contrastive reconstruction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10721–10740, 2023.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32142–32159, 2022.

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. MambaAD: Exploring state space models for multi-class unsupervised anomaly detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 71162–71187, 2024.

Archibald Vivian Hill. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *The Journal of Physiology*, 40(Suppl):iv–vii, 1910. doi: 10.1113/jphysiol.1910.sp001386.

Archibald Vivian Hill. The combinations of haemoglobin with oxygen and with carbon monoxide. I. *Biochemical Journal*, 7(5):471–480, 1913. doi: 10.1042/bj0070471.

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Interpreting and unifying outlier scores. In *Proceedings of the 2011 SIAM International Conference on Data Mining (SDM)*, pp. 13–24, 2011.

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Part I*, pp. 68–85. Springer, 2015.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3787–3798, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Norman Mu and Justin Gilmer. MNIST-C: A robustness benchmark for computer vision. arXiv preprint arXiv:1906.02337, 2019.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015. doi: 10.1609/aaai.v29i1.9602.

Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 353–362, 2019. doi: 10.1145/3292500.3330871.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021. doi: 10.1145/3439950.

Guansong Pang, Chunhua Shen, Huidong Jin, and Anton van den Hengel. Deep weakly-supervised anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1795–1807, 2023. doi: 10.1145/3580305.3599302.

Lorenzo Perini, Vincent Vercruyssen, and Jesse Davis. Quantifying the confidence of anomaly detectors in their example-wise predictions. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Part III*, pp. 227–243. Springer, 2020.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, pp. 151. BMVA Press, 2018.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402, 2018.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkgH0TEYwH.

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. *Proceedings of the IEEE*, 109(5):756–795, 2021. doi: 10.1109/JPROC.2021.3052449.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017.

Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J. Pencina, and Ewout W. Steyerberg. A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176, 2016.

Yuki Yamanaka, Tomoharu Iwata, Hiroshi Takahashi, Masanori Yamada, and Sekitoshi Kanai. Autoencoding binary classifiers for supervised anomaly detection. In *Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, volume 11671 of *Lecture Notes in Computer Science*, pp. 647–659. Springer, 2019. doi: 10.1007/978-3-030-29911-8_50.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4571–4584, 2022.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002. doi: 10.1145/775047.775151.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1_53.

Yue Zhao and Maciej K. Hryniewicki. XGBOD: Improving supervised outlier detection with unsupervised representation learning. In *International Joint Conference on Neural Networks*, pp. 1–8. IEEE, 2018. doi: 10.1109/IJCNN.2018.8489605.

Yingjie Zhou, Xucheng Song, Yanru Zhang, Fanxing Liu, Ce Zhu, and Lingqiao Liu. Feature encoding with autoencoders for weakly supervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2454–2465, 2022. doi: 10.1109/TNNLS.2021.3086137.

A Training and inference procedure

Algorithm 1 gives the procedure of Section 4.2. Of the two training protocols, CRETAP-JOINT minimizes equation 11 from initialization and CRETAP-PRE first minimizes the mean error of the samples with $y_i = 0$ over θ alone, then equation 11; both start from (k_0, n_0) and train (k, n) by BCE alone, and the protocol is chosen per dataset for tabular and embedding data, while images use CRETAP-JOINT (Appendix H). One Adam optimizer over (θ, κ, ν) without weight decay is shared by all stages; in the reconstruction stage of CRETAP-PRE, κ and ν belong to the optimizer but receive no gradient, so they keep their initial values. The hyperparameters of Algorithm 1 are k_0, n_0, ϵ, the learning rate, the batch size, the numbers of epochs E_{pre} and E_{BCE}, and the protocol of each dataset; training details are in Appendix H.

Targets that are all zero. When both target classes are present, samples with $y_i = 0$ and $y_i = 1$ supply opposing signals. When the available targets are all zero, the loop remains well defined but receives only the nominal-side signal: by equation 10, the objective depends on the error only through $(e + \epsilon)/k$ and n, so it can be decreased either by lowering the errors or by raising k, and nothing ties k to the scale of the nominal errors. That regime therefore keeps $\sigma(u(x))$ in $(0, 1)$, but its calibration is neither guaranteed nor inferred from that range.

B Design alternatives to the Hill function

This section expands the comparisons that follow Proposition 1.

Platt scaling and ranking. Proposition 1(i) shows that the Hill function is Platt scaling (Platt, 1999) in log-error space; the novelty of CRETAP therefore lies not in the functional form but in training it jointly with the reconstruction network. Proposition 1(ii) implies that reporting anomaly probabilities instead of errors leaves every ranking metric unchanged, and that an operating point chosen on the probability scale can be communicated as an explicit error threshold.

Algorithm 1 CRETAP training and inference. The reconstruction network is a free choice; the aggregation, the Hill function, the BCE stage, and the attribution maps are the method. $\mathbb{I}[\cdot]$ is the indicator function.

Require: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where the targets may contain only zeros; reconstruction network ψ_θ (for images, $d = C \cdot H \cdot W$ and ψ_θ reshapes internally, returning a flattened $\hat{x}$); protocol $\in \{\text{CRETAP-JOINT}, \text{CRETAP-PRE}\}$; epochs $E_{\text{pre}}, E_{\text{BCE}}$; batch size; learning rate; $k_0, n_0 > 0$; $\epsilon > 0$; operating point $\tau \in (0, 1)$

initialize θ; $\kappa \leftarrow \log k_0$; $\nu \leftarrow \log n_0$
create one Adam optimizer over (θ, κ, ν) with the learning rate and no weight decay, shared by every stage below
if protocol $=$ CRETAP-PRE **then** ▷ reconstruction stage
  for E_{pre} epochs, each mini-batch $\mathcal{X}_0$ of samples with $y_i = 0$ **do**
    Adam step on ∇_θ of the mean of $e(x_i)$ over $x_i \in \mathcal{X}_0$ ▷ κ, ν get no gradient
  end for
end if
for E_{BCE} epochs, each mini-batch $\mathcal{S} \subseteq \mathcal{D}$ **do** ▷ BCE stage
  compute $r_j(x_i)$ and $e(x_i)$ by equation 9, and $u(x_i) = \exp(\nu)[\log(e(x_i) + \epsilon) - \kappa]$ by equation 10, for $(x_i, y_i) \in \mathcal{S}$
  Adam step on $\nabla_{\theta,\kappa,\nu}$ of the mean of $\ell\big(y_i, \sigma(u(x_i))\big)$ over $\mathcal{S}$, computed stably from u (`BCEWithLogits`) ▷ reaches θ through e
end for
$k \leftarrow \exp(\kappa)$; $n \leftarrow \exp(\nu)$
inference: set ψ_θ to evaluation mode (dropout disabled) and run one forward pass without gradients
compute $r_j(x)$, $e(x)$, $u(x) = n[\log(e(x) + \epsilon) - \log k]$, and the anomaly probability $\sigma(u(x))$; $\hat{y}_\tau(x) \leftarrow \mathbb{I}[\sigma(u(x)) \geq \tau]$ ▷ equation 6
residual map $r_j(x)/d$, with $\sum_j r_j(x)/d = e(x)$, and attribution $b_j(x) \leftarrow r_j(x)/(d\,k)$, with $\sum_j b_j(x) = e(x)/k$ ▷ b: equation 16
images only (Section 6.2): $\tilde{b}_z(x) \leftarrow \sum_{j \in z} b_j(x)$ for each pixel z, so $\sum_z \tilde{b}_z(x) = e(x)/k$
return e, $\sigma(u)$, $\hat{y}_\tau$, r/d, b, k, n, and $\tilde{b}$ for images

Symmetry and the Weibull CDF. Proposition 1(iii) makes k exactly the error level at which a sample is equally likely to be nominal or anomalous. This does not hold for asymmetric functions on $[0, \infty)$: for the Weibull CDF $1 - \exp(-(e/k)^n)$, the probability at $e = k$ is $1 - \exp(-1) \approx 0.63$. In distributional terms, $h_{k,n}$ is the CDF of a log-logistic distribution with median k, so $p_{k,n}(x)$ can be regarded as the probability that the error exceeds a random tolerance whose median is k.

Units of the error. Proposition 1(iv) covers the arbitrary choices that make raw errors non-comparable: sum versus mean aggregation ($c = 1/d$), squared versus root errors ($a = 1/2$, e.g., MSE versus RMSE), and a change of measurement units that rescales all residuals by a common factor. A logistic function of the raw error is closed under affine transforms of e but not under power transforms.

Bounded influence. By Proposition 1(v), no single sample has a gradient in $\log e$ larger than n in magnitude. With $p = h_{k,n}(e)$, the per-sample BCE is

$$\ell(0, p) = \log\big(1 + (e/k)^n\big), \qquad \ell(1, p) = \log\big(1 + (k/e)^n\big), \tag{18}$$

so BCE contracts nominal errors below k and pushes anomalous errors above it, and the BCE of a $y = 0$ sample grows only logarithmically in e. A sample with $y = 0$ and a very large error, such as an unlabeled anomaly in the training set, therefore has bounded influence, whereas under the Weibull CDF its BCE $(e/k)^n$ and the gradient are unbounded.

C PROOFS

C.1 PROOF OF PROPOSITION 1

Proof. (i) For $e > 0$, dividing the numerator and denominator of equation 4 by e^n gives $h_{k,n}(e) = \big[1 + (k/e)^n\big]^{-1} = \big[1 + \exp(-n[\log e - \log k])\big]^{-1}$; the odds follow from $1 - h_{k,n}(e) = k^n/(k^n + e^n)$.

(ii) $h_{k,n}$ is continuous with $h_{k,n}(0) = 0$, $\lim_{e\to\infty} h_{k,n}(e) = 1$, and $h'_{k,n}(e) = nk^n e^{n-1}/(k^n + e^n)^2 > 0$ for $e > 0$. Solving $(e/k)^n = \tau/(1-\tau)$ yields the inverse $h^{-1}_{k,n}(\tau) = k[\tau/(1-\tau)]^{1/n}$ and hence equation 6; a strictly increasing map preserves both strict order and ties, which gives the claim on rankings.

(iii) By (i), the odds at kt and k/t are t^n and t^{-n}, so $h_{k,n}(kt) = t^n/(1+t^n)$ and $h_{k,n}(k/t) = 1/(1+t^n)$; setting $t = 1$ gives the first claim.

(iv) $\big(T(e)/T(k)\big)^{n/a} = \big(e^a/k^a\big)^{n/a} = (e/k)^n$, so both sides of equation 7 have the same odds by (i) when $e > 0$, and both vanish at $e = 0$. The map $(k, n) \mapsto (T(k), n/a)$ is a bijection of $(0,\infty)^2$, which gives the second statement.

(v) Let $u = n[\log e - \log k]$ be the logit, so that $p = \sigma(u)$ and $\partial\ell/\partial u = p - y$; equation 8 follows from the chain rule and the bound from $|p - y| \le 1$. Writing $u = \alpha \log e + \beta$ with $(\alpha, \beta) = (n, -n \log k)$, ℓ is the BCE of $\sigma(u)$, which is convex in u, composed with a map that is affine in (α, β), hence convex. □

C.2 Proof of Proposition 2

Proof. (i) By Proposition 1(ii), $h_{k,n}$ is injective, so $p_{k,n}(x)$ and $e(x)$ generate the same σ-algebra and $\Pr\big(y = 1 \mid p_{k,n}(x)\big) = \eta\big(e(x)\big)$ almost surely; substitute into equation 2.

(ii) Write $p = p_{k,n}(x)$ and $\eta = \eta(e(x))$. Each event $\{p \in B_m\}$ is determined by $e(x)$, so the tower property and Jensen's inequality give $\big|\mathbb{E}[y - p \mid p \in B_m]\big| = \big|\mathbb{E}[\eta - p \mid p \in B_m]\big| \le \mathbb{E}\big[|\eta - p| \,\big|\, p \in B_m\big]$; weighting by $\Pr(p \in B_m)$ and summing yields $\mathrm{ECE}_{\mathcal{B}} \le \mathrm{CE}$ by (i). Conditionally on $e(x)$, the BCE decomposes as $\mathbb{E}[\ell(y,p) \mid e(x)] = \mathcal{H}(\eta) + \mathrm{KL}\big(\mathrm{Ber}(\eta) \,\|\, \mathrm{Ber}(p)\big)$, where $\mathcal{H}$ is the binary entropy. The first term does not depend on p and the second vanishes at $g = \eta$, so $R^\star_e = \mathbb{E}[\mathcal{H}(\eta)]$ and equation 13 follows. Pinsker's inequality (Cover & Thomas, 2006) gives $|\eta - p| \le \sqrt{\frac{1}{2}\mathrm{KL}(\mathrm{Ber}(\eta) \,\|\, \mathrm{Ber}(p))}$, and Jensen's inequality for the concave square root gives the upper bound in equation 14.

(iii) By Proposition 1(v), $\partial\hat{R}/\partial \log k = -\frac{n}{N}\sum_i (p_i - y_i)$ and $\partial\hat{R}/\partial n = \frac{1}{N}\sum_i (p_i - y_i)\log\big(e(x_i)/k\big)$. Setting both to zero and using $n > 0$ gives equation 15. The single-bin ECE is $\big|\frac{1}{N}\sum_i (y_i - p_i)\big| = 0$. Since $\sigma^{-1}(p_i) = n \log e(x_i) - n \log k$ is affine in $\log e(x_i)$, equation 15 implies $\sum_i (p_i - y_i) = \sum_i (p_i - y_i)\,\sigma^{-1}(p_i) = 0$, so the gradient of the recalibration BCE vanishes at $(\gamma, \delta) = (1, 0)$; that objective is convex in (γ, δ), so $(1, 0)$ is a global minimizer.

(iv) By Proposition 1(iv), corresponding parameters yield identical anomaly probabilities, which have identical calibration errors. □

C.3 Proof of Corollary 1

Proof. The first equality in equation 17 follows from the definition of b_j in equation 16 and from equation 9, the second is Proposition 1(i), and the threshold statement is equation 6 divided by k. Invariance follows from Proposition 1(iv) with $a = 1$. □

D Calibration: canonical link, diagnostics, and prior shift

Why the logistic form matters. Proposition 2(iii) is specific to the Hill function. By Proposition 1(i), the Hill function is σ, the inverse of the canonical (logit) link of the Bernoulli likelihood, applied to an affine function of $\log e$, so its first-order conditions are the unweighted moment conditions equation 15, which coincide with the calibration intercept and slope conditions of Cox (1958) and with weak calibration in the hierarchy of Van Calster et al. (2016). For a non-canonical link, such as the complementary log-log link, whose inverse applied to $n\,[\log e - \log k]$ gives the Weibull CDF, stationarity yields weighted moment conditions, and mean calibration is not guaranteed.

