The Threshold Paradox: Why Calibrating on the Test Set Introduces Bias in Anomaly Detection

Aurélie Cools[a], Sédrick Stassin[b] and Sidi Ahmed Mahmoudi[c]
Department of Informatics, Software and AI, University of Mons, Belgium

Keywords: Deep Learning, Computer Vision, Anomaly Detection, Industry 4.0, Threshold Calibration, KDE.

Abstract: Anomaly detection plays a central role in quality control systems within Industry 4.0. While recent deep learning approaches report promising AUROC scores, current evaluation protocols calibrate detection thresholds using test data annotations—a practice that violates unsupervised learning principles. We demonstrate through concrete examples and detailed analysis that standard benchmarks often include annotation errors, and test-calibrated methods systematically adapt to these errors, masking real defects. This creates a performance overestimation compared to deployment conditions. In this paper, we propose a hybrid approach combining an extension of Dinomaly, a self-supervised anomaly detection method, with reconstruction capabilities and a statistical Kernel Density Estimation-based threshold calibration using only normal training data. Our method achieves a recall of 98.1% on MVTec AD without test supervision, matching the state of the art while revealing annotation inconsistencies. This ensures truly unsupervised evaluation aligned with industrial constraints.

1 INTRODUCTION

Industry 4.0 integrates automation, artificial intelligence, and IoT to optimize production and maintenance (Rüßmann et al., 2015). In this context, anomaly detection—identifying deviations from normal behavior (Chandola et al., 2009)—is essential for detecting manufacturing defects before they affect product quality or disrupt operations. Vision-based quality control plays a key role in this transformation (Bergmann et al., 2021). Despite advances in deep learning, anomaly detection remains challenging due to class imbalance and the diversity of potential defects. Supervised methods require large numbers of labeled anomalies, which is impractical in industrial settings where defects are rare and often unknown in advance. This has motivated one-class and self-supervised approaches trained only on normal data.

Most critically, current evaluation protocols contain a methodological flaw: detection thresholds are often calibrated using annotated test data (Bergmann et al., 2020; Kim, 2022; Perini et al., 2023; Gungor and al., 2025). This assumes knowledge of the test distribution and requires anomaly labels at inference time—contradicting the unsupervised nature of the task. In this work, our contributions are as follows:

- We demonstrates empirically that test-dependent threshold calibration masks annotation errors and introduces methodological bias;
- We propose an approach combining reconstruction with KDE-based calibration performed solely on training data;
- We eliminate the bias of test-calibrated thresholding by enabling label-free inference;
- We reveal substantial performance gaps between train- and test-calibrated settings, showing that current protocols overestimate real-world performance.

The remainder of this paper is organized as follows. Section 2 reviews related work, followed by Section 3. Section 4 presents the proposed approach. Section 5 reports the results, which are discussed in Section 6. Section 7 concludes the paper.

2 RELATED WORK

In this section, we first review relevant anomaly detection approaches, ranging from reconstruction-

[a] https://orcid.org/0000-0002-2656-351X
[b] https://orcid.org/0000-0001-5179-9623
[c] https://orcid.org/0000-0002-1530-9524

Cools, A., Stassin, S. and Mahmoudi, S. A.
The Threshold Paradox: Why Calibrating on the Test Set Introduces Bias in Anomaly Detection.
DOI: 10.5220/0014392600004084

In *Proceedings of the 21st International Conference on Computer Vision Theory and Applications (VISAPP 2026) - Volume 1*, pages 475-482
ISBN: 978-989-758-804-4; ISSN: 2184-4321

based models to feature- and diffusion-based frameworks. We then discuss the critical but often overlooked issue of threshold calibration, which directly affects the validity of reported performance in unsupervised settings.

