GRADPAO: GRADIENT-BASED PSEUDO-ANOMALY OPTIMIZATION FOR TABULAR ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular anomaly detection (TAD) is an important task since many real-world machine learning applications operate on tabular data, and it remains challenging due to the scarcity of labeled samples and the ambiguous definition of anomalies. Despite the rapid progress of tabular foundation models (TFMs) on predictive tasks, building a foundation model for TAD is especially difficult since its anomaly prior must cover diverse data types and define what counts as abnormal before any target dataset is seen. We therefore ask how a pretrained TFM can be turned into an anomaly detector, rather than developing another foundation model for TAD. The obstacle is that in-context learning of TFMs operates on a labeled context set, while only normal samples are available in TAD, so no anomaly class exists for a test query to be assigned to. To fill this gap, we propose **Grad**ient-Based **P**seudo **A**nomaly **O**ptimization for tabular anomaly detection (GradPAO), a novel framework that supplies the missing classes by defining pseudo-anomalies as trainable parameters and optimizing them through the foundation model. GradPAO first places the pseudo-anomalies using the inter-sample attention of the foundation model, and then optimizes them to be separable from the normal samples and diverse among themselves. Extensive experiments on two benchmarks with 57 and 690 datasets demonstrate that GradPAO achieves state-of-the-art performance.

1 INTRODUCTION

Anomaly detection, which aims to identify samples that deviate from normal patterns, is a fundamental task in machine learning. It plays a vital role in numerous domains, including finance (Al-Hashedi & Magalingam, 2021), healthcare (Fernando et al., 2021), manufacturing (Esmaeili et al., 2023), and cybersecurity (Malaiya et al., 2019). Tabular anomaly detection (TAD) is particularly important, since many real-world machine learning applications operate on tabular data. However, it remains a challenging problem due to the scarcity of labeled data and the ambiguous notion of abnormal patterns. Because anomalies are rare and costly to annotate, detection models are often trained on normal samples alone, a setting known as one-class classification. This difficulty is compounded in the tabular domain, where features are heterogeneous and do not have shared structure across datasets. Many machine learning methods and deep neural networks have been proposed to address these challenges (Ramaswamy et al., 2000; Breunig et al., 2000; Liu et al., 2008; Ruff et al., 2018; Graham et al., 2023; Livernoche et al., 2024). However, these methods must be trained separately for each dataset, preventing direct reuse of a pretrained cross-dataset prior and incurring additional per-dataset training cost.

Meanwhile, tabular foundation models (TFMs) have emerged as a new paradigm for tabular predictive tasks such as multiclass classification and regression (Hollmann et al., 2023; Qu et al., 2026; Zhang et al., 2025). Given a labeled context set, they make predictions on a downstream task through in-context learning without any model update, and have shown remarkable performance. Inspired by this paradigm, recent work has extended it to TAD by pretraining foundation models for anomaly detection. FoMo-0D (Shen et al., 2025) is the first foundation model for TAD, pretrained on synthetic datasets with labeled normal and abnormal samples to classify each query as normal or abnormal against a normal-only context set. At test time, its predicted probability serves as the anomaly score without any model update. OutFormer (Ding et al., 2026b) advances this approach with a richer synthetic data prior. Both, however, require large-scale pretraining dedicated to TAD, and the anomalies they can recognize are largely influenced by a prior that is defined before any target dataset is seen.

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.

As studied by Ding et al. (2026b), this paradigm becomes suboptimal when pretraining data fail to cover diverse anomaly types. Consequently, the development of foundation models for TAD remains challenging and has progressed slowly, whereas TFMs for predictive tasks have been actively studied. This gap leads us to raise a research question: *can we leverage TFMs validated on predictive tasks for TAD, rather than developing a foundation model solely for TAD?* A naive approach is to use a TFM as a reconstruction-based detector: treat one column of the test input as the target, predict it from a normal-only context set, and aggregate reconstruction errors over columns. However, this often shows suboptimal performance since averaging errors across columns can dilute the abnormal signal, and repeating the procedure for every column is computationally expensive (see Section 4.2).

This limitation motivates us to consider other ways of leveraging a pretrained TFM. The in-context learning of a TFM operates on a **labeled** context set, and the key obstacle to directly applying a TFM to TAD is the absence of anomaly classes in the context set. Thus, if we construct anomaly classes and include them in the context set with the normal samples, we can obtain the anomaly score of a test query as the probability that a TFM does not assign it to the normal class. This approach directly exploits the classification capability of a TFM and does not require multiple inference over the columns to aggregate reconstruction errors. Constructing pseudo-anomalies with heuristic rules, however, does not ensure that the resulting anomalies serve as effective anomaly classes.

To address this difficulty and construct effective pseudo-anomalies, we propose **Grad**ient-Based **P**seudo **A**nomaly **O**ptimization for tabular anomaly detection (GradPAO). Our key idea is to define pseudo-anomalies as trainable parameters and optimize them by gradient descent through the pretrained foundation model. Our GradPAO consists of two stages.

- **Attention-based pseudo-anomaly initialization** provides starting points for pseudo-anomalies from the structure of the normal samples captured by the foundation model. A poor starting point can make the subsequent optimization difficult, so we cluster the normal samples under the inter-sample attention of the foundation model and displace the least attended rows of each cluster outward from its centroid. Displaced rows from the same cluster form one pseudo-anomaly class to represent diverse abnormal patterns, and they are defined as trainable parameters for the subsequent optimization (see Section 3.2).
- **Pseudo-anomaly optimization** optimizes these initial points into effective anomaly classes to make them separable from the normal samples and diverse among themselves. For separability, we update only the pseudo-anomalies by gradient descent through the frozen TFM to minimize the cross-entropy objective of classifying the normal samples and the pseudo-anomalies into their own classes. This objective alone, however, leaves the pseudo-anomalies free to collapse onto a single point. Thus, we add a diversity objective on the anomaly-to-normal attention scores to encourage pseudo-anomalies to attend to normal samples in distinct ways (see Section 3.3).

GradPAO differs from prior work in both where the anomaly classes come from and how they are obtained. For FoMo-0D (Shen et al., 2025) and OutFormer (Ding et al., 2026b), what counts as abnormal is defined by the pretraining prior rather than by the target dataset. GradPAO instead constructs anomaly classes from the normal samples of the target dataset using a foundation model. Generating additional samples from a target dataset has also been studied in tabular data synthesis (Kotelnikov et al., 2023; Zhang et al., 2024), but these methods cannot directly provide anomaly classes. They learn the distribution of the target dataset, mainly to augment training data for downstream classification, so they can only generate normal-like samples from a normal-only dataset. Thus, constructing anomaly classes from normal samples remains underexplored. We address this gap by optimizing pseudo-anomalies through the foundation model to construct the missing anomaly classes from the normal-only dataset without an additional generator.

Extensive experiments on two benchmarks with 57 and 690 datasets, ADBench (Han et al., 2022) and OddBench (Ding et al., 2026a), show that GradPAO achieves the best average rank and rAUC among all compared methods, spanning classical, deep learning, and TAD foundation model baselines. This demonstrates that a tabular foundation model pretrained for predictive tasks, once supplied with optimized anomaly classes, can be turned into an effective tabular anomaly detector.

In summary, the contributions of this paper are as follows:

- We identify that the key obstacle to using a tabular foundation model as an anomaly detector is the absence of anomaly classes, and show that supplying effective anomaly classes can turn the tabular foundation model into a strong tabular anomaly detector.

- We propose GradPAO, a novel framework that constructs pseudo-anomaly classes for each target dataset through the frozen foundation model. Its two stages place the pseudo-anomalies from the inter-sample attention and optimize them with separability and diversity objectives, requiring no generator, no update to the foundation model, and no large-scale pretraining dedicated to TAD.
- Extensive experiments on ADBench and OddBench, with 57 and 690 datasets, demonstrate that GradPAO achieves state-of-the-art performance, supported by ablation studies and analyses of how the optimization shapes the pseudo-anomalies.

2 Related Work

2.1 Tabular Anomaly Detection

Most work on tabular anomaly detection assumes a one-class classification setting, where detection models are often trained on normal samples alone, since collecting and annotating anomalies are costly. Under this setting, existing methods differ in how they define an anomaly score or how they learn normal patterns. Classical methods measure statistical deviation from the normal samples through distance to neighbors, local density, or the difficulty of isolating a sample by random splits (Ramaswamy et al., 2000; Breunig et al., 2000; Liu et al., 2008). Deep learning methods instead learn the notion of normality by mapping normal samples into a compact region (Ruff et al., 2018), by solving self-supervised tasks on them (Bergman & Hoshen, 2020; Shenkar & Wolf, 2022), or by modeling them through a diffusion process (Graham et al., 2023; Livernoche et al., 2024). Despite their diverse modeling strategies, all of these methods must be trained from scratch on each dataset. There is no pretrained prior to build upon, so nothing learned on one dataset transfers to another, and each new dataset requires selecting a method and tuning its hyperparameters without labeled data to validate against.

2.2 Tabular Foundation Models

Tabular foundation models (TFMs) such as TabPFN (Hollmann et al., 2023) and TabICLv2 (Qu et al., 2026) are prior-data fitted networks, pretrained on large synthetic datasets drawn from a prior over tabular predictive tasks (*e.g.*, classification and regression) so that a single forward pass performs approximate Bayesian inference in context. Given a context set of labeled rows and a query row, they infer the input-label relationship without any per-dataset training, so the context set is where task-specific information enters the model. To handle the heterogeneity of tabular data, these models attend both across columns within a row and across rows within a table, and the resulting inter-sample attention encodes which rows the model regards as related.

Inspired by the success of TFMs on predictive tasks, recent work has extended this paradigm to TAD. FoMo-0D (Shen et al., 2025) is pretrained only on synthetic data sampled from Gaussian Mixture Models and predicts a test query's anomaly score from a normal-only context set. OutFormer (Ding et al., 2026b) advances this approach by diversifying the synthetic data prior with structural causal models and copulas, and by introducing a self-evolving curriculum that adaptively balances training across these heterogeneous priors. These methods require large-scale pretraining on diverse types of data dedicated solely to TAD, and the anomalies they can recognize are shaped by a prior that is designed before any target dataset is seen. GradPAO instead reuses a TFM already pretrained for predictive tasks and requires no pretraining dedicated to TAD. It constructs the pseudo-anomaly classes from the normal samples of the target dataset by defining the pseudo-anomalies as trainable parameters and optimizing them. Once the pseudo-anomalies are constructed, a test query is scored in a single forward pass with a context set of normal samples and optimized pseudo-anomalies.