Tightness and scope of the bound. The bound equation 14 vanishes for some (k, n) if and only if η coincides almost surely with a Hill function, i.e., has a logit affine in $\log e$ with positive slope. The bound is stated for CE, which upper-bounds every binned ECE, because binning systematically underestimates the calibration error of continuous-output models (Kumar et al., 2019).

Complete separation. If the training errors of the two classes do not overlap, the BCE has no minimizer (Albert & Anderson, 1984): when the anomalous errors all exceed the nominal ones,

n grows without bound and the anomaly probabilities become overconfident, and if training stops early or n is regularized, equation 15 holds only approximately.

Calibration diagnostics. The two parameters map onto the two classical calibration diagnostics: if $\eta = h_{k^\star,n^\star}$ for some $k^\star, n^\star > 0$, then, as functions of e,

$$\sigma^{-1}(\eta) = \frac{n^\star}{n}\,\sigma^{-1}(h_{k,n}) + n^\star \log\frac{k}{k^\star}, \tag{19}$$

so $n > n^\star$ corresponds to overconfidence (slope below one) and $k > k^\star$ to a systematic underestimation of the anomaly probability (positive intercept). Temperature scaling (Guo et al., 2017), which divides the logit by a positive constant, amounts to dividing n by that constant, and a prior shift is absorbed by k alone, as shown next.

Prior shift and class reweighting. Suppose the anomaly rate differs between the distribution on which the Hill function was trained (ρ_{tr}) and deployment (ρ_{te}), with the class-conditional error distributions unchanged (a prior shift). The standard adjustment of a probabilistic classifier to a new anomaly rate (Saerens et al., 2002) multiplies the odds by $\omega = \frac{\rho_{\mathrm{te}}/(1-\rho_{\mathrm{te}})}{\rho_{\mathrm{tr}}/(1-\rho_{\mathrm{tr}})}$. By Proposition 1(i), the adjusted anomaly probability is again a Hill function with the same shape n and the learned boundary k_{tr} rescaled to

$$k_{\mathrm{te}} = k_{\mathrm{tr}}\,\omega^{-1/n}, \tag{20}$$

so no retraining is needed. The same correction applies when training uses class weights or class-balanced sampling: the moment conditions equation 15 then hold for the reweighted training set, the anomaly probabilities are calibrated to the reweighted anomaly rate, and equation 20 maps them back to the true rate.

E Details on CRETAP Score

Completeness and units. Writing $\varphi(p) = [p/(1-p)]^{1/n}$, equation 17 states that the attributions sum to $\varphi\big(p_{k,n}(x)\big) - \varphi(0)$, the change in a monotone transform of the odds relative to a perfectly reconstructed baseline. Each CRETAP Score $s_j = d\,b_j = r_j/k$ is the residual of element j in units of k: $s_j > 1$ means that the residual of element j exceeds k, the error level at which a sample is equally likely to be nominal or anomalous. Because s_j does not scale with d, CRETAP Scores are comparable across datasets of different dimensionality. A single attribution satisfies $b_j \le \sum_{j'=1}^{d} b_{j'}$, with equality only when element j carries all of the error; the sum can therefore exceed 1 while every b_j is below 1, so a region can support a positive decision even when no single element would trigger it alone.

Residual map. The residual map $r(x)/d$, whose entries sum to $e(x)$, is exactly complete with respect to the error; this is the definition of a mean, and we claim no more for it. The attribution map b is the CRETAP-specific one: its entries are dimensionless multiples of k, so, for $\sigma(u(x))$ of equation 10, by equation 6 they sum to a quantity that crosses $1 - \epsilon/k$ exactly when $\sigma(u(x)) \ge \frac{1}{2}$. We therefore report b numerically: its entries are on one scale for every sample scored by the same trained model.

CRETAP Scores for tabular data and images. The CRETAP Score of Section 4.4 and its average over the elements are

$$s_j(x) = d\,b_j(x) = \frac{r_j(x)}{k} = \frac{\big(x_j - \hat{x}_j\big)^2}{k}, \qquad \frac{1}{d}\sum_{j=1}^{d} s_j(x) = \frac{e(x)}{k}, \tag{21}$$

so the CRETAP Scores of one tabular sample, whose columns are single-channel elements, average to $e(x)/k$, the quantity that the decision at $\tau = \frac{1}{2}$ thresholds (up to ϵ/k). For an image with C channels on an $H \times W$ grid, the figures shade, at each pixel z, the residual summed over the channels of that pixel in units of k,

$$\tilde{s}_z(x) = \frac{1}{k}\sum_{j\in z} r_j(x) = d\,\tilde{b}_z(x), \qquad \frac{1}{HW}\sum_{z} \tilde{s}_z(x) = \frac{C\,e(x)}{k}, \tag{22}$$

the channel sum of the element-wise CRETAP Scores of equation 21. The residual is taken in the scale the network sees, and the denominator k is one number per trained model, so nothing in the

denominator is computed from the sample displayed. Because CRETAP Scores span several orders of magnitude, the figures use logarithmic color scales: all image figures share one scale, 10^{-2} to 10^{4}, while each tabular figure uses its own, because feature counts and error ranges differ by orders of magnitude between tabular datasets; Appendix H.4 gives the drawing procedure for both.

Cost relative to post-hoc explainers. The saving is one backward pass relative to single-gradient attribution such as Grad-CAM (Selvaraju et al., 2017), one forward and one backward pass per integration step relative to Integrated Gradients (Sundararajan et al., 2017), and one forward pass per masked input relative to perturbation methods that evaluate masked inputs besides the original (Zeiler & Fergus, 2014; Petsiuk et al., 2018); relative to a conventional residual map, which costs the same, the difference is the unit and not the speed. We therefore position the attribution map as an alternative route to attribution, available whenever the reconstruction is input-aligned, rather than as a claim of superior attribution quality.

Gradient weighting at the residual layer. Gradient weighting would add nothing at this layer: from equation 9 and equation 10,

$$\frac{\partial u(x)}{\partial r_j(x)} = \frac{n}{d\left[e(x)+\epsilon\right]}, \qquad j = 1, \ldots, d, \tag{23}$$

which is identical for every j, so gradient weighting at the residual layer can only rescale the attribution map uniformly. This is specific to the residual layer and implies nothing about Grad-CAM at an arbitrary hidden layer.

F Results outside the primary scope

Table 2 applies the protocol unchanged outside the primary scope. On all 57 datasets CRETAP does not differ significantly from any member of the boosting group on AUROC, but CatBoost is 0.94 points higher on average with 12 wins against 38 losses, and on AUPR the gap to CatBoost is significant; CatBoost alone holds all three frontiers. Above 5% contamination CRETAP falls to sixth on both detection metrics, significantly behind CatBoost, LightGBM, and RF, and its ECE is significantly worse than CatBoost's.

The pattern follows from the assumption the method is built on. Reconstruction learns a model of nominal structure, and as anomalies become common that structure is itself contaminated: five of the eight datasets on which CRETAP trails the best boosting model by more than two AUROC points have anomaly ratios between 12.5% and 34.9%. The two industrial datasets in ADBench, MVTec-AD (23.7%) and `fault` (34.7%), fall in this group, where CRETAP trails by 3.2 and 5.3 AUROC points. The framework is a proposal for the low-contamination regime, not a general-purpose replacement for gradient boosting on tabular data.

Sensitivity to the scope boundary. The nine embedding datasets other than MVTec-AD are fixed by construction at 4.95–5.00%, so they fall on one side of the 5% median together. Moving the boundary of the primary scope changes the two-axis result: CRETAP lies on the AUPR–ECE frontier at all 13 boundaries tested between 5% and 11.3%, and at none of the 15 boundaries below 5%, where the embedding datasets pinned at 4.95–5.00% leave the scope (Table 3).

Table 2: The protocol of Section 5 outside the primary scope. Entries are for CRETAP; detection ranks (1 = best) are among 29 methods, ECE ranks among the eight natively probabilistic detectors. A model is listed under Pareto if it lies on at least one of the three frontiers (Section 5 and Appendix G.1).

Scope	AUROC (rank)	AUPR (rank)	ECE (rank)	Pareto
Anomaly ratio $\leq 5\%$ (29 sets)	94.11 (4)	71.19 (3)	0.0073 (2)	CRETAP, CatBoost, XGBOD
All 57 sets	95.06 (4)	79.54 (7)	0.0156 (3)	CatBoost only
Anomaly ratio $> 5\%$ (28 sets)	96.04 (6)	88.18 (6)	0.0241 (4)	CatBoost only

Table 3: Sensitivity of the AUPR–ECE frontier to the scope boundary ($\gamma_l = 100\%$, ten-bin ECE). A boundary is placed midway between each pair of adjacent anomaly ratios among the 57 datasets and is evaluated when at least ten datasets fall inside the scope, which leaves 43 of the 52 candidate boundaries and drops the nine lowest, whose scopes hold one to nine datasets. The 28 boundaries at or below 11.3% are listed, each by the largest anomaly ratio it admits; 15 of them lie below 5% and 13 between 5% and 11.3%. The row at 5.0000% is the primary scope, and frontier members are ordered by mean ECE.

Scope: anomaly ratio	# datasets	AUPR–ECE frontier
$\leq 2.4431\%$	10	RF, CatBoost
$\leq 2.4655\%$	11	CatBoost
$\leq 2.7248\%$	12	CatBoost
$\leq 2.8758\%$	13	CatBoost
$\leq 2.9044\%$	14	CatBoost
$\leq 3.0444\%$	15	XGBOD, CatBoost
$\leq 3.1679\%$	16	XGBOD, CatBoost
$\leq 3.4341\%$	17	XGBOD, CatBoost
$\leq 4.0541\%$	18	CatBoost
$\leq 4.2056\%$	19	CatBoost
$\leq 4.4843\%$	20	CatBoost
$\leq 4.9516\%$	21	CatBoost
$\leq 4.9881\%$	22	CatBoost
$\leq 4.9953\%$	23	XGBOD, CatBoost
$\leq 4.9971\%$	24	XGBOD, CatBoost
$\leq 5.0000\%$	29	XGBOD, CRETAP, CatBoost
$\leq 5.3331\%$	30	XGBOD, CRETAP, CatBoost
$\leq 5.9274\%$	31	XGBOD, CRETAP, CatBoost
$\leq 6.2041\%$	32	XGBOD, CRETAP, CatBoost
$\leq 6.2500\%$	33	XGBOD, CRETAP, CatBoost
$\leq 7.1511\%$	34	XGBOD, CRETAP, CatBoost
$\leq 7.4167\%$	35	XGBOD, CRETAP, CatBoost
$\leq 7.7519\%$	36	XGBOD, CRETAP, CatBoost
$\leq 9.1176\%$	37	XGBOD, CRETAP, CatBoost
$\leq 9.2069\%$	38	XGBOD, CRETAP, CatBoost
$\leq 9.4567\%$	39	XGBOD, CRETAP, CatBoost
$\leq 9.6122\%$	40	XGBOD, CRETAP, CatBoost
$\leq 11.2654\%$	41	XGBOD, CRETAP, CatBoost

G Additional results

G.1 Detection results of all methods

Secondary frontiers. Besides the primary AUPR–ECE frontier, we compute two secondary Pareto frontiers among the eight natively probabilistic detectors. One pairs the average per-dataset AUROC rank with the average per-dataset ECE rank, where ranks are taken among the eight models on each dataset, so that a few datasets cannot dominate the average; on the 29 primary datasets only CRETAP and CatBoost lie on it. The other pairs mean AUROC with mean ECE; there CRETAP is dominated, as XGBOD leads on both AUROC (94.61 vs. 94.11) and ECE (0.0067 vs. 0.0073), and CatBoost and XGBOD form the frontier.

The AUROC and AUPR ranks quoted in Section 6.1 are ranks of the mean among all 29 evaluated methods (Tables 4–6). By average per-dataset rank among the 28 methods that enter the statistical tests (DAGMM has no result on three datasets and is left out), CRETAP is second on AUROC (5.74, behind CatBoost at 4.59) and third on AUPR (also 5.74, behind CatBoost at 4.34 and SVM at 4.55), and no member of the boosting group differs significantly from it on either metric. On 26 of the 29 datasets CRETAP is within one point of, equal to, or ahead of the best boosting model on that dataset by AUROC, and it leads all four on 9.

Table 4: Detection of all 29 methods on the 29 datasets with anomaly ratio $\leq 5\%$ (primary scope) ($\gamma_l = 100\%$, three seeds). Means are over datasets of means over three seeds, $\pm$ std over seeds; the rank in parentheses is the rank of the mean among the 29 methods (1 = best). Rank is the average per-dataset rank among the 28 methods that enter the tests. Verdict is the ADBench critical-difference outcome against CRETAP (all-pairs Wilcoxon with Holm correction, $\alpha = 0.05$): W, CRETAP significantly better; L, significantly worse; –, no significant difference. DAGMM has no result on three of the primary datasets, so its mean is taken over the datasets on which it ran and is ranked among the 29 by that mean, but it is excluded from the per-dataset ranks and from the tests. Higher AUPR and AUROC are better, and a lower rank is better. Rows are ordered by mean AUPR.

Model	AUPR	Rank	Verdict	AUROC	Rank	Verdict
SVM	74.26 ± 2.00 (1)	4.55	–	93.76 ± 0.87 (5)	6.24	–
CatBoost	73.20 ± 0.85 (2)	4.34	–	94.69 ± 0.26 (1)	4.59	–
CRETAP	71.19 ± 1.73 (3)	5.74	–	94.11 ± 0.92 (4)	5.74	–
XGBoost	70.72 ± 1.86 (4)	6.09	–	94.52 ± 0.40 (3)	6.02	–
DeepSAD	69.76 ± 1.11 (5)	6.36	–	92.82 ± 0.25 (9)	8.16	–
MLP	69.43 ± 2.01 (6)	7.40	–	93.27 ± 1.01 (6)	7.95	–
XGBOD	69.40 ± 1.15 (7)	7.10	–	94.61 ± 0.27 (2)	6.16	–
LightGBM	69.33 ± 1.75 (8)	6.19	–	91.30 ± 1.68 (12)	7.12	–
RF	68.61 ± 0.96 (9)	7.28	–	92.54 ± 0.24 (10)	8.31	–
DevNet	65.53 ± 1.57 (10)	8.98	–	92.99 ± 0.95 (7)	7.78	–
PReNet	65.34 ± 1.58 (11)	9.40	–	92.46 ± 1.31 (11)	9.14	–
FEAWAD	64.04 ± 4.55 (12)	9.91	W	92.97 ± 0.39 (8)	8.91	–
NB	44.76 ± 2.11 (13)	16.38	W	83.29 ± 0.84 (13)	16.09	W
PCA	29.76 ± 1.30 (14)	19.72	W	76.49 ± 0.55 (17)	19.00	W
CBLOF	29.40 ± 0.51 (15)	17.67	W	79.22 ± 1.34 (14)	16.76	W
IForest	29.24 ± 0.54 (16)	19.34	W	77.35 ± 0.61 (16)	18.03	W
REPEN	28.87 ± 3.10 (17)	18.26	W	77.68 ± 1.90 (15)	17.47	W
HBOS	27.82 ± 0.74 (18)	20.24	W	75.86 ± 1.28 (18)	20.57	W
GANomaly	26.28 ± 3.04 (19)	20.45	W	72.38 ± 1.43 (22)	20.12	W
ECOD	26.01 ± 1.09 (20)	20.43	W	75.78 ± 0.76 (19)	20.59	W
COPOD	23.79 ± 0.92 (21)	21.40	W	75.77 ± 0.96 (20)	20.45	W
OCSVM	23.21 ± 1.48 (22)	21.38	W	73.33 ± 1.19 (21)	20.90	W
KNN	21.92 ± 1.95 (23)	18.79	W	71.30 ± 0.77 (23)	19.00	W
LODA	21.06 ± 1.93 (24)	21.59	W	69.77 ± 0.23 (25)	22.86	W
SOD	16.22 ± 0.97 (25)	20.90	W	71.30 ± 0.46 (24)	20.59	W
COF	15.99 ± 0.56 (26)	20.62	W	68.59 ± 1.80 (26)	20.34	W
DAGMM	12.94 ± 2.33 (27)	–	–	64.39 ± 1.89 (28)	–	–
LOF	12.67 ± 0.59 (28)	20.72	W	66.09 ± 1.15 (27)	20.90	W
DeepSVDD	11.26 ± 1.37 (29)	24.76	W	55.70 ± 3.47 (29)	26.24	W

Table 5: Detection of all 29 methods on the 57 datasets; columns and conventions as in Table 4.