2.1 Anomaly Detection Approaches

Early statistical methods (Chandola et al., 2009) laid the foundations of anomaly detection by modeling normal data distributions. Reconstruction-based approaches using autoencoders (Zhou and Paffenroth, 2017) and VAEs (An and Cho, 2015) build on this principle, assuming that anomalies produce higher reconstruction errors. More recently, anomaly detection has increasingly relied on feature-based approaches that leverage powerful pretrained networks. PatchCore (Roth et al., 2022) constructs a memory bank of normal patch features and assigns anomaly scores via k-nearest neighbor search, while PaDiM (Defard et al., 2021) models the distribution of image features using multivariate Gaussians. Similarly, SPADE (Cohen and Hoshen, 2020) employs nearest-neighbor search in feature space to detect local deviations. A complementary line of work introduces synthetic anomaly generation to compensate for the scarcity of real defect samples. CutPaste (Li et al., 2021) artificially creates anomalies through patch manipulation, while DRAEM (Zavrtanik et al., 2021) combines a reconstruction module with a discriminative network trained on synthetic defects. Building on this idea, the NAS framework (Schlüter et al., 2022) generates realistic anomalies using procedural techniques such as Perlin noise, allowing models to generalize to more diverse failure modes. In parallel, generative modeling has gained momentum in anomaly detection. DDAD (Mousakhan et al., 2024) leverages denoising diffusion models to learn the distribution of normal images, while AnoDDPM (Wyatt et al., 2022) applies denoising diffusion probabilistic models to detect deviations. CFlow-AD (Gudovskiy et al., 2022) adopts conditional normalizing flows to achieve real-time detection with competitive accuracy. Most recently, Dinomaly (Guo et al., 2025) has emerged as a state-of-the-art method by combining self-supervised feature learning with Vision Transformers, offering both high detection accuracy and strong generalization across industrial categories.

2.2 The Threshold Calibration Problem

Despite methodological advances, most works rely on test-calibrated thresholding. CFlow-AD explicitly maximizes the F1-score on the test set, PatchCore optimizes the precision-recall curves using test annotations, and DRAEM selects thresholds based on test performance. This practice contradicts the principles of unsupervised learning, as it requires access to test set labels for threshold selection. Consequently, it introduces supervision into what should be an unsupervised evaluation protocol, leading to the biased evaluation we demonstrate in Section 5. To the best of our knowledge, no prior work has investigated the fundamental issue of test-calibrated thresholding in anomaly detection evaluation. While Fourure et al. (Fourure et al., 2021) demonstrates artificial F1 inflation and Bergmann et al. (Bergmann et al., 2020) briefly mentions the importance of threshold selection; neither study addresses the systematic use of test set annotations for threshold optimization that we expose in this work.

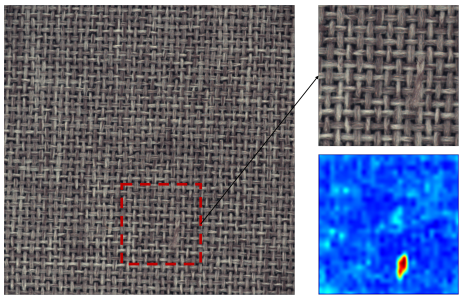
In summary, although recent methods have improved anomaly detection performance, most still rely on test labels for threshold selection. This contradicts the unsupervised setting and leads to overly optimistic results that do not reflect deployment conditions. This motivates our work, which calibrates thresholds using only normal training data while maintaining competitive performance.

3 PROBLEM STATEMENT

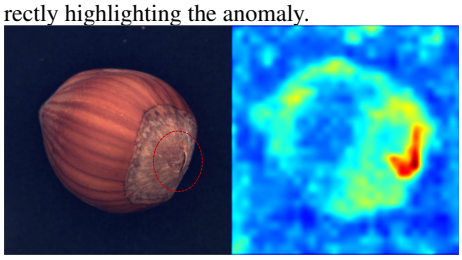
In this section, we highlight the fundamental problems of current evaluation protocols for anomaly detection. For all experiments presented in this study, we use the MVTec AD dataset (Bergmann et al., 2019), a standard benchmark widely used for visual anomaly detection. MVTec AD contains 15 object and texture categories: 10 object classes and 5 texture classes. Although deep learning methods achieve high AUROC scores on this benchmark, these results often rely on thresholds calibrated directly on the test set. This practice introduces a severe methodological bias and creates a discrepancy between reported performance and real deployment conditions.