2.3 Tabular Data Synthesis

Tabular data synthesis has been proposed to mitigate the scarcity of training samples in the tabular domain. Most work learns the distribution of a target dataset with a generative model to improve on downstream predictive tasks such as classification and regression (Xu et al., 2019; Kotelnikov et al., 2023; Zhang et al., 2024). SPADA (Yang et al., 2025) further uses large language models to understand the structure of the target dataset. Although these methods have improved performance on predictive tasks, they cannot be directly applied to anomaly detection. Their goal is to learn the

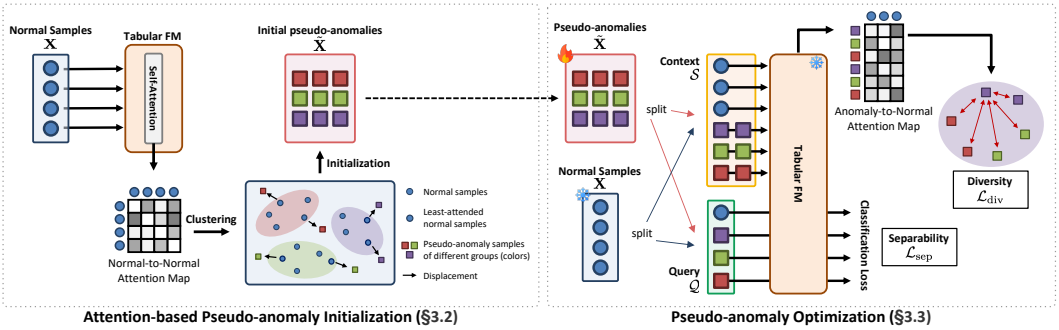


Figure 1: Overview of GradPAO.

distribution of the training data, so they cannot produce unseen anomaly classes from a normal-only dataset. In contrast, GradPAO constructs anomaly classes from a normal-only target dataset by optimizing pseudo-anomalies through the frozen foundation model, requiring no additional generator.

3 METHOD

3.1 OVERVIEW

We begin by providing an overview of our framework and its motivation. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be the normal-only train set of a target dataset, $\mathbf{x}_q \in \mathbb{R}^d$ be a test query, and f_ϕ be a frozen TFM that performs in-context classification over a label set $\mathcal{Y} = \{0, 1, \ldots, C\}$. Given a labeled context set $\mathcal{S} = \{(\mathbf{x}'_j,\ y_j)\}_{j=1}^{n'}$ with labels $y_j \in \mathcal{Y}$ and $\mathbf{x}_q$, f_ϕ applies inter-sample attention among the context samples and lets $\mathbf{x}_q$ attend to them, producing a predictive distribution over $\mathcal{Y}$, denoted by $p_\phi(y \mid \mathbf{x}_q, \mathcal{S})$. The predicted label is the most probable class,

$$\hat{y}_q = \arg\max_{y \in \mathcal{Y}} p_\phi(y \mid \mathbf{x}_q, \mathcal{S}). \tag{1}$$

In tabular anomaly detection, however, every row in $\mathbf{X}$ carries the same label (*i.e.*, normal) and no anomaly class exists for the query to be assigned to, so f_ϕ cannot be directly applied with the normal-only train set as the context set.

To fill this gap, we propose **Grad**ient-Based **P**seudo **A**nomaly **O**ptimization for tabular anomaly detection (GradPAO). Our key idea is to define pseudo-anomalies as trainable parameters and optimize them by gradient descent through the pretrained foundation model. Once they are obtained, we can construct a labeled context set from the normal samples and the pseudo-anomalies, and classify a test query against it with f_ϕ directly.

We organize the pseudo-anomalies into C groups and assign a different anomaly class to each group so that they represent diverse abnormal patterns. Given the pseudo-anomalies $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_{c,l} \in \mathbb{R}^d \mid 1 \le c \le C,\ 1 \le l \le k\}$ with k rows per group, we construct the context set by labeling every normal row as the normal class 0 and every pseudo-anomaly row as its own group index c:

$$\mathcal{S} \ = \ \{(\mathbf{x}_i,\ 0)\}_{i=1}^{n} \ \cup \ \{(\tilde{\mathbf{x}}_{c,l},\ c) \mid 1 \le c \le C,\ 1 \le l \le k\}. \tag{2}$$

Then, the anomaly score is the probability that f_ϕ does not assign a test query to the normal class,

$$s(\mathbf{x}_q) \ = \ 1 - p_\phi(y = 0 \mid \mathbf{x}_q, \mathcal{S})\,, \tag{3}$$

which requires one forward pass of f_ϕ. The pseudo-anomalies $\tilde{\mathbf{X}}$ are constructed once for each target dataset, and all test queries share the same context set $\mathcal{S}$.

To construct effective pseudo-anomalies $\tilde{\mathbf{X}}$, we introduce a two-stage framework: (*i*) **attention-based pseudo-anomaly initialization** (Section 3.2), which provides the starting points for pseudo-anomalies using the inter-sample attention of f_ϕ, and (*ii*) **pseudo-anomaly optimization** (Section 3.3), which optimizes them to be separable from the normal samples and diverse among themselves. We describe each stage in the following sections, with the overall pipeline in Figure 1 and detailed algorithms in Appendix A.1.

3.2 Attention-based Pseudo-anomaly Initialization

We now describe how the pseudo-anomalies $\tilde{\mathbf{X}}$ are initialized from the structure of the normal samples captured by the inter-sample attention of f_ϕ. This initialization plays a crucial role since a poor starting point can make the subsequent optimization difficult. To capture the structure of the normal samples for this initialization, we leverage the inter-sample attention of f_ϕ rather than distances in the input space. Since heterogeneous tabular columns differ in scale and semantics, the distance in the input space can be an unreliable signal for this structure. In contrast, the inter-sample attention reflects which samples the foundation model relates to one another, as f_ϕ is pretrained to model dependencies across rows and columns.

Formally, given the normal-only train set $\mathbf{X} \in \mathbb{R}^{n \times d}$, we run a single forward pass of f_ϕ and extract the inter-sample attention map $\mathbf{M} \in \mathbb{R}^{n \times n}$ from an inter-sample attention layer, where $\mathbf{M}_{ij}$ is the attention that row i pays to row j. We symmetrize it as $\bar{\mathbf{M}} = (\mathbf{M} + \mathbf{M}^\top)/2$, convert it into a distance $\mathbf{D} = 1 - \bar{\mathbf{M}}/\max(\bar{\mathbf{M}})$, and group the normal samples into C clusters by agglomerative clustering under $\mathbf{D}$, whose index sets we denote by $\mathcal{G}_1, \ldots, \mathcal{G}_C$. As each cluster leads to one anomaly class, we fix $C = 9$, since most TFMs support at most 10 classes and one of them is reserved for the normal class (Hollmann et al., 2023; Qu et al., 2026; Zhang et al., 2025).

Within each cluster, we select the rows that receive the least attention and displace them outward from the cluster centroid. The attention that row j receives is $\rho_j = \sum_i \mathbf{M}_{ij}$, and a row with low ρ_j is one that the normal samples rarely attend to, *i.e.*, one that the foundation model relates weakly to the rest of the normal samples. Let $\mathbf{x}_{c,1}, \ldots, \mathbf{x}_{c,k}$ denote the k rows with the lowest ρ in cluster c. We then move each selected row outward along the direction from its cluster centroid $\bar{\mathbf{x}}_c$ to place it beyond the extent of its cluster,

$$\tilde{\mathbf{x}}_{c,l} \xleftarrow{\text{init}} \mathbf{x}_{c,l} + \alpha\, r_c \frac{\mathbf{x}_{c,l} - \bar{\mathbf{x}}_c}{\|\mathbf{x}_{c,l} - \bar{\mathbf{x}}_c\|}, \quad \bar{\mathbf{x}}_c = \frac{1}{|\mathcal{G}_c|} \sum_{i \in \mathcal{G}_c} \mathbf{x}_i, \quad r_c = \max_{i \in \mathcal{G}_c} \|\mathbf{x}_i - \bar{\mathbf{x}}_c\|, \tag{4}$$

where α scales the displacement, set to $\alpha = 1.0$ by default. The resulting $\tilde{\mathbf{X}} \in \mathbb{R}^{Ck \times d}$ is defined as a set of trainable parameters, and each $\tilde{\mathbf{x}}_{c,l}$ takes the anomaly class label c of its own cluster. We validate the effectiveness of the initialization in Section 4.3.2 and provide its procedure in Algorithm 1.

3.3 Pseudo-anomaly Optimization

We elaborate on how the initialized $\tilde{\mathbf{X}}$ is optimized into effective anomaly classes. We optimize $\tilde{\mathbf{X}}$ through f_ϕ itself, so that pseudo-anomalies are shaped by the very model that will consume them at test time, since initialization alone cannot guarantee that the resulting rows serve as useful anomaly classes in context. For optimization, we consider two objectives: (*i*) a separability objective $\mathcal{L}_{\text{sep}}$ that separates each pseudo-anomaly from the normal samples and (*ii*) a diversity objective $\mathcal{L}_{\text{div}}$ that keeps the pseudo-anomalies diverse. Based on these two objectives, we define the following optimization problem:

$$\tilde{\mathbf{X}}^* = \arg\min_{\tilde{\mathbf{X}}} \; \mathbb{E}_{(\mathcal{S},\mathcal{Q})} \left[\mathcal{L}_{\text{sep}} + \lambda\, \mathcal{L}_{\text{div}} \right], \tag{5}$$

where λ balances separability and diversity, and $(\mathcal{S}, \mathcal{Q})$ is a pair of context and query sets.

We solve Eq. (5) by gradient descent on $\tilde{\mathbf{X}}$ with f_ϕ frozen, drawing a new pair of $\mathcal{S}$ and $\mathcal{Q}$ at each gradient step. Each pair is obtained by randomly splitting the normal samples and the pseudo-anomalies of each class into two disjoint parts, so that both $\mathcal{S}$ and $\mathcal{Q}$ contain normal samples and pseudo-anomalies of every class, labeled as in Eq. (2). We assign more pseudo-anomalies to $\mathcal{S}$ than to $\mathcal{Q}$, since all pseudo-anomalies serve as context at test time (see Appendix A.2).

Separability $\mathcal{L}_{\text{sep}}$. To make the pseudo-anomalies separable from the normal samples, we optimize $\tilde{\mathbf{X}}$ so that f_ϕ classifies both the normal samples and the pseudo-anomalies into their correct classes. Given a pair of $\mathcal{S}$ and $\mathcal{Q}$ randomly split from $\mathbf{X}$ and $\tilde{\mathbf{X}}$, the cross-entropy objective over $\mathcal{Q}$ is

$$\mathcal{L}_{\text{sep}} = -\frac{1}{|\mathcal{Q}|} \sum_{(\mathbf{x},y) \in \mathcal{Q}} \log p_\phi(y \mid \mathbf{x}, \mathcal{S}). \tag{6}$$

Minimizing this objective makes the pseudo-anomalies separable from the normal samples. However, since it only requires each query to be assigned to its correct class, the pseudo-anomalies of the

same class can collapse onto a single point during optimization. The optimized pseudo-anomalies may then fail to represent diverse abnormal patterns (see Section 4.4).

Diversity $\mathcal{L}_{\text{div}}$. To prevent the pseudo-anomalies from collapsing, we introduce an attention-based diversity objective computed on the context set $\mathcal{S}$. A repulsion term in the input space could also prevent this collapse, but distance measured from heterogeneous features can be suboptimal as a diversity measure (see Section 4.3.3). Thus, we measure diversity with the attention of f_ϕ over $\mathcal{S}$. Let $\mathbf{A} \in \mathbb{R}^{n_A \times n_N}$ be the attention map from the n_A pseudo-anomaly rows to the n_N normal rows of $\mathcal{S}$, pooled over the heads and layers of f_ϕ. Each row of $\mathbf{A}$ is normalized to sum to one, so $\mathbf{A}_{m,:}$ is the distribution over normal samples that the m-th pseudo-anomaly row attends to. The Gram matrix $\mathbf{G} = \mathbf{A}\mathbf{A}^\top$ then measures how much the attention distributions of each pair of pseudo-anomalies overlap, and we penalize its off-diagonal mass,

$$\mathcal{L}_{\text{div}} = \frac{1}{n_A(n_A-1)} \sum_{m \neq m'} \mathbf{G}_{m,m'}. \tag{7}$$

Collapsed pseudo-anomalies share nearly identical attention distributions over the normal samples and thus incur a large penalty, whereas pseudo-anomalies attending to different parts of the normal samples receive a much smaller penalty.