Model	AUPR	Rank	Verdict	AUROC	Rank	Verdict
CatBoost	82.53 ± 0.45 (1)	3.37	L	96.00 ± 0.19 (1)	3.55	–
XGBoost	80.72 ± 1.07 (2)	5.31	–	95.69 ± 0.23 (2)	5.26	–
LightGBM	80.21 ± 0.84 (3)	4.89	–	94.16 ± 0.84 (7)	5.36	–
SVM	80.18 ± 1.18 (4)	5.87	–	94.28 ± 0.56 (6)	7.11	W
XGBOD	79.96 ± 0.67 (5)	5.80	–	95.64 ± 0.20 (3)	5.39	–
RF	79.82 ± 0.60 (6)	5.45	–	94.82 ± 0.16 (5)	5.91	–
CRETAP	79.54 ± 0.93 (7)	5.79	–	95.06 ± 0.46 (4)	5.82	–
DeepSAD	77.26 ± 0.72 (8)	7.00	–	93.86 ± 0.20 (8)	7.90	W
MLP	76.16 ± 0.87 (9)	7.46	–	93.73 ± 0.43 (9)	7.83	–
FEAWAD	70.68 ± 2.31 (10)	10.09	W	92.67 ± 0.32 (10)	9.61	W
PReNet	69.47 ± 0.83 (11)	10.13	W	91.24 ± 0.59 (12)	9.91	W
DevNet	68.97 ± 0.87 (12)	10.19	W	91.33 ± 0.45 (11)	9.50	W
NB	54.83 ± 0.67 (13)	14.68	W	84.98 ± 0.32 (13)	14.39	W
IForest	34.71 ± 0.59 (14)	19.30	W	73.87 ± 0.56 (14)	18.26	W
REPEN	34.44 ± 1.05 (15)	18.55	W	73.76 ± 0.72 (15)	18.20	W
PCA	34.23 ± 0.89 (16)	20.21	W	71.73 ± 0.34 (20)	19.77	W
GANomaly	34.00 ± 2.28 (17)	19.35	W	71.19 ± 1.40 (21)	19.50	W
CBLOF	33.71 ± 0.87 (18)	18.85	W	72.85 ± 1.01 (16)	18.42	W
HBOS	33.14 ± 0.52 (19)	20.02	W	71.79 ± 0.82 (19)	20.25	W
ECOD	32.18 ± 0.49 (20)	20.32	W	72.01 ± 0.42 (18)	20.05	W
COPOD	31.66 ± 0.17 (21)	20.80	W	72.17 ± 0.53 (17)	19.84	W
OCSVM	29.93 ± 0.41 (22)	21.65	W	68.62 ± 0.65 (23)	21.58	W
KNN	28.83 ± 0.89 (23)	19.58	W	69.30 ± 0.62 (22)	19.37	W
LODA	27.62 ± 1.35 (24)	22.07	W	64.36 ± 0.42 (25)	23.30	W
SOD	23.99 ± 0.44 (25)	21.39	W	68.03 ± 0.42 (24)	21.02	W
DAGMM	21.85 ± 0.86 (26)	–	–	62.98 ± 0.62 (27)	–	–
COF	21.47 ± 0.53 (27)	22.18	W	63.01 ± 1.06 (26)	22.11	W
LOF	20.31 ± 0.18 (28)	21.96	W	62.11 ± 1.00 (28)	21.88	W
DeepSVDD	19.25 ± 1.30 (29)	23.75	W	56.08 ± 2.56 (29)	24.89	W

Table 6: Detection of all 29 methods on the 28 datasets with anomaly ratio $> 5\%$; columns and conventions as in Table 4.

Model	AUPR	Rank	Verdict	AUROC	Rank	Verdict
CatBoost	92.18 ± 0.16 (1)	2.36	L	97.35 ± 0.12 (1)	2.48	L
LightGBM	91.48 ± 0.16 (2)	3.55	L	97.11 ± 0.09 (3)	3.54	L
RF	91.43 ± 0.25 (3)	3.55	L	97.18 ± 0.09 (2)	3.43	L
XGBoost	91.09 ± 0.27 (4)	4.50	–	96.91 ± 0.07 (4)	4.48	–
XGBOD	90.89 ± 0.25 (5)	4.45	–	96.71 ± 0.12 (5)	4.61	–
CRETAP	88.18 ± 0.20 (6)	5.84	–	96.04 ± 0.06 (6)	5.89	–
SVM	86.32 ± 0.41 (7)	7.23	–	94.82 ± 0.26 (8)	8.00	W
DeepSAD	85.02 ± 0.53 (8)	7.66	–	94.93 ± 0.15 (7)	7.64	W
MLP	83.12 ± 0.44 (9)	7.52	–	94.21 ± 0.18 (9)	7.71	–
FEAWAD	77.55 ± 0.40 (10)	10.27	W	92.37 ± 0.25 (10)	10.32	W
PReNet	73.74 ± 0.07 (11)	10.89	W	89.98 ± 0.19 (11)	10.71	W
DevNet	72.53 ± 0.19 (12)	11.45	W	89.61 ± 0.10 (12)	11.29	W
NB	65.26 ± 0.98 (13)	12.91	W	86.73 ± 0.23 (13)	12.64	W
GANomaly	42.00 ± 2.31 (14)	18.21	W	69.96 ± 1.37 (15)	18.86	W
IForest	40.37 ± 0.85 (15)	19.25	W	70.26 ± 0.51 (14)	18.50	W
REPEN	40.20 ± 1.63 (16)	18.86	W	69.69 ± 0.81 (16)	18.96	W
COPOD	39.82 ± 0.62 (17)	20.18	W	68.45 ± 0.20 (17)	19.21	W
PCA	38.87 ± 0.47 (18)	20.71	W	66.79 ± 0.23 (21)	20.57	W
HBOS	38.64 ± 0.50 (19)	19.79	W	67.58 ± 0.49 (19)	19.93	W
ECOD	38.58 ± 0.42 (20)	20.21	W	68.10 ± 0.07 (18)	19.50	W
CBLOF	38.17 ± 1.56 (21)	20.07	W	66.24 ± 0.85 (22)	20.14	W
OCSVM	36.88 ± 0.70 (22)	21.93	W	63.74 ± 0.09 (24)	22.29	W
KNN	35.99 ± 0.67 (23)	20.39	W	67.23 ± 0.55 (20)	19.75	W
LODA	34.41 ± 1.11 (24)	22.57	W	58.76 ± 0.80 (26)	23.75	W
SOD	32.04 ± 0.63 (25)	21.89	W	64.66 ± 0.44 (23)	21.46	W
DAGMM	31.50 ± 1.05 (26)	–	–	61.46 ± 0.84 (25)	–	–
LOF	28.22 ± 0.96 (27)	23.25	W	57.99 ± 0.97 (27)	22.89	W
DeepSVDD	27.52 ± 1.58 (28)	22.71	W	56.48 ± 1.73 (29)	23.50	W
COF	27.15 ± 0.60 (29)	23.79	W	57.22 ± 0.35 (28)	23.93	W

G.2 Calibration on every dataset

Table 7 gives the test ECE of the eight natively probabilistic detectors on each of the 57 datasets, in the primary scope ($\leq 5\%$ anomalies, 29 datasets) and above it (28 datasets), with the mean ECE, the average rank, and the number of datasets on which each model has the lowest ECE for each block and for all 57. Over the 29 primary datasets, the per-dataset ECE rank of CRETAP ranges from first to seventh, with a median of second.

Reliability diagrams. ECE summarizes a reliability diagram: predicted probabilities are grouped into ten equal-width bins, the mean predicted probability of each bin is compared with the observed anomaly frequency in it, and ECE is the sample-weighted mean of the absolute differences. Because anomalies are rare, most test samples fall into the lowest bin, which therefore carries most of the weight, and the high-probability bins hold few samples. Figure 3 shows the diagrams on one dataset per domain; they are datasets on which CRETAP has the lowest ECE and thus favorable cases.

Anomalies available for post-hoc calibration. Splitting 10% of the training data for validation leaves 15 of the 29 datasets with fewer than ten validation anomalies on seed average—a median of 7.3 over the 29—and `fraud` and `smtp` have a single validation anomaly. A post-hoc calibrator for the baselines would have to be fit on such a split, which CRETAP itself uses for model selection and early stopping. The comparisons of Section 6.1 are between uncalibrated outputs, and post-hoc calibration of the baselines would narrow the gap wherever enough held-out anomalies exist to fit it.

Beyond binned ECE. At anomaly ratios of at most 5%, a constant predictor at the base rate already has an ECE close to zero, and the lowest bin carries most of the weight, so ECE says little about calibration near the decision boundary k (Tables 8 and 9).

Table 7: Test ECE (↓; 10 equal-width bins, no post-hoc calibration, fully labeled regime) of the eight natively probabilistic detectors on every dataset, averaged over three seeds (and over the physical datasets of an image or text dataset), rounded to four decimals. Bold marks the lowest ECE of each dataset before rounding. Datasets are ordered by anomaly ratio (%). The average rank is taken within each block.