3.1 Concrete Evidence: Annotation Errors in Standard Benchmarks

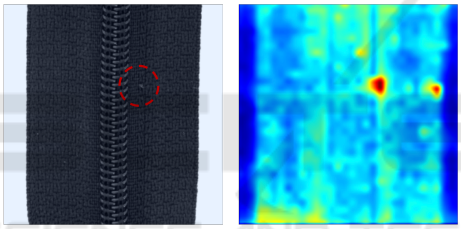
We identify several annotation errors in the MVTec AD test set that demonstrate the limitations of test-calibrated evaluation. This is particularly concerning because this test set is routinely used to tune detection thresholds in many recent methods, including PatchCore (Roth et al., 2022), DRAEM (Zavrtanik et al., 2021), and Dinomaly (Guo et al., 2025). Figure 1 il-



(a) Carpet class: Visible thread defect. The right panel shows our reconstruction map correctly highlighting the anomaly.

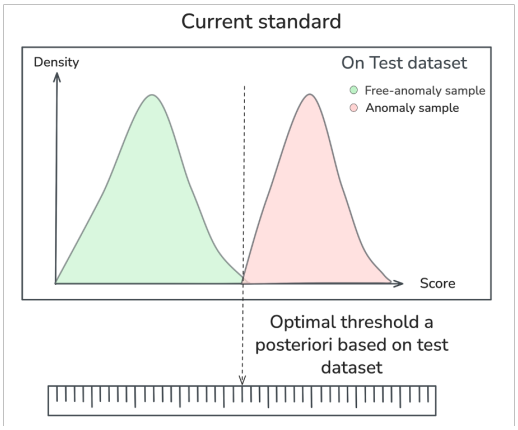


(b) Hazelnut class: Broken shell with visible crack. Our proposed method detects this obvious defect.

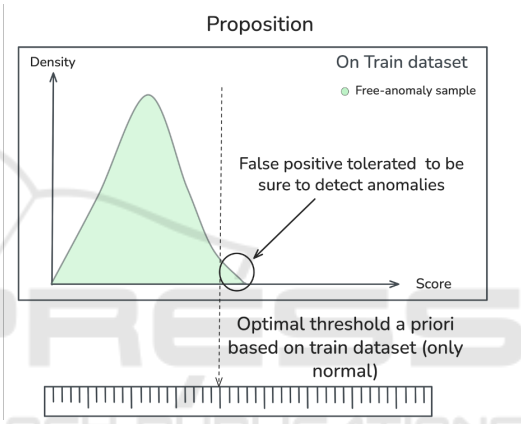


(c) Zipper class: Misaligned teeth clearly visible.

Figure 1: Annotation errors in MVTec AD: defective samples labeled as "good".



(a) Current standard: test-calibrated threshold selection adapts to annotation errors



(b) Our approach: KDE calibration on normal training data preserves anomaly sensitivity

Figure 2: Methodological comparison between test-calibrated approaches (a) and our proposed hybrid method (b).

lustrates concrete examples of obvious defects incorrectly labeled as "good" in the dataset.

When threshold selection relies on such mislabeled test data, test-calibrated methods adapt to annotation noise rather than detecting true anomalies. This results in two critical issues:

- **Performance Inflation:** Metrics are artificially boosted by aligning with erroneous labels
- **Safety Risks:** True defects remain undetected as the system is calibrated to disregard them

3.2 Current Methodological Contradiction

The vast majority of published methods use test set annotations to determine optimal thresholds. This practice directly corresponds to the workflow illustrated in Fig. 2a, where threshold selection depends on labeled test anomalies.

This dual use of test labels—for both calibration and evaluation—creates a feedback loop that inflates reported performance and conceals dataset quality issues. In contrast, our proposed approach (Fig. 2b) eliminates reliance on the test set entirely. We adopt an a priori threshold selection based solely on normal training data using Kernel Density Estimation (KDE) (Silverman, 2018; Scott, 2015), a non-parametric statistical method for estimating probability distributions. By modeling the distribution of anomaly scores from normal samples and setting the threshold at a high percentile of this distribution, we establish a decision boundary without any exposure to test annotations. This decoupling preserves sensitivity to genuine anomalies while avoiding the annotation bias shown in Figure 2a, and ensures that evaluation conditions match real deployment scenar-

ios where anomaly labels are not available.