With both objectives combined, Eq. (5) keeps the pseudo-anomalies separable from the normal samples and distinct from one another. The optimized $\tilde{\mathbf{X}}^*$ then forms the context set in Eq. (2) with the normal samples, and a test query is scored by Eq. (3). We validate the effectiveness of the optimization in Section 4.3.3 and provide detailed procedures in Algorithms 2 and 3.

4 Experiment

We design our experiments to answer the following questions:

- Does our framework outperform existing methods on both benchmarks? (§4.2)
- How does each component and design choice contribute to the overall performance? (§4.3)
- How does our optimization shape the pseudo-anomalies? (§4.4)

4.1 Setup

We evaluate GradPAO on two benchmarks: (*i*) ADBench (Han et al., 2022), a widely used benchmark for tabular anomaly detection with 57 datasets, and (*ii*) OddBench from MacrOData (Ding et al., 2026a), a large-scale benchmark with 690 datasets. We follow the evaluation pipeline of the work that provides its public baseline results, OutFormer (Ding et al., 2026b) for ADBench and MacrOData (Ding et al., 2026a) for OddBench, and all baseline results are taken directly from these two works. Although the two works do not use the same baseline set, both include classical methods (Ramaswamy et al., 2000; Breunig et al., 2000; Liu et al., 2008), deep learning methods (Shenkar & Wolf, 2022; Bergman & Hoshen, 2020; Livernoche et al., 2024), foundation models for TAD (Shen et al., 2025; Ding et al., 2026b), and TabPFN-OD, which repurposes TabPFN (Hollmann et al., 2023) as a reconstruction-based detector. For additional baselines, see Appendix A.3. Following the protocol shared by both pipelines, the train set contains only normal samples, and the test set contains the remaining normal samples and all abnormal samples. For model comparison, we report ranking-based scores, the average rank (Avg. Rank), rescaled AUCPR (rAUC), ELO (Elo, 1967), and win rate (WinRate), since raw AUCPR is not directly comparable across datasets.

We implement GradPAO on top of a frozen TabICLv2 (Qu et al., 2026) and only update the pseudo-anomalies. For each dataset, we initialize $C = 9$ pseudo-anomaly classes with $k = 8$ rows each, and optimize them with Adam (Kingma & Ba, 2015) for 10 epochs at a learning rate of 0.1 with a diversity weight of $\lambda = 20$. Appendices A.2 and A.4 provide implementation details and a hyperparameter sensitivity analysis, indicating that GradPAO is robust to optimization hyperparameters.

4.2 Main Result

Main results are summarized in Table 1. GradPAO achieves state-of-the-art performance on both benchmarks across every metric. As shown in Table 1a, GradPAO outperforms the second-best

Table 1: Performance summary on ADBench and OddBench in AUCPR. The two benchmarks provide different baseline sets of 12 and 14 models in their official results. Detailed results are provided in Appendix A.6 for ADBench and in the supplementary material for 690 OddBench datasets.

(a) ADBench

Model	Rank (↓)	ELO (↑)	Win (↑)	rAUC (↑)
kNN	6.175	1047	0.5541	0.6543
LOF	6.974	1002	0.5110	0.5596
IForest	9.877	790.1	0.2712	0.3483
ICL	6.079	1071	0.5826	0.6572
DeepSVDD	10.07	794.8	0.2463	0.3364
GOAD	9.921	798.8	0.2303	0.3485
DDPM	7.465	943.1	0.4386	0.5557
DTE-NP	5.623	1091	0.6265	0.7012
DTE-C	7.202	988.2	0.4846	0.5481
FoMo-0D	6.535	1039	0.5365	0.6577
TabPFN-OD	5.693	1096	0.6243	0.6734
OutFormer	5.263	1128	0.6535	0.7228
GradPAO	**4.123**	**1211**	**0.7405**	**0.7817**
Δ vs. 2nd	**+1.14**	**+82.6**	**+0.0870**	**+0.0589**

(b) OddBench

Model	Rank (↓)	ELO (↑)	Win (↑)	rAUC (↑)
kNN	7.576	1026	0.5235	0.5188
LOF	7.804	1004	0.5195	0.5087
IForest	9.239	915.8	0.4135	0.4054
CBLOF	8.888	941.2	0.4385	0.4617
OCSVM	7.361	1029	0.5549	0.5386
EGMM	6.739	1078	0.5956	0.5803
ICL	8.788	943.6	0.4288	0.4512
GOAD	11.78	745.7	0.2166	0.2299
DTE-NP	6.766	1072	0.5872	0.5493
DTE-C	8.507	975.2	0.4598	0.4737
NPT-AD	11.12	781.2	0.2631	0.2605
FoMo-0D	7.203	1053	0.5457	0.5587
TabPFN-OD	6.896	1070	0.5709	0.5628
OutFormer	6.345	1114	0.6220	0.5850
GradPAO	**4.225**	**1250**	**0.7788**	**0.7756**
Δ vs. 2nd	**+2.12**	**+136.9**	**+0.1568**	**+0.1906**

model OutFormer (Ding et al., 2026b) on ADBench by a margin of 1.14 in average rank, 82.6 points in ELO, 0.0870 in win rate, and 0.0589 in rAUC. As shown in Table 6 in Appendix A.6, the baselines show large performance variance across the ADBench datasets, while GradPAO maintains mostly high performance, indicating a stable detection capability. As shown in Table 1b, GradPAO outperforms OutFormer (Ding et al., 2026b) on OddBench by a larger margin of 2.12 in average rank, 136.9 points in ELO, 0.1568 in win rate, and 0.1906 in rAUC, even though OddBench compares a broader set of baselines. GradPAO also outperforms TabPFN-OD by 0.1083 in rAUC on ADBench, and by 0.2128 in rAUC on OddBench. Since GradPAO and TabPFN-OD operate on top of a pretrained TFM and differ mainly in how anomaly scores are derived, this gap shows that using a pretrained TFM as a classifier against optimized pseudo-anomalies is more effective than using it as a reconstruction-based detector. These results demonstrate that GradPAO generalizes robustly across diverse tabular anomaly detection tasks.

4.3 Ablation

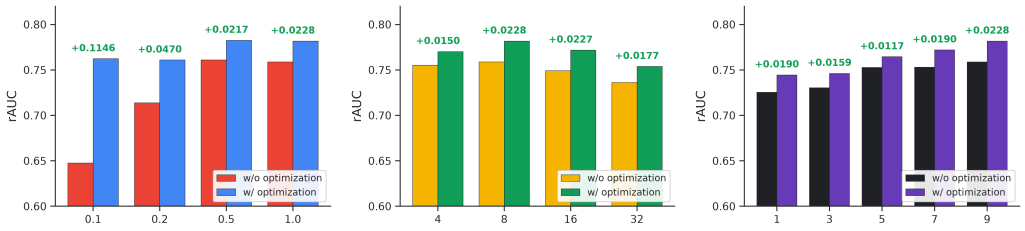
In this section, we conduct ablation studies on each component of GradPAO, covering the choice of the frozen backbone, the pseudo-anomaly initialization, and the pseudo-anomaly optimization. Ablation studies are mainly conducted on ADBench (Han et al., 2022), since evaluating every variant over the 690 datasets of OddBench (Ding et al., 2026a) is computationally intensive.

4.3.1 Backbone

We test whether our framework generalizes to other backbones by replacing TabICLv2 (Qu et al., 2026) with TabPFNv3 (Grinsztajn et al., 2026), another widely used tabular foundation model. As shown in Table 2, TabPFNv3 achieves an average rank of 4.421 and an rAUC of 0.7708 on ADBench, and 4.350 and 0.7688 on OddBench, within 0.011 rAUC of the default backbone on both benchmarks. It also outperforms the strongest baseline, OutFormer (Ding et al., 2026b), on both benchmarks, so our framework does not depend on a specific tabular foundation model. We use TabICLv2 as the default backbone since it has fewer parameters and a lower computational cost than TabPFNv3.

Table 2: Backbone ablation.

Backbone	Rank (↓)	rAUC (↑)
ADBench		
TabICLv2	4.123	0.7817
TabPFNv3	4.421	0.7708
OddBench		
TabICLv2	4.225	0.7756
TabPFNv3	4.350	0.7688



(a) Displacement Scale α (b) Number of pseudo-anomalies k (c) Number of clusters C

Figure 2: Initialization hyperparameter sensitivity

4.3.2 Attention-based Pseudo-anomaly Initialization

We now study the design of the initialization stage, covering both the choice of initialization method and its associated hyperparameters. Table 3 compares our initialization with the synthetic anomaly types defined in ADBench, where each variant replaces only the initialization stage and is followed by the same optimization stage. Each type models the relationships among the raw features and constructs synthetic anomalies that break them, and we refer to ADBench (Han et al., 2022) for the detailed definition of each type. Among them, the global variant performs best at an rAUC of 0.7539, followed by the cluster, dependency, and local variants at 0.7444, 0.7432, and 0.7315. Our attention-based initialization reaches 0.7817, outperforming the strongest ADBench variant by 0.0278.

Table 3: Initialization ablation.

Init Method	rAUC (↑)
Global	0.7539
Cluster	0.7444
Local	0.7315
Dependency	0.7432
GradPAO	**0.7817**

Figure 2 reports the sensitivity to the three initialization hyperparameters, the displacement scale α, the number of pseudo-anomalies k per cluster, and the number of clusters C. Optimization improves on the initialization-only variant at every value of α and k. Without optimization, a small α leaves the pseudo-anomalies too close to the normal samples and degrades performance sharply, whereas the optimized variant varies little across α. For the number of pseudo-anomalies per cluster, performance peaks at $k = 8$, although every optimized variant still outperforms all baselines in rAUC. In contrast, performance generally rises with C, indicating that covering more distinct regions of the normal samples matters more than placing more pseudo-anomalies within the same regions.