Dataset	Ratio	CRETAP	NB	MLP	RF	LGBM	XGB	CatB	XGBOD
Anomaly ratio $\leq 5\%$ (29 datasets, primary scope)									
smtp	0.0315	0.0002	0.0700	**0.0001**	0.0002	0.0236	0.0003	0.0004	0.0004
fraud	0.1727	**0.0005**	0.0109	0.0006	0.0010	0.0048	0.0011	0.0011	0.0011
http	0.3896	**0.0000**	0.0001	0.0013	0.0001	0.0005	0.0003	0.0004	0.0003
cover	0.9603	0.0026	0.0049	0.0020	0.0026	**0.0014**	0.0015	0.0014	0.0014
satimage-2	1.2235	0.0024	0.0024	0.0022	0.0023	0.0016	0.0015	**0.0015**	0.0016
speech	1.6549	0.0104	0.0290	0.0135	**0.0072**	0.0162	0.0153	0.0123	0.0125
celeba	2.2443	0.0104	0.2163	0.0127	0.0071	0.0121	0.0146	0.0107	**0.0057**
pendigits	2.2707	0.0017	0.0086	0.0021	0.0042	0.0015	0.0012	**0.0007**	0.0019
mammography	2.3250	0.0072	0.0365	0.0063	0.0064	0.0095	0.0085	0.0064	**0.0045**
backdoor	2.4431	**0.0018**	0.0103	0.0026	0.0052	0.0022	0.0021	0.0019	0.0084
thyroid	2.4655	0.0061	0.0070	0.0067	0.0058	0.0023	0.0023	0.0022	**0.0020**
WDBC	2.7248	0.0003	0.0001	0.0046	0.0021	**0.0000**	0.0033	0.0007	0.0034
optdigits	2.8758	**0.0011**	0.2243	0.0016	0.0079	0.0018	0.0018	0.0016	0.0020
Waveform	2.9044	0.0180	0.0664	**0.0090**	0.0122	0.0203	0.0181	0.0150	0.0114
ALOI	3.0444	0.0063	0.8417	0.0043	0.0094	0.0211	0.0231	0.0095	**0.0039**
musk	3.1679	0.0000	**0.0000**	0.0002	0.0000	0.0000	0.0008	0.0001	0.0008
vowels	3.4341	**0.0089**	0.0193	0.0208	0.0197	0.0107	0.0105	0.0094	0.0129
Lymphography	4.0541	0.0011	0.0398	0.0046	0.0019	**0.0000**	0.0036	0.0006	0.0029
glass	4.2056	0.0155	0.1017	0.0200	0.0069	**0.0033**	0.0061	0.0042	0.0093
WBC	4.4843	0.0129	0.0803	0.0228	0.0082	**0.0001**	0.0084	0.0018	0.0072
20news	4.9516	0.0202	0.6689	0.0251	**0.0165**	0.0393	0.0351	0.0278	0.0248
FashionMNIST	4.9881	0.0100	0.3114	0.0148	0.0149	0.0158	0.0138	0.0106	**0.0082**
SVHN	4.9953	0.0139	0.7767	0.0412	0.0139	0.0384	0.0379	0.0170	**0.0093**
CIFAR10	4.9971	0.0174	0.4082	0.0390	0.0168	0.0382	0.0346	0.0189	**0.0128**
MNIST-C	5.0000	**0.0034**	0.3879	0.0085	0.0190	0.0113	0.0114	0.0064	0.0055
agnews	5.0000	**0.0104**	0.3316	0.0162	0.0214	0.0304	0.0289	0.0198	0.0106
amazon	5.0000	**0.0105**	0.9265	0.0179	0.0108	0.0400	0.0411	0.0249	0.0107
imdb	5.0000	0.0094	0.4040	0.0140	0.0135	0.0388	0.0411	0.0227	**0.0084**
yelp	5.0000	**0.0101**	0.7278	0.0189	0.0160	0.0349	0.0342	0.0220	0.0111
Mean ECE		0.0073	0.2315	0.0115	0.0087	0.0145	0.0139	0.0087	0.0067
Average rank		2.93	7.28	4.93	4.17	4.79	5.07	3.45	3.38
Lowest ECE (datasets)		9	1	2	2	5	0	2	8
Anomaly ratio $> 5\%$ (28 datasets)									
Wilt	5.3331	**0.0055**	0.1274	0.0222	0.0149	0.0111	0.0095	0.0071	0.0093
donors	5.9274	0.0001	0.0074	0.0016	0.0014	**0.0000**	0.0007	0.0005	0.0022
census	6.2041	0.0166	0.4486	0.0465	0.0114	0.0186	0.0184	0.0116	**0.0075**
letter	6.2500	0.0253	0.0665	0.0213	0.0383	0.0384	0.0334	0.0281	**0.0197**
shuttle	7.1511	0.0017	0.0071	0.0020	0.0015	0.0010	0.0006	**0.0006**	0.0011
annthyroid	7.4167	0.0186	0.0455	0.0421	0.0101	0.0107	0.0109	**0.0068**	0.0077
wine	7.7519	0.0070	0.0000	0.0099	0.0019	**0.0000**	0.0037	0.0008	0.0042
Stamps	9.1176	0.0185	0.0249	0.0313	0.0130	**0.0013**	0.0077	0.0065	0.0132
mnist	9.2069	0.0069	0.0637	0.0065	0.0393	0.0074	**0.0062**	0.0077	0.0104
PageBlocks	9.4567	0.0100	0.0844	0.0125	**0.0088**	0.0180	0.0162	0.0112	0.0094
cardio	9.6122	0.0146	0.0622	0.0137	0.0178	0.0168	0.0145	0.0139	**0.0102**
campaign	11.2654	0.0155	0.1545	0.0564	**0.0115**	0.0288	0.0420	0.0239	0.0133
vertebral	12.5000	0.0597	0.1891	0.0492	0.0205	**0.0046**	0.0074	0.0114	0.0193
Hepatitis	16.2500	0.0016	0.4747	0.0162	0.0015	**0.0000**	0.0044	0.0015	0.0136
InternetAds	18.7182	0.0296	0.1582	0.0400	0.0644	**0.0227**	0.0235	0.0258	0.0312
landsat	20.7148	0.0273	0.3213	0.0404	0.0370	0.0209	0.0302	**0.0154**	0.0209
skin	20.7539	0.0039	0.0375	0.0090	0.0081	0.0034	0.0032	**0.0030**	0.0037
Cardiotocography	22.0435	0.0232	0.1022	0.0272	0.0269	0.0300	0.0249	**0.0185**	0.0186
MVTec-AD	23.6998	0.0407	0.1910	0.0295	0.0450	0.0206	**0.0196**	0.0242	0.0304
WPBC	23.7374	0.0353	0.2551	0.0700	0.0364	**0.0055**	0.0098	0.0234	0.0564
satellite	31.6395	0.0245	0.2239	0.0354	0.0286	0.0243	0.0350	**0.0163**	0.0166
yeast	34.1644	0.0581	0.4082	0.0486	**0.0401**	0.0990	0.1230	0.0425	0.0470
fault	34.6728	0.0602	0.4483	**0.0393**	0.0528	0.0676	0.0878	0.0502	0.0402
Pima	34.8958	0.0664	0.1201	0.0730	0.0501	0.0735	0.0752	**0.0429**	0.0613
breastw	34.9927	0.0176	0.0426	0.0264	0.0196	0.0089	0.0101	**0.0067**	0.0117
magic.gamma	35.1630	0.0196	0.2180	0.0237	0.0303	0.0195	0.0464	**0.0138**	0.0230
Ionosphere	35.8974	0.0371	0.1420	0.0558	0.0285	**0.0079**	0.0118	0.0195	0.0113
SpamBase	39.9097	0.0298	0.2044	**0.0174**	0.0483	0.0224	0.0229	0.0206	0.0243
Mean ECE		0.0241	0.1653	0.0310	0.0253	0.0208	0.0250	0.0162	0.0192
Average rank		4.43	7.75	5.54	4.68	3.46	4.18	2.39	3.57
Lowest ECE (datasets)		1	0	2	3	8	2	9	3
All 57 datasets									
Mean ECE		0.0156	0.1990	0.0211	0.0169	0.0176	0.0193	0.0124	0.0129
Average rank		3.67	7.51	5.23	4.42	4.14	4.63	2.93	3.47
Lowest ECE (datasets)		10	1	4	5	13	2	11	11

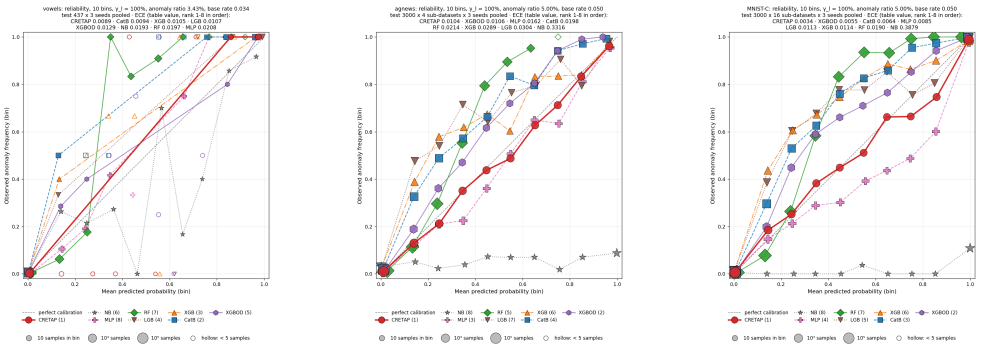

Figure 3: Reliability diagrams of the eight natively probabilistic detectors (10 bins, test data of three seeds pooled, fully labeled regime) on `vowels` (tabular), `agnews` (text embedding), and `MNIST-C` (image embedding). The dashed diagonal is perfect calibration; marker size follows the number of samples in a bin, and hollow markers are bins with fewer than five samples. The three datasets are among the nine on which CRETAP has the lowest ECE and show favorable cases; the ECE of every model on every dataset is in Table 7.

Table 8: Behavior at the operating point $\tau = \frac{1}{2}$ and scores that are not binned, on the 29 primary datasets ($\gamma_l = 100\%$, means over datasets of means over three seeds). For CRETAP the threshold $p \geq \frac{1}{2}$ is the reconstruction-error rule $e \geq k$; the baselines use $p \geq \frac{1}{2}$ on their own probabilities. The last row is a constant predictor that always outputs the base rate of its training split. Recall and FPR are fractions at that threshold; BCE, Brier and ECE are lower-is-better, the Brier score being the mean squared difference between the probability and the label. ECE is over 15 bins, equal-width and equal-mass.

Model	Recall	FPR	BCE	Brier	ECE-15	ECE-15m
CRETAP	0.5626	0.0025	0.0544	0.0137	0.0080	0.0074
CatBoost	0.5830	0.0009	0.0546	0.0139	0.0089	0.0091
XGBOD	0.5242	0.0011	0.0564	0.0144	0.0075	0.0075
XGBoost	0.5806	0.0013	0.0722	0.0146	0.0142	0.0123
LightGBM	0.5634	0.0021	0.1048	0.0157	0.0147	0.0135
RF	0.5252	0.0005	0.0709	0.0152	0.0094	0.0085
MLP	0.5554	0.0042	0.0731	0.0152	0.0123	0.0108
NB	0.8060	0.2375	5.1747	0.2286	0.2317	0.2304
Constant base rate	0.0000	0.0000	0.1353	0.0304	0.0004	0.0000

Table 9: CRETAP on each of the 29 primary datasets: the learned boundary k and slope n (means over three seeds and, for an image or text dataset, over its physical datasets) and the recall and false-positive rate at $e \geq k$. Datasets are ordered by the anomaly ratio of their test split.

Dataset	Test anomalies (%)	k	n	Recall	FPR
smtp	0.04	1.0862	2.8183	0.1667	0.0000
fraud	0.14	0.9918	4.0705	0.1333	0.0000
http	0.38	0.5283	4.3417	1.0000	0.0000
cover	1.02	1.3153	1.8138	0.9060	0.0013
satimage-2	1.21	0.9666	2.4156	0.8095	0.0000
speech	1.63	0.9931	4.0339	0.0926	0.0015
celeba	2.26	0.9615	4.3411	0.0879	0.0033
pendigits	2.28	1.0953	1.3145	0.9716	0.0005
mammography	2.32	1.1551	2.6376	0.4586	0.0031
thyroid	2.47	2.5355	1.1265	0.8214	0.0027
backdoor	2.52	0.9365	2.9323	0.9072	0.0003
WDBC	2.78	0.9107	2.6438	1.0000	0.0000
optdigits	2.88	1.1575	3.2270	0.9778	0.0002
Waveform	2.90	1.0401	0.5827	0.0667	0.0003
musk	3.16	0.9630	3.7958	1.0000	0.0000
ALOI	3.27	0.9742	1.0174	0.0244	0.0003
vowels	3.43	1.3386	0.8710	0.8889	0.0008
glass	4.00	0.9835	8.0030	0.9333	0.0046
Lymphography	4.11	1.0717	3.1838	1.0000	0.0000
WBC	4.22	0.8590	2.7085	0.6524	0.0012
20news	4.86	1.0670	5.4907	0.1855	0.0045
SVHN	5.00	1.0846	2.8580	0.1825	0.0056
MNIST-C	5.00	1.0999	3.9875	0.8422	0.0021
agnews	5.00	1.0105	3.3075	0.5444	0.0092
amazon	5.00	0.9756	6.9203	0.1778	0.0075
imdb	5.00	0.9837	7.6307	0.1156	0.0029
yelp	5.00	0.9963	7.8576	0.3311	0.0077
CIFAR10	5.00	1.0600	2.5270	0.2882	0.0083
FashionMNIST	5.01	1.2414	4.1242	0.7498	0.0055

G.3 Gap between training and test calibration

For each dataset we measure the 10-bin ECE on the training data and on the test data and take the absolute difference $|\mathrm{ECE}_{\text{train}} - \mathrm{ECE}_{\text{test}}|$; a small value means that the probabilities are calibrated to the same degree on data the model was fitted on and on held-out data. Tables 10 and 11 report its mean and median over datasets, together with the test and training ECE themselves, and Figures 4 and 5 plot the two ECE values per dataset. A small gap alone does not indicate good calibration: NB has a small gap because it is poorly calibrated on both. Training is early-stopped, so Proposition 2(iii), which holds at a stationary point, is not tested directly by these gaps (Table 12).

G.4 Robustness to ADBench noise

ADBench defines three corruptions—duplicating anomalies, adding irrelevant features, and flipping labels (label contamination, called label noise in Section 6.1)—applied to tabular data only; the label-noise runs are the one exception to the fully labeled regime of Section 5: duplicated anomalies resamples the anomalies with replacement to $\lfloor m\,t \rfloor$ instances, separately in each split; irrelevant features appends $\lfloor r/(1-r)\,d \rfloor$ uniform columns before the split; and label contamination flips a fraction r of the training labels in both directions, so a share of the true anomalies is flipped to $y = 0$. Table 13 reports the 20 low-contamination tabular datasets. CRETAP is re-tuned under each variant with the protocol of Appendix H.2, except that under label contamination it is trained without early stopping, because corrupted validation labels corrupt the stopping signal; this choice was made after observing the results.

Under anomaly duplication and under label contamination, the degradation of CRETAP is not significantly different from that of any boosting model. Under duplication it is significantly smaller than that of seven unsupervised detectors (IForest, KNN, ECOD, COPOD, CBLOF, COF, and SOD), and its ECE rises significantly less than XGBOD's ($+0.0196$ vs. $+0.0633$). Label contamination is

Table 10: Gap between training and test calibration for the eight natively probabilistic detectors (29 datasets with anomaly ratio $\leq 5\%$, fully labeled regime, 10-bin ECE, mean of three seeds per dataset). For each dataset we take $|\mathrm{ECE}_{\text{train}} - \mathrm{ECE}_{\text{test}}|$ and report its mean and median over datasets, with the rank among the eight in parentheses (1 = smallest); rows are sorted by the mean. Test and training ECE are the dataset averages. The training ECE of CRETAP is measured on the samples that received gradients (90% of the training split); that of the baselines on the full training split. The row of CRETAP is in bold.

	$\lvert\mathrm{ECE}_{\text{train}} - \mathrm{ECE}_{\text{test}}\rvert \downarrow$			
Model	mean (rank)	median (rank)	test ECE (rank)	train ECE
CRETAP	**0.0017 (1)**	**0.0011 (2)**	**0.0073 (2)**	**0.0070**
CatBoost	0.0036 (2)	0.0013 (3)	0.0087 (3)	0.0086
MLP	0.0047 (3)	0.0008 (1)	0.0115 (5)	0.0070
NB	0.0048 (4)	0.0028 (5)	0.2315 (8)	0.2291
RF	0.0067 (5)	0.0044 (7)	0.0087 (4)	0.0126
XGBOD	0.0079 (6)	0.0026 (4)	0.0067 (1)	0.0128
XGBoost	0.0118 (7)	0.0051 (8)	0.0139 (6)	0.0020
LightGBM	0.0124 (8)	0.0033 (6)	0.0145 (7)	0.0023

Table 11: Gap between training and test calibration for the eight natively probabilistic detectors (all 57 datasets, fully labeled regime, 10-bin ECE, mean of three seeds per dataset). For each dataset we take $|\mathrm{ECE}_{\text{train}} - \mathrm{ECE}_{\text{test}}|$ and report its mean and median over datasets, with the rank among the eight in parentheses (1 = smallest); rows are sorted by the mean. Test and training ECE are the dataset averages. The training ECE of CRETAP is measured on the samples that received gradients (90% of the training split); that of the baselines on the full training split. The row of CRETAP is in bold.

	$\lvert\mathrm{ECE}_{\text{train}} - \mathrm{ECE}_{\text{test}}\rvert \downarrow$			
Model	mean (rank)	median (rank)	test ECE (rank)	train ECE
CRETAP	**0.0045 (1)**	**0.0020 (2)**	**0.0156 (3)**	**0.0130**
MLP	0.0065 (2)	0.0015 (1)	0.0211 (7)	0.0156
NB	0.0071 (3)	0.0039 (4)	0.1990 (8)	0.1945
CatBoost	0.0100 (4)	0.0027 (3)	0.0124 (1)	0.0183
RF	0.0114 (5)	0.0062 (6)	0.0169 (4)	0.0213
XGBoost	0.0126 (6)	0.0075 (7)	0.0193 (6)	0.0071
LightGBM	0.0128 (7)	0.0078 (8)	0.0176 (5)	0.0081
XGBOD	0.0177 (8)	0.0041 (5)	0.0129 (2)	0.0283

compared only among the 15 methods that use labels (CRETAP and 14 baselines); there CRETAP degrades significantly less than FEAWAD at 10% and more than GANomaly at 25%, and its ECE rises significantly more than that of XGBoost and LightGBM at 10% ($+0.0944$ vs. $+0.0620$ and $+0.0694$).