3.3 Quantitative Evidence of Bias

Beyond qualitative examples, we conduct two systematic experiments on MVTec AD to demonstrate how AUROC-based test calibration overfits the test distribution.

3.3.1 First Experiment: Effect of Test-Set Availability

Objective and Methodology. We measure how partial access to test-set anomaly annotations biases threshold estimation. In real deployments, anomaly data is scarce and incomplete. We progressively vary the proportion of anomalies available during calibration (10%, 20%, 30%, up to 100%), fix the threshold by maximizing AUROC on this subset, then evaluate on the full test set. This protocol is applied with 10 independent runs per proportion for each class.

Results and Impact. Threshold values fluctuate strongly with the number of available anomaly samples, as shown in Fig. 3 for the leather class. Rather than converging to a stable value reflecting the true normal/anomalous boundary, the threshold varies dramatically depending on which samples are included. This reveals that test-calibrated methods depend on label availability rather than intrinsic anomaly characteristics. At 100%, no unseen anomalies remain to validate generalization—a fundamental flaw. The practical consequences are severe (Fig. 4): with only 10% of anomalies in the leather class, 30 out of 92 defects remain undetected—nearly one third. The system learns to ignore anomalies outside the calibration subset, undermining real-world reliability. This pattern extends across all 15 MVTec AD classes. The screw class is particularly affected: at 10%, 58 out of 119 anomalies are misclassified (±49%)—nearly half the defects ignored. Even at 30%, 21 anomalies remain undetected. These results confirm that test-calibrated thresholds become erratic when anomaly labels are scarce—precisely the condition defining real-world deployments.

3.3.2 Second Experiment: Robustness to Missing Anomaly Types

Objective and Methodology. We investigate whether test-calibrated thresholds generalize to unseen anomaly types. Production systems encounter diverse defect manifestations, and not all types may be present during initial calibration. We exclude one or more anomaly types from the test set during calibration, fix the AUROC-optimal threshold on the reduced dataset, then evaluate on the complete test set. We examine two cases: (A) excluding a single anomaly type, and (B) excluding multiple variants of the same anomaly family.

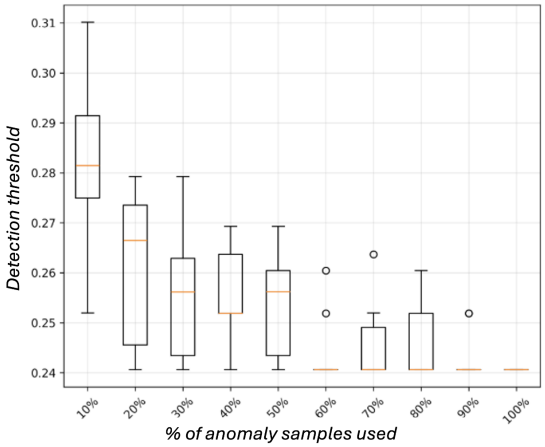


Figure 3: Threshold instability as a function of available anomaly samples in the leather class (MVTec AD).

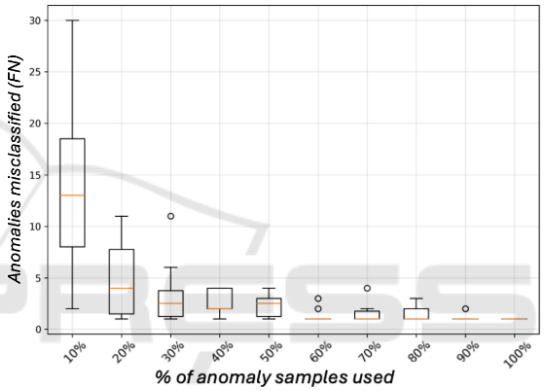


Figure 4: False negatives as a function of available anomaly samples in the leather class (MVTec AD). Total anomalies: 92.