4.3.3 Pseudo-anomaly Optimization

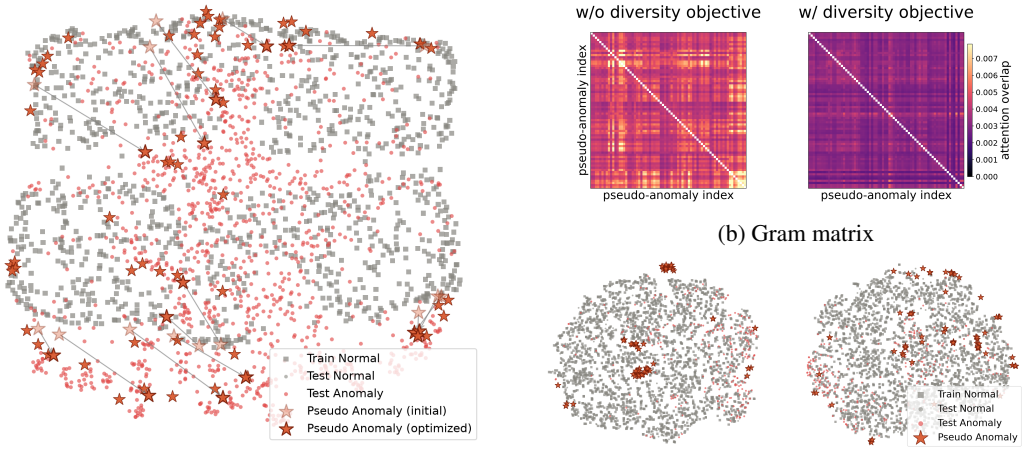
We provide an ablation on the optimization, covering the contribution of the optimization and the diversity objective design. For the diversity objective, we consider three designs, (*i*) contrastive objective in the input space, (*ii*) contrastive objective in the embedding space, and (*iii*) our attention-based diversity objective. As shown in Table 4, the initialization-only variant (1st row) yields an rAUC of 0.7589. Adding the optimization (2nd row) improves it to 0.7687, but the gain remains limited as the pseudo-anomalies tend to collapse after optimization (see Section 4.4). Introducing a diversity objective mitigates this collapse, with the input space and embedding space designs reaching an rAUC of 0.7761 and 0.7712, respectively. Our attention-based diversity objective further improves the rAUC to 0.7817, suggesting that applying the constraint in the attention space can be beneficial. This can be attributed to where each design applies its constraint, since our attention-based diversity objective enforces diversity on the attention of the foundation model rather than geometric separation in the input or embedding space.

Table 4: Optimization contribution.

Separability	Diversity	rAUC (↑)
✗	✗	0.7589
✓	✗	0.7687
✓	Input	0.7761
✓	Embedding	0.7712
✓	Attention	**0.7817**

4.4 Analysis

In this section, we provide a comprehensive analysis of how the optimization stage shapes the pseudo-anomalies, using visualizations of the pseudo-anomalies before and after optimization (Fig-



(a) Optimization visualization

(b) Gram matrix

(c) w/o diversity

(d) w/ diversity

Figure 3: Visualization of optimization and the contribution of the diversity objective.

ure 3a), the Gram matrix of the pseudo-anomaly to normal attention of the context set (Figure 3b), and visualizations of the optimized pseudo-anomalies with and without the diversity objective (Figure 3c and Figure 3d). We begin by showing how the pseudo-anomalies move during the optimization stage. As shown in Figure 3a, the attention-based initialization sometimes places the starting points near the normal samples. The optimization then moves these points away from the normal samples, and some optimized pseudo-anomalies lie near regions occupied by real anomalies in the visualization. This indicates that the optimization can make the pseudo-anomalies serve as effective anomaly classes, which the initialization alone does not guarantee.

We next analyze how the diversity objective keeps the optimized pseudo-anomalies diverse. Figure 3b compares the Gram matrix with and without the diversity objective. Without the diversity objective, the off-diagonal entries often take large values, whereas applying the diversity objective makes them small. This indicates that each pseudo-anomaly attends to a distinct subset of the normal samples rather than to the same ones. We further provide a t-SNE visualization of the real normal samples, the real anomalies, and the optimized pseudo-anomalies in the input space (Figure 3c and Figure 3d). Without the diversity objective, the pseudo-anomalies are located in a few narrow regions and leave most of the normal boundary uncovered (Figure 3c). In contrast, applying the diversity objective makes them spread across diverse regions and reach the regions where the real anomalies exist (Figure 3d). These results show that the two terms of our objective act as intended: the cross-entropy objective makes the pseudo-anomalies separable from the normal samples, and the diversity objective keeps them distinct from one another.

5 Conclusion

In this paper, we propose GradPAO, a novel framework that turns a tabular foundation model pretrained for predictive tasks into a tabular anomaly detector. Our key idea begins with why such a foundation model cannot be directly applied to tabular anomaly detection. It was pretrained to perform in-context multi-class classification against a labeled context set, whereas tabular anomaly detection provides normal samples alone and no anomaly classes. To fill this gap, we define the pseudo-anomalies as trainable parameters and optimize them with a cross-entropy objective and a diversity objective through the very foundation model that will consume them at test time. On two benchmarks with 57 and 690 datasets, GradPAO outperforms classical, deep learning, and foundation model baselines, and our ablations and analysis show that both stages and the attention-based diversity design contribute to this gain. Rather than developing another foundation model dedicated to tabular anomaly detection, we leverage tabular foundation models already pretrained for predictive tasks, and we hope this direction can serve as a foundation for future work on reusing them for tasks their pretraining did not cover.

AI USE STATEMENT

We used generative AI tools to assist with the implementation of the proposed method and to edit the manuscript for clarity and readability. All AI-assisted code and text were reviewed by the authors. The authors take full responsibility for the final content of this paper, including all claims, experimental results, and code.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. The study did not involve human participants or animal experiments. All datasets are publicly available benchmarks for tabular anomaly detection and were used in accordance with their respective licenses. The datasets contain no personally identifiable information, and the experiments pose no foreseeable privacy or security risks.

REPRODUCIBILITY STATEMENT

We have made a concerted effort to ensure the reproducibility of the results reported in this paper. The source code is provided in the supplementary material, and detailed information on the experimental setup, including training procedures, model configurations, and hyperparameters, is provided. All experiments were conducted on NVIDIA RTX 4090 GPUs.

REFERENCES

Khaled Gubran Al-Hashedi and Pritheega Magalingam. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40: 100402, 2021.

Lion Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *International Conference on Management of Data*, 2000.

Xueying Ding, Simon Klüttermann, Haomin Wen, Yilong Chen, and Leman Akoglu. Macrodata: New benchmarks of thousands of datasets for tabular outlier detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026a.

Xueying Ding, Haomin Wen, Simon Klüttermann, and Leman Akoglu. From zero to hero: Advancing zero-shot foundation models for tabular outlier detection. In *International Conference on Machine Learning (ICML)*, 2026b.

Arpad E. Elo. The proposed uscf rating system, its development, theory, and applications. *Chess Life*, 22:242–247, 1967.

Fatemeh Esmaeili, Erica Cassie, Hong Phan T. Nguyen, Natalie O. V. Plank, Charles P. Unsworth, and Alan Wang. Anomaly detection for sensor signals utilizing deep learning autoencoder-based neural networks. *Bioengineering*, 10, 2023.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection – a survey. *ACM computing surveys (CSUR)*, 54, 2021.

Michael Glodek, Martin Schels, and Friedhelm Schwenker. Ensemble gaussian mixture models for probability density estimation. *Comput. Stat.*, 28(1):127–138, February 2013. ISSN 0943-4062. doi: 10.1007/s00180-012-0374-5. URL https://doi.org/10.1007/s00180-012-0374-5.

Mark S. Graham, Walter H. L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2948–2957, 2023. doi: 10.1109/CVPRW59228.2023.00296.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, 2022.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recogn. Lett.*, 24(9–10):1641–1650, jun 2003. ISSN 0167-8655.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pp. 17564–17579, 2023.

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

Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection. In *International Conference on Learning Representations*, 2024.

Ritesh K Malaiya, Donghwoon Kwon, Sang C Suh, Hyunjoo Kim, Ikkyun Kim, and Jinoh Kim. An Empirical Evaluation of Deep Learning for Network Anomaly Detection. *IEEE Access*, 7: 140806–140817, 2019.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. In *International Conference on Machine Learning (ICML)*, 2026.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000.

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*, pp. 4393–4402, 2018.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, 1999.

Yuchen Shen, Haomin Wen, and Leman Akoglu. FoMo-0D: A foundation model for zero-shot tabular outlier detection. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *International Conference on Learning Representations*, 2022.

Hugo Thimonier, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. Beyond individual input for deep anomaly detection on tabular data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48097–48123. PMLR, 21–27 Jul 2024.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

Shuo Yang, Zheyu Zhang, Bardh Prenkaj, and Gjergji Kasneci. Doubling your data in minutes: Ultra-fast tabular data generation via llm-induced dependency graphs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10348–10369, 2025.

Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *International Conference on Learning Representations*, 2024.

Xiyuan Zhang, Danielle C. Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W. Mahoney, Cuixiong Hu, Huzefa Rangwala, George Karypis, and Bernie Wang. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In *Advances in Neural Information Processing Systems*, 2025.

A Appendix

A.1 Algorithms

We provide the detailed procedure of the attention-based initialization in Algorithm 1, the pseudo-anomaly optimization in Algorithm 2, and the attention-based diversity objective in Algorithm 3.

Algorithm 1: Attention-based Pseudo-anomaly Initialization

Input: Normal samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, frozen TFM f_ϕ, number of clusters C, rows per cluster k, displacement scale α

$\mathbf{M} \leftarrow \mathrm{Attn}(f_\phi, \mathbf{X})$; // inter-sample attention, $\mathbf{M} \in \mathbb{R}^{n \times n}$

$\bar{\mathbf{M}} \leftarrow (\mathbf{M} + \mathbf{M}^\top)/2$;

$\mathbf{D} \leftarrow 1 - \bar{\mathbf{M}} / \max(\bar{\mathbf{M}})$;

$\{\mathcal{G}_1, \ldots, \mathcal{G}_C\} \leftarrow \mathrm{Clustering}(\mathbf{D}, C)$;

$\rho_j \leftarrow \sum_i \mathbf{M}_{ij}$; // attention received by row j

for *cluster* $c = 1, \ldots, C$ **do**

  $\mathbf{x}_{c,1}, \ldots, \mathbf{x}_{c,k} \leftarrow$ the k rows in $\mathcal{G}_c$ with the smallest ρ;

  $\bar{\mathbf{x}}_c \leftarrow \frac{1}{|\mathcal{G}_c|} \sum_{i \in \mathcal{G}_c} \mathbf{x}_i, \quad r_c \leftarrow \max_{i \in \mathcal{G}_c} \|\mathbf{x}_i - \bar{\mathbf{x}}_c\|$;

  for $l = 1, \ldots, k$ **do**

    $\hat{\mathbf{v}}_{c,l} \leftarrow (\mathbf{x}_{c,l} - \bar{\mathbf{x}}_c) \,/\, \|\mathbf{x}_{c,l} - \bar{\mathbf{x}}_c\|$;

    $\tilde{\mathbf{x}}_{c,l} \leftarrow \mathbf{x}_{c,l} + \alpha\, r_c\, \hat{\mathbf{v}}_{c,l}$;

return $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_{c,l} \mid 1 \le c \le C,\ 1 \le l \le k\}$

Algorithm 2: Pseudo-anomaly Optimization

Input: Normal samples $\mathbf{X}$, initialized $\tilde{\mathbf{X}}$ from Algorithm 1, frozen TFM f_ϕ, steps T, learning rate η, weight λ

for $t = 1, \ldots, T$ **do**

  $\mathbf{X}_\mathcal{S}, \mathbf{X}_\mathcal{Q} \leftarrow \mathrm{Split}(\mathbf{X}), \quad \tilde{\mathbf{X}}_\mathcal{S}, \tilde{\mathbf{X}}_\mathcal{Q} \leftarrow \mathrm{Split}(\tilde{\mathbf{X}})$; // class-wise, $|\tilde{\mathbf{X}}_\mathcal{S}| > |\tilde{\mathbf{X}}_\mathcal{Q}|$