Irrelevant features are a genuine weakness. At 50% added noise features CRETAP loses 3.6 AUROC points against 1.2–1.6 for the boosting group. The 20-dataset test does not resolve this difference, but on all 47 tabular datasets the same protocol gives -3.3 against -0.7 to -1.0 for the boosting group and -1.2 for RF, and there CRETAP degrades significantly more than CatBoost, XGBoost, XGBOD, LightGBM, RF, and NB. Re-tuning under each variant does not remove the effect. A feature-weighted aggregation would be the natural remedy, and we have not evaluated one. These results do not make robustness a reason to select CRETAP; they identify the conditions under which it should not be selected.

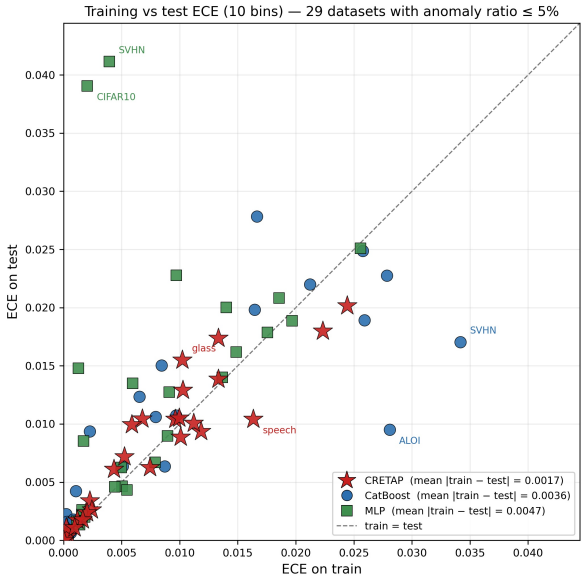

Figure 4: Training versus test ECE (10 bins) on the 29 datasets with anomaly ratio $\leq 5\%$ for CRETAP, CatBoost, and MLP, the three models ranked in the top three by both the mean and the median of $|\mathrm{ECE}_{\mathrm{train}} - \mathrm{ECE}_{\mathrm{test}}|$ on the 29 datasets (Table 10). The three models share one pair of axes. Each point is one dataset (mean of three seeds); the dashed line marks equal training and test ECE, and points above it are worse calibrated on test data. The two datasets with the largest gap are labeled per model.

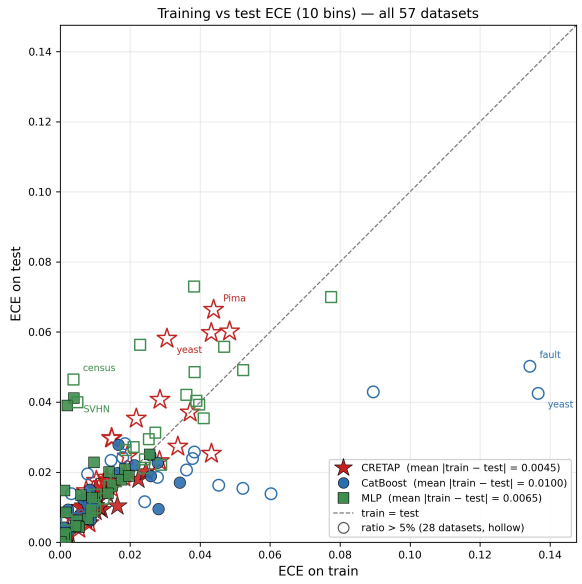

Figure 5: Training versus test ECE (10 bins) on all 57 datasets for the same three models as Figure 4, kept for comparison with it. Marker shape denotes the model (star: CRETAP, circle: CatBoost, square: MLP); filled markers are the 29 datasets with anomaly ratio $\leq 5\%$ and hollow markers the 28 above it; the dashed line marks equal training and test ECE. The axis range is set for this scope, so it differs from Figure 4.

Table 12: CRETAP on each of the 29 primary datasets, beyond the binned gap of Table 10 (means over three seeds and, for an image or text dataset, over its physical datasets). Calibration-in-the-large is the mean predicted probability minus the observed anomaly rate, so 0 means the overall level is right. The calibration slope is the coefficient of a logistic regression of the label on $\mathrm{logit}(p)$, so a well-calibrated model has slope 1. Separation is the AUROC of the reconstruction error e on the training split, and e_1/e_0 the ratio of the median error of the two classes there. On the eight datasets whose separation is 1.0000 the two classes are linearly separable in $\mathrm{logit}(p)$, so the slope has no finite maximum-likelihood estimate and the value shown is only where the iteration stopped; excluding them, the median test slope is 1.14 rather than 1.40. Epoch is the epoch early stopping selected, with the epoch reached in parentheses. Rows are ordered by separation.

	In-large		Slope		Training error		
Dataset	Train	Test	Train	Test	Separation	e_1/e_0	Epoch
http	-0.0000	-0.0000	3.47	3.20	1.0000	275	34 (39)
WDBC	+0.0000	+0.0001	4.45	4.25	1.0000	443	45 (49)
Lymphography	+0.0005	+0.0004	5.00	4.82	1.0000	72.5	44 (46)
musk	+0.0000	+0.0000	1.95	1.95	1.0000	15,890	30 (30)
pendigits	+0.0001	+0.0002	50.39	37.22	1.0000	29,202	9 (14)
optdigits	-0.0002	-0.0005	3.52	5.49	1.0000	129	20 (25)
cover	+0.0010	+0.0013	1.55	1.32	0.9999	3,083	28 (33)
glass	-0.0017	+0.0042	16.53	12.02	0.9998	6.80	43 (45)
vowels	+0.0033	+0.0034	1.78	1.81	0.9993	30,900	15 (20)
backdoor	-0.0002	-0.0002	1.81	1.73	0.9992	1,565	10 (15)
WBC	-0.0066	-0.0087	2.77	1.40	0.9989	98.1	32 (37)
thyroid	-0.0003	-0.0001	1.19	1.24	0.9982	10,944	15 (20)
satimage-2	-0.0017	-0.0024	1.19	1.69	0.9967	157	8 (13)
speech	+0.0073	+0.0041	2.68	0.86	0.9939	6.16	28 (33)
FashionMNIST	+0.0007	-0.0010	3.51	0.79	0.9921	9,083	13 (18)
celeba	-0.0032	-0.0045	1.48	0.78	0.9840	7.02	21 (26)
Waveform	+0.0048	+0.0044	2.11	1.72	0.9817	1,147	17 (22)
fraud	-0.0003	-0.0004	2.39	1.89	0.9705	8.23	12 (17)
MNIST-C	+0.0001	-0.0011	4.41	3.76	0.9690	3,954	20 (24)
agnews	+0.0031	+0.0023	1.16	1.04	0.9658	32.9	36 (39)
mammography	-0.0007	-0.0025	0.98	0.95	0.9578	19.5	16 (20)
yelp	+0.0012	-0.0006	1.22	1.14	0.9505	1.76	29 (34)
CIFAR10	+0.0001	-0.0032	1.34	0.83	0.9504	51.0	8 (13)
amazon	+0.0051	+0.0020	1.17	1.05	0.9319	1.73	28 (33)
smtp	-0.0000	-0.0001	7.04	2.82	0.9259	207	20 (25)
imdb	-0.0045	-0.0053	1.25	1.02	0.9222	1.55	33 (38)
SVHN	+0.0000	-0.0020	1.33	0.82	0.9144	39.2	14 (19)
20news	+0.0076	+0.0065	1.97	1.23	0.8786	4.89	31 (35)
ALOI	-0.0002	-0.0016	1.52	0.69	0.8204	3.23	27 (32)
Median	+0.0000	-0.0001	1.95	1.40	0.9921	98.1	21 (26)

Table 13: Mean AUROC (↑, %) on the 20 low-contamination tabular datasets under the three ADBench noise protocols, with the change from each model's own clean baseline in parentheses. CRETAP was re-tuned under the same validation protocol for every variant, except that under label contamination it was trained without early stopping (see text).

Variant	CRETAP	CatBoost	XGBoost	XGBOD	LightGBM	RF
Clean baseline	95.9	96.9	96.7	96.9	92.0	95.8
Anomaly duplication 6x	95.4 (−0.5)	96.7 (−0.2)	96.5 (−0.2)	86.7 (−10.3)	95.4 (+3.4)	96.1 (+0.3)
Irrelevant features 50%	92.3 (−3.6)	95.3 (−1.6)	95.3 (−1.4)	95.7 (−1.2)	90.6 (−1.4)	94.1 (−1.7)
Label contamination 10%	92.1 (−3.7)	92.8 (−4.1)	91.5 (−5.2)	93.5 (−3.4)	91.8 (−0.2)	93.4 (−2.4)
Label contamination 25%	87.4 (−8.5)	89.0 (−7.9)	85.1 (−11.6)	86.9 (−10.0)	86.1 (−6.0)	88.6 (−7.2)

G.5 Effect of the two excluded baselines

Adding ResNet and FTTransformer at the values published with ADBench and comparing on AUROC and AUPR changes the AUPR ordering on the 29 and the 57 datasets. ResNet attains a higher mean AUPR than CRETAP on the 29-dataset primary scope of Section 6.1 (73.45 against 71.19) and on the full 57 datasets (81.00 against 79.54), so CRETAP moves from third to fourth and from seventh to eighth by mean AUPR (Appendix F). The AUROC orderings are unchanged in both scopes, no significance verdict for CRETAP changes, and neither model differs significantly from CRETAP on either metric. Above 5% contamination, where Appendix F also reports ranks, the two models move CRETAP from sixth to seventh by mean AUROC and from sixth to eighth by mean AUPR; there the published values cover 26 of the 28 datasets, and among the models significantly ahead of CRETAP, RF drops out on AUROC and XGBOD enters on AUPR.

H Implementation details

H.1 Datasets and baselines

Datasets. Table 7 lists the 57 ADBench datasets and their anomaly ratios, computed over the full ADBench release of each dataset before any split. All experiments use the fully labeled regime, i.e., ADBench's labeled-anomaly ratio of 100%: every anomaly in a training split is labeled. Each image or text dataset of ADBench is a collection of physical datasets; results on it are averaged over its physical datasets, and CRETAP selects one configuration per physical dataset. The image datasets are the ResNet-18 embeddings and the text datasets the BERT embeddings of the ADBench release; the ViT and RoBERTa variants it also ships are not used. Sizes and dimensionalities are those listed by ADBench (Han et al., 2022). We draw the data through ADBench's own generator with its default settings, so its size bounds apply—datasets below 1,000 samples are resampled to 1,000 and those above 10,000 are subsampled to 10,000 (Han et al., 2022)—and split it 70/30 into a training and a test set.

Baselines. The 28 baselines are the ADBench roster without ResNet and FTTransformer: PCA, OCSVM, LOF, CBLOF, COF, HBOS, KNN, SOD, COPOD, ECOD, DeepSVDD, DAGMM, LODA and IForest (unsupervised); GANomaly, DeepSAD, REPEN, DevNet, PReNet, FEAWAD and XGBOD (semi-supervised); NB, SVM, MLP, RF, LightGBM, XGBoost and CatBoost (supervised). Every baseline follows the settings stated in the ADBench paper, and is run with the same three seeds and the same data splits as CRETAP; REPEN, for instance, is trained for the 30 epochs the paper specifies. No baseline is tuned or calibrated post hoc. SVM is scored by its decision function and enters the detection comparison only: its own output is a decision-function margin, and the probabilities scikit-learn reports for it come from an internal Platt calibration, so including them would put a calibrated output against uncalibrated ones. Every model in this paper is trained on CPU. The probabilities are `predict_proba`$[:, 1]$ of the fitted classifier for NB, MLP, RF, CatBoost, XGBoost and LightGBM, and, for XGBOD, PyOD's `decision_function`, which returns the `predict_proba`$[:, 1]$ of its internal XGBoost classifier. Runs use Python 3.10.20 with scikit-learn 1.0.2, XGBoost 1.6.1, LightGBM 4.7.0, CatBoost 1.2.10, PyOD 1.0.0, PyTorch 1.13.1, TensorFlow 2.10.0 with Keras 2.10.0, NumPy 1.23.5, SciPy 1.10.1 and pandas 2.3.3.

Difference from the published ADBench settings. ADBench uses only one seed for its image and text datasets (Han et al., 2022). We instead evaluate them with the same three-seed configuration used for the 47 tabular datasets.

H.2 Training and model selection

Table 14 lists the settings of the two instantiations. Every configuration is chosen on validation data only, and the test split plays no part in the selection; because one configuration per dataset (per physical dataset for the image and text collections) is scored on the validation splits of three seeds whose partitions differ, a validation sample of one seed can be a test sample of another. The validation split also serves early stopping, so CRETAP is fit on the remaining 90% of each training split, whereas the baselines are fit on the full training split. Baselines are run at the configurations published with ADBench. All implementations evaluate the logit as $\exp(\nu)[\log(e+\epsilon) - \kappa]$ with $\epsilon = 10^{-6}$, so the computed probability of a perfect reconstruction is $\sigma(n[\log\epsilon - \log k]) > 0$ rather than $h_{k,n}(0) = 0$, and the decision at $\tau = 1/2$ is $e \geq k - \epsilon$. The configuration selected for each physical dataset, with the (k, n) it learned and the epoch at which it stopped, is in Tables 16

and 17.The three seeds are 1, 2 and 3, and the TPE sampler is seeded with 2026. The mean validation AUROC averages only the seeds whose validation split contains both classes; at $\gamma_l = 100\%$ all three do, on all 121 physical datasets.

Table 14: Training and selection settings. Both instantiations share the Hill function, the BCE objective, a single Adam optimizer, $k_0 = 1$, and no weight decay on (κ, ν).

	Tabular and embedding inputs	Images
Network	fully connected: $\mathrm{Linear}(d, d_{\mathrm{hid}})$, activation, dropout, $\mathrm{Linear}(d_{\mathrm{hid}}, d_{\mathrm{lat}})$; mirrored decoder	convolutional: one strided convolution (window $>$ stride) and a 1×1 convolution to d_{lat} channels; mirrored transposed decoder
Normalization layers	none	none
Input scaling	min–max, fitted on the training split	$x/255$, then min–max per pixel or standardization per channel, fitted on the nominal fit images
Validation data	10% of the training split, stratified	30% of the training pool, stratified
Protocol	CRETAP-JOINT or -PRE (searched; 20 reconstruction epochs)	CRETAP-JOINT
Batch size	64	16–64 (MVTec AD), 128 (MNIST-C)
Epochs	at most 30 or 50 (searched)	at most 200 (MVTec AD), 50 (MNIST-C)
Early stopping	validation BCE, patience 5 (not applied when the validation split contains no labeled anomaly)	validation BCE, patience 10, after a minimum of 10 epochs
Weight decay	none	network weights only, $\{0, 10^{-5}, 10^{-4}\}$ (MNIST-C: 10^{-5})
Search	Optuna TPE, 60 trials, mean validation AUROC over three seeds	Optuna TPE, 16 trials (MVTec AD) or 12 (MNIST-C), two objectives: mean validation AP and mean validation defect specificity over the three splits (selection: see text)
Search space	Table 15	see text

Table 15: Hyperparameter search space for the fully connected instantiation, selected per dataset on the validation split.