Results and Impact: (A) Single Missing Type. Held-out anomaly types are severely under-detected, confirming that AUROC-based calibration overfits the observed distribution.

In the bottle class (Fig. 5), all anomaly types included during calibration yield a perfect confusion matrix. However, excluding only the ”contamination” sub-class causes five immediate misses. A single missing defect type breaks the illusion of reliability. This pattern occurs across all 15 classes, demonstrating that test-calibrated thresholds adapt only to seen cases, leaving unseen variations undetected.

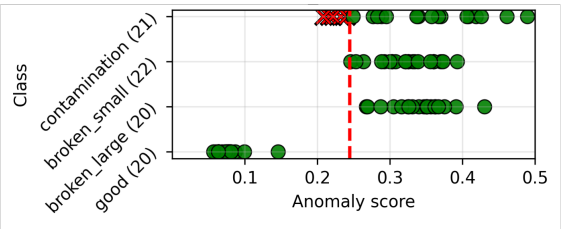


Figure 5: Bottle class (MVTec AD): excluding one anomaly type distorts calibration. Green: correct; red: misclassified; dashed red: threshold without contamination.

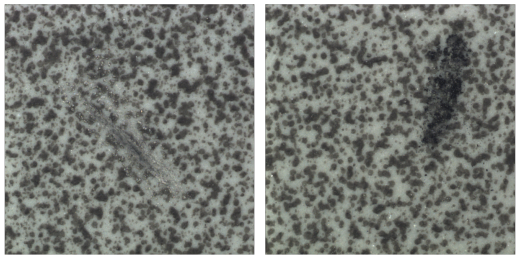


Figure 6: Tile class (MVTec AD): two anomaly variants of the same family (surface defects).

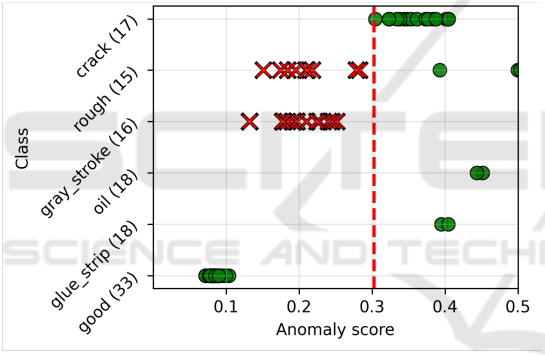


Figure 7: Tile class (MVTec AD): excluding two surface defect variants. Green: correct; red: misclassified; red line: threshold without selected types.

Results and Impact: (B) Multiple Related Variants. The failure becomes dramatic when multiple variants of the same anomaly family are absent—a realistic scenario since defects manifest with subtle variations. The tile class demonstrates this effect (Fig. 6), where two distinct variants both correspond to surface damage. In such cases, AUROC-based calibration adapts only to the remaining anomaly types, leading to thresholds that no longer reflect the full range of possible defects observed at deployment.

When both variants are excluded from calibration, the consequences are catastrophic (Fig. 7): 28 out of 30 anomalies are missed, while inclusion yields perfect detection. This demonstrates the fundamental fragility of test-based threshold calibration: it adapts narrowly to observed manifestations but fails catastrophically when relevant variations are absent.

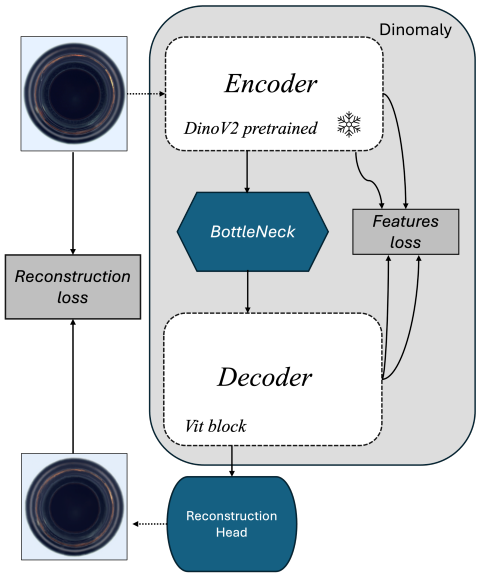


Figure 8: Hybrid architecture: frozen DINOv2 encoder, trainable ViT decoder with reconstruction head, and KDE-based score calibration.

