SEMI-SUPERVISED TABULAR ANOMALY DETECTION WITH CONTAMINATION-TOLERANT IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) enables pretrained tabular models to adapt to new tasks without task-specific parameter updates by treating training instances as context. However, directly applying ICL to semi-supervised tabular anomaly detection is challenging because unlabeled contextual data may be contaminated by hidden anomalies, while only a few labeled anomalies are available to provide reliable supervision. Moreover, open-world unlabeled data may contain unknown anomaly types that are not represented by the labeled anomalies. To address these challenges, we propose TabPFN-AD, a framework that iteratively constructs a more accurate in-context set under the semi-supervised tabular anomaly detection setting. TabPFN-AD leverages the inter-instance attention of a pretrained tabular foundation model to evaluate the degree of anomalousness at both the instance and group levels. By mining both known and unknown anomalies from unlabeled data, it maximizes the utility of limited labeled supervision to improve detection performance. Experimental results demonstrate that TabPFN-AD achieves state-of-the-art performance on 17 benchmark datasets and exhibits strong robustness under scenarios where unlabeled data is mixed with a high proportion of potential anomalies and labeled anomalies are extremely scarce.

1 INTRODUCTION

Tabular anomaly detection aims to identify rare or abnormal instances in structured data and plays an important role in applications such as financial fraud detection (Chen et al., 2024; Zhang et al., 2022), network intrusion detection (Moustafa & Slay, 2015; Yin et al., 2023), and electronic health-record auditing (Röchner & Rothlauf, 2023; Kuwaye & Cho, 2024). Existing tabular anomaly detection methods (Wang et al., 2020; Liu et al., 2021; Bergman & Hoshen, 2020; Xu et al., 2023a) predominantly follow an unsupervised paradigm, detecting anomalies without explicit supervision under the assumptions that most instances are normal and that the distribution of anomalies differs from that of normal ones. With the increasing availability of small numbers of labeled anomalies in real-world applications, semi-supervised anomaly detection methods (Wei et al., 2024; Ruff et al., 2020; Pang et al., 2023) have emerged as a promising alternative by using limited anomaly supervision to guide the detection process. Nevertheless, existing semi-supervised methods typically require task-specific model training, limiting their efficiency and adaptability in real-world applications.

Table 1: AUPRC on the UNSW-NB15 dataset by varying contamination ratios and labeled anomaly counts. Here, ↑/↓ indicate performance gain/loss of TabPFN relative to TabPFN-AD.

Contamination Ratio			
Method	0.01	0.05	0.1
TabPFN	0.770 ↓8.4%	0.668 ↓19.6%	0.346 ↓59.2%
TabPFN-AD	**0.841**	**0.831**	**0.849**
# Labeled Anomalies			
Method	5	30	100
TabPFN	0.683 ↓14.3%	0.802 ↓7.7%	0.886 ↓1.2%
TabPFN-AD	**0.797**	**0.869**	**0.897**

In-context learning (ICL) has recently emerged as a promising paradigm for adapting pretrained models to new tasks without task-specific parameter updates. Following this paradigm, tabular

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.

foundation models treat labeled training instances as contextual instances for test instances and perform prediction in a single forward pass. Among them, TabPFN (Hollmann et al., 2023; 2025) is a representative model equipped with feature-wise and instance-wise attention. This naturally raises an important question: *can ICL be effectively applied to semi-supervised tabular anomaly detection?*

In real-world scenarios, constructing a reliable context set for anomaly detection faces two critical challenges. The first is the scarcity of labeled anomalies. Only a few labeled anomalies are typically available, providing insufficient supervision to characterize anomalous patterns. As shown in Table 1, reducing the number of labeled anomalies from 100 to 5 causes a relative AUPRC drop of 22.9% for TabPFN. The second challenge is anomaly contamination in unlabeled data. Unlabeled instances are commonly treated as normal despite containing hidden anomalies, some of which may represent previously novel anomaly types that differ from the labeled anomalies. Treating these hidden anomalies as normal corrupts the contextual information used for prediction. As the contamination ratio increases from 1% to 10%, the AUPRC of TabPFN decreases by 55.1%.

To address these challenges, we propose TabPFN-AD, a framework that constructs a more accurate context for semi-supervised tabular anomaly detection by identifying and relabeling potential anomalies in unlabeled data. By leveraging the instance-wise attention mechanism in TabPFN to mine potential anomalies that deviate from normal patterns, these mined anomalies are assigned pseudo-labels and iteratively refined, providing progressively more accurate contextual information for detection. This anomaly mining mechanism comprises two complementary aspects: First, instance-level assessment utilizes in-degree and out-degree characteristics in sparse attention graphs to identify anomalies structurally similar to labeled anomalies; second, group-level assessment discovers novel anomaly types through subgraph compactness analysis, capturing patterns that deviate from both normal data and known anomalies. As shown in Table 1, TabPFN-AD demonstrates superior robustness over TabPFN under both challenging settings. Specifically, as the contamination ratio increases from 1% to 10%, TabPFN-AD maintains a stable AUPRC (0.841–0.849), while TabPFN suffers a 55.1% relative drop. Furthermore, with only five labeled anomalies, TabPFN-AD achieves a 16.7% relative improvement in AUPRC over TabPFN.

The main contributions of this paper are summarized as follows:

- We propose TabPFN-AD, a framework that mines potential anomalies from unlabeled data to iteratively construct a more accurate in-context set for semi-supervised anomaly detection.
- We introduce complementary anomaly assessment mechanisms to mine potential known and unknown anomalies from unlabeled data for enhanced contextual detection.
- Extensive experiments on 17 benchmark datasets demonstrate that TabPFN-AD significantly outperforms both existing tabular foundation models and semi-supervised methods. Moreover, it remains robust under high contamination ratios and extremely limited labeled anomalies, mitigating a critical weakness of existing tabular foundation models.

2 Related Work

Tabular Anomaly Detection Tabular data presents unique challenges for anomaly detection: complex nonlinear feature relationships, mixed numerical and categorical attributes, and the absence of spatial or sequential structures. Early methods such as Isolation Forest (Liu et al., 2008), LOF (Breunig et al., 2000), and OC-SVM (Tax & Duin, 2004) rely on manual feature engineering and statistical assumptions and struggle to capture complex patterns in high-dimensional data. As such, deep learning approaches have emerged, primarily through boundary-based methods (Ruff et al., 2018; Goyal et al., 2020) that map normal data to compact decision boundaries and self-supervised learning (Pang et al., 2018; Xu et al., 2023c; Ye et al., 2025a) that constructs pseudo-tasks favoring normal instances. Semi-supervised methods (Ruff et al., 2020; Pang et al., 2019) leverage limited labeled anomalies for precise boundaries, while reinforcement learning approaches (Pang et al., 2021; Shou et al., 2025) address partially labeled anomaly categories and unlabeled data contamination through simulated environments. Despite their effectiveness, these methods require large training data and extensive hyperparameter tuning.

LLMs for Anomaly Detection Large language models (LLMs), owing to their strong contextual understanding and reasoning capabilities, have recently been applied to anomaly detection across

multiple modalities, including images (Zhang et al., 2025a; Gu et al., 2024), time series (He et al., 2026; Liu et al., 2025), and tabular data (Li et al., 2024; Tsai et al., 2025). However, existing studies predominantly focus on image and time-series anomalies, while the application of LLMs to tabular anomaly detection remains in its early stages, with only a few methods specifically designed for tabular inputs. Li et al. (2024) explore zero-shot batch-level anomaly detection by prompting an LLM to evaluate individual dimensions and aggregate the resulting scores across dimensions. Although it demonstrates the feasibility of LLMs for zero-shot tabular anomaly detection, its formulation provides limited adaptation to dataset-specific distributions and does not exploit the supervision available from labeled anomalies. AnoLLM (Tsai et al., 2025) converts tabular instances into textual representations and derives anomaly scores from their negative log-likelihood. Nevertheless, representing numerical tables as text may not fully preserve their native structural and numerical relationships. Moreover, its reliance on autoregressive language-model processing introduces substantial computational overhead during both training and inference.

Tabular Foundation Model Tabular foundation models are general models pre-trained on massive, diverse tabular datasets with cross-dataset generalization capabilities, enabling rapid adaptation to downstream tasks via in-context learning. The most representative model is TabPFN (Hollmann et al., 2023; 2025), which transforms tabular classification into an in-context learning problem by pre-training a transformer on synthetic data to approximate Bayesian inference. This enables single forward-pass inference on small-scale datasets without hyperparameter tuning while achieving satisfactory classification accuracy. However, its quadratic attention complexity and low pre-training data quality hinder its applicability to larger-scale scenarios. Further studies have thus proposed various improvements: TabFlex (Zeng et al., 2025) employs linear attention to enhance large-scale data processing capabilities; TabForestPFN (den Breejen et al., 2024) and TabDPT (Ma et al., 2025) improve pre-training data quality; TabICL (Qu et al., 2025) achieves a balance between performance and scalability. Despite its success in classification, TabPFN's application to anomaly detection remains underexplored. The scarcity of labeled anomalies and contamination of unlabeled data significantly undermine its in-context learning effectiveness. To bridge this gap, we propose TabPFN-AD, which adapts the tabular foundation model TabPFN for anomaly detection via an instance-wise attention mechanism and iterative pseudo-label refinement.

3 METHOD

This section introduces TabPFN-AD, a semi-supervised tabular anomaly detection framework that leverages TabPFN's instance-wise attention to mine potential anomalies from unlabeled data and iteratively construct more reliable contextual information for in-context prediction. An overview of the TabPFN-AD framework is shown in Figure 1. Specifically, TabPFN-AD consists of three key components: (i) attention-head selection using the labeled anomaly relevance criterion C_1 and the self-similarity criterion C_2; (ii) complementary anomaly mining comprising an instance-level score based on out- and in-degree characteristics of instances and a group-level score via subgraph compactness analysis to mine potential anomalies; and (iii) iterative pseudo-label refinement to progressively enhance label confidence.

3.1 PRELIMINARIES

We consider a tabular anomaly detection task with an input space $\mathcal{X} \in \mathbb{R}^m$. Given a training set $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$ such that $|\mathcal{D}_L| \ll |\mathcal{D}_U|$, where $\mathcal{D}_L = \left\{\boldsymbol{x}^{(i)} \mid y^{(i)} = 1\right\}_{i=1}^{n_l}$ contains only labeled anomalous instances and $\mathcal{D}_U = \{\boldsymbol{x}^{(i)}\}_{i=n_l+1}^{n}$ is a predominantly normal, unlabeled set with a small fraction of anomalous instances. Given a test instance $\boldsymbol{x}_{\text{test}}$, the goal is to predict its anomaly probabilities $p(\cdot|\boldsymbol{x}_{\text{test}}, \mathcal{D})$.

TabPFN (Hollmann et al., 2023; 2025) is a pretrained tabular foundation model that performs in-context prediction without parameter updates. Given labeled context instances and unlabeled queries, its inter-instance attention allows each query to aggregate contextual information and predict its label in a single forward pass. Therefore, prediction quality depends strongly on context label quality. For a head $h \in \mathcal{H}$ and N instances, the attention map is denoted by $\mathcal{A}^h = [a_{ij}^h] \in \mathbb{R}^{N \times N}$, where a_{ij}^h is the attention from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$. Different heads capture distinct inter-instance relationships.