Variable	Type	Search space
activation	categorical	ReLU, sigmoid
latent width d_{lat}	categorical	$\{16, 32, 64, 128\}$
hidden width d_{hid}	categorical	$\{32, 64, 128, 256, 512\}$
dropout	categorical	$\{0.0, 0.1, \ldots, 0.6\}$
learning rate	log-uniform	$[10^{-4}, 10^{-2}]$
protocol	categorical	CRETAP-JOINT, CRETAP-PRE
epochs (upper bound)	categorical	$\{30, 50\}$
n_0	categorical	$\{0.5, 1, 2, 4, 8\}$

For MVTec AD the convolutional search covers window/stride $\{8/4, 12/4, 16/8, 24/8, 32/8\}$, padding $\{0, (k-s)/2, k-s\}$, $d_{\mathrm{hid}} \in \{64, 128\}$, $d_{\mathrm{lat}} \in [4, 256]$, activation (ReLU, $\tanh$, sigmoid), dropout $[0, 0.5]$, learning rate $[5 \cdot 10^{-4}, 3 \cdot 10^{-3}]$, batch size, weight decay, $n_0 \in \{1, 2, 4\}$, and the floor $\{0.01, 0.02, 0.05\}$ on the divisor of the input scaling; MNIST-C uses a reduced space with windows $\{3/1, 4/1, 4/2\}$, padding $\{0, (k-s)/2\}$, $d_{\mathrm{lat}} \in \{16, 32, 64\}$, activation (ReLU, $\tanh$, sigmoid), dropout $\{0, 0.3\}$ and learning rate $[5 \cdot 10^{-4}, 1.5 \cdot 10^{-3}]$, with per-channel standardization, $d_{\mathrm{hid}} = 128$ and $n_0 = 4$ fixed. The input normalization is set per dataset. Within a dataset the trial kept is the one whose mean validation AP is within 0.02 of the best, that among those is within 0.01 of the highest mean validation clump—the share of the attribution of an anomalous image lying in connected components of at least 16 pixels (4 for MNIST-C)—and that then has the highest mean validation defect specificity, the attribution mass inside the defect mask minus the mass the average

nominal map places there; ties go to the finer grid. The defect masks enter the selection only, never the loss. For MNIST-C the selection is restricted to the nine trials with stride 1, fixed before it: a stride of 2 is 7.1% of the 28-pixel width against 1.6% and 3.1% for MVTec AD, and 88% of the MNIST-C trials reach a validation AP of at least 0.999, so the clump criterion, higher on average at stride 2 (0.82) than at stride 1 (0.78), would otherwise decide the choice.

H.3 IMAGE DATA

MVTec AD. All 15 categories and every image of the release (Bergmann et al., 2019a): 5,354 images, of which 3,629 nominal images form the training folder and 1,725 the test folder, 1,258 of them defective (12.8–32.5% of the images of each category). These are exactly the rows of the fifteen ADBench MVTec-AD sets, and the labels (`good` folders nominal, defect folders anomalous) agree with ADBench on every row. The files were obtained from the HuggingFace mirror `TheoM55/mvtec_anomaly_detection` (commit `6363bbd` at the time of download); all 6,612 image and mask files match the SHA-256 hashes published with the mirror. Images are converted to RGB (replicating the grayscale `grid`, `screw`, `zipper`) and resized to 256×256 (bilinear); defect masks are resized with nearest-neighbor interpolation and drawn as outlines in the figures. Every defective image comes from the test folder of the release.

MNIST-C. From the official Zenodo record (`doi:10.5281/zenodo.3239543`, MD5 verified). MNIST-C carries digit labels only; the nominal/anomalous distinction and the composition are those of ADBench (Mu & Gilmer, 2019; Han et al., 2022): per corruption, 10,000 rows from the MNIST training split, 9,500 uncorrupted (`identity`) and 500 (5%) corrupted. We take the 28×28 image of the same MNIST index for every ADBench row. The mask of an anomalous image is the set of pixels that differ from its `identity` copy, derived by us rather than distributed. We use the five corruptions whose corrupted pixels cover on average less than 20% of the image (4.8–18.5%) (`dotted_line`, `zigzag`, `impulse_noise`, `spatter`, `shot_noise`); the others move the digit or alter the whole image.

Splits. Because every defective MVTec AD image lies in the test folder and CRETAP trains on labeled anomalies, the release partition is not used: each dataset is pooled and split 70/30 into a training pool and a test set, and the pool 70/30 into fit and validation sets, stratified by defect type (MVTec AD) or label (MNIST-C). The clean and corrupted copies of the same digit (60–78 per corruption) are kept on the same side. The three seeds differ only in this partition, and normalization coefficients come from the nominal fit images only. The reconstruction network is the convolutional autoencoder of Table 14, trained by CRETAP-JOINT, with one configuration per dataset.

H.4 FIGURES

Image figures. Rows show the input ($x/255$), the reconstruction mapped back to pixel values and clipped to $[0, 1]$, and the CRETAP Score $\tilde{s}_z$ of equation 22 on one logarithmic color scale, 10^{-2} to 10^4, shared by all datasets, with the color bar marking $\tilde{s}_z = 1$. We have randomly selected seven of the 20 image datasets and they are shown. The four test images of seed 1 per dataset are chosen for illustration from the correctly classified ones: the nominal images with the second- and third-lowest error and anomalies from the upper half by error (Table 18).

Tabular figures. The heat map shows up to 40 random nominal and 40 random anomalous test samples, sorted by e within each class, with the row mean e/k as the last column, on a logarithmic color scale set per dataset (first percentile of the cells to the larger of the largest cell and 1). The bars are column-wise means over all test samples of each class on a linear axis. When bar heights span too wide a range, the axis is split by a wavy break at a height fixed per dataset by one rule: the lower part holds every nominal bar and the anomalous bars of columns whose anomalous mean is below three times the nominal mean, and the axis is split only when the largest bar exceeds three times that height. The split changes no value. On `mammography` (Figure 2), the bars average all 2,931 nominal and all 69 anomalous test samples, and the break lies between 1.0 and 13.3.

H.5 STATISTICAL PROCEDURE

We reuse the procedure ADBench applies to its critical difference diagrams: a Friedman test over all compared models, then a Wilcoxon signed-rank test per model pair at the dataset level, then a Holm correction applied to all pairs at once—378 pairs among 28 detection models, 28 pairs among

Table 16: The configuration selected for each of the 121 physical datasets at $\gamma_l = 100\%$, with the Hill parameters it learned and the epoch at which it stopped; Table 17 continues it. The eight configuration columns are the values Optuna selected from the space of Table 15: Joint and Pre abbreviate CRETAP-JOINT and CRETAP-PRE, Act. and Drop. are the activation and the dropout rate, and Max ep. is the searched upper bound on the number of epochs. The boundary k, the slope n and the stopping epoch are means over the three seeds, and Stop gives the epoch early stopping selected with the epoch reached in parentheses. Datasets follow the order of Table 7, and within an image or text dataset its physical datasets are in alphabetical order.

	Selected configuration								Learned		
Physical dataset	Protocol	Act.	d_{lat}	d_{hid}	Drop.	lr	Max ep.	n_0	k	n	Stop
Anomaly ratio $\leq 5\%$ (29 datasets, primary scope)											
smtp	Joint	sigmoid	128	64	0.2	0.002449	50	2	1.0862	2.8183	20 (25)
fraud	Joint	sigmoid	32	256	0.2	0.000680	50	4	0.9918	4.0705	12 (17)
http	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.5283	4.3417	34 (39)
cover	Pre	sigmoid	16	128	0.1	0.002198	50	0.5	1.3153	1.8138	28 (33)
satimage-2	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9666	2.4156	8 (13)
speech	Pre	ReLU	32	512	0.1	0.000292	50	4	0.9931	4.0339	28 (33)
celeba	Pre	ReLU	16	128	0.1	0.000741	50	4	0.9615	4.3411	21 (26)
pendigits	Joint	ReLU	16	512	0.0	0.001691	30	1	1.0953	1.3145	9 (14)
mammography	Joint	ReLU	32	512	0.1	0.004558	30	4	1.1551	2.6376	16 (20)
backdoor	Joint	ReLU	64	64	0.0	0.007321	30	4	0.9365	2.9323	10 (15)
thyroid	Pre	ReLU	64	64	0.1	0.007318	50	0.5	2.5355	1.1265	15 (20)
WDBC	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9107	2.6438	45 (49)
optdigits	Joint	sigmoid	128	64	0.2	0.002449	50	2	1.1575	3.2270	20 (25)
Waveform	Pre	ReLU	64	512	0.0	0.000503	50	0.5	1.0401	0.5827	17 (22)
ALOI	Pre	ReLU	128	512	0.3	0.000416	50	1	0.9742	1.0174	27 (32)
musk	Joint	ReLU	64	64	0.0	0.007321	30	4	0.9630	3.7958	30 (30)
vowels	Pre	ReLU	128	512	0.0	0.004345	50	0.5	1.3386	0.8710	15 (20)
Lymphography	Joint	sigmoid	128	64	0.2	0.002449	50	2	1.0717	3.1838	44 (46)
glass	Joint	ReLU	64	128	0.0	0.000343	50	8	0.9835	8.0030	43 (45)
WBC	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.8590	2.7085	32 (37)
20news_0	Joint	ReLU	64	128	0.0	0.000813	50	8	0.9844	7.8579	31 (35)
20news_1	Joint	ReLU	128	128	0.0	0.000433	50	8	0.9921	7.9948	28 (33)
20news_2	Joint	ReLU	64	256	0.0	0.000395	50	8	0.9934	7.8903	21 (24)
20news_3	Pre	sigmoid	16	128	0.1	0.001341	50	0.5	1.3282	0.7020	40 (43)
20news_4	Joint	ReLU	64	128	0.0	0.000464	50	0.5	1.1071	0.5985	38 (40)
20news_5	Joint	sigmoid	64	128	0.0	0.001090	50	8	0.9965	7.9006	30 (35)
FashionMNIST_0	Pre	ReLU	16	256	0.1	0.001242	50	4	0.9949	3.9000	4 (9)
FashionMNIST_1	Pre	ReLU	32	512	0.4	0.001065	50	8	0.9984	7.6064	10 (15)
FashionMNIST_2	Pre	ReLU	64	256	0.0	0.000292	30	8	0.9982	7.9329	7 (12)
FashionMNIST_3	Joint	sigmoid	32	512	0.2	0.002820	50	0.5	1.6812	1.1192	11 (16)
FashionMNIST_4	Joint	ReLU	32	256	0.1	0.000208	50	2	0.9794	2.0204	18 (23)
FashionMNIST_5	Pre	sigmoid	128	256	0.0	0.001593	50	0.5	1.5877	1.0286	16 (21)
FashionMNIST_6	Pre	ReLU	32	128	0.1	0.001093	50	0.5	1.4785	0.8265	12 (17)
FashionMNIST_7	Pre	sigmoid	16	128	0.3	0.000526	50	0.5	1.7074	1.0027	38 (41)
FashionMNIST_8	Joint	ReLU	32	512	0.1	0.000522	50	8	0.9906	7.9305	6 (11)
FashionMNIST_9	Pre	ReLU	16	512	0.3	0.000566	50	8	0.9972	7.8752	5 (10)
SVHN_0	Joint	sigmoid	128	256	0.2	0.000789	50	0.5	1.4965	0.8817	21 (26)
SVHN_1	Pre	ReLU	64	128	0.2	0.001244	30	4	0.9950	3.5882	9 (14)
SVHN_2	Joint	ReLU	32	64	0.4	0.002373	30	2	0.9560	1.7823	8 (13)
SVHN_3	Pre	ReLU	64	512	0.5	0.000388	50	8	0.9960	7.5240	8 (13)
SVHN_4	Pre	sigmoid	16	128	0.1	0.001311	50	0.5	1.3117	0.8208	11 (16)
SVHN_5	Joint	sigmoid	16	512	0.3	0.000167	50	2	0.9944	2.0218	23 (28)
SVHN_6	Pre	ReLU	64	512	0.5	0.000401	30	8	0.9965	7.6034	9 (14)
SVHN_7	Joint	sigmoid	128	256	0.3	0.000515	30	1	1.0399	1.1864	25 (28)
SVHN_8	Joint	sigmoid	64	32	0.0	0.000579	30	2	0.9854	2.0459	14 (19)
SVHN_9	Joint	sigmoid	16	256	0.5	0.001518	30	1	1.0748	1.1259	15 (20)
CIFAR10_0	Pre	ReLU	64	32	0.0	0.000232	30	4	0.9894	4.0110	15 (20)
CIFAR10_1	Joint	ReLU	64	64	0.1	0.005896	30	2	0.9533	1.8566	2 (7)
CIFAR10_2	Pre	ReLU	128	256	0.5	0.001469	30	2	0.9897	1.8464	9 (14)
CIFAR10_3	Pre	ReLU	32	128	0.0	0.002599	50	1	1.0352	1.0412	3 (8)
CIFAR10_4	Pre	ReLU	32	512	0.3	0.000826	50	4	0.9972	3.9061	3 (8)
CIFAR10_5	Joint	sigmoid	16	64	0.1	0.005553	50	2	1.0173	1.9252	6 (11)
CIFAR10_6	Joint	sigmoid	16	256	0.2	0.000745	30	4	0.9942	3.9955	9 (14)
CIFAR10_7	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9838	1.9654	7 (12)
CIFAR10_8	Pre	ReLU	128	512	0.5	0.000993	50	4	0.9919	3.7598	8 (13)
CIFAR10_9	Joint	sigmoid	128	512	0.6	0.002152	50	0.5	1.6481	0.9630	13 (18)

Table 17: The configuration selected for each physical dataset, continuing Table 16; columns and conventions as there.