  $\mathcal{S} \leftarrow \{(\mathbf{x}_i, 0)\}_{\mathbf{x}_i \in \mathbf{X}_\mathcal{S}} \cup \{(\tilde{\mathbf{x}}_{c,l}, c)\}_{\tilde{\mathbf{x}}_{c,l} \in \tilde{\mathbf{X}}_\mathcal{S}}$;

  $\mathcal{Q} \leftarrow \{(\mathbf{x}_i, 0)\}_{\mathbf{x}_i \in \mathbf{X}_\mathcal{Q}} \cup \{(\tilde{\mathbf{x}}_{c,l}, c)\}_{\tilde{\mathbf{x}}_{c,l} \in \tilde{\mathbf{X}}_\mathcal{Q}}$;

  $\mathbf{Z}, \mathbf{M}^\mathcal{S} \leftarrow f_\phi(\mathcal{Q} \mid \mathcal{S})$; // logits and context attention

  $\mathcal{L} \leftarrow \mathcal{L}_{\text{sep}}(\mathbf{Z}, \mathcal{Q}) + \lambda\, \mathcal{L}_{\text{div}}(\mathbf{M}^\mathcal{S})$; // Eq. (6), Algorithm 3

  $\tilde{\mathbf{X}} \leftarrow \tilde{\mathbf{X}} - \eta \nabla_{\tilde{\mathbf{X}}} \mathcal{L}$;

return $\tilde{\mathbf{X}}^* = \tilde{\mathbf{X}}$

Algorithm 3: Diversity Objective $\mathcal{L}_{\text{div}}$

Input: Context attention $\mathbf{M}^\mathcal{S} \in \mathbb{R}^{(n_N + n_A) \times (n_N + n_A)}$ pooled over the heads and layers of f_ϕ, with the n_N normal rows of $\mathcal{S}$ ordered before the n_A pseudo-anomaly rows

$\mathbf{B} \leftarrow \mathbf{M}^\mathcal{S}[\, n_N{:},\ {:}n_N \,]$; // pseudo-anomaly → normal attention

$\mathbf{A}_{m,:} \leftarrow \mathbf{B}_{m,:} \big/ \sum_j \mathbf{B}_{m,j}$;

$\mathbf{G} \leftarrow \mathbf{A}\mathbf{A}^\top$; // $\mathbf{G} \in \mathbb{R}^{n_A \times n_A}$

$\mathcal{L}_{\text{div}} \leftarrow \frac{1}{n_A(n_A - 1)} \sum_{m \neq m'} \mathbf{G}_{m,m'}$;

return $\mathcal{L}_{div}$

A.2 Implementation Details

We use the official implementations and pretrained checkpoints of TabICLv2 and TabPFNv3 and keep all of their parameters frozen. We use $C = 9$ clusters, $k = 8$ pseudo-anomalies per cluster, and a displacement scale of $\alpha = 1.0$. The pseudo-anomalies are optimized as free parameters in the input space with Adam at a learning rate of $1\mathrm{e}{-1}$, a diversity weight of $\lambda = 20$, and 10 epochs. Each step samples 512 normal rows for the context set and 512 for the query set, scaled down

proportionally when fewer than 1024 normal rows are available, and splits the pseudo-anomalies of each class between the two sets in a 3:1 ratio. Since pseudo-anomalies are only used in the context set at test time, we assign more pseudo-anomalies to the context set during optimization. At test time, the context set combines all normal training rows with all Ck optimized pseudo-anomalies, and every test query is scored in a single forward pass. All reported results are averaged over 5 seeds.

Following the tabular foundation model literature (Zhang et al., 2025; Ding et al., 2026b), we report four metrics computed across datasets rather than within a single dataset, the average rank of each model, an ELO rating and a win rate from pairwise comparisons between models, and the rescaled AUC (rAUC) that normalizes each model's error between the best and the worst model on each dataset. We refer to OutFormer (Ding et al., 2026b) for the exact definitions.

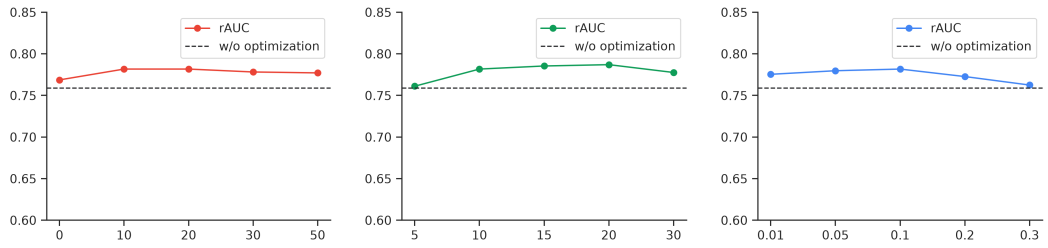
A.3 Baselines

We describe the baselines of the two benchmarks, taken from the public results of each, OutFormer (Ding et al., 2026b) for ADBench and MacrOData (Ding et al., 2026a) for OddBench. Both benchmarks share the same protocol, where the train set contains only normal samples and the test set contains the remaining normal samples and all abnormal samples. Ten baselines appear in both public results, k-th Nearest Neighbors (kNN) (Ramaswamy et al., 2000), Local Outlier Factor (LOF) (Breunig et al., 2000), Isolation Forest (IForest) (Liu et al., 2008), Internal Contrastive Learning (ICL) (Shenkar & Wolf, 2022), Classification-Based Anomaly Detection using General Data (GOAD) (Bergman & Hoshen, 2020), Diffusion Time Estimation-Non-Parametric (DTE-NP) and Diffusion Time Estimation-Categorical (DTE-C) (Livernoche et al., 2024), Zero-shot Tabular Outlier Detection (FoMo-0D) (Shen et al., 2025), TabPFN-OD, and OutFormer (Ding et al., 2026b). On ADBench, the baselines additionally include Deep Support Vector Data Description (DeepSVDD) (Ruff et al., 2018) and Denoising Diffusion Probabilistic Models (DDPM) (Graham et al., 2023), for 12 baselines in total, and the OddBench results add Cluster-Based Local Outlier Factor (CBLOF) (He et al., 2003), One-Class Support Vector Machine (OCSVM) (Schölkopf et al., 1999), Ensemble Gaussian Mixture Model (EGMM) (Glodek et al., 2013), and Non-Parametric Transformer Anomaly Detection (NPT-AD) (Thimonier et al., 2024), for 14 baselines in total.

Among them, we further describe the methods that are most closely related to GradPAO. TabPFN-OD adapts TabPFN (Hollmann et al., 2023), a prior-data fitted network pretrained for tabular predictive tasks, to anomaly detection by reconstructing each feature from the remaining ones in context and scoring samples by their reconstruction error. OutFormer is a transformer-based tabular anomaly detector pretrained across a large collection of synthetic datasets, enabling zero-shot detection on unseen tables. NPT-AD reconstructs each sample by attending to the training samples with a non-parametric transformer and scores it by the resulting reconstruction error, sharing the reconstruction view of TabPFN-OD without any pretraining across datasets.

A.4 Hyperparameter Sensitivity

We provide a hyperparameter analysis for the three optimization hyperparameters, the diversity weight λ (Figure 4a), the number of epochs (Figure 4b), and the learning rate (Figure 4c). Every tested value improves over the initialization-only variant, which reaches an rAUC of 0.7589. For λ, performance is stable across $\lambda \in \{10, 20, 30, 50\}$ with the best result at $\lambda = 20$, whereas $\lambda = 0$ removes the diversity objective and falls back to an rAUC of 0.7687. For the number of epochs, rAUC rises up to 20 epochs and decreases at 30, while the optimization cost grows linearly with the number of epochs. For the learning rate, rAUC peaks at $1e{-1}$ and decreases on both sides of it, staying within 0.0063 of the peak at $1e{-2}$ but dropping by 0.0192 at $3e{-1}$. Throughout the paper we use $\lambda = 20$, 10 epochs, and lr of $1e{-1}$ as a single default for every dataset, choosing 10 epochs over 20 for half the optimization cost, and this analysis shows that the performance is not sensitive to that choice.



(a) λ (b) Epoch (c) Learning rate

Figure 4: Optimization hyperparameter sensitivity.

A.5 Runtime Analysis

In this section, we analyze the time GradPAO spends on each ADBench dataset, measured on a single NVIDIA RTX 4090 GPU and summarized in Figure 5. Since the foundation model stays frozen and only the Ck pseudo-anomaly rows are updated, the optimization is inexpensive, with a median of 17 seconds and a third quartile of 100 seconds. The datasets beyond the third quartile are the largest ones, where the number of optimization steps grows with the number of normal samples, reaching 29 minutes on the largest one. Scoring the entire test set takes 4 seconds for the median dataset, since each query is answered by a single forward pass. This optimization is performed once for each target dataset, and the resulting context set is then reused for every test query without any further update.

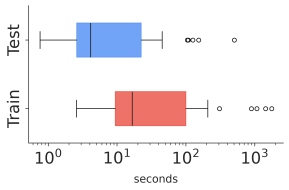


Figure 5: Runtime of GradPAO on ADBench.

A.6 Extended Results

In this section, we provide extended results for the main experiment. Table 5 summarizes performance on both benchmarks in AUCROC, and Tables 6 and 7 report per-dataset results on ADBench in AUCPR and AUCROC, averaged over 5 seeds.

GradPAO also leads in AUCROC, achieving the best average rank, ELO, and win rate on both benchmarks. On ADBench, GradPAO reaches an average rank of 4.710, an ELO of 1166, and a win rate of 0.7083, while OutFormer reaches a higher rAUC of 0.8365 compared with 0.8268, a gap that is small relative to the margins GradPAO holds on the other three metrics. On OddBench with 690 datasets, GradPAO is the best on all four metrics, improving over the second-best model, DTE-NP, by 1.259 in average rank, 90.7 in ELO, 0.1125 in win rate, and 0.0891 in rAUC.

Table 5: Performance summary on ADBench and OddBench in AUCROC. The two benchmarks provide different baseline sets of 12 and 14 models in their official results. Detailed results are provided in Appendix A.6 for ADBench and in the supplementary material for 690 OddBench datasets.