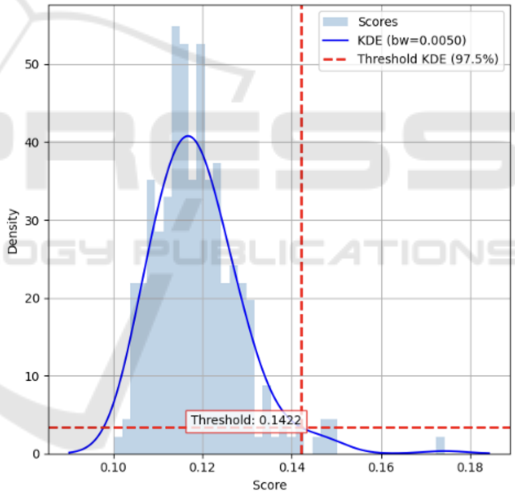


Figure 9: Score distributions (Screw class) showing heavy-tailed behavior motivating KDE calibration.

4 PROPOSED HYBRID METHOD

Anomaly detection in industrial visual inspection requires both expressive feature representations and principled threshold calibration.

We therefore propose a hybrid method that combines both perspectives. As shown in Fig. 8, a frozen DINOv2 encoder extracts high-level features and a trainable ViT-based decoder reconstructs normal patterns for pixel-level anomaly localization. The resulting anomaly scores are calibrated using Kernel Density Estimation (KDE), which is well-suited to the

non-Gaussian and heavy-tailed score distributions observed in practice (Fig. 9). Calibration is performed exclusively on normal training data: 75% is used for model training and 25% is reserved for KDE-based threshold estimation, ensuring complete separation between training, calibration, and evaluation.

5 TESTS OF OUR METHOD

5.1 Experimental Setup

Similarly to previous experiments, we evaluate our method on MVTec AD (Bergmann et al., 2019).

Comparisons. We compare four approaches: Dinomaly (test-calib), the original state-of-the-art baseline; Ours (test-calib), our hybrid extension of Dinomaly evaluated with the same AUROC-based calibration for fair comparison; Dinomaly (KDE), the original baseline with KDE calibration applied; and Ours (KDE), our main contribution relying on statistical KDE-based calibration. In all cases, the same hyperparameters were used across the entire dataset to ensure methodological fairness.

Implementation Details. All experiments were implemented in PyTorch 2.1.1 using PyTorch Lightning. Training was conducted on a NVIDIA RTX 3090 GPU with 24 GB of memory. Input images were resized to 448×448 pixels with a batch size of 8. Models were trained for a total of 5000 iterations. Optimization was performed with StableAdamW, using a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}.

5.2 Main Results

Test-calibrated results (the first two rows of Table 1) give the illusion of superior performance, with both Dinomaly and our method reaching nearly perfect F1-scores.

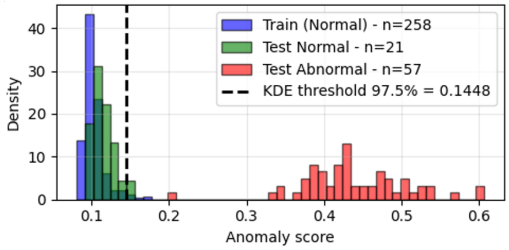
Table 1: Performance comparison between test-calibrated and KDE-calibrated methods on MVTec AD: Dinomaly baseline vs. our adapted model (image-level).

Method	Acc.	Recall	F1
Dinomaly (test-calib)	0.98	0.98	0.99
Ours (test-calib)	0.97	0.97	0.98
Dinomaly (KDE)	0.79	1.00	0.87
Ours (KDE)	0.95	0.98	0.96

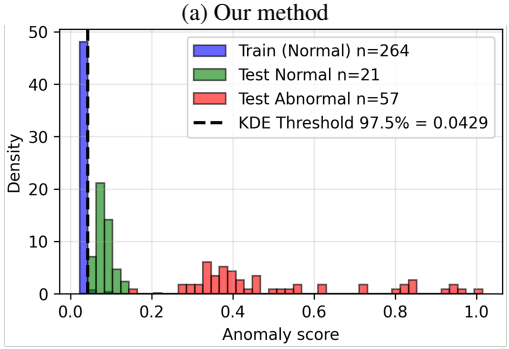
All values rounded to 2 decimal places.