	Selected configuration								Learned		
Physical dataset	Protocol	Act.	d_{lat}	d_{hid}	Drop.	lr	Max ep.	n_0	k	n	Stop
Anomaly ratio $\leq 5\%$ (primary scope), continued											
MNIST-C_brightness	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9565	2.6008	12 (17)
MNIST-C_canny_edges	Joint	ReLU	128	64	0.3	0.002076	30	8	0.9654	7.5264	7 (12)
MNIST-C_dotted_line	Pre	ReLU	64	512	0.5	0.000255	50	8	0.9942	7.9188	16 (21)
MNIST-C_fog	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9581	3.1780	35 (38)
MNIST-C_glass_blur	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9514	2.9817	24 (29)
MNIST-C_identity	Pre	ReLU	32	64	0.4	0.003492	30	4	0.9587	3.1704	7 (12)
MNIST-C_impulse_noise	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9567	3.4518	50 (50)
MNIST-C_motion_blur	Joint	sigmoid	128	64	0.2	0.004908	50	8	0.9890	7.1736	13 (18)
MNIST-C_rotate	Joint	sigmoid	32	256	0.0	0.003100	30	1	1.2283	1.5592	17 (22)
MNIST-C_scale	Pre	sigmoid	16	128	0.1	0.002198	50	0.5	1.9866	1.3692	17 (22)
MNIST-C_shear	Pre	ReLU	64	512	0.3	0.000228	50	4	0.9932	4.0011	11 (16)
MNIST-C_shot_noise	Joint	sigmoid	16	256	0.0	0.003471	30	2	1.0246	2.2741	9 (14)
MNIST-C_spatter	Joint	sigmoid	64	64	0.1	0.000612	30	0.5	1.7268	1.1685	25 (29)
MNIST-C_stripe	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9538	3.3843	50 (50)
MNIST-C_translate	Joint	ReLU	32	256	0.3	0.000410	50	4	0.9725	4.0560	13 (18)
MNIST-C_zigzag	Joint	ReLU	64	128	0.0	0.000343	50	8	0.9832	7.9868	14 (19)
agnews_0	Joint	ReLU	64	32	0.0	0.000326	50	1	1.0372	1.3776	49 (50)
agnews_1	Joint	ReLU	128	32	0.0	0.000313	50	8	0.9926	8.4235	30 (35)
agnews_2	Joint	ReLU	64	64	0.0	0.000824	50	1	1.0109	1.3698	37 (42)
agnews_3	Joint	ReLU	128	512	0.0	0.000392	30	2	1.0011	2.0591	28 (30)
amazon	Joint	ReLU	64	256	0.0	0.001598	50	8	0.9756	6.9203	28 (33)
imdb	Joint	ReLU	64	256	0.0	0.000472	50	8	0.9837	7.6307	33 (38)
yelp	Pre	sigmoid	16	512	0.1	0.000418	50	8	0.9963	7.8576	29 (34)
Anomaly ratio $> 5\%$ (28 datasets)											
Wilt	Joint	ReLU	32	64	0.0	0.006360	50	4	0.9514	3.7027	20 (25)
donors	Pre	ReLU	64	64	0.1	0.007318	50	0.5	3.4513	2.0146	23 (28)
census	Pre	ReLU	64	512	0.5	0.000630	50	4	0.9874	3.7700	4 (9)
letter	Pre	ReLU	32	512	0.1	0.000682	50	0.5	1.0949	0.6695	32 (37)
shuttle	Pre	ReLU	16	512	0.1	0.001339	30	0.5	1.8290	0.9794	17 (21)
annthyroid	Joint	ReLU	64	64	0.0	0.007321	30	4	0.9791	2.2652	13 (18)
wine	Joint	sigmoid	128	64	0.2	0.002449	50	2	0.9015	2.7497	37 (41)
Stamps	Joint	ReLU	64	128	0.0	0.000343	50	8	0.9868	7.9296	42 (44)
mnist	Joint	ReLU	16	512	0.0	0.006167	30	0.5	2.6109	1.4730	7 (12)
PageBlocks	Pre	ReLU	16	512	0.3	0.001427	50	0.5	1.3424	0.8189	18 (23)
cardio	Joint	ReLU	64	64	0.0	0.007321	30	4	0.9447	3.7764	9 (14)
campaign	Joint	sigmoid	32	128	0.1	0.003063	50	0.5	2.7833	1.5136	16 (21)
vertebral	Pre	ReLU	128	64	0.1	0.003949	50	0.5	1.1330	0.8313	29 (34)
Hepatitis	Joint	ReLU	64	128	0.0	0.000343	50	8	0.9871	8.0233	50 (50)
InternetAds	Pre	ReLU	128	128	0.4	0.004182	30	4	0.9794	3.7063	3 (8)
landsat	Joint	ReLU	64	512	0.1	0.000336	50	0.5	1.0122	0.7759	34 (39)
skin	Joint	ReLU	64	64	0.0	0.006357	30	4	1.0099	3.9366	11 (16)
Cardiotocography	Pre	ReLU	64	512	0.4	0.001515	50	0.5	1.2810	0.7648	23 (28)
MVTec-AD_bottle	Pre	sigmoid	32	512	0.0	0.002286	50	0.5	1.0102	0.7007	26 (31)
MVTec-AD_cable	Pre	ReLU	32	256	0.2	0.000732	50	1	0.9628	1.1013	44 (47)
MVTec-AD_capsule	Joint	ReLU	64	256	0.0	0.000454	50	1	0.9814	1.0327	42 (45)
MVTec-AD_carpet	Pre	ReLU	128	256	0.0	0.002108	50	2	0.9823	2.0020	23 (28)
MVTec-AD_grid	Pre	ReLU	64	512	0.0	0.002411	30	2	0.9936	1.9614	12 (17)
MVTec-AD_hazelnut	Pre	ReLU	32	64	0.0	0.005938	50	2	0.9476	2.1314	35 (40)
MVTec-AD_leather	Joint	sigmoid	64	128	0.0	0.000707	50	1	0.9124	1.2055	50 (50)
MVTec-AD_metal_nut	Joint	ReLU	64	64	0.0	0.003848	30	4	0.9658	4.0335	30 (30)
MVTec-AD_pill	Joint	ReLU	32	256	0.0	0.001832	30	4	0.9838	3.8697	26 (28)
MVTec-AD_screw	Joint	ReLU	64	512	0.0	0.001709	30	1	0.9921	1.0419	15 (20)
MVTec-AD_tile	Pre	ReLU	32	512	0.0	0.000694	30	1	0.9806	1.0214	26 (29)
MVTec-AD_toothbrush	Pre	sigmoid	16	128	0.1	0.002198	50	0.5	0.8252	0.9360	44 (47)
MVTec-AD_transistor	Pre	sigmoid	16	128	0.1	0.002198	50	0.5	1.2320	1.0194	48 (50)
MVTec-AD_wood	Pre	ReLU	128	512	0.2	0.000577	50	1	0.9886	1.0406	34 (39)
MVTec-AD_zipper	Joint	sigmoid	16	256	0.0	0.003471	30	2	0.9858	2.2991	29 (30)
WPBC	Joint	ReLU	32	512	0.3	0.001559	50	8	0.9828	7.6739	32 (37)
satellite	Joint	ReLU	128	512	0.1	0.000497	50	4	0.9762	3.9307	24 (29)
yeast	Pre	ReLU	32	512	0.2	0.002962	50	2	1.0043	1.9028	13 (18)
fault	Pre	ReLU	64	64	0.0	0.002818	30	0.5	1.0210	0.6908	21 (26)
Pima	Pre	ReLU	64	256	0.1	0.003490	30	1	0.9772	1.0128	23 (25)
breastw	Joint	ReLU	64	512	0.0	0.001753	50	8	0.9888	7.7410	19 (24)
magic.gamma	Pre	ReLU	64	128	0.1	0.001926	50	8	0.9022	5.8646	30 (35)
Ionosphere	Pre	ReLU	16	512	0.3	0.001111	50	0.5	1.0517	0.6598	37 (40)
SpamBase	Joint	ReLU	128	512	0.0	0.004442	30	1	1.0263	1.0099	10 (15)

the eight natively probabilistic detectors. DAGMM, which has no result on three of the 29 primary datasets, is left out of the detection tests because the Friedman test needs a result for every compared model on every dataset; all rank statistics that enter a test are therefore among 28 methods. The correction is conservative at this many pairs, so throughout we read "no significant difference" as a failure to detect a difference and never as evidence of equality. All tests use $\alpha = 0.05$. The Friedman test rejects in every case: on the 29 primary datasets $\chi^2 = 538.6$ for AUROC and 557.6 for AUPR over the 28 detection models and 63.9 for the ten-bin ECE over the eight natively probabilistic detectors, all with $p < 10^{-10}$, and the values on the 57 and the 28 datasets are larger still. AUPR is the average precision of `sklearn`, without interpolation. On the primary scope, the Holm-corrected p-values of the calibration comparisons in which CRETAP is significantly better are below 10^{-4} against NB, 0.0054 against MLP, 0.0441 against LightGBM, and 0.0309 against XGBoost; against CatBoost, XGBOD, and RF the test does not separate the models.

I ATTRIBUTION MAPS: ADDITIONAL RESULTS

I.1 IMAGE DATA

Figures 6–11 show the other six image datasets of Section 6.2 in the format of Figure 1, and Table 18 lists, for each of the seven, the learned boundary, the test split, detection, and the e/k and within-class rank of every displayed image.

Table 18: The seven image datasets shown, seed 1: learned boundary k, test split, median e/k of the nominal and anomalous test images, detection (AUROC and AUPR, %, ↑), and the e/k and within-class rank by error of the displayed images (rank 1 = lowest error).

		Test		Median e/k				Displayed: e/k (rank)	
Dataset	k	nominal	anom.	nominal	anom.	AUROC	AUPR	nominal	anomalous
`bottle`	0.954	69	19	0.076	1.16	99.1	97.1	0.045 (2), 0.047 (3)	38.9 (19/19), 3.42 (17/19)
`cable`	1.02	85	28	0.578	0.877	76.6	63.6	0.499 (2), 0.500 (3)	41.8 (28/28), 25.7 (27/28)
`hazelnut`	1.01	130	21	0.198	1.31	99.4	96.2	0.039 (2), 0.040 (3)	25.7 (21/21), 14.9 (20/21)
`leather`	1.69	83	28	0.141	16.8	99.6	98.9	0.097 (2), 0.100 (3)	390 (28/28), 66.5 (24/28)
`wood`	1.15	80	18	0.046	5.69	99.0	96.9	0.026 (2), 0.031 (3)	115 (18/18), 25.1 (13/18)
`dotted_line`	0.971	2,850	150	0.143	2.17	100.0	100.0	0.065 (2), 0.066 (3)	3.14 (149/150), 3.08 (148/150)
`spatter`	0.943	2,850	150	0.089	6.25	100.0	99.9	0.046 (2), 0.047 (3)	17.4 (150/150), 16.8 (149/150)

CRETAP Scores. On nominal images the map stays pale: at most 7% of the pixels on the displayed nominal images other than `cable` reach $\tilde{s}_z \geq 1$, the level at which one pixel's error alone equals the boundary, and over all nominal test images the median fraction is 0.6–12% per dataset, with `cable` the exception (46% on the displayed nominal images, 50% over all). On anomalous images the largest CRETAP Scores, 10^3 and above, fall inside the annotated defect on the five MVTec AD datasets: the contamination and the broken rim of `bottle`, the missing cables of `cable`, the print and the cracked shell of `hazelnut`, the color and glue defects of `leather`, and the color defects of `wood`. MNIST-C ships no defect annotation, so for `dotted_line` and `spatter` we can only say that the largest scores coincide with the corrupted pixels, the added dots and the gray spatter. The map is not confined to the defect. Elevated CRETAP Scores also appear outside the annotation: on the background and along the left edge of the printed `hazelnut`, around the cracked one, and in a patch near the center of the broken `bottle`; and in `leather` and `wood` as scattered square blocks, together with a strip along the lower border of one `leather` image. These blocks are square, like the blocky high values at the defects, a shape we attribute to the window and stride of the convolution. On `cable`, whose nominal images sit at $e/k \approx 0.5$, the whole map of a nominal image is shaded rather than pale.

Reconstructed images. On nominal images the reconstruction largely follows the input. At the defect it departs strongly from the input, and values outside $[0, 1]$ are clipped for display. This follows from the objective rather than from rendering: CRETAP has no reconstruction term, and for an anomalous sample the BCE gradient with respect to $\log e$ is $n(p-1) < 0$, so training raises the error of anomalous samples and leaves the reconstruction at a defect free to move away from the input. The direction differs by dataset. The defect is reconstructed below zero in every channel and appears black in `bottle` and `hazelnut`, and in `spatter`, where the gray spatter is erased, together with part of the digit in one image; it is reconstructed above one in every channel and appears white in `cable`; and the channels move in opposite directions, giving saturated green in

leather and yellow in wood. In dotted_line the dots are redrawn one pixel to the side, dark on the original dots and bright on their neighbors. cable is an exception on nominal images as well: its reconstruction carries a global color offset, its red channel lying below zero on average, so even nominal reconstructions appear green. The decision depends on the size of the error rather than on the appearance of the reconstruction, and the map still outlines the missing cables, but the reconstructed image should not be read as a copy of the input.

I.2 Tabular data: concentration of the attributions

Definition. On the 20 low-contamination tabular datasets we reuse, without retraining, the fitted CRETAP models of Section 6.1 at seed 1; the test probabilities recomputed from the stored weights were checked to be bit-identical to the stored outputs. From the attribution map $b(x)$ of every test sample we take the within-sample share $q_j = b_j / \sum_{j'} b_{j'}$ (identical for b and for the CRETAP Scores s_j) and the three columns with the largest share. The *concentration* of a class is the mean top-3 overlap between two of its samples, relative to chance,

$$\mathrm{conc} = \frac{\sum_{j=1}^{d} c_j(c_j - 1)}{3\, N_{\mathrm{c}}(N_{\mathrm{c}} - 1)} \Big/ \frac{3}{d}, \tag{24}$$

where c_j is the number of the class's N_{c} test samples whose top three include column j, and $3/d$ is the value expected when every sample's three columns are drawn at random; a concentration of 1 means no recurring columns, and larger values mean that the samples of the class return to the same columns. The definition was fixed before the results were examined.

Results. Table 19 gives the concentration for each of the 20 datasets; Figure 2 shows the map of one of them. In nominal samples the error spreads over many columns: the top three columns of two nominal samples coincide 1.7 times as often as a random choice of columns would (median over the 20 datasets). In anomalous samples the error returns to the same columns, at 4.7 times chance. Anomalous concentration exceeds nominal concentration on 16 of the 19 datasets on which both are defined; the exceptions are ALOI (3.5 against 3.9) and speech (27.3 against 36.8), http is tied at 1.0, the ceiling for three features, and smtp has a single test anomaly. Because the chance value $3/d$ shrinks with the number of features, the ratio is larger on wide datasets—1.9 on the six-feature thyroid and mammography, above 50 on musk ($d = 166$) and backdoor ($d = 196$)—so values are best compared between the two classes of one dataset rather than across datasets.