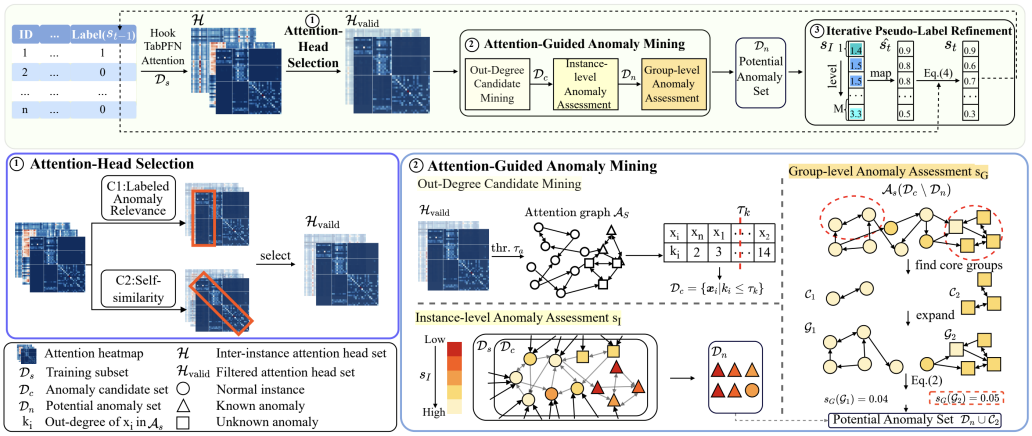

Figure 1: Overall framework of TabPFN-AD.

3.2 Attention-Head Selection

To mine potential anomalies from unlabeled data, we first select attention heads that exhibit same-class associations, characterized by stronger normal-to-normal and anomaly-to-anomaly associations than cross-class associations. To satisfy TabPFN's context-size constraint and reduce attention extraction costs, we construct $\mathcal{D}_s = \mathcal{D}_L \cup \mathcal{D}_{U,s}$, where $\mathcal{D}_{U,s} \subseteq \mathcal{D}_U$ is randomly sampled from the unlabeled data and all labeled anomalies are retained.

On this subset, we introduce two complementary criteria, as illustrated in Figure 1: the *labeled-anomaly relevance criterion* C_1 and the *self-similarity criterion* C_2. Criterion C_1 evaluates whether an attention head captures strong associations among the available labeled anomalies, thereby excluding heads without evident anomaly-related structures. Criterion C_2 requires sufficiently concentrated diagonal attention, which serves as an empirical indicator of similarity preservation. An attention head is retained only when both criteria are satisfied.

Let $\mathcal{H}$ denote the set of all attention heads under consideration, and let $C_1(h)$ and $C_2(h)$ be the binary indicators of the two criteria for head $h \in \mathcal{H}$. The selected head set is defined as

$$\mathcal{H}_{\text{valid}} = \{h \in \mathcal{H} \mid C_1(h) \wedge C_2(h)\}. \tag{1}$$

The exact definitions of C_1 and C_2, together with qualitative visualizations, cross-dataset consistency analysis, and ablation results, are provided in Appendix A.1.

3.3 Attention-Guided Anomaly Mining

Having identified attention heads that capture same-class associations through the aforementioned attention head selection, we now proceed to mine potential known and unknown anomalies from unlabeled data on these filtered heads via TabPFN's inter-instance attention mechanism. Specifically, this component comprises three modules: First, the anomaly candidate set $\mathcal{D}_c$ is constructed by leveraging the out-degree differences between normal and anomalous instances. Second, an instance-level anomaly score based on both out- and in-degree characteristics of instances is designed to better distinguish the two classes, thereby further filtering potential anomalies from $\mathcal{D}_c$. Finally, motivated by the observation that anomalies tend to form more compact subgroups in the attention space, a group-level anomaly score is designed to capture unknown anomalies missed by the single instance-level score.

Anomaly Candidate Construction Recall that the selected attention heads capture same-class associations. Consequently, anomalous instances tend to concentrate their attention on their few same-class peers, whereas normal instances distribute their attention across a much larger set of normal instances. As illustrated in Figure 2(a), this hypothesis holds: anomalous instances display pronounced attention peaks over a small subset, whereas normal instances yield flat distributions. Motivated by this disparity, we convert the attention map into a directed sparse graph $\mathcal{A}_s$ to preserve

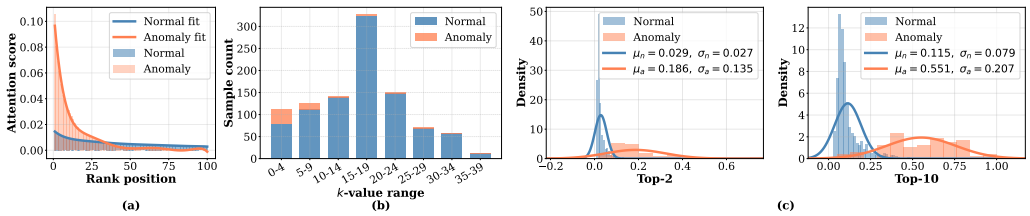

Figure 2: Empirical analysis of the attention characteristics underlying the instance- and group-level anomaly scores on NB15. (a) Descending ranked top-100 class-averaged attention scores with polynomial fits (solid lines). (b) Binned instance counts by out-degree k for normal and anomalous instances. (c) Histograms and Gaussian density fits of top-k attention score sums for normal (blue) and anomaly (red) instances. Each subplot shows the fitted Gaussian parameters (μ_n, σ_n) for normal data and (μ_a, σ_a) for anomalies.

its essential structural information. Specifically, given a threshold τ_a, we retain for each instance the top-k outgoing edges, where k is the smallest number such that their cumulative attention reaches τ_a. Since anomalous instances concentrate high attention scores on fewer peers, they require fewer edges (smaller k) to reach τ_a. Conversely, normal instances with flatter distributions need more edges (larger k). As shown in Figure 2(b), bins with smaller out-degrees k contain a notably higher proportion of anomalies. This phenomenon verifies that anomalous instances inherently exhibit low out-degrees, offering a simple yet effective criterion for potential anomaly detection.

Then, we construct the candidate anomaly set $\mathcal{D}_c$ by selecting instances whose out-degree is less than or equal to a predefined threshold τ_k, i.e., $k \leq \tau_k$. Empirically, $\mathcal{D}_c$ achieves significantly higher anomaly proportions than the original unlabeled data across datasets while covering most anomalies in $\mathcal{D}_s$. However, out-degree alone is insufficient to obtain a pure potential anomaly set, as $\mathcal{D}_c$ still contains normal instances (see Appendix C.1 for dataset-wise visualizations). To address this, we propose a dual-perspective anomaly assessment mechanism to evaluate an instance's degree of anomalousness at both the instance and group levels to identify potential anomalies within $\mathcal{D}_c$.

Instance-Level Anomaly Assessment The out-degree reflects an instance's active tendency to seek similar peers, and pointing to others indicates that it regards those instances as references. In the constructed graph $\mathcal{A}_s$, each instance only preserves a small set of high-weight outgoing edges. Consequently, being pointed to by others (i.e., in-degree) becomes a non-trivial indicator of an instance's representativeness, offering complementary discriminative clues to distinguish normal and anomalous instances: (1) **Distribution difference:** Compared with $\mathcal{D}_c$, the complementary set $\mathcal{D}_s \setminus \mathcal{D}_c$ contains substantially more normal instances and only a small number of anomalies. As a result, normal instances in $\mathcal{D}_c$ receive many in-edges from these external normal instances, yielding substantially higher in-degrees from the full subset $\mathcal{D}_s$ than from the candidate set alone (i.e., $d_{in}^{\mathcal{D}_s} \gg d_{in}^{\mathcal{D}_c}$). In contrast, anomalous instances in $\mathcal{D}_c$ primarily receive in-edges from other anomalies within $\mathcal{D}_c$, resulting in $d_{in}^{\mathcal{D}_s} \approx d_{in}^{\mathcal{D}_c}$. (2) **Quantity difference:** A large in-degree implies rich shared features and a wide reference frequency, which is a typical trait of normal instances. On the contrary, anomalous instances lack sufficient similar peers, resulting in much lower overall in-degree counts. Specifically, anomalies obtain far fewer in-degree neighbors within $\mathcal{D}_s$ compared with normal instances. By leveraging both out- and in-degree characteristics, we design an instance-level anomaly score s_I to measure individual anomaly degrees:

$$s_I = d_{out} \cdot \left(\frac{d_{in}^{\mathcal{D}_s} + \epsilon}{d_{in}^{\mathcal{D}_c} + \epsilon} \right) \cdot \exp\left(\frac{d_{in}^{\mathcal{D}_s}}{\lambda} \right), \tag{2}$$

where λ controls the strength of the exponential scaling, ϵ prevents division by zero, the second term captures the distribution difference, and the third term captures the quantity difference described above. In this way, smaller s_I values indicate stronger anomalous characteristics.

All instances in $\mathcal{D}_c$ are sorted by s_I in ascending order. The cutoff is set to the potential anomaly count $|\mathcal{D}_n|$ from the previous iteration, with instances ranked above this threshold forming a new potential anomaly set $\mathcal{D}_n$. For the initial iteration, $|\mathcal{D}_n| = 0.3 \times |\mathcal{D}_c|$.

Group-Level Anomaly Assessment The assessment mechanism based on s_I preferentially captures homogeneous anomalies, which form strong attention connections with labeled anomalies and present small out-degrees, thus making them easily identifiable. In contrast, unknown anomalies have no corresponding labeled anomalies and may instead direct their attention toward other anomalies of the same unknown type. When these instances form mutually connected subgroups, their individual in- and out-degree patterns become less distinguishable from those of normal instances, preventing s_I from assigning them high anomaly rankings. Nevertheless, despite their scarcity, unknown anomalies with high mutual attention tend to form tightly connected clusters under heads that capture same-class associations. We therefore introduce the group-level score s_G to identify them by measuring subgroup compactness.

To mine potential anomalies in $\mathcal{D}_c \setminus \mathcal{D}_n$, a two-stage subgraph construction is employed. First, instance pairs with bidirectional edges are selected as initial core groups $\mathcal{C}$. Second, instances with in-degree or out-degree connections to these cores are expanded to form the peripheral set $\mathcal{P}$, yielding the induced subgraph $\mathcal{G} = \mathcal{I}(\mathcal{C} \cup \mathcal{P})$, where $\mathcal{I}(\cdot)$ extracts the subgraph induced by a given vertex set. The group-level anomaly score is then defined as the average attention within the induced subgroup:

$$s_G(\mathcal{G}) = \frac{1}{|\mathcal{G}|^2} \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{G}} a_{ij}. \tag{3}$$

This score measures the overall attention concentration within the subgroup. A higher s_G indicates that the subgroup exhibits stronger internal attention concentration relative to its size, which is used as an indicator of subgroup compactness in the attention space.

To ensure effective anomalous subgroup identification via $s_G(\mathcal{G})$, we characterize the discrimination boundary between anomalous and normal subgroups. Let $\mathcal{G}_a = \mathcal{I}(\mathcal{C}_a \cup \mathcal{P}_a)$ denote an anomalous subgroup with core size $c_a = |\mathcal{C}_a|$ and peripheral size $p_a = |\mathcal{P}_a|$, giving $|\mathcal{G}_a| = c_a + p_a$. Similarly, $\mathcal{G}_n = \mathcal{I}(\mathcal{C}_n \cup \mathcal{P}_n)$ denotes a normal subgroup with $c_n = |\mathcal{C}_n|$, $p_n = |\mathcal{P}_n|$, and $|\mathcal{G}_n| = c_n + p_n$. We consider the following two cases.

Case 1: When the anomalous subgroup has a smaller scale ($|\mathcal{G}_a| < |\mathcal{G}_n|$) and a higher within-subgroup attention concentration, we have $s_G(\mathcal{G}_a) > s_G(\mathcal{G}_n)$.

Case 2: When the normal subgroup has a smaller scale, we have $\rho = |\mathcal{G}_a|^2/|\mathcal{G}_n|^2 = (c_a+p_a)^2/(c_n+p_n)^2 > 1$. In this case, the anomalous subgroup must exhibit sufficiently stronger internal attention concentration to overcome the larger normalization term and achieve a higher group-level score.