We report image-level metrics, as the primary objective in industrial anomaly detection is to decide whether an image contains a defect; localization only becomes meaningful once reliable detection is achieved. However, this apparent performance comes at the cost of methodological integrity, since calibration relies on test annotations. When switching to KDE calibration—chosen because training score distributions deviate from Gaussianity and exhibit heavy tails—the weaknesses of test-calibrated thresholds become evident. The third row shows that Dinomaly (KDE) collapses in accuracy (0.79) despite perfect recall, reflecting an inability to distinguish normal from anomalous samples without guidance from test labels. By contrast, the fourth row shows that our hybrid method maintains strong and balanced performance under KDE calibration (0.95 accuracy, 0.98 recall, 0.96 F1). The threshold is obtained from the KDE-estimated training distribution by selecting the 97.5th percentile. This highlights the stabilizing role of the reconstruction module, which regularizes score distributions and enables KDE to operate effectively. Our approach not only improves over Dinomaly in the unsupervised setting but also aligns with deployment reality, where no anomaly labels are available for threshold selection.

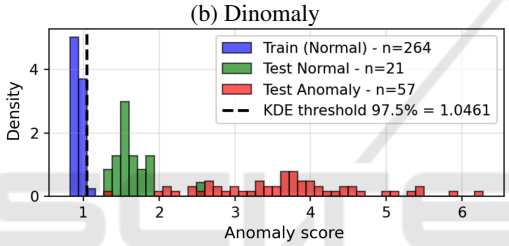
We also applied KDE-based calibration to PatchCore, which builds a memory bank from a representative subset of training patches selected using a greedy coreset strategy. While PatchCore achieves excellent separation between normal and anomalous samples when calibrated on the test set—yielding near-perfect AUROC scores—this result is misleading. Anomaly scores of training samples are systematically lower than those of normal test samples, since the training images themselves constitute the memory bank while the test normals do not. As a consequence, when thresholds are fixed using the training set, many test normals are incorrectly classified as anomalies. This exposes a fundamental limitation of applying density-based calibration on top of PatchCore: performance appears perfect under test calibration but collapses under a realistic unsupervised setting. The score distributions in Figure 10 confirm this analysis. Our method (Fig. 10a) shows clear separation between training normal scores and test normal scores, with both distributions well separated from anomalies, enabling reliable KDE-based threshold selection. In contrast, both Dinomaly (Fig. 10b) and PatchCore (Fig. 10c) exhibit systematically lower training scores compared to test normals, confirming that the training distribution cannot be directly used to set a reliable threshold without overfitting to test data.



(a) Our method



(b) Dinomaly



(c) PatchCore

Figure 10: Score distributions for the grid class (MVTec AD) across the compared methods.