(a) ADBench

Model	Rank (↓)	ELO (↑)	Win (↑)	rAUC (↑)
kNN	5.965	1069	0.5804	0.7727
LOF	6.833	1013	0.5212	0.6627
IForest	9.079	859.9	0.3121	0.5581
ICL	6.465	1042	0.5548	0.7594
DeepSVDD	10.52	757.1	0.1857	0.3346
GOAD	10.06	784.1	0.2222	0.4103
DDPM	8.026	915.7	0.4196	0.6526
DTE-NP	5.193	1110	0.6477	0.8102
DTE-C	6.807	1019	0.5183	0.7172
FoMo-0D	6.658	1020	0.5095	0.7374
TabPFN-OD	5.693	1100	0.6228	0.8026
OutFormer	4.991	1143	0.6974	**0.8365**
GradPAO	**4.710**	**1166**	**0.7083**	0.8268
Δ vs. 2nd	**+0.281**	**+23.2**	**+0.0109**	-0.0097

(b) OddBench

Model	Rank (↓)	ELO (↑)	Win (↑)	rAUC (↑)
kNN	6.938	1064	0.5745	0.7188
LOF	8.025	998.6	0.4989	0.6630
IForest	8.053	999.2	0.4912	0.6733
CBLOF	8.547	964.1	0.4561	0.6586
OCSVM	8.133	991.5	0.4807	0.6894
EGMM	6.628	1076	0.6046	0.7222
ICL	9.336	912.9	0.3878	0.6129
GOAD	12.28	695.4	0.1930	0.2854
DTE-NP	6.130	1116	0.6261	0.7444
DTE-C	7.906	1011	0.5016	0.6794
NPT-AD	11.13	784.9	0.2694	0.4109
FoMo-0D	7.813	1010	0.5089	0.6675
TabPFN-OD	6.961	1064	0.5761	0.7262
OutFormer	6.533	1106	0.6084	0.7285
GradPAO	**4.870**	**1206**	**0.7386**	**0.8335**
Δ vs. 2nd	**+1.259**	**+90.7**	**+0.1125**	**+0.0891**

Table 6: Extended Results (AUCPR)

Dataset	kNN	LOF	IForest	ICL	DeepSVDD	GOAD	DDPM	DTE-NP	DTE-C	FoMo-0D	TabPFN-OD	OutFormer	GradPAO
hepatitis	0.9031±0.0436	0.4367±0.1079	0.5536±0.0609	0.9983±0.0038	0.9873±0.0105	0.6577±0.0543	0.9514±0.0134	0.8232±0.1026	0.9582±0.0385	0.9945±0.0110	**0.9988**±0.0023	0.9957±0.0086	0.9937±0.0111
MVTec-AD	0.7576±0.0271	0.7579±0.0295	0.7000±0.0307	**0.8946**±0.0223	0.8378±0.0332	0.7262±0.0274	0.7366±0.0272	0.8294±0.0268	0.8511±0.0296	0.7325±0.0179	0.8891±0.0102	0.8378±0.0054	0.8556±0.0067
wine	0.9511±0.0181	0.8995±0.0264	0.6712±0.0762	0.9826±0.0389	0.7856±0.0959	0.7010±0.0629	0.9765±0.0072	0.9680±0.0558	0.9985±0.0023	0.9962±0.0077	**0.9987**±0.0027	0.9971±0.0058	0.9906±0.0116
lymphography	0.9917±0.0094	0.8416±0.0427	0.9438±0.0328	**1.0000**±0.0000	0.9682±0.0369	0.9876±0.0088	0.9930±0.0102	0.9934±0.0093	0.8677±0.0924	0.9915±0.0126	0.9935±0.0130	0.9976±0.0049	0.9980±0.0041
wpbc	0.4611±0.0274	0.4120±0.0262	0.4073±0.0315	**0.8931**±0.0537	0.7488±0.0558	0.3888±0.0382	0.5461±0.0375	0.6902±0.1391	0.6035±0.0488	0.6736±0.0434	0.8920±0.0496	0.8134±0.0406	0.8828±0.0402
glass	0.4232±0.0847	0.3812±0.0991	0.2137±0.0358	**0.9235**±0.0832	0.5235±0.2104	0.1833±0.0723	0.3121±0.1130	0.3738±0.1551	0.4151±0.0591	0.8360±0.0463	0.8413±0.0757	0.7176±0.1113	0.8393±0.1050
wbc	0.9201±0.0396	0.2489±0.0349	0.9424±0.0395	0.9512±0.0514	0.5651±0.1112	0.9195±0.0326	0.9384±0.0245	0.9610±0.0217	0.2997±0.0310	0.9646±0.0249	0.9698±0.0271	0.9487±0.0544	**0.9965**±0.0031
vertebral	0.2611±0.0249	0.3387±0.0430	0.2075±0.0198	0.5875±0.0730	0.2342±0.0317	0.2139±0.0498	0.3584±0.0927	0.2521±0.0890	0.3514±0.0539	0.6939±0.0397	0.7713±0.0835	0.6628±0.0389	**0.7717**±0.0384
stamps	0.7168±0.0835	0.6484±0.0817	0.5884±0.0684	0.7954±0.0544	0.4262±0.0994	0.4957±0.1772	0.6474±0.1287	0.8247±0.0408	0.5765±0.0840	0.8937±0.0350	0.8460±0.0454	0.9095±0.0510	**0.9177**±0.0470
ionosphere	0.9795±0.0069	0.9458±0.0157	0.9170±0.0189	0.9906±0.0032	0.9809±0.0081	0.9317±0.0264	0.9643±0.0050	0.9822±0.0102	0.9683±0.0038	0.9775±0.0076	**0.9915**±0.0022	0.9846±0.0043	0.9898±0.0053
wdbc	0.8203±0.0328	0.9364±0.0306	0.7198±0.0857	0.9560±0.0612	0.8432±0.0889	0.7877±0.0426	0.8430±0.0444	0.9047±0.0791	0.6882±0.1242	0.9382±0.0830	0.8501±0.1018	0.9040±0.0816	**0.9739**±0.0286
20news	0.1347±0.0052	0.1504±0.0067	0.1156±0.0043	0.1462±0.0088	0.1290±0.0190	0.1149±0.0051	0.1148±0.0044	0.1561±0.0135	0.1733±0.0236	0.1351±0.0084	**0.2096**±0.0369	0.1586±0.0148	0.1699±0.0054
breastw	0.9892±0.0032	**0.8801**±0.1009	0.9949±0.0012	0.9679±0.0161	0.9601±0.0128	0.9877±0.0037	0.9860±0.0051	0.9919±0.0014	0.8825±0.0106	0.9903±0.0034	0.9946±0.0020	0.9923±0.0023	**0.9955**±0.0017
pima	0.7537±0.0299	0.6840±0.0379	0.7365±0.0207	0.7863±0.0194	0.5975±0.0175	0.6515±0.0380	0.7118±0.0206	0.7970±0.0244	0.6798±0.0286	0.7927±0.0174	**0.8054**±0.0167	0.8005±0.0302	0.8013±0.0203
vowels	0.3021±0.0000	0.3309±0.0000	0.1197±0.0105	0.2739±0.0575	0.1688±0.0193	0.2094±0.0243	0.4272±0.0480	0.3159±0.0159	0.3810±0.0489	0.2410±0.0000	0.1906±0.0089	**0.4678**±0.0000	0.3520±0.0307
yeast	0.4826±0.0000	0.4894±0.0000	0.4678±0.0037	0.4955±0.0136	0.4921±0.0388	0.5077±0.0218	0.5105±0.0165	0.4812±0.0049	0.4974±0.0072	0.5100±0.0000	0.4973±0.0008	**0.5210**±0.0000	0.5117±0.0054
letter	0.0870±0.0000	0.1126±0.0000	0.0822±0.0025	**0.1280**±0.0117	0.0893±0.0052	0.0813±0.0005	0.0953±0.0051	0.0857±0.0013	0.0895±0.0013	0.0933±0.0000	0.0944±0.0002	0.0945±0.0000	0.0895±0.0002
cardio	0.7722±0.0000	0.7015±0.0000	0.7863±0.0269	0.4791±0.1147	0.3889±0.0552	**0.8479**±0.0057	0.6928±0.0091	0.7741±0.0085	0.6929±0.0127	0.7892±0.0000	0.5514±0.0188	0.7185±0.0000	0.6144±0.0273
fault	0.6198±0.0000	0.5044±0.0000	0.5919±0.0202	0.6318±0.0070	0.5546±0.0148	0.6214±0.0082	**0.6475**±0.0069	0.6217±0.0014	0.6393±0.0072	0.6308±0.0000	0.6293±0.0036	0.6164±0.0000	0.5699±0.0131
internetads	0.4922±0.0000	0.5043±0.0000	0.2920±0.0174	0.6003±0.0139	0.5156±0.0475	0.4743±0.0086	0.4770±0.0016	0.5130±0.0202	0.5522±0.0365	0.3829±0.0522	0.4590±0.0380	**0.6952**±0.0170	0.6876±0.0155
cardiotocography	0.5743±0.0000	0.5732±0.0000	0.6285±0.0342	0.4866±0.0384	0.4578±0.0510	**0.6752**±0.0082	0.5131±0.0198	0.5868±0.0114	0.5334±0.0126	0.6355±0.0000	0.4704±0.0033	0.5689±0.0000	0.4824±0.0088
musk	1.0000±0.0000	1.0000±0.0000	0.4039±0.2610	0.9679±0.0161	0.9601±0.0128	0.9877±0.0037	0.9860±0.0051	0.9919±0.0014	0.8825±0.0106	0.9903±0.0034	0.9946±0.0020	0.9923±0.0023	**0.9955**±0.0017
waveform	0.2700±0.0000	**0.3066**±0.0000	0.1053±0.0076	0.1863±0.0408	0.1152±0.0339	0.0886±0.0064	0.0931±0.0105	0.2787±0.0000	0.0999±0.0113	0.0995±0.0000	0.2082±0.0039	0.1365±0.0005	0.2776±0.0086
speech	0.0280±0.0000	0.0315±0.0000	0.0325±0.0100	0.0338±0.0050	0.0338±0.0038	0.0281±0.0031	0.0300±0.0029	0.0317±0.0000	0.0285±0.0012	0.0294±0.0031	0.0281±0.0032	0.0275±0.0003	0.0276±0.0008
thyroid	0.8094±0.0000	0.6057±0.0000	0.7966±0.0562	0.5151±0.1275	0.6906±0.0810	0.8009±0.0089	0.8222±0.0087	0.8103±0.0031	0.8167±0.0097	0.6700±0.0000	0.8160±0.0014	0.7013±0.0033	**0.8235**±0.0106
spambase	0.8332±0.0000	0.7271±0.0000	**0.8826**±0.0132	0.8678±0.0059	0.7526±0.0244	0.8209±0.0021	0.7289±0.0042	0.8365±0.0058	0.8380±0.0052	0.8099±0.0000	0.8592±0.0006	0.7968±0.0007	0.8628±0.0122