For a peripheral instance $\boldsymbol{x}_i$, let $A_i = \sum_{\boldsymbol{x}_j \in \mathcal{G}} a_{ij}$ denote its attention contribution within subgroup $\mathcal{G}$. We assume that the peripheral attention sums are i.i.d. within each subgroup, with $A_i^{(a)} \sim \mathcal{N}(\mu_a, \sigma_a^2)$ and $A_i^{(n)} \sim \mathcal{N}(\mu_n, \sigma_n^2)$. For each core node i, the implementation retains the minimum number k_i of complete outgoing edges whose cumulative attention reaches τ_a. Since the last retained edge is not clipped, the retained attention mass can be written as $S_i = \sum_{\ell=1}^{k_i} a_{i,(\ell)} = \tau_a + \omega_i$, where $0 \le \omega_i < a_{i,(k_i)}$. We denote the total threshold overshoots by $\Omega_a = \sum_{i \in \mathcal{C}_a} \omega_i$ and $\Omega_n = \sum_{i \in \mathcal{C}_n} \omega_i$.

Let $\widetilde{p}_a = p_a - \sqrt{2p_a\sigma_a^2 \log(1/\delta)}/\mu_a$ and $\widetilde{p}_n = p_n + \sqrt{2p_n\sigma_n^2 \log(1/\delta)}/\mu_n$. According to the Gaussian tail bound (see the detailed derivation in Appendix A.3), if

$$\mu_a > \frac{(\rho c_n - c_a)\tau_a + \rho\Omega_n - \Omega_a}{\widetilde{p}_a} + \frac{\rho\widetilde{p}_n}{\widetilde{p}_a}\mu_n, \tag{4}$$

then, with probability at least $1 - 2\delta$, $s_G(\mathcal{G}_a) > s_G(\mathcal{G}_n)$.

This result provides a sufficient, rather than necessary, separation condition. It explicitly accounts for the threshold overshoot introduced by retaining the complete last edge. Figure 2(c) provides empirical support for a key premise of the condition: anomalous instances exhibit substantially larger top-k attention sums than normal instances, and the gap increases with k. However, the top-k comparison alone does not verify every term in Eq. 4; direct verification additionally requires subgroup sizes and the realized overshoot terms Ω_a and Ω_n.

Finally, the core set instances $\mathcal{C}$ from the subgroup with the highest s_G score are added to the potential anomaly set: $\mathcal{D}_n \leftarrow \mathcal{D}_n \cup \mathcal{C}$.

Table 2: Summary of the performance of each model in terms of mean AUPRC, average rank across datasets, and count of best-performing (top-1) results. The best and second-best results are highlighted in **bold** and underlined, respectively.

	Unsupervised			Semi-Supervised							Foundation Model		
Method	AnoLLM	DRL	Disent-AD	XGBoost	DeepSAD	DevNet	RoSAS	FeaWAD	PReNet	READ	LimiX	TabPFN	TabPFN-AD
Avg. AUPRC	0.4892	0.3824	0.5555	0.7089	0.7151	0.7599	0.7489	0.6522	0.7505	0.7696	0.8002	0.7465	**0.8469**
Avg. Rank	10.88	11.94	10.53	8.06	6.18	5.12	6.35	9.06	5.00	5.29	3.47	5.59	**1.53**
Rank-1 Count	0	0	0	1	1	1	2	0	3	1	2	3	**12**

3.4 Iterative Pseudo-Label Refinement

We assign anomaly pseudo-labels to the mined set $\mathcal{D}_n$ to correct contaminated contextual labels. The refined labels are fed back into TabPFN to update attention scores for subsequent mining and are eventually used as context for test-time prediction.

At iteration t, base pseudo-labels $\hat{s}_t$ are assigned to instances in $\mathcal{D}_n^t$ according to their instance-level score rankings. For an instance selected in the current iteration, its base and historical pseudo-labels are combined through temporal smoothing; otherwise, its historical pseudo-label gradually decays:

$$s_t(\boldsymbol{x}_i) = \begin{cases} \alpha s_{t-1}(\boldsymbol{x}_i) + (1-\alpha)\hat{s}_t(\boldsymbol{x}_i), & \boldsymbol{x}_i \in \mathcal{D}_n^t, \\ \beta s_{t-1}(\boldsymbol{x}_i), & \boldsymbol{x}_i \in \mathcal{D}_{U,s} \setminus \mathcal{D}_n^t, \end{cases} \tag{5}$$

where $\alpha, \beta \in (0,1)$ are the smoothing and decay coefficients. All unlabeled instances are initialized with $s_0(\boldsymbol{x}_i) = 0$, while labeled anomalies retain their original labels. Through multiple refinement rounds, the procedure expands the coverage of anomaly patterns beyond single-shot mining while improving pseudo-label reliability. To prevent erroneous pseudo-labels from being repeatedly reinforced, refinement terminates when the mined anomaly set becomes stable or the maximum number of iterations is reached. The base-label assignment and stopping criteria are detailed in Appendix B.2.

4 Experiments

4.1 Experimental Setup

Datasets We evaluated the proposed method on seventeen datasets: twelve from the ODDS library spanning chemistry, healthcare, and astronautics, and five from other sources commonly used for anomaly detection. Dataset details and partitions are provided in Appendix B.1.

Baselines We compared against ten established approaches and two tabular foundation models (TabPFN (Hollmann et al., 2025) and LimiX (Zhang et al., 2025b)). Unsupervised baselines include AnoLLM (Tsai et al., 2025), DRL (Ye et al., 2025a), and Disent-AD (Ye et al., 2025b). Semi-supervised methods comprise DeepSAD (Ruff et al., 2020), DevNet (Pang et al., 2019), RoSAS (Xu et al., 2023b), FeaWAD (Zhou et al., 2021), PReNet (Pang et al., 2023), READ (Shou et al., 2025), and XGBoost (Chen & Guestrin, 2016).

Implementation Details All experiments follow a semi-supervised setting with 10 labeled anomalies and a large volume of unlabeled data with 5% anomaly contamination. For baseline implementation, the DeepOD library is used for DeepSAD, DevNet, RoSAS, FeaWAD, and PReNet; all other baselines use official implementations with identical data splits. Trainable models undergo grid search on validation sets and are averaged over ten runs. All experiments run on NVIDIA(R) A800 40GB GPUs. Hyperparameter settings are detailed in Appendix B.2.

4.2 Overall Detection Performance

As shown in Table 2, we summarize the average AUPRC, average rank, and top-1 counts of the competing methods and TabPFN-AD across 17 datasets. Semi-supervised methods with limited

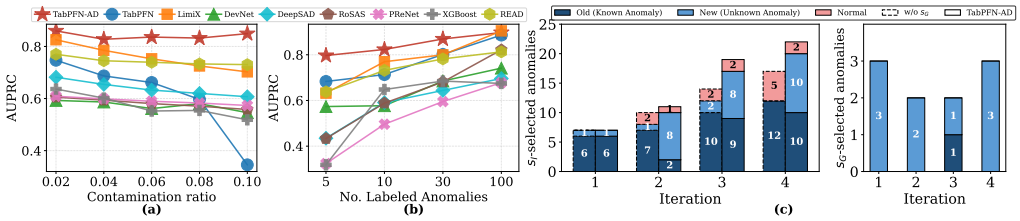

Figure 3: Robustness and anomaly-mining behavior of TabPFN-AD under challenging semi-supervised settings. (a) AUPRC under varying contamination ratios. (b) AUPRC under varying numbers of labeled anomalies. (c) Left: composition of instances selected by s_I in TabPFN-AD and its variant w/o s_G. Right: composition of subgraphs detected by s_G across iterations. Dark blue, light blue, and red denote known anomalies, unknown anomalies, and normal instances, respectively.

Table 3: Component ablation study on instance-level score s_I and group-level score s_G.

	TabPFN	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	TabPFN-AD
$s_I(R) = (d_{in}^{\mathcal{D}_s} + \epsilon)/(d_{in}^{\mathcal{D}_c} + \epsilon)$	×	✓	×	×	✓	✓	×	✓	✓
$s_I(E) = d_{out}$	×	×	✓	×	✓	×	✓	✓	✓
$s_I(A) = \exp(d_{in}^{\mathcal{D}_s}/\lambda)$	×	×	×	✓	×	✓	✓	✓	✓
s_G	×	×	×	×	×	×	×	×	✓
AUPRC	0.7465	0.8218	0.7993	0.8205	0.8287	0.8246	0.8294	0.8311	**0.8469**

labeled anomalies substantially outperform unsupervised baselines, highlighting the value of leveraging even a small amount of labeled data. The pretrained TabPFN baseline also achieves competitive performance, underscoring the transfer capability of tabular foundation models. Our proposed TabPFN-AD achieves an average AUPRC of 0.847 and an average rank of 1.53, corresponding to a relative improvement of 13.4% over TabPFN. Moreover, TabPFN-AD ranks first on 12 out of 17 datasets, demonstrating consistent advantages across diverse datasets. We further conduct paired Wilcoxon signed-rank tests on per-dataset AUPRC results, with detailed results reported in Table 10 in the appendix; the improvement over TabPFN remains statistically significant after correction.

4.3 Robustness against Contamination Ratio and Number of Labeled Anomalies

We evaluated model robustness by varying the unlabeled data contamination ratio and the number of labeled anomalies. As shown in Figure 3(a), most methods degrade as the contamination ratio increases from 0.02 to 0.1. Notably, LimiX and TabPFN, along with semi-supervised methods DeepSAD and XGBoost, suffer significant performance drops. In contrast, TabPFN-AD and READ, which employ preference-oriented sampling for potential anomaly mining, demonstrate superior contamination resilience. However, while READ achieves efficient single-time training, it requires extensive hyperparameter tuning. As the contamination ratio rises, the performance gap between TabPFN-AD and other baselines widens, evidencing our superior robustness. Regarding the labeled anomaly utilization experiment, as shown in Figure 3(b), the performance of most methods improves as more labeled anomalies become available. TabPFN-AD significantly outperforms all baseline models even with only 5 labeled instances, achieving the largest performance margin among all methods in this extremely scarce label scenario. Specifically, with 5 labeled anomalies, TabPFN-AD attains an AUPRC of 0.7968, outperforming TabPFN by 16.6%. This consistent advantage stems from effectively mining and utilizing potential anomalies in unlabeled data.

4.4 Ablation Study and Parameter Sensitivity

Component Ablation Table 3 presents the ablation results for the instance-level anomaly assessment s_I and group-level anomaly assessment s_G across all datasets. For s_I, we observe that each individual component consistently improves upon the original TabPFN baseline. This stems from inter-instance attention and iterative pseudo-label refinement, which effectively improves the utiliza-

tion of unlabeled data. Furthermore, the best performance is achieved when all three components are used together, confirming their synergistic complementarity.

For s_G, introducing it improved the average AUPRC by 1.90%. Since most datasets have only binary labels without fine-grained anomaly categories, we conducted supplementary experiments on the NB15 dataset with multiple anomaly patterns to verify its effectiveness. As shown in the right panel of Figure 3(c), s_G consistently identifies anomalous subgroups across iterations, with a high proportion of unknown anomalies in these subgroups. In the left panel, we compare the composition of instances selected by s_I in TabPFN-AD and its variant without s_G. Notably, the integration of s_G enhances the iterative pseudo-label updating process, which in turn enables s_I to better detect unknown anomalies. This demonstrates the synergistic interaction between the two scores: s_G discovers novel anomaly patterns, while the refined pseudo-labels improve the overall coverage of diverse anomalies. A detailed ablation on the attention-head selection criteria C_1 and C_2 is provided in Table 5 in the appendix.

Parameter Sensitivity In addition to component ablations, we analyze the sensitivity of TabPFN-AD to α, β, the initial candidate ratio $|\mathcal{D}_n|/|\mathcal{D}_s|$, the coupled scaling of τ_k and λ, and the attention threshold τ_a. Due to space limitations, we report the full results in Appendix C.3. Overall, TabPFN-AD performs stably across these parameter settings, showing that the proposed method is not overly sensitive to hyperparameter choices.