6 CRITICAL ANALYSIS AND DISCUSSION

Our experiments provide clear evidence of the methodological flaws introduced by AUROC-based calibration. As shown in Figure 3 and 4, reported performance systematically increases with greater access to test labels—even when only a fraction of anomalies is used. This confirms that test-set calibration inflates metrics rather than reflecting genuine improvements. The problem is amplified when anomaly classes are withheld: AUROC calibration then fails to detect them, adapting to the observed distribution instead of generalizing to unseen cases. This is particularly problematic in industrial or safety-critical contexts, where detecting novel defects is essential. In contrast, our KDE-based approach remains stable regardless of test data availability or anomaly distribution. Because it relies exclusively on normal training data, it requires no test-time supervision and better reflects the unsupervised setting. Rather than concealing annotation errors—as test-calibrated thresholds do—our method exposes them, highlighting limitations of existing benchmarks. From an industrial perspective, the impact is significant: calibrating on only a fraction of test anomalies can leave up to 30% of defects undetected, meaning defective components may still reach end users, with potentially serious consequences in sectors such as automotive, aerospace, or healthcare. Thus, abandoning test-based calibration is not only scientifically justified but also necessary for reliable deployment in Industry 4.0 environments. These findings call current evaluation protocols into question: if thresholds are tuned on the same data used for evaluation, method comparison becomes unreliable and inflated metrics create an illusion of progress while masking weaknesses in models and datasets. By breaking this feedback loop, our hybrid approach offers a principled alternative that preserves methodological integrity and yields more trustworthy insight into anomaly detection performance. While we focus on KDE as a non-parametric estimator suited to heavy-tailed scores, other approaches such as extreme-value theory or adaptive quantile estimators could also be explored. Our goal is not to exhaustively benchmark calibration strategies, but to advocate for test-independent thresholding that better reflects deployment reality. In future work, we plan broader validation across datasets, ablation experiments, deeper hyperparameter tuning, and comparison with reconstruction-based methods.

7 CONCLUSION

We have shown that current evaluation protocols in anomaly detection suffer from a methodological contradiction: thresholds are often calibrated on the test set itself, introducing bias, inflating performance, and masking annotation errors. Our contribution addresses this issue by extending the state-of-the-art Dinomaly with a hybrid design that couples its neural reconstruction architecture with a statistical KDE-based threshold calibration. This approach not only improves Dinomaly's effectiveness (achieving 98.1% recall) but also preserves methodological integrity by separating calibration from test annotations. Importantly, it exposes annotation errors rather than adapting to them, offering a more faithful assessment of both model capability and dataset quality. Moreover, the proposed thresholding strategy is entirely independent of the test set, improves precision by avoiding adaptation to mislabeled anomalies, and generalizes across MVTec AD classes without requiring class-

specific hyperparameter tuning. We therefore advocate for a paradigm shift: anomaly detection research must abandon test-calibrated thresholds and instead adopt thresholding strategies derived solely from normal data, in order to respect the unsupervised setting and better reflect deployment reality.

ACKNOWLEDGEMENT

The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247.

REFERENCES

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18.

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., and Steger, C. (2021). The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2019). Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2020). Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Cohen, N. and Hoshen, Y. (2020). Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*.

Defard, T., Setkov, A., Loesch, A., and Audigier, R. (2021). Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pages 475–489. Springer.

Fourure, D., Javaid, M. U., Posocco, N., and Tihon, S. (2021). Anomaly detection: How to artificially increase your f1-score with a biased evaluation protocol. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 3–18. Springer.

Gudovskiy, D., Ishizaka, S., and Kozuka, K. (2022). Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 98–107.

Gungor, O. and al. (2025). A robust framework for evaluation of unsupervised time-series anomaly detection. In Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, C.-L., Bhattacharya, S., and Pal, U., editors, *Pattern Recognition*, pages 48–64, Cham. Springer Nature Switzerland.

Guo, J., Lu, S., Zhang, W., Chen, F., Li, H., and Liao, H. (2025). Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20405–20415.

Kim, S. e. a. (2022). Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 7194–7201.

Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. (2021). Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674.

Mousakhan, A., Brox, T., and Tayyub, J. (2024). Anomaly detection with conditioned denoising diffusion models. In *DAGM German Conference on Pattern Recognition*, pages 181–195. Springer.

Perini, L., Bürkner, P.-C., and Klami, A. (2023). Estimating the contamination factor's distribution in unsupervised anomaly detection. In *International Conference on Machine Learning*, pages 27668–27679. PMLR.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. (2022). Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328.

Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., and Harnisch, M. (2015). Industry 4.0: The future of productivity and growth in manufacturing industries. *Boston consulting group*, 9(1):54–89.

Schlüter, H. M., Tan, J., Hou, B., and Kainz, B. (2022). Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489. Springer.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Wyatt, J., Leach, A., Schmon, S. M., and Willcocks, C. G. (2022). Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 650–656.

Zavrtanik, V., Kristan, M., and Skočaj, D. (2021). Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339.

Zhou, C. and Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674.