wilt	0.1225±0.0000	0.1574±0.0000	0.0881±0.0051	0.2894±0.0331	0.0708±0.0017	0.1086±0.0137	0.1724±0.0046	0.1220±0.0160	0.2541±0.0136	**0.7746**±0.0000	0.2403±0.0009	0.5606±0.0053	0.4943±0.0534
SVHN	0.1534±0.0000	0.1597±0.0000	0.1385±0.0049	0.1560±0.0031	0.1240±0.0095	0.1493±0.0018	0.1508±0.0004	0.1553±0.0000	0.1547±0.0047	0.1467±0.0033	0.1567±0.0050	**0.1671**±0.0007	0.1598±0.0007
optdigits	0.2911±0.0000	0.4363±0.0000	0.1541±0.0321	**0.5094**±0.0859	0.0453±0.0104	0.0776±0.0115	0.2556±0.0425	0.3175±0.0486	0.1534±0.0217	0.3188±0.0000	0.2843±0.0180	0.1737±0.0009	0.2974±0.0474
CIFAR10	0.1962±0.0000	0.2217±0.0000	0.1636±0.0064	0.1739±0.0059	0.1403±0.0108	0.1940±0.0040	0.1955±0.0008	0.1991±0.0000	0.1968±0.0086	0.1745±0.0052	0.1858±0.0055	0.2188±0.0011	**0.2268**±0.0089
pageblocks	0.6760±0.0000	**0.7107**±0.0000	0.4342±0.0090	0.6811±0.0230	0.5205±0.0389	0.6210±0.0115	0.6210±0.0095	0.6745±0.0098	0.6642±0.0123	0.6267±0.0000	0.7010±0.0062	0.6381±0.0043	0.6953±0.0063
satimage-2	**0.9669**±0.0000	0.8846±0.0000	0.9453±0.0055	0.9470±0.0119	0.7628±0.0821	0.9589±0.0010	0.8805±0.0510	0.9616±0.0000	0.6821±0.0315	0.9278±0.0000	0.5524±0.0154	0.9218±0.0017	0.9602±0.0033
FashionMNIST	0.5915±0.0000	0.6361±0.0000	0.4473±0.0188	0.6308±0.0096	0.4510±0.0204	0.5659±0.0040	0.5714±0.0012	0.5979±0.0000	0.5501±0.0104	0.5219±0.0278	0.5841±0.0151	0.6417±0.0031	**0.6703**±0.0037
landsat	0.5485±0.0000	**0.6137**±0.0000	0.4731±0.0350	0.5312±0.0157	0.4943±0.0240	0.3121±0.0080	0.3483±0.0091	0.5452±0.0405	0.3675±0.0123	0.5824±0.0000	0.4218±0.0013	0.3862±0.0007	0.5394±0.0083
satellite	0.8601±0.0000	0.8586±0.0000	0.8235±0.0088	0.8762±0.0024	0.8110±0.0197	0.7896±0.0046	0.8509±0.0018	0.8583±0.0072	0.8479±0.0035	**0.8805**±0.0000	0.8372±0.0005	0.8041±0.0005	0.8403±0.0023
pendigits	**0.9699**±0.0000	0.7855±0.0000	0.5879±0.0521	0.6641±0.0758	0.0934±0.0778	0.3335±0.0285	0.6114±0.0473	0.9189±0.0330	0.4844±0.0592	0.6633±0.0000	0.6158±0.0087	0.8307±0.0064	0.9642±0.0199
annthyroid	0.6807±0.0000	0.5353±0.0000	0.5902±0.0539	0.4583±0.0216	0.2783±0.0593	0.5874±0.0500	0.6288±0.0259	0.6815±0.0038	**0.8288**±0.0059	0.4881±0.0000	0.6843±0.0008	0.6912±0.0034	0.6491±0.0415
mnist	0.7272±0.0000	0.7097±0.0000	0.5415±0.0653	0.6845±0.0163	0.4600±0.0955	0.6509±0.0057	0.6242±0.0439	0.7368±0.0131	0.5626±0.0326	0.5797±0.0000	**0.7690**±0.0142	0.6123±0.0034	0.6832±0.0153
MNIST-C	0.4620±0.0000	0.5189±0.0000	0.3280±0.0243	0.5147±0.0110	0.3144±0.0304	0.4123±0.0035	0.4178±0.0012	0.4742±0.0000	0.4716±0.0136	0.3840±0.0334	0.5126±0.0133	0.5219±0.0047	**0.6115**±0.0018
agnews	0.1668±0.0000	**0.2586**±0.0000	0.1194±0.0032	0.1542±0.0038	0.1022±0.0185	0.1242±0.0031	0.1185±0.0003	0.1735±0.0000	0.1397±0.0297	0.1516±0.0317	0.2397±0.0089	0.2050±0.0020	0.2365±0.0049
amazon	0.1169±0.0000	0.1104±0.0000	0.1107±0.0019	0.1019±0.0046	0.1024±0.0127	0.1194±0.0021	0.1076±0.0081	0.1173±0.0000	0.1115±0.0065	0.1110±0.0146	0.1206±0.0013	0.1148±0.0011	0.1096±0.0028
imdb	0.0892±0.0000	0.0903±0.0000	0.0897±0.0017	0.1024±0.0013	0.0968±0.0147	0.0880±0.0011	0.0870±0.0002	0.0897±0.0000	0.0890±0.0052	**0.1053**±0.0236	0.0925±0.0032	0.0903±0.0017	0.0986±0.0028
yelp	0.1603±0.0000	0.1614±0.0000	0.1315±0.0029	0.1040±0.0009	0.1002±0.0101	0.1313±0.0039	0.1280±0.0002	0.1635±0.0000	0.1300±0.0157	0.1390±0.0300	**0.1666**±0.0020	0.1581±0.0022	0.1372±0.0017
mammography	0.4127±0.0000	0.3407±0.0000	0.3794±0.0321	0.1711±0.0375	0.2754±0.1140	0.2782±0.0384	0.1993±0.0392	0.4209±0.0086	0.3980±0.0426	0.3628±0.0126	**0.4743**±0.0032	0.3786±0.0063	0.4255±0.0676
magic.gamma	0.8586±0.0000	0.8636±0.0000	0.8027±0.0107	0.8133±0.0057	0.6954±0.0073	0.7613±0.0242	0.8797±0.0084	0.8615±0.0074	0.8968±0.0050	0.8754±0.0018	0.8854±0.0005	**0.9007**±0.0007	0.8981±0.0066
campaign	0.4904±0.0000	0.4024±0.0000	0.4573±0.0185	0.4890±0.0098	0.3695±0.1270	0.2309±0.0718	0.4887±0.0029	**0.4995**±0.0067	0.4690±0.0071	0.3424±0.0047	0.4323±0.0090	0.4767±0.0017	0.4518±0.0371
shuttle	0.9786±0.0000	0.9975±0.0000	0.9861±0.0034	0.9972±0.0014	0.9803±0.0013	0.6016±0.2692	0.9791±0.0026	0.9814±0.0048	0.9403±0.0011	0.9936±0.0017	0.9834±0.0006	0.9806±0.0033	**0.9993**±0.0003
aloi	0.0602±0.0000	0.0654±0.0000	0.0582±0.0010	0.0550±0.0013	0.0623±0.0035	0.0570±0.0024	0.0597±0.0006	0.0605±0.0006	0.0576±0.0004	0.0693±0.0005	0.0589±0.0002	0.0635±0.0002	**0.0704**±0.0022
smtp	0.5053±0.0592	0.4809±0.0738	0.0110±0.0012	0.0381±0.0383	0.3073±0.2282	0.3240±0.0848	0.4081±0.1336	0.5020±0.0639	0.5037±0.0615	0.3818±0.0990	0.3683±0.0484	0.5001±0.0579	**0.6585**±0.0128
backdoor	0.4654±0.0141	0.5347±0.0260	0.0937±0.0148	0.8918±0.0104	0.8477±0.0279	0.0631±0.0194	0.1420±0.0063	0.4570±0.1250	0.6274±0.0229	0.4386±0.1497	0.2084±0.0198	0.7197±0.0132	**0.9160**±0.0042
celeba	0.1192±0.0050	0.0361±0.0015	0.1170±0.0135	0.0974±0.0068	0.0709±0.0432	0.0401±0.0121	**0.1803**±0.0260	0.1065±0.0049	0.1419±0.0231	0.0839±0.0038	0.1380±0.0073	0.0664±0.0026	0.0467±0.0071
skin	**0.9824**±0.0041	0.6168±0.0185	0.6458±0.0109	0.3246±0.0100	0.4299±0.0328	0.4218±0.0184	0.7637±0.0579	0.9478±0.0233	0.6908±0.0050	0.5104±0.0172	0.6880±0.0913	0.8831±0.0043	0.5814±0.0270
fraud	0.3868±0.0719	0.5509±0.0819	0.1822±0.0366	0.5388±0.0878	0.4833±0.1713	0.2944±0.2466	**0.6921**±0.0346	0.4210±0.0763	0.6214±0.1992	0.5513±0.1377	0.4463±0.0773	0.2252±0.0233	0.4079±0.2755
cover	0.5579±0.0374	0.8292±0.0219	0.0866±0.0153	0.3448±0.1639	0.0269±0.0153	0.0109±0.0018	0.7327±0.0336	0.5997±0.1057	0.6373±0.1233	0.7257±0.0838	0.0269±0.0064	0.8415±0.0300	**0.8444**±0.0599
census	**0.2168**±0.0063	0.1371±0.0042	0.1419±0.0073	0.2120±0.0061	0.1535±0.0107	0.0869±0.0099	0.1967±0.0060	0.2105±0.0067	0.1794±0.0109	0.1601±0.0297	0.1752±0.0253	0.1840±0.0121	0.2041±0.0000
http	**1.0000**±0.0000	0.9712±0.0392	0.5343±0.1208	0.7082±0.4206	0.3609±0.3185	0.6838±0.0801	0.9996±0.0006	0.9710±0.0404	0.5545±0.0561	0.9058±0.0501	0.8048±0.0738	0.9893±0.0081	0.7673±0.0119
donors	0.8909±0.0994	0.6339±0.0189	0.4051±0.0359	0.9835±0.0087	0.4275±0.2748	0.0900±0.0193	0.2666±0.0291	0.8555±0.0456	0.7133±0.0389	0.9843±0.0048	0.5582±0.0508	0.8112±0.0095	**1.0000**±0.0000
rAUC	0.6543	0.5596	0.3483	0.6572	0.3364	0.3485	0.5557	0.7012	0.5481	0.6577	0.6734	0.7228	**0.7817**
WinRate	0.5541	0.5110	0.2712	0.5826	0.2463	0.2303	0.4386	0.6265	0.4846	0.5365	0.6243	0.6535	**0.7405**
Avg. Rank	6.1754	6.9737	9.8772	6.0789	10.0702	9.9211	7.4649	5.6228	7.2018	6.5351	5.6930	5.2632	**4.1228**