Additional Analyses We further compare TabPFN-AD with naive iterative refinement, showing that its gains are not solely attributable to iterative pseudo-label augmentation (Appendix C.4). We also observe only minor performance differences between using the sampled subset and the full dataset (Appendix C.5). Finally, TabPFN-AD remains computationally efficient by using a frozen backbone and a fixed configuration across datasets (Appendix C.6).

4.5 Generalization across TabPFN Versions

All preceding experiments use the TabPFN V2 model as the default backbone model. To evaluate whether the proposed framework generalizes to an improved TabPFN checkpoint, we further apply TabPFN-AD to TabPFN V3 without modifying the anomaly-mining or pseudo-label refinement strategy. Table 4 reports results on the five datasets showing the largest AUPRC improvements from TabPFN V2 to V3, while complete AUPRC and AUROC results across all 17 datasets are provided in Appendix C.7. The results show that TabPFN-AD is compatible with both TabPFN V2 and V3 models and can benefit from the improvements of the backbones while retaining the benefits of its anomaly-mining strategy.

Table 4: AUPRC comparison on the five datasets with the largest performance improvements from TabPFN V2 to V3. The last row reports the average over all 17 datasets.

Dataset	TabPFN		TabPFN-AD	
	V2	V3	V2	V3
LetterRecognition	0.417	0.847	0.584	0.862
ForestCover	0.599	0.960	0.993	0.998
Census	0.146	0.453	0.464	0.523
Credit	0.456	0.702	0.724	0.771
Nuclear	0.688	0.852	0.687	0.918
Avg.	0.747	0.854	0.847	0.888

5 Conclusion

In this paper, we propose TabPFN-AD, a framework that mines potential anomalies from unlabeled data to iteratively construct a more accurate in-context set for semi-supervised tabular anomaly detection. By leveraging instance-level and group-level scoring mechanisms, TabPFN-AD effectively mines potential anomalies from unlabeled data and iteratively refines pseudo-labels, addressing the challenges of limited labeled anomalies and unknown anomaly types. Extensive experiments on 17 benchmark datasets demonstrate that TabPFN-AD achieves state-of-the-art performance and exhibits strong robustness under high contamination ratios and scarce labeled anomalies.

Despite these promising results, iterative pseudo-label refinement introduces additional adaptation overhead, and the framework's generality across backbone architectures has thus far been evaluated only using different versions of TabPFN. Future work will extend the framework to other tabular foundation models and improve its computational efficiency.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improving the clarity of the manuscript. All technical content, experimental design, results, and conclusions were developed and verified by the authors.

REPRODUCIBILITY STATEMENT

We provide the implementation of TabPFN-AD, detailed experimental settings, and scripts for reproducing the experiments. Due to data redistribution and storage considerations, the repository includes one example dataset, while the remaining datasets are publicly available from their original sources. Their sources, data splits, and preprocessing procedures are documented in the appendix. The code is available at `https://anonymous.4open.science/r/TabPFN-AD-91DD`

REFERENCES

Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *The Eighth International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1lK_lBtvS.

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

Chiao-Ting Chen, Chi Lee, Szu-Hao Huang, and Wen-Chih Peng. Credit card fraud detection via intelligent sampling and self-supervised learning. *ACM Transactions on Intelligent Systems and Technology*, 15(2):35:1–35:29, 2024.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Felix den Breejen, Sangmin Bae, Stephen Cha, and Se-Young Yun. Fine-tuned in-context learning transformers are excellent tabular data classifiers. *arXiv:2405.13396*, 2024. URL https://arxiv.org/abs/2405.13396.

Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. DROCC: Deep robust one-class classification. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3711–3721, 2020.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1932–1940, 2024.

Zelin He, Sarah Alnegheimish, and Matthew Reimherr. Harnessing vision-language models for time series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21690–21698, 2026.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Daren Kuwaye and Hyunkeun Ryan Cho. Anomaly detection and correction in dense functional data within electronic medical records. *Statistics in Medicine*, 43(24):4768–4777, 2024.

Aodong Li, Yunhan Zhao, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Anomaly detection of tabular data using LLMs. *arXiv:2406.16308*, 2024. URL https://arxiv.org/abs/2406.16308.

Boyang Liu, Ding Wang, Kaixiang Lin, Pang-Ning Tan, and Jiayu Zhou. RCA: A deep collaborative autoencoder approach for anomaly detection. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pp. 1505–1511, 2021.

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.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Large language models can deliver accurate and interpretable time series anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 4623–4634, 2025.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 172692–172722, 2025. doi: 10.52202/085713-5748.

Nour Moustafa and Jill Slay. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6, 2015.

Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2041–2050, 2018.

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

Guansong Pang, Anton Van Den Hengel, Chunhua Shen, and Longbing Cao. Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1298–1308, 2021.

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

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 50817–50847, 2025.

Philipp Röchner and Franz Rothlauf. Unsupervised anomaly detection of implausible electronic health records: A real-world evaluation in cancer registries. *BMC Medical Research Methodology*, 23(1):125, 2023.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4393–4402, 2018.

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

Hongzhe Shou, Guanyu Lu, Martin Pavlovski, and Fang Zhou. READ: Robust and efficient anomaly detection under data contamination and limited supervision. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 2586–2596, 2025.

David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54(1): 45–66, 2004.

Che-Ping Tsai, Ganyu Teng, Phillip Wallis, and Wei Ding. AnoLLM: Large language models for tabular anomaly detection. In *International Conference on Learning Representations*, volume 2025, pp. 7562–7592, 2025.

Hu Wang, Guansong Pang, Chunhua Shen, and Congbo Ma. Unsupervised representation learning by predicting random distances. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, pp. 2950–2956, 2020.

Rulan Wei, Zewei He, Martin Pavlovski, and Fang Zhou. GAD: A generalized framework for anomaly detection at different risk levels. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2513–2522, 2024.

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

Hongzuo Xu, Yijie Wang, Guansong Pang, Songlei Jian, Ning Liu, and Yongjun Wang. RoSAS: Deep semi-supervised anomaly detection with contamination-resilient continuous supervision. *Information Processing & Management*, 60(5):103459, 2023b.

Hongzuo Xu, Yijie Wang, Juhui Wei, Songlei Jian, Yizhou Li, and Ning Liu. Fascinating supervisory signals and where to find them: Deep anomaly detection with scale learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 38655–38673, 2023c.

Hangting Ye, He Zhao, Wei Fan, Mingyuan Zhou, Dandan Guo, and Yi Chang. DRL: Decomposed representation learning for tabular anomaly detection. In *International Conference on Learning Representations*, volume 2025, pp. 24554–24589, 2025a.

Jianan Ye, Zhaorui Tan, Yijie Hu, Xi Yang, Guangliang Cheng, and Kaizhu Huang. Disentangling tabular data towards better one-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13061–13068, 2025b.

Yuhua Yin, Julian Jang-Jaccard, Wen Xu, Amardeep Singh, Jinting Zhu, Fariza Sabrina, and Jin Kwak. IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. *Journal of Big Data*, 10(1):15, 2023.

Yuchen Zeng, Tuan Dinh, Wonjun Kang, and Andreas C. Mueller. TabFlex: Scaling tabular learning to millions with linear attention. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 74051–74079, 2025.

Jiangning Zhang, Haoyang He, Xuhai Chen, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. GPT-4V-AD: Exploring grounding potential of VQA-oriented GPT-4V for zero-shot anomaly detection. In *Human Activity Recognition and Anomaly Detection — 4th International Workshop, DL-HAR 2024, and First International Workshop, ADFM 2024, Held in Conjunction with IJCAI 2024, Jeju, South Korea, August 3–9, 2024, Revised Selected Papers*, pp. 3–16, 2025a.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, et al. Limix: Unleashing structured-data modeling capability for generalist intelligence. *arXiv:2509.03505*, 2025b. URL `https://arxiv.org/abs/2509.03505`.

Yan-Feng Zhang, Hong-Liang Lu, Hong-Fan Lin, Xue-Chen Qiao, and Hao Zheng. The optimized anomaly detection models based on an approach of dealing with imbalanced dataset for credit card fraud detection. *Mobile Information Systems*, 2022(1):8027903, 2022.

Yingjie Zhou, Xucheng Song, Yanru Zhang, Fanxing Liu, Ce Zhu, and Lingqiao Liu. Feature encoding with autoencoders for weakly supervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2454–2465, 2021.

A Additional Methodological Details

A.1 Analysis of Attention-Head Selection

Selection Criteria Different attention heads in TabPFN capture different types of inter-instance relationships, and not all of them provide informative associations for anomaly detection. We therefore apply two complementary criteria, denoted by C_1 and C_2, to identify attention heads with reliable anomaly-related and local structural patterns.

For computational efficiency, a subset $\mathcal{D}_s$ is randomly sampled from the training set $\mathcal{D}$ while retaining all available labeled anomalies, i.e., $\mathcal{D}_L \subseteq \mathcal{D}_s$. For each attention head $h \in \mathcal{H}$, we extract its instance-level attention matrix $\mathcal{A}_s^h \in \mathbb{R}^{|\mathcal{D}_s| \times |\mathcal{D}_s|}$, where a_{ij}^h denotes the attention weight from instance $\boldsymbol{x}_i$ to instance $\boldsymbol{x}_j$.

The first criterion $C_1(h)$, termed *labeled-anomaly relevance*, evaluates whether a head captures meaningful associations among the available labeled anomalies. For each instance $\boldsymbol{x}_i \in \mathcal{D}_s$, we compute its total attention toward the labeled anomalies as $A_{\text{ano}}^h(i) = \sum_{\boldsymbol{x}_j \in \mathcal{D}_L} a_{ij}^h$. Let $r_h(i)$ denote the rank of $\boldsymbol{x}_i$ when all instances in $\mathcal{D}_s$ are sorted in descending order according to $A_{\text{ano}}^h(i)$. Criterion $C_1(h)$ is defined as

$$C_1(h) = \mathbb{I}\left[\max_{\boldsymbol{x}_i \in \mathcal{D}_L} r_h(i) \leq \left\lceil \frac{|\mathcal{D}_s|}{2} \right\rceil \ \wedge \ \min_{\boldsymbol{x}_i \in \mathcal{D}_L} r_h(i) \leq |\mathcal{D}_L|\right], \tag{6}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. The first condition requires all labeled anomalies to appear in the upper half of the ranking, ensuring consistently strong anomaly-related associations. The second requires at least one labeled anomaly to appear among the top $|\mathcal{D}_L|$ positions, preventing the selection of heads whose anomaly-attention scores are only marginally above the median.

The second criterion $C_2(h)$, termed *self similarity*, uses diagonal attention as an empirical indicator of whether a head preserves local similarity structures. Under uniform attention, each entry has an expected weight of $1/|\mathcal{D}_s|$. We therefore define $C_2(h)$ as

$$C_2(h) = \mathbb{I}\left[a_{ii}^h > \frac{1}{|\mathcal{D}_s|}, \quad \forall\, \boldsymbol{x}_i \in \mathcal{D}_s\right], \tag{7}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. This condition retains heads whose attention is consistently more concentrated on self-similar instances than would be expected under uniform attention.

The final set of selected heads is obtained by requiring both criteria:

$$\mathcal{H}_{\text{valid}} = \{h \in \mathcal{H} \mid C_1(h) \wedge C_2(h)\}. \tag{8}$$

Qualitative Visualization To illustrate why attention-head selection is necessary, we visualize the instance-level attention matrices of three representative heads from the last instance-level attention block of TabPFN in Figure 4. The visualization contains 10 labeled anomalies, 10 unlabeled anomalies, and 20 unlabeled normal instances. Ground-truth labels are used only to organize and interpret this post-hoc visualization; they are not used by the attention-head selection procedure.