Table 19: Concentration of the top-3 columns relative to chance (equation 24; 1 = chance, larger = more recurring columns) on the 20 low-contamination tabular datasets, seed 1, every test sample. `smtp` has one test anomaly, so its anomalous concentration is undefined; with $d = 3$, `http` and `smtp` are at the ceiling of 1.0.

		Test samples		Concentration ($\times$)	
Dataset	d	nominal	anomalous	nominal	anomalous
`http`	3	2,986	14	1.0	1.0
`smtp`	3	2,999	1	1.0	–
`mammography`	6	2,931	69	1.8	1.9
`thyroid`	6	1,104	28	1.2	1.9
`glass`	7	287	13	1.6	2.0
`WBC`	9	290	10	1.5	1.8
`cover`	10	2,971	29	1.2	2.9
`vowels`	12	422	15	1.3	3.7
`pendigits`	16	2,014	47	1.1	3.6
`Lymphography`	18	287	13	2.2	5.2
`Waveform`	21	1,003	30	1.6	4.7
`ALOI`	27	2,906	94	3.9	3.5
`fraud`	29	2,996	4	7.6	8.1
`WDBC`	30	292	8	1.8	10.0
`satimage-2`	36	1,720	21	5.7	7.6
`celeba`	39	2,933	67	1.3	6.1
`optdigits`	64	1,520	45	2.1	12.6
`musk`	166	890	29	4.6	55.3
`backdoor`	196	2,924	76	15.6	54.2
`speech`	400	1,088	18	36.8	27.3
Median				1.7	4.7

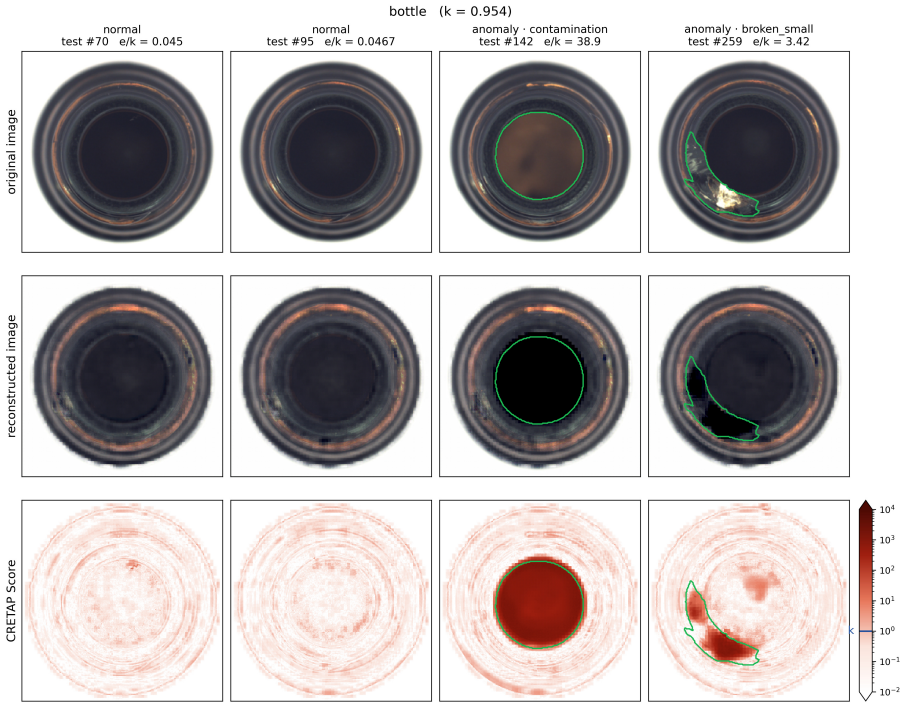

Figure 6: MVTec AD `bottle`, seed 1, in the format of Figure 1. Green outline: annotated defect.

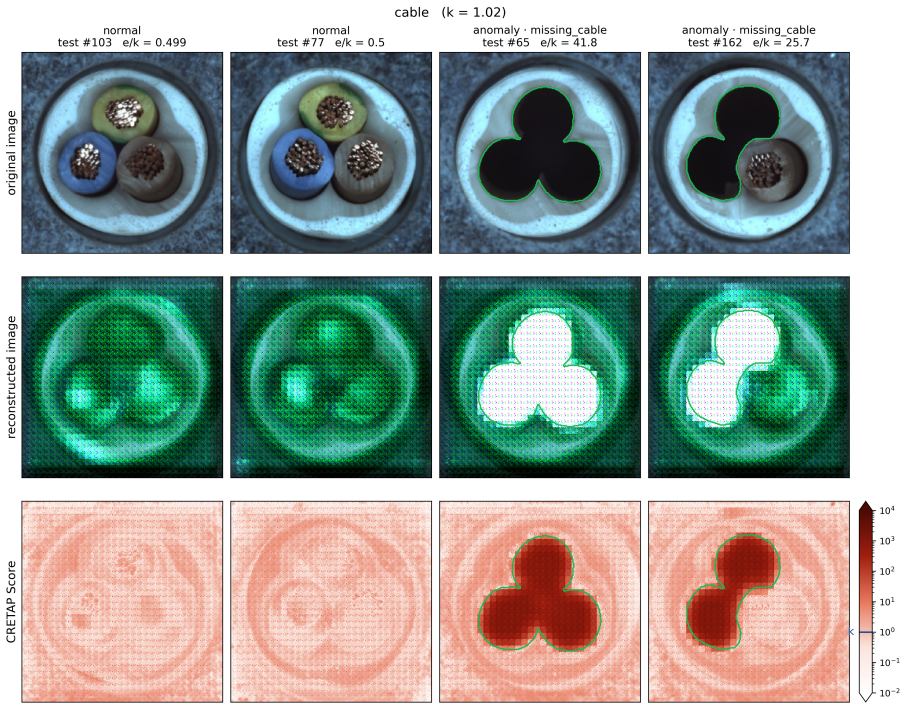

Figure 7: MVTec AD `cable`, seed 1, in the format of Figure 1. Green outline: annotated defect. The reconstruction carries a global color offset (Appendix I.1).

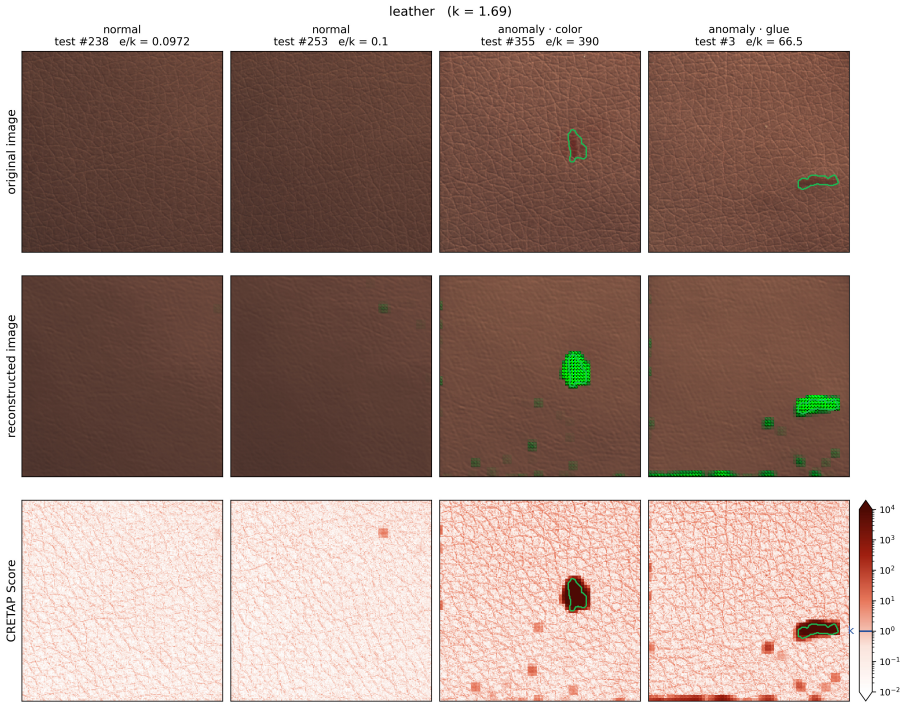

Figure 8: MVTec AD `leather`, seed 1, in the format of Figure 1. Green outline: annotated defect. The green patches in the reconstructed image are clipped values, not annotations.

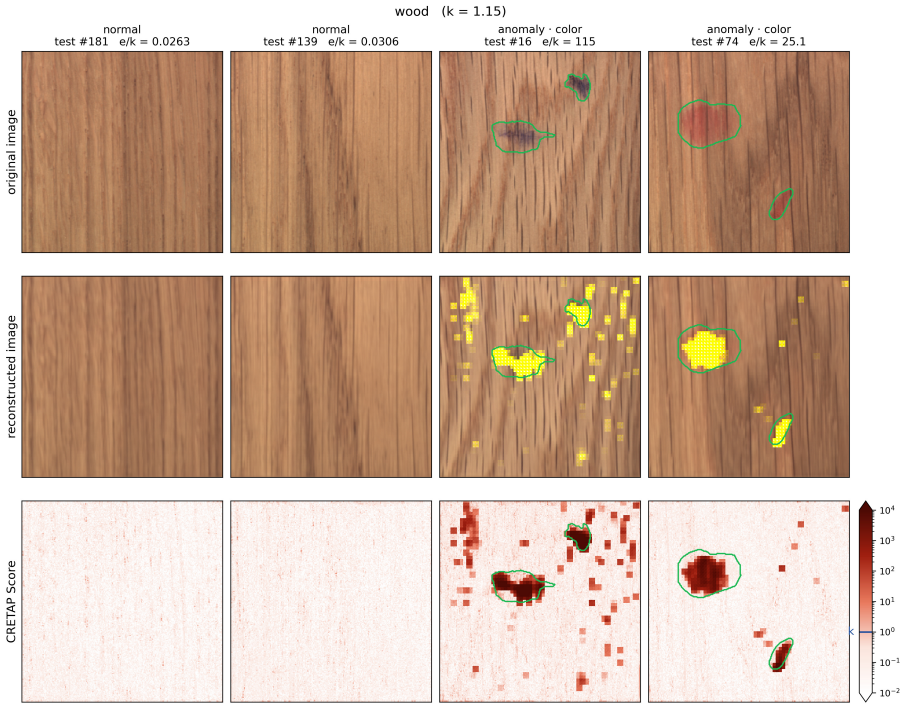

Figure 9: MVTec AD `wood`, seed 1, in the format of Figure 1. Green outline: annotated defect.

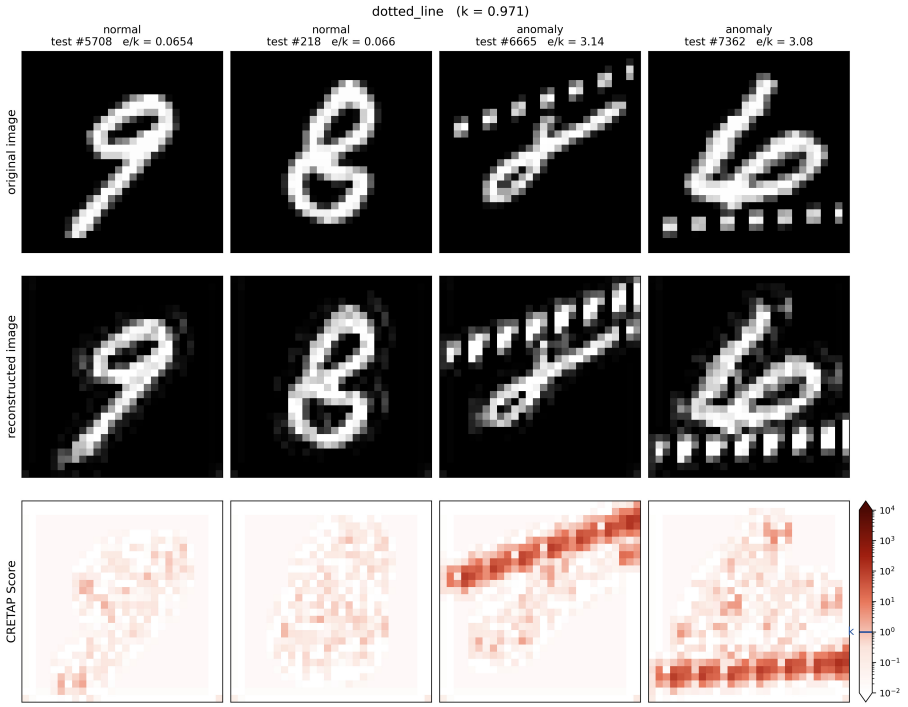

Figure 10: MNIST-C `dotted_line`, seed 1, in the format of Figure 1. No outline is drawn for MNIST-C.

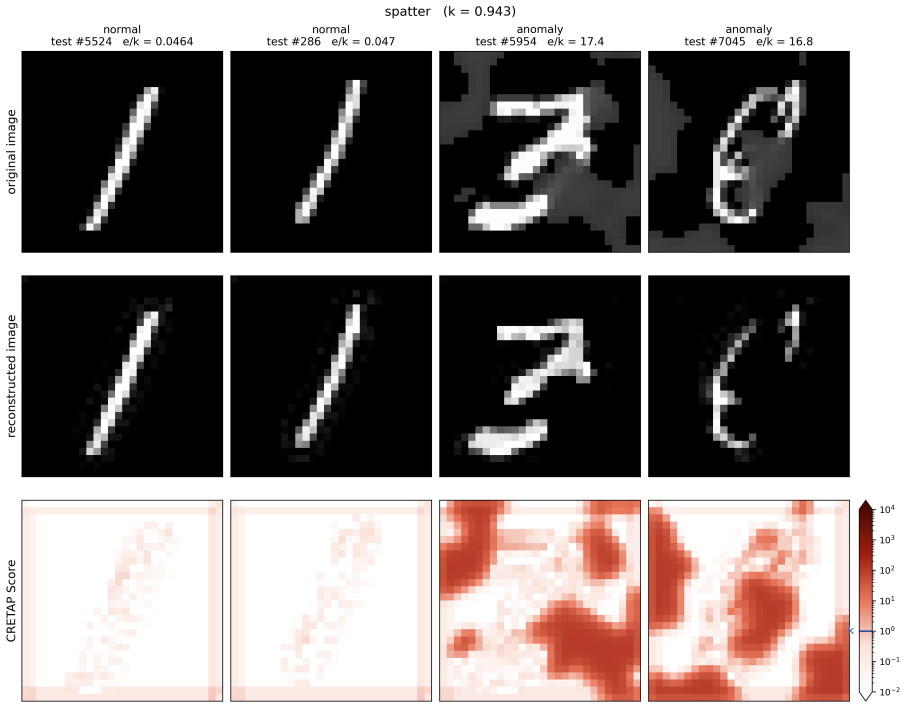

Figure 11: MNIST-C `spatter`, seed 1, in the format of Figure 1. No outline is drawn for MNIST-C.