Table 7: Extended Results (AUCROC)

Dataset	kNN	LOF	IForest	ICL	DeepSVDD	GOAD	DDPM	DTE-NP	DTE-C	FoMo-0D	TabPFN-OD	OutFormer	GradPAO
hepatitis	0.9646±0.0146	0.6692±0.0701	0.8269±0.0275	0.9994±0.0013	0.9957±0.0024	0.8450±0.0325	0.9774±0.0118	0.9322±0.0390	0.9878±0.0088	0.9988±0.0024	**0.9996**±0.0009	0.9989±0.0022	0.9985±0.0024
MVTec-AD	0.8151±0.0182	0.8036±0.0214	0.7738±0.0195	**0.9475**±0.0084	0.8955±0.0219	0.7706±0.0210	0.7800±0.0199	0.8966±0.0151	0.8939±0.0200	0.7531±0.0126	0.9422±0.0046	0.8630±0.0029	0.9023±0.0047
wine	0.9919±0.0022	0.9836±0.0038	0.9392±0.0179	0.9987±0.0029	0.9216±0.0434	0.9411±0.0186	0.9961±0.0009	0.9944±0.0097	0.9997±0.0004	0.9994±0.0012	0.9997±0.0005	0.9995±0.0010	0.9986±0.0017
lymphography	0.9993±0.0008	0.9821±0.0075	0.9945±0.0032	**1.0000**±0.0000	0.9973±0.0030	0.9989±0.0008	0.9994±0.0009	0.9993±0.0009	0.9899±0.0040	0.9993±0.0011	0.9995±0.0011	0.9998±0.0004	0.9998±0.0004
wpbc	0.6367±0.0234	0.5736±0.0201	0.5633±0.0272	**0.9661**±0.0117	0.8267±0.0547	0.5139±0.0574	0.6649±0.0317	0.8316±0.1346	0.6888±0.0270	0.7610±0.0211	0.9650±0.0153	0.9073±0.0161	0.9577±0.0139
glass	0.9204±0.0112	0.8882±0.0198	0.8109±0.0256	**0.9944**±0.0058	0.8367±0.1620	0.5903±0.1264	0.6667±0.1363	0.8964±0.0354	0.9242±0.0230	0.9830±0.0053	0.9898±0.0057	0.9450±0.0289	0.9861±0.0117
wbc	0.9912±0.0019	0.8052±0.0447	0.9941±0.0037	0.9966±0.0036	0.9144±0.0364	0.9914±0.0020	0.9920±0.0039	0.9954±0.0028	0.8053±0.0710	0.9961±0.0031	0.9965±0.0037	0.9955±0.0050	**0.9996**±0.0003
vertebral	0.5767±0.0358	0.6430±0.0131	0.4564±0.0383	0.7921±0.0636	0.4979±0.0917	0.4673±0.0474	0.7060±0.0000	0.5430±0.1547	1.0641±0.0000	0.8720±0.0227	**0.9530**±0.0182	0.8457±0.0256	0.9431±0.0128
stamps	0.9589±0.0144	0.9374±0.0236	0.9347±0.0142	0.9668±0.0109	0.7109±0.0368	0.8146±0.1536	0.9184±0.0415	0.9787±0.0037	0.9160±0.0204	0.9752±0.0039	0.9773±0.0059	0.9859±0.0057	**0.9896**±0.0048
ionosphere	0.9744±0.0098	0.9429±0.0220	0.9121±0.0137	0.9898±0.0032	0.9720±0.0126	0.9154±0.0305	0.9460±0.0085	0.9777±0.0139	0.9542±0.0058	0.9659±0.0125	**0.9904**±0.0024	0.9795±0.0086	0.9899±0.0044
wdbc	0.9905±0.0027	0.9962±0.0021	0.9873±0.0064	0.9978±0.0027	0.9931±0.0040	0.9896±0.0025	0.9930±0.0017	0.9952±0.0038	0.9848±0.0077	0.9977±0.0025	0.9935±0.0037	0.9964±0.0026	**0.9987**±0.0014
20news	0.5736±0.0066	0.6019±0.0062	0.5491±0.0117	0.6130±0.0105	0.5564±0.0417	0.5492±0.0104	0.5488±0.0052	0.5997±0.0074	0.6430±0.0317	0.5693±0.0146	**0.6738**±0.0039	0.6543±0.0081	0.6506±0.0034
breastw	0.9905±0.0026	0.8891±0.0680	0.9950±0.0008	0.9828±0.0045	0.9696±0.0092	0.9886±0.0033	0.9870±0.0043	0.9928±0.0012	0.9278±0.0175	0.9915±0.0023	0.9951±0.0013	0.9939±0.0012	**0.9961**±0.0010
pima	0.7694±0.0187	0.7053±0.0219	0.7426±0.0163	0.7968±0.0306	0.5799±0.0285	0.6231±0.1370	0.7027±0.0228	0.8150±0.0257	0.6988±0.0195	0.8046±0.0148	**0.8281**±0.0111	0.8130±0.0115	0.8069±0.0231
vowels	0.8221±0.0000	0.8630±0.0000	0.6183±0.0062	0.8510±0.0213	0.5573±0.0443	0.6849±0.0371	0.8638±0.0194	0.8142±0.0144	0.8693±0.0225	0.8129±0.0000	0.7610±0.0058	**0.8714**±0.0000	0.8347±0.0135
yeast	0.4474±0.0000	0.4579±0.0000	0.4180±0.0075	0.4898±0.0239	0.4762±0.0596	**0.5253**±0.0388	0.4913±0.0285	0.4458±0.0033	0.4708±0.0110	0.4991±0.0000	0.4892±0.0011	0.5115±0.0000	0.4978±0.0087
letter	0.3543±0.0000	0.4483±0.0000	0.3204±0.0064	0.4248±0.0017	0.3205±0.0305	0.3108±0.0055	0.3438±0.0098	0.3438±0.0012	0.3438±0.0095	0.4090±0.0000	0.4217±0.0017	0.3833±0.0000	0.3725±0.0037
cardio	0.9200±0.0000	0.9221±0.0000	0.9332±0.0143	0.8001±0.0212	0.6543±0.0437	**0.9601**±0.0027	0.8694±0.0196	0.9180±0.0059	0.8726±0.0100	0.9427±0.0000	0.7626±0.0034	0.9126±0.0000	0.8908±0.0055
fault	0.5873±0.0000	0.4742±0.0000	0.5586±0.0203	0.6063±0.0050	0.5431±0.0162	0.5889±0.0061	0.6109±0.0116	0.5864±0.0065	0.5946±0.0145	0.6182±0.0000	0.5668±0.0052	0.5836±0.0000	0.5178±0.0151
internetads	0.6808±0.0000	0.7172±0.0000	0.4787±0.0211	0.7220±0.0055	0.7296±0.0324	0.6565±0.0021	0.6576±0.0006	0.6996±0.0222	0.7757±0.0154	0.5564±0.0712	0.6069±0.0122	0.7698±0.0114	0.7380±0.0163
cardiotocography	0.6211±0.0000	0.6449±0.0000	0.7424±0.0288	0.5420±0.0180	0.4775±0.0859	**0.7606**±0.0140	0.5454±0.0296	0.6376±0.0188	0.6013±0.0259	0.7215±0.0000	0.4738±0.0056	0.6229±0.0000	0.5499±0.0228
musk	1.0000±0.0000	1.0000±0.0000	0.9058±0.0620	0.9937±0.0063	0.9999±0.0001	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000	0.9962±0.0043	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000
waveform	0.7521±0.0000	0.7600±0.0000	0.7229±0.0152	0.6868±0.0366	0.5994±0.0258	0.6499±0.0310	0.6217±0.0264	0.7448±0.0000	0.6521±0.0105	0.7102±0.0000	0.7152±0.0010	0.7263±0.0007	**0.8061**±0.0042
speech	0.3636±0.0000	0.3753±0.0000	0.3770±0.0172	0.4886±0.0272	**0.4888**±0.0288	0.3663±0.0114	0.3696±0.0086	0.4137±0.0000	0.3817±0.0057	0.4010±0.0175	0.3718±0.0122	0.3739±0.0024	0.3792±0.0091
thyroid	0.9868±0.0000	0.9272±0.0000	**0.9896**±0.0020	0.9540±0.0098	0.8877±0.0369	0.9515±0.0085	0.9795±0.0022	0.9863±0.0004	0.9374±0.0014	0.9684±0.0000	0.9761±0.0006	0.9734±0.0005	0.9872±0.0013
spambase	0.8336±0.0000	0.7323±0.0000	0.8518±0.0169	0.8353±0.0045	0.7024±0.0502	0.8178±0.0036	0.6454±0.0080	0.8374±0.0069	0.8301±0.0041	0.8321±0.0000	**0.8554**±0.0088	0.7973±0.0009	0.8373±0.0157
wilt	0.6366±0.0000	0.6881±0.0000	0.4797±0.0112	0.7642±0.0346	0.3441±0.0170	0.5138±0.0533	0.7166±0.0081	0.6291±0.0559	0.8510±0.0113	**0.9713**±0.0000	0.7786±0.0006	0.9274±0.0007	0.9333±0.0060
SVHN	0.6169±0.0000	0.6382±0.0000	0.5899±0.0088	0.6170±0.0050	0.5392±0.0305	0.6075±0.0049	0.6137±0.0008	0.6213±0.0000	0.6291±0.0110	0.5993±0.0064	0.6244±0.0038	**0.6445**±0.0004	0.6376±0.0016
optdigits	0.9372±0.0000	0.9665±0.0000	0.8107±0.0328	**0.9718**±0.0080	0.3945±0.1865	0.6746±0.0494	0.9076±0.0212	0.9428±0.0167	0.8238±0.0268	0.8528±0.0000	0.9189±0.0079	0.8557±0.0010	0.9199±0.0189
CIFAR10	0.6753±0.0000	0.7030±0.0000	0.6404±0.0093	0.6360±0.0083	0.5612±0.0220	0.6753±0.0072	0.6791±0.0013	0.6782±0.0000	0.6853±0.0159	0.6453±0.0062	0.6554±0.0086	0.7098±0.0020	**0.7155**±0.0018
pageblocks	0.8965±0.0000	0.9130±0.0000	0.8264±0.0089	0.8839±0.0077	0.7839±0.0153	0.8805±0.0119	0.8693±0.0042	0.8932±0.0026	0.8989±0.0055	0.8658±0.0000	**0.9211**±0.0013	0.9088±0.0017	0.9024±0.0030
satimage-2	0.9971±0.0000	0.9938±0.0000	0.9912±0.0021	0.9948±0.0022	0.9294±0.0284	0.9899±0.0006	0.9962±0.0016	0.9967±0.0000	0.9934±0.0007	0.9840±0.0000	0.9668±0.0015	**0.9975**±0.0001	0.9957±0.0006
FashionMNIST	0.8987±0.0000	**0.9160**±0.0000	0.8415±0.0107	0.9056±0.0027	0.7545±0.0222	0.8803±0.0024	0.8852±0.0007	0.9014±0.0000	0.9021±0.0055	0.8666±0.0066	0.8981±0.0062	0.9121±0.0003	0.9124±0.0009
landsat	0.6825±0.0000	0.6658±0.0000	0.5880±0.0221	0.6513±0.0044	0.5944±0.0141	0.4052±0.0232	0.5137±0.0100	0.6820±0.0175	0.5279±0.0163	**0.6953**±0.0000	0.5905±0.0014	0.5209±0.0011	0.6368±0.0035
satellite	0.8224±0.0000	0.8030±0.0000	0.7746±0.0148	**0.8515**±0.0048	0.7619±0.0266	0.6876±0.0092	0.7770±0.0053	0.8211±0.0067	0.7861±0.0070	0.8498±0.0000	0.8002±0.0003	0.7377±0.0005	0.7762±0.0040
pendigits	**0.9987**±0.0000	0.9905±0.0000	0.9702±0.0083	0.9671±0.0083	0.4629±0.1175	0.8997±0.0203	0.9811±0.0023	0.9961±0.0017	0.9729±0.0068	0.9811±0.0000	0.9713±0.0014	0.9931±0.0003	0.9985±0.0009
annthyroid	0.9281±0.0000	0.8863±0.0000	0.9028±0.0152	0.8111±0.0107	0.5501±0.0362	0.8101±0.0516	0.8882±0.0134	0.9290±0.0025	**0.9752**±0.0015	0.8517±0.0000	0.9174±0.0003	0.9445±0.0013	0.9244±0.0153
mnist	0.9385±0.0000