As shown in Figure 4(a), labeled and unlabeled anomalies exhibit strong mutual attention, while normal instances form a distinct attention pattern. Consequently, the unlabeled anomalies attend more strongly to the labeled anomalies than the normal instances do, making this head informative for potential anomaly identification. In contrast, the heads shown in Figures 4(b) and 4(c) do not exhibit the same class-consistent structure. In these heads, normal instances may assign stronger attention to labeled anomalies than to instances within their own group, producing potentially misleading associations. This comparison motivates filtering the available attention heads before using their inter-instance relationships for anomaly mining.

Consistency across Datasets In the reported experiments, attention-head selection is performed independently for each dataset using its corresponding sampled training subset and available labeled anomalies. Despite these dataset-specific inputs, all 17 datasets selected the same head set. Importantly, this consistency is observed after independently applying the selection procedure and is not assumed in advance when designing the experimental protocol.

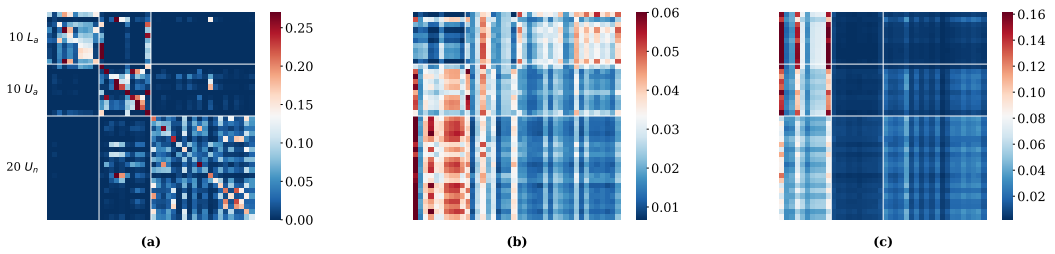

Figure 4: Instance-level attention heatmaps of three representative heads from the last instance-level attention block of TabPFN. The 40 instances are ordered as 10 labeled anomalies (L_a), 10 unlabeled anomalies (U_a), and 20 unlabeled normal instances (U_n). Ground-truth labels are used only to organize and interpret this post-hoc visualization.

Table 5: Ablation of the attention-head selection criteria C_1 and C_2. Results are averaged over all 17 datasets and obtained with the same sampling seed.

Setting	AUPRC	AUROC
w/o C_1, w/o C_2	0.768	0.922
w/ C_1, w/o C_2	0.798	0.946
w/o C_1, w/ C_2	<u>0.837</u>	**0.958**
w/ C_1, w/ C_2	**0.858**	**0.958**

This observation suggests that the two criteria primarily identify stable attention behavior of the fixed pretrained TabPFN backbone rather than highly dataset-specific patterns. It also indicates that, in practical deployment with the same backbone, the selected head set may potentially be determined once and subsequently reused. Nevertheless, the consistency observed on the 17 evaluated datasets does not guarantee that exactly the same heads will be selected for every unseen dataset.

Ablation of Selection Criteria We further evaluated the individual contributions of C_1 and C_2. Criterion C_1 evaluates whether a head captures associations involving the available labeled anomalies, whereas C_2 uses self-attention concentration as an empirical indicator of local similarity preservation. As shown in Table 5, introducing C_1 or C_2 individually improves the average AUPRC over the unfiltered setting by 0.030 and 0.069, respectively, indicating that both criteria help identify informative attention heads. Combining the two criteria yields the best performance, improving the average AUPRC by 0.090 over the unfiltered setting and by 0.021 over using C_2 alone. These results show that C_2 provides the stronger individual contribution, while C_1 offers complementary anomaly-related information.

A.2 Adaptive Update of τ_k and λ

The out-degree threshold τ_k and scaling parameter λ are adaptively updated during iterative refinement. At iteration t, the out-degree threshold is defined as

$$\tau_k^t = \text{Quantile}_{\rho_t}\left(\{k_i^t\}_{\boldsymbol{x}_i \in \mathcal{D}_s}\right), \tag{9}$$

where ρ_t denotes the proportion of low-out-degree instances retained as candidates. Initially, $\rho_0 = 50\%$. For each subsequent iteration, it is updated according to the change observed in the preceding iteration:

$$\rho_t = \rho_{t-1}\left(1 - \frac{\Delta_{t-1}}{|\mathcal{D}_n^{t-1}|}\right), \tag{10}$$

where Δ_{t-1} denotes the number of instances whose pseudo-labels change during iteration $t-1$, and $|\mathcal{D}_n^{t-1}|$ is the size of the corresponding potential anomaly set. The resulting ρ_t is then used to determine τ_k^t from the current out-degree distribution.

Meanwhile, the scaling parameter is computed based on the global in-degree distribution as

$$\lambda_t = \text{Quantile}_{p_t}\left(\left\{d_{\text{in}}^{\mathcal{D}_s,t}(\boldsymbol{x}_i)\right\}_{\boldsymbol{x}_i \in \mathcal{D}_s}\right), \tag{11}$$

where p_t and ρ_t satisfy

$$p_t \rho_t = p_0 \rho_0, \qquad p_0 = 10\%, \quad \rho_0 = 50\%. \tag{12}$$

Thus, their product is fixed at 5%. As refinement proceeds, a decreasing ρ_t imposes a stricter out-degree threshold and increases the anomaly concentration in the candidate set $\mathcal{D}_c$. Accordingly, p_t increases to relax the in-degree scaling and avoid excessively suppressing potential anomalies with relatively large in-degrees.

A.3 Theoretical Derivation of the Group-Level Score

This section derives a probabilistic sufficient condition under which an anomalous subgroup has a higher group-level score than a normal subgroup, where the score measures normalized within-subgroup attention concentration in the sparse attention graph.

A.3.1 Problem Setup

Consider an anomalous subgroup $\mathcal{G}_a$ and a normal subgroup $\mathcal{G}_n$ in the sparse attention graph $\mathcal{A}_s$. Let $|\mathcal{G}_a| = c_a + p_a$ and $|\mathcal{G}_n| = c_n + p_n$, where c_a, c_n and p_a, p_n denote the numbers of core and peripheral nodes, respectively. The group-level score, which measures normalized within-subgroup attention concentration, is defined as

$$s_G(\mathcal{G}) = \frac{1}{|\mathcal{G}|^2} \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{G}} a_{ij}. \tag{13}$$

For each core node $\boldsymbol{x}_i$, let its outgoing attention weights be sorted in descending order as $a_{i,(1)} \geq a_{i,(2)} \geq \cdots \geq a_{i,(n)}$. The sparse graph retains the minimum number of outgoing edges whose cumulative attention reaches the threshold τ_a:

$$k_i = \min \left\{ k : \sum_{\ell=1}^{k} a_{i,(\ell)} \geq \tau_a \right\}. \tag{14}$$

Because the complete last edge is retained rather than clipped, the total retained outgoing attention of core node i is $S_i = \sum_{\ell=1}^{k_i} a_{i,(\ell)} = \tau_a + \omega_i$, where $0 \leq \omega_i < a_{i,(k_i)}$. For subgroup $\mathcal{G}_g$, $g \in \{a, n\}$, define its total core overshoot as $\Omega_g = \sum_{i \in \mathcal{C}_g} \omega_i$. By construction of the induced subgroup, all retained outgoing edges of its core nodes are contained in $\mathcal{G}_g$. Their exact contribution to the internal attention is therefore $c_g \tau_a + \Omega_g$.

For each peripheral node $\boldsymbol{x}_i$, let

$$A_i = \sum_{\boldsymbol{x}_j \in \mathcal{G}_g} a_{ij}$$

denote its total attention contribution within its subgroup. We assume that the peripheral attention sums are independent and identically distributed within each subgroup, with

$$A_i^{(a)} \sim \mathcal{N}(\mu_a, \sigma_a^2), \qquad A_i^{(n)} \sim \mathcal{N}(\mu_n, \sigma_n^2).$$

For $\delta \in (0, 1)$, define

$$\widetilde{p}_a = p_a - \frac{\sqrt{2 p_a \sigma_a^2 \log(1/\delta)}}{\mu_a}, \qquad \widetilde{p}_n = p_n + \frac{\sqrt{2 p_n \sigma_n^2 \log(1/\delta)}}{\mu_n}.$$

We assume that $p_a, p_n > 0$, $\mu_a, \mu_n > 0$, and $\widetilde{p}_a > 0$.

Case 2: $|\mathcal{G}_a| > |\mathcal{G}_n|$. Suppose that the anomalous subgroup is larger than the normal subgroup, and define

$$\rho = \frac{|\mathcal{G}_a|^2}{|\mathcal{G}_n|^2} = \frac{(c_a + p_a)^2}{(c_n + p_n)^2} > 1.$$

Conditioned on the realized core overshoots Ω_a and Ω_n, if

$$\mu_a > \frac{(\rho c_n - c_a)\tau_a + \rho \Omega_n - \Omega_a}{\widetilde{p}_a} + \frac{\rho \widetilde{p}_n}{\widetilde{p}_a} \mu_n, \tag{15}$$

then

$$s_G(\mathcal{G}_a) > s_G(\mathcal{G}_n)$$

with probability at least $1 - 2\delta$, where the probability is taken over the peripheral attention sums.

A.3.2 DETAILED DERIVATION

For $g \in \{a, n\}$, the group-level score can be decomposed into the contributions of its core and peripheral nodes:

$$s_G(\mathcal{G}_g) = \frac{c_g\tau_a + \Omega_g + \sum_{i\in\mathcal{P}_g} A_i^{(g)}}{(c_g + p_g)^2}. \tag{16}$$

Since the sum of the p_a anomalous peripheral contributions follows $\mathcal{N}(p_a\mu_a, p_a\sigma_a^2)$, the Gaussian tail bound gives

$$\sum_{i\in\mathcal{P}_a} A_i^{(a)} \geq p_a\mu_a - \sqrt{2p_a\sigma_a^2\log(1/\delta)} = \widetilde{p}_a\mu_a \tag{17}$$

with probability at least $1-\delta$. Similarly,

$$\sum_{i\in\mathcal{P}_n} A_i^{(n)} \leq p_n\mu_n + \sqrt{2p_n\sigma_n^2\log(1/\delta)} = \widetilde{p}_n\mu_n \tag{18}$$

with probability at least $1-\delta$.

By the union bound, both events hold simultaneously with probability at least $1-2\delta$. Substituting them into Eq. equation 16 yields

$$s_G(\mathcal{G}_a) \geq \frac{c_a\tau_a + \Omega_a + \widetilde{p}_a\mu_a}{(c_a + p_a)^2}, \qquad s_G(\mathcal{G}_n) \leq \frac{c_n\tau_a + \Omega_n + \widetilde{p}_n\mu_n}{(c_n + p_n)^2}.$$

A sufficient condition for $s_G(\mathcal{G}_a) > s_G(\mathcal{G}_n)$ is therefore

$$\frac{c_a\tau_a + \Omega_a + \widetilde{p}_a\mu_a}{(c_a + p_a)^2} > \frac{c_n\tau_a + \Omega_n + \widetilde{p}_n\mu_n}{(c_n + p_n)^2}.$$

Using $\rho = (c_a + p_a)^2/(c_n + p_n)^2$, this inequality is equivalent to

$$c_a\tau_a + \Omega_a + \widetilde{p}_a\mu_a > \rho\left(c_n\tau_a + \Omega_n + \widetilde{p}_n\mu_n\right).$$

Rearranging gives

$$\widetilde{p}_a\mu_a > (\rho c_n - c_a)\tau_a + \rho\Omega_n - \Omega_a + \rho\widetilde{p}_n\mu_n.$$

Finally, since $\widetilde{p}_a > 0$, division by $\widetilde{p}_a$ yields

$$\mu_a > \frac{(\rho c_n - c_a)\tau_a + \rho\Omega_n - \Omega_a}{\widetilde{p}_a} + \frac{\rho\widetilde{p}_n}{\widetilde{p}_a}\mu_n,$$

which is the sufficient condition stated in Eq. equation 15.

A.3.3 POST-HOC AUDIT OF GROUP-LEVEL MINING

We conduct a post-hoc audit to examine the behavior of the group-level anomaly assessment procedure. This audit is performed only after the experiments: ground-truth labels are used to evaluate and categorize candidate subgroups and their core sets, but are not used in subgroup construction, group scoring, pseudo-label updating, hyperparameter selection, or prediction.

Across the 17 datasets with the same sampling seed, the method constructs 237 candidate subgroups in total. At each group-level mining event, the candidate subgroup with the highest group-level score s_G is selected, and its core instances are added to the potential anomaly set. This procedure results in 33 selected subgroups across all datasets and refinement iterations.

We first evaluate whether the selected subgroups are enriched with true anomalies. For each subgroup, we compute its anomaly fraction, defined as the proportion of ground-truth anomalies among its instances. As shown in Table 6, the mean anomaly fraction increases from 46.3% over all candidate subgroups to 83.4% over the selected highest-s_G subgroups. This substantial increase indicates that ranking candidates by s_G effectively prioritizes subgroups with stronger anomaly concentration.

We further evaluate the anomaly purity of the instances actually added by the group-level component. Among the 96 core-set instances added across all group-level mining events, 89 are ground-truth

Table 6: Post-hoc audit of group-level anomaly assessment on 17 datasets with the same sampling seed. Cases 1 and 2 correspond to $|\mathcal{G}_a| \leq |\mathcal{G}_n|$ and $|\mathcal{G}_a| > |\mathcal{G}_n|$, respectively. Pairwise separation is evaluated over candidate subgroups independently of the subgroup selected by the algorithm.

Audit statistic	Result
Total number of candidate subgroups	237
Mean anomaly fraction among all candidate subgroups	46.3%
Mean anomaly fraction among selected subgroups	83.4%
True-anomaly precision among selected core instances	89/96 (92.7%)
Case 1 pairs satisfying $s_G(\mathcal{G}_a) > s_G(\mathcal{G}_n)$	41/41
Case 2 pairs satisfying $s_G(\mathcal{G}_a) > s_G(\mathcal{G}_n)$	11/15

anomalies, corresponding to an anomaly precision of 92.7%. This result provides direct evidence that the core sets selected for pseudo-label updating predominantly contain true anomalies. Note that ground-truth labels are used only to compute this post-hoc statistic and do not affect subgroup or instance selection.

We next examine whether s_G assigns higher scores to anomalous subgroups than to normal subgroups. This pairwise analysis is conducted over the candidate subgroup pool and is independent of which subgroup is ultimately selected by the algorithm. Using ground-truth labels only for this post-hoc analysis, a candidate subgroup is categorized as anomalous if more than half of its instances are anomalies, whereas a candidate subgroup is categorized as normal only if it contains no anomaly.

To control for differences in the refinement context, anomalous and normal candidate subgroups are paired only when they originate from the same dataset, refinement iteration, and attention head. A candidate subgroup may participate in multiple eligible pairs, producing 56 subgroup pairs in total.

We divide these pairs according to their relative subgroup sizes. Case 1 contains pairs in which the anomalous subgroup is not larger than the normal subgroup,

$$|\mathcal{G}_a| \leq |\mathcal{G}_n|,$$

whereas Case 2 contains pairs in which the anomalous subgroup is larger than the normal subgroup,

$$|\mathcal{G}_a| > |\mathcal{G}_n|.$$

The 56 pairs consist of 41 Case 1 pairs and 15 Case 2 pairs. For each pair, s_G is considered to achieve the expected separation when

$$s_G(\mathcal{G}_a) > s_G(\mathcal{G}_n).$$

Here, $s_G(\cdot)$ denotes the normalized within-subgroup attention concentration used by TabPFN-AD. As shown in Table 6, all 41 Case 1 pairs satisfy the expected separation. This result suggests that when the anomalous subgroup is not larger than the normal subgroup, its internal attention concentration is sufficient to produce a higher normalized group score.

Case 2 is more challenging because the anomalous subgroup has a larger normalization denominator in Eq. equation 3. In this regime, 11 of the 15 pairs satisfy the expected separation. The remaining four pairs show that an anomalous subgroup does not necessarily receive a higher s_G than every normal subgroup. This observation is consistent with the theoretical analysis, which provides a sufficient rather than necessary condition for Case 2 separation. The Case 2 result should therefore be interpreted as a conditional characterization rather than a universal guarantee.

A.4 Base pseudo-label assignment

At iteration t, instances in $\mathcal{D}_n^t$ are sorted in ascending order according to their instance-level scores s_I. They are then partitioned into M ordered levels according to the unique values of s_I, such that instances with the same score are assigned to the same level. Let $l(\boldsymbol{x}_i) \in \{1, \ldots, M\}$ denote the level index of $\boldsymbol{x}_i$, where a smaller index corresponds to a smaller instance-level score and hence a higher degree of anomalousness. Its base pseudo-label is defined as

$$\hat{s}_t(\boldsymbol{x}_i) = 0.9 - 0.4 * \frac{l(\boldsymbol{x}_i) - 1}{M - 1}. \tag{19}$$

Thus, the base pseudo-labels decrease linearly from 0.9 to 0.5 as the level index increases.

A.5 Stopping Strategy for Pseudo-Label Refinement

To prevent early pseudo-label errors from being repeatedly reinforced, we adopt a label-free stopping strategy based on two conditions: the stability of the mined anomaly set and a maximum number of refinement iterations. First, let $\mathcal{D}_n^t$ and $\mathcal{D}_n^{t-1}$ denote the potential anomaly sets obtained in two consecutive iterations. Refinement is terminated when their symmetric difference satisfies

$$\left|\mathcal{D}_n^t \setminus \mathcal{D}_n^{t-1}\right| + \left|\mathcal{D}_n^{t-1} \setminus \mathcal{D}_n^t\right| \leq \epsilon_s, \qquad \epsilon_s = 2. \tag{20}$$

This condition indicates that the mined anomaly set has become stable. Second, the refinement process is limited to a maximum of five iterations. Together with temporal smoothing and pseudo-label decay, these stopping conditions mitigate the accumulation of unreliable pseudo-labels without requiring additional validation labels.

B Detailed Experimental Setup

B.1 Datasets

Table 7 summarizes the 17 benchmark datasets used in our experiments, comprising 12 datasets from the Outlier Detection DataSets (ODDS[1]) repository, selected for their sufficient anomaly counts to simulate anomaly contamination in unlabeled data, and 5 additional datasets: NB15[2], NSL-KDD[3], Census[4], Nuclear and Credit[5] from other public sources. For each dataset, we report the number of numerical and categorical features, along with the anomaly and normal instance counts in the training, validation, and test splits.

Table 7: Summary of the 17 benchmark datasets. The train/validation/test columns report anomaly and normal instance counts in *anomaly*/normal format.

Dataset	Number	Category	Train	Validation	Test
Annthyroid	6	0	*267*/4883	*107*/963	*160*/682
Breast	30	0	*23*/247	*20*/53	*20*/53
Cardio	21	0	*60*/950	*40*/360	*60*/337
Census	13	20	*6000*/113810	*2800*/28000	*3000*/30000
Credit	29	0	*344*/6346	*30*/10000	*99*/12000
ForestCover	10	0	*1373*/25897	*549*/4941	*824*/7416
Letter Recognition	32	0	*60*/950	*15*/135	*25*/225
Mammography	6	0	*130*/2280	*52*/468	*72*/702
Musk	166	0	*60*/950	*30*/270	*30*/270
NB15	39	3	*840*/15770	*767*/11904	*4001*/18001
NSL-KDD	38	3	*900*/16910	*707*/12302	*1200*/15376
Nuclear	12	0	*28*/342	*10*/50	*15*/100
Optdigits	64	0	*75*/1235	*30*/270	*27*/405
Pendigits	16	0	*75*/1235	*30*/270	*45*/405
Satellite	36	0	*160*/2850	*40*/360	*60*/540
Shuttle	9	0	*1500*/28310	*400*/3600	*800*/7200
Thyroid	6	0	*60*/950	*10*/90	*20*/180

B.2 Implementation Details

For TabPFN-AD, the key hyperparameters are set as follows: the attention threshold $\tau_a = 0.05 \times (t+1)$, smoothing term $\epsilon = 10^{-6}$, momentum coefficient $\alpha = 0.9$, and decay coefficient $\beta = 0.9$.

[1] http://odds.cs.stonybrook.edu/

[2] https://research.unsw.edu.au/projects/unsw-nb15-dataset

[3] https://www.unb.ca/cic/datasets/nsl.html

[4] https://archive.ics.uci.edu/ml/datasets/census+income

[5] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

C Additional Experimental Results

C.1 Anomaly Candidate Set Analysis

Figure 5 visualizes the composition of the anomaly candidate sets $\mathcal{D}_c$ constructed via out-degree thresholding across all 17 datasets. For each dataset, instances with out-degree $k \leq \tau_k$ are selected to form $\mathcal{D}_c$. The candidate sets achieve substantially higher anomaly proportions compared to the original unlabeled data, while preserving the majority of anomalies from the training subset $\mathcal{D}_s$. Nevertheless, normal instances remain present in $\mathcal{D}_c$, indicating that out-degree thresholding alone is insufficient to obtain a pure anomaly set.

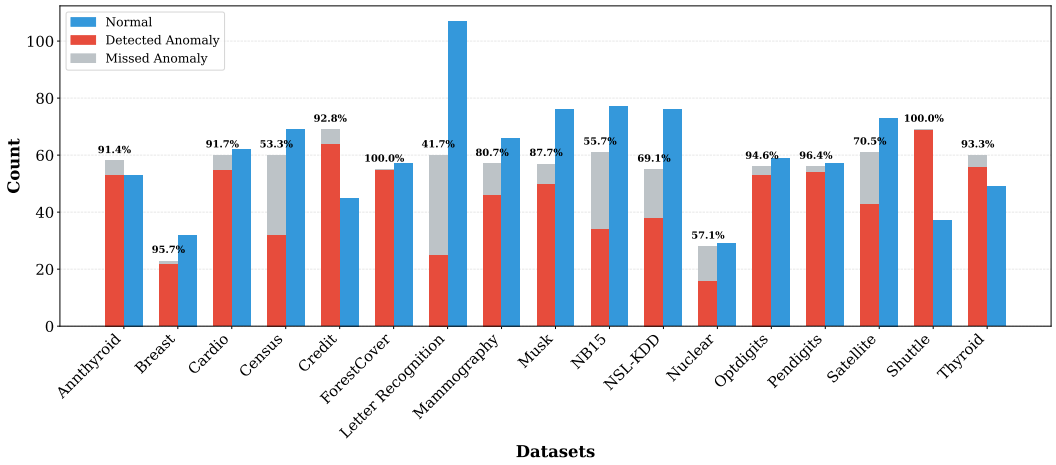

Figure 5: Distribution of normal instances (blue) and detected anomalies (red) within each dataset's anomaly candidate set $\mathcal{D}_c$; gray bars represent anomalies in training subset $\mathcal{D}_s$ that were undetected and thus excluded from $\mathcal{D}_c$. The percentage indicates the recall rate of anomalies in $\mathcal{D}_s$.

C.2 Detection Performance

This section presents the complete experimental results omitted from the main paper due to space constraints. Table 8 and Table 9 report the detailed AUPRC and AUROC performance (mean and standard deviation) of all methods across the 17 benchmark datasets, respectively. "OOM" indicates that the method encountered out-of-memory errors and could not complete the experiment on the corresponding dataset. To determine whether the gains of TabPFN-AD are consistent across datasets rather than being driven by a few favorable cases, we conduct two-sided paired Wilcoxon signed-rank tests on the per-dataset AUPRC results. As reported in Table 10, TabPFN-AD records 15/1/1 wins/ties/losses against LimiX and 14/1/2 against TabPFN. After Holm correction across the two comparisons, the adjusted p-values are 1.6×10^{-3} and 1.9×10^{-3}, respectively. Both improvements remain statistically significant at the 0.01 level, indicating that the advantage of TabPFN-AD is broadly consistent across the evaluated datasets.

C.3 Parameter Sensitivity

We evaluate the sensitivity of TabPFN-AD to four groups of hyperparameters and threshold configurations on all 17 benchmark datasets. Unless otherwise specified, the sampling seed is fixed to 0, and all other experimental settings remain unchanged. Results are reported as the mean test AUPRC and AUROC across all 17 datasets.

Sensitivity to α and β. We evaluate different combinations of the smoothing coefficient α and decay coefficient β. Relative to the default setting $(\alpha, \beta) = (0.90, 0.90)$, the best combination improves the mean AUPRC by only 0.003, while the mean AUROC remains unchanged. Across all evaluated combinations, the maximum decreases in AUPRC and AUROC are only 0.002 and 0.001,

Table 8: AUPRC performance (mean ± std.) of all methods across 17 benchmark datasets. Best results in each category are **bolded**; second-best are underlined.

	Unsupervised			Semi-Supervised							Foundation Models		
Dataset	**AnoLLM**	**DRL**	**Disent-AD**	**XGBoost**	**DeepSAD**	**DevNet**	**RoSAS**	**FeaWAD**	**PReNet**	**READ**	**Limix**	**TABPFN**	**Ours**
Annthyroid	0.631±0.002	0.428±0.024	0.673±0.034	0.815±0.039	0.742±0.051	0.726±0.035	0.851±0.039	0.640±0.040	0.769±0.071	0.824±0.089	0.923	0.914	**0.936**±0.002
Breast	0.725±0.012	0.537±0.055	0.670±0.006	0.954±0.034	0.965±0.021	0.996±0.004	**0.998**±0.004	0.979±0.011	**0.998**±0.004	0.990±0.023	**0.998**	0.987	0.992±0.013
Cardio	0.696±0.004	0.423±0.018	0.754±0.002	0.872±0.060	0.923±0.017	0.935±0.033	0.861±0.097	0.815±0.074	0.924±0.027	0.922±0.036	0.936	0.917	**0.937**±0.005
Census	0.135±0.001	0.231±0.046	0.310±0.012	0.305±0.054	0.396±0.061	0.426±0.069	0.377±0.081	0.396±0.075	0.416±0.071	0.389±0.077	0.315	0.146	**0.464**±0.011
Credit	0.404±0.001	0.387±0.133	0.526±0.091	0.424±0.103	0.242±0.075	0.686±0.036	0.679±0.048	0.690±0.042	0.680±0.062	0.651±0.090	0.681	0.456	**0.724**±0.005
ForestCover	0.298±0.003	0.148±0.024	0.729±0.106	0.871±0.025	0.968±0.015	0.974±0.007	0.946±0.019	0.562±0.130	0.980±0.008	0.985±0.020	0.924	0.599	**0.993**±0.002
Letter	0.142±0.002	0.438±0.063	0.123±0.004	0.326±0.052	0.474±0.094	0.309±0.080	0.412±0.087	0.186±0.046	0.390±0.098	0.347±0.077	0.500	0.417	**0.584**±0.023
Mammography	0.476±0.003	0.307±0.036	0.599±0.049	0.583±0.076	0.670±0.056	0.717±0.061	0.655±0.056	0.624±0.104	0.723±0.051	0.754±0.048	0.726	**0.805**	0.780±0.006
Musk	OOM	0.366±0.107	0.995±0.001	**1.000**±0.000	**1.000**±0.000	**1.000**±0.000	**1.000**±0.000	0.996±0.010	**1.000**±0.000	0.993±0.037	**1.000**	**1.000**	**1.000**±0.000
NB15	0.459±0.002	0.427±0.039	0.440±0.057	0.669±0.073	0.657±0.093	0.638±0.088	0.615±0.091	0.489±0.037	0.621±0.110	0.710±0.112	0.764	0.668	**0.817**±0.012
NSL-KDD	0.601±0.001	0.474±0.037	0.553±0.018	0.515±0.047	0.630±0.077	0.615±0.086	0.572±0.041	0.594±0.063	0.579±0.022	0.637±0.079	0.649	0.552	**0.745**±0.025
Nuclear	0.196±0.001	0.371±0.077	0.368±0.027	0.476±0.105	0.514±0.111	0.367±0.021	0.380±0.112	0.270±0.059	0.446±0.088	0.437±0.082	0.599	**0.688**	0.687±0.018
Optdigits	0.152±0.005	0.060±0.007	0.074±0.007	0.935±0.045	0.775±0.026	0.995±0.004	0.972±0.012	0.710±0.268	**0.998**±0.002	0.835±0.134	0.943	0.913	0.978±0.037
Pendigits	0.542±0.003	0.140±0.012	0.519±0.006	0.917±0.042	0.985±0.009	0.987±0.006	0.965±0.022	0.821±0.202	0.989±0.004	0.960±0.052	0.971	0.974	**0.991**±0.004
Satellite	0.647±0.001	0.494±0.033	0.642±0.013	0.576±0.089	0.717±0.017	0.689±0.044	0.683±0.050	0.545±0.094	0.704±0.035	0.730±0.054	0.784	0.763	**0.819**±0.012
Shuttle	0.943±0.001	0.735±0.173	0.931±0.009	0.914±0.025	0.954±0.058	0.991±0.001	0.955±0.061	0.924±0.030	0.642±0.112	**0.995**±0.009	0.990	0.976	0.994±0.005
Thyroid	0.781±0.001	0.534±0.027	0.537±0.006	0.899±0.016	0.545±0.079	0.868±0.027	0.810±0.045	0.847±0.046	0.900±0.020	0.925±0.037	0.901	0.916	**0.956**±0.004
Average	0.489	0.382	0.556	0.709	0.715	0.760	0.749	0.652	0.751	0.770	0.800	0.747	**0.847**

Table 9: AUROC performance (mean ± std.) of all methods across 17 benchmark datasets. Best results in each category are **bolded**; second-best are underlined.

	Unsupervised			Semi-Supervised							Foundation Models		
Dataset	**AnoLLM**	**DRL**	**Disent-AD**	**XGBoost**	**DeepSAD**	**DevNet**	**RoSAS**	**FeaWAD**	**PreNet**	**READ**	**Limix**	**TABPFN**	**Ours**
Annthyroid	0.872±0.001	0.748±0.023	0.897±0.024	0.943±0.025	0.930±0.027	0.830±0.052	0.930±0.031	0.828±0.022	0.914±0.044	0.919±0.038	0.984	0.985	**0.988**±0.001
Breast	0.890±0.003	0.782±0.028	0.856±0.002	0.964±0.029	0.983±0.012	**0.999**±0.001	**0.999**±0.002	0.991±0.005	**0.999**±0.002	0.980±0.025	0.996	0.994	0.991±0.007
Cardio	0.886±0.003	0.716±0.016	0.958±0.001	0.968±0.015	0.974±0.007	0.980±0.012	0.906±0.125	0.934±0.054	0.975±0.010	0.958±0.021	0.985	0.984	**0.989**±0.007
Census	0.693±0.001	0.634±0.064	0.780±0.021	0.784±0.028	0.851±0.024	0.861±0.026	0.806±0.052	0.849±0.046	0.852±0.023	0.746±0.076	0.819	0.716	**0.862**±0.001
Credit	0.941±0.000	0.922±0.014	0.946±0.007	0.878±0.047	0.881±0.023	0.936±0.012	0.881±0.016	0.932±0.018	0.913±0.024	0.906±0.030	0.949	0.947	**0.952**±0.015
ForestCover	0.842±0.002	0.647±0.051	0.968±0.017	0.951±0.018	0.995±0.003	0.997±0.001	0.990±0.005	0.897±0.102	0.998±0.001	0.992±0.008	0.992	0.808	**0.999**±0.001
Letter	0.590±0.002	0.856±0.021	0.539±0.008	0.736±0.055	0.848±0.032	0.694±0.086	0.708±0.050	0.585±0.053	0.692±0.070	0.684±0.096	0.881	0.759	**0.921**±0.003
Mammography	0.831±0.001	0.766±0.036	0.895±0.020	0.890±0.024	0.914±0.013	0.915±0.022	0.827±0.029	0.898±0.043	0.901±0.031	0.916±0.022	0.910	0.940	**0.943**±0.012
Musk	OOM	0.791±0.085	0.999±0.000	**1.000**±0.000	**1.000**±0.000	**1.000**±0.000	**1.000**±0.000	0.999±0.001	**1.000**±0.000	0.981±0.029	**1.000**	**1.000**	**1.000**±0.000
NB15	0.770±0.001	0.774±0.028	0.796±0.024	0.838±0.045	0.877±0.026	0.776±0.037	0.796±0.040	0.809±0.023	0.840±0.048	0.830±0.055	0.928	0.899	**0.931**±0.003
NSL-KDD	0.933±0.000	0.913±0.007	0.921±0.018	0.872±0.041	0.943±0.014	0.843±0.069	0.855±0.061	0.861±0.047	0.861±0.039	0.865±0.042	0.928	0.895	**0.947**±0.003
Nuclear	0.448±0.002	0.707±0.050	0.623±0.025	0.744±0.062	0.796±0.048	0.630±0.032	0.659±0.118	0.497±0.071	0.682±0.094	0.707±0.069	**0.869**	0.846	0.839±0.004
Optdigits	0.794±0.005	0.451±0.044	0.596±0.031	0.989±0.011	0.985±0.010	**1.000**±0.000	0.992±0.007	0.952±0.078	**1.000**±0.000	0.938±0.038	0.994	0.995	0.997±0.002
Pendigits	0.889±0.004	0.537±0.032	0.913±0.002	0.986±0.009	**0.998**±0.001	0.997±0.003	0.990±0.009	0.963±0.051	**0.998**±0.002	0.975±0.013	0.996	**0.998**	0.997±0.004
Satellite	0.803±0.001	0.828±0.007	0.806±0.034	0.864±0.029	0.889±0.014	0.853±0.017	0.862±0.038	0.703±0.026	0.858±0.020	0.856±0.042	0.910	0.904	**0.915**±0.020
Shuttle	0.997±0.000	0.952±0.071	0.996±0.001	0.927±0.030	0.990±0.019	0.995±0.001	0.972±0.038	0.991±0.005	0.725±0.110	0.990±0.003	**0.999**	0.999	**0.999**±0.001
Thyroid	0.980±0.000	0.947±0.009	0.930±0.008	0.986±0.006	0.908±0.035	0.986±0.002	0.962±0.021	0.977±0.007	0.986±0.004	0.975±0.027	0.994	0.991	**0.996**±0.006

respectively, indicating that TabPFN-AD is insensitive to these two coefficients within the evaluated ranges.

Sensitivity to the initial candidate ratio. We evaluate the initial candidate ratio $|\mathcal{D}_n|/|\mathcal{D}_s|$ over the range from 0.20 to 0.50. Relative to the default ratio of 0.30, the mean AUPRC and AUROC decrease by at most 0.009 and 0.004, respectively. This limited degradation indicates that TabPFN-AD remains stable across different initial candidate-set sizes.

Sensitivity to τ_k and λ. We jointly scale τ_k and λ using a common scaling factor s: Decreasing s from 1.0 to 0.8 leaves the mean AUPRC unchanged and decreases the mean AUROC by 0.003. Increasing s to 1.2 improves the mean AUPRC by 0.006 while decreasing the mean AUROC by only 0.001. These small changes demonstrate robustness to moderate joint variations in τ_k and λ.

$$\tau_k^{(s)} = s\tau_k, \qquad \lambda^{(s)} = s\lambda. \tag{21}$$

Sensitivity to the attention threshold τ_a. The main experiments use a single global attention-threshold schedule shared across all datasets, although τ_a can be configured separately for each dataset. To evaluate the potential benefit of dataset-specific selection, we consider

$$\tau_a(t) = \text{slope} \times t + \text{bias}, \tag{22}$$

where $\text{bias} \in \{0.1, 0.2, 0.3\}$ and $\text{slope} \in \{0.01, 0.05\}$. The dataset-specific configuration is selected through grid search on validation performance and is used only in this sensitivity analysis.

Dataset-specific selection improves the mean AUPRC and AUROC over the global configuration by only 0.009 and 0.002, respectively. Given these limited gains, all main experiments use the global configuration to avoid dataset-specific threshold tuning.

Table 10: Paired significance tests on per-dataset AUPRC results. Holm correction is applied across the two comparisons.

Comparison	**W/T/L**	W	**Raw** p	**Holm-adjusted** p
TabPFN-AD vs. LimiX	15/1/1	3	7.8×10^{-4}	1.6×10^{-3}
TabPFN-AD vs. TabPFN	14/1/2	8	1.9×10^{-3}	1.9×10^{-3}

Table 11: Sensitivity to α and β.

α	β	**AUPRC**	**AUROC**
0.90	0.90	0.858	**0.958**
0.90	0.99	**0.861**	**0.958**
0.99	0.90	0.857	0.957
0.99	0.99	0.856	0.957

C.4 Comparison with Naive Iterative Refinement

Since TabPFN-AD iteratively updates pseudo-labels, a natural question is whether similar gains can be achieved by simply adding the same iterative procedure to existing tabular foundation models. To examine this possibility, we construct naive iterative variants of TabPFN and LimiX. At each refinement round, the in-context training subset contains all available labeled anomalies together with randomly sampled unlabeled instances and is used to predict anomaly scores for the remaining unlabeled instances. The top 0.1% highest-scoring instances are then assigned anomaly pseudo-labels and added to the labeled anomaly set. This procedure is repeated for 10 rounds, with the test AUPRC recorded after each round.

Figure 6 reports the results across all 17 datasets. Naive iterative refinement improves TabPFN and LimiX only on a subset of datasets. Moreover, the best-performing refinement round varies considerably across datasets, revealing no consistent improvement pattern or universally reliable stopping point. In contrast, TabPFN-AD outperforms the naive iterative variants on almost all datasets. These results demonstrate that simple iterative pseudo-label augmentation is insufficient to reliably identify potential anomalies. The attention-guided instance-level and group-level scores in TabPFN-AD are therefore important for controlling pseudo-label quality during iterative refinement.

C.5 Full Dataset Utilization

The main experiments perform refined anomaly mining on a sampled training subset because the context size of TabPFN is limited. To investigate whether TabPFN-AD can also exploit information outside this subset, we conduct an additional experiment on five large-scale datasets. Potential anomalies are mined from the remaining training data and added to the labeled anomaly set. A balanced normal subset is then constructed through stratified sampling based on the predicted anomaly scores.

As shown in Table 15, extending TabPFN-AD to the full training set improves its average AUPRC from 0.8023 to 0.8162, corresponding to a relative improvement of approximately 1.7%. In contrast, directly applying TabPFN to the full training data results in lower performance, likely because the scarce labeled anomalies are overwhelmed by the severe class imbalance. LimiX benefits from its internal data-selection strategy on some datasets but remains less accurate than TabPFN-AD on

Table 12: Sensitivity to the initial candidate ratio $|\mathcal{D}_n|/|\mathcal{D}_s|$.

$\lvert\mathcal{D}_n\rvert/\lvert\mathcal{D}_s\rvert$	**AUPRC**	**AUROC**
0.20	0.854	0.956
0.30	**0.858**	**0.958**
0.40	0.849	0.956
0.50	0.853	0.954

Table 13: Sensitivity to the coupled scaling of τ_k and λ defined in Eq. equation 21.

Scaling factor s	AUPRC	AUROC
0.8	0.858	0.955
1.0	0.858	**0.958**
1.2	**0.864**	0.957

Table 14: Sensitivity to global and dataset-specific configurations of the attention threshold τ_a. Dataset-specific configurations are selected through grid search on validation performance and are evaluated only in this sensitivity analysis. All other experimental results use the fixed global configuration.

τ_a configuration	AUPRC	AUROC
Global	0.858	0.958
Dataset-specific	**0.867**	**0.960**
Absolute difference	+0.009	+0.002

average. These results indicate that TabPFN-AD can effectively exploit information beyond the initially sampled subset without directly exposing TabPFN to the full imbalanced training set.

Table 15: AUPRC under subset-based and full-data utilization settings on five large-scale datasets. "Subset" uses the initially sampled subset, whereas "All" additionally mines anomaly candidates from the remaining training data.

Dataset	**Subset** (1000, 5%)			**All**		
	LimiX	TabPFN	TabPFN-AD	LimiX	TabPFN	TabPFN-AD
NB15	0.7598	0.7052	**0.8308**	0.7636	0.6676	**0.8339**
NSL-KDD	0.5803	0.6609	**0.7335**	0.6491	0.5518	**0.7658**
Census	0.2629	0.3965	**0.4580**	0.3151	0.1461	**0.4902**
Shuttle	**0.9960**	0.9850	0.9948	0.9763	0.9899	**0.9993**
ForestCover	0.9693	0.9656	**0.9944**	0.9235	0.5986	**0.9920**
Average	0.7137	0.7426	**0.8023**	0.7255	0.5908	**0.8162**

C.6 Computational Efficiency

Table 16 reports the estimated total runtime per dataset after accounting for hyperparameter tuning, together with the average AUPRC across all 17 datasets. For methods using grid search, the reported runtime aggregates the fit/adaptation and prediction costs over all evaluated configurations. LimiX, TabPFN, and TabPFN-AD instead use one fixed configuration across all datasets.

TabPFN-AD achieves the highest average AUPRC with an estimated total runtime of 121.03 seconds. Compared with training-based deep baselines, it avoids gradient-based backbone optimization and extensive dataset-specific model selection, resulting in a lower total runtime than most of these methods while achieving substantially better performance. LimiX and TabPFN remain faster because they perform direct inference without data-dependent adaptation, whereas the additional cost of TabPFN-AD arises from iterative anomaly mining and pseudo-label refinement. Overall, the results show that TabPFN-AD provides a favorable trade-off between computational cost and detection performance.

C.7 Performance across TabPFN Versions

The main experiments use TabPFN V2 as the default backbone. To evaluate the compatibility of TabPFN-AD with newer TabPFN checkpoints, we further apply the complete framework to TabPFN V3 without modifying the proposed anomaly mining and pseudo-label refinement procedures. The



Figure 6: AUPRC over 10 iterations of naive iterative refinement across all 17 datasets. Solid lines denote iterative application of TabPFN (blue) and LimiX (orange), while dashed lines indicate the reference performance of TabPFN, LimiX, and TabPFN-AD.

Table 16: Estimated total runtime per dataset after accounting for hyperparameter tuning, together with the average AUPRC across all 17 datasets. Methods marked as fixed use one configuration across all datasets.

Category	Method	Total Runtime incl. Tuning (s)	Avg. AUPRC
Unsupervised	AnoLLM†	2554.12	0.4893
	DRL	25317.50	0.3821
	Disent-AD	42575.00	0.5555
Semi-supervised	XGBoost	13.89	0.7088
	DeepSAD	737.50	0.7149
	DevNet	1540.00	0.7599
	RoSAS	3702.50	0.7488
	FeaWAD	78.75	0.6522
	PReNet	8425.00	0.7505
	READ	687.59	0.7697
Foundation model	LimiX (fixed)	1.97	0.8002
	TabPFN (fixed)	43.88	0.7465
	TabPFN-AD (fixed)	121.03	**0.8469**

For methods using grid search, the estimated total runtime includes the fit/adaptation and prediction costs accumulated over 125 hyperparameter configurations. Fixed methods are evaluated using one configuration shared across all datasets. For TabPFN-AD, adaptation corresponds to iterative pseudo-label refinement, while the pretrained backbone remains frozen. †AnoLLM uses one fixed configuration because extensive hyperparameter search is computationally impractical.

attention-head selection procedure is performed independently for each dataset. Interestingly, it consistently selects the same head for both V2 and V3 across all 17 datasets.

Table 17: AUPRC performance of TabPFN and TabPFN-AD using V2 and V3 checkpoints across 17 datasets. Results are reported as mean ± standard deviation when repeated runs are available. † indicates that all ten runs produced identical results.

Dataset	TabPFN V2	TabPFN-AD V2	TabPFN V3	TabPFN-AD V3
Annthyroid	$0.914^{\dagger}$	0.936 ± 0.002	$0.922^{\dagger}$	0.931 ± 0.008
Breast	$0.987^{\dagger}$	0.992 ± 0.013	$0.988^{\dagger}$	0.998 ± 0.004
Cardio	$0.917^{\dagger}$	0.937 ± 0.005	$0.948^{\dagger}$	0.960 ± 0.007
Census	$0.146^{\dagger}$	0.464 ± 0.011	$0.453^{\dagger}$	0.523 ± 0.012
Credit	$0.456^{\dagger}$	0.724 ± 0.005	$0.702^{\dagger}$	0.771 ± 0.007
ForestCover	$0.599^{\dagger}$	0.993 ± 0.002	$0.960^{\dagger}$	0.998 ± 0.001
LetterRecognition	$0.417^{\dagger}$	0.584 ± 0.023	$0.847^{\dagger}$	0.862 ± 0.004
Mammography	$0.805^{\dagger}$	0.780 ± 0.006	$0.805^{\dagger}$	0.812 ± 0.017
Musk	$1.000^{\dagger}$	1.000 ± 0.000	$1.000^{\dagger}$	1.000 ± 0.000
NB15	$0.668^{\dagger}$	0.817 ± 0.012	$0.762^{\dagger}$	0.813 ± 0.008
NSL-KDD	$0.552^{\dagger}$	0.745 ± 0.025	$0.654^{\dagger}$	0.773 ± 0.005
Nuclear	$0.688^{\dagger}$	0.687 ± 0.018	$0.852^{\dagger}$	0.918 ± 0.002
Optdigits	$0.913^{\dagger}$	0.978 ± 0.037	$1.000^{\dagger}$	1.000 ± 0.000
Pendigits	$0.974^{\dagger}$	0.991 ± 0.004	$0.962^{\dagger}$	0.984 ± 0.007
Satellite	$0.763^{\dagger}$	0.819 ± 0.012	$0.782^{\dagger}$	0.847 ± 0.010
Shuttle	$0.976^{\dagger}$	0.994 ± 0.005	$0.951^{\dagger}$	0.978 ± 0.013
Thyroid	$0.916^{\dagger}$	0.956 ± 0.004	$0.927^{\dagger}$	0.928 ± 0.007
Average	$0.747^{\dagger}$	0.847	$0.854^{\dagger}$	$\mathbf{0.888}$

Table 18: AUROC performance of TabPFN and TabPFN-AD using V2 and V3 checkpoints across 17 datasets. Results are reported as mean ± standard deviation when repeated runs are available. † indicates that all ten runs produced identical results.

Dataset	TabPFN V2	TabPFN-AD V2	TabPFN V3	TabPFN-AD V3
Annthyroid	$0.985^{\dagger}$	0.988 ± 0.001	$0.988^{\dagger}$	0.989 ± 0.002
Breast	$0.994^{\dagger}$	0.991 ± 0.007	$0.999^{\dagger}$	0.999 ± 0.002
Cardio	$0.984^{\dagger}$	0.989 ± 0.007	$0.991^{\dagger}$	0.991 ± 0.001
Census	$0.716^{\dagger}$	0.862 ± 0.001	$0.883^{\dagger}$	0.900 ± 0.008
Credit	$0.947^{\dagger}$	0.952 ± 0.015	$0.962^{\dagger}$	0.930 ± 0.003
ForestCover	$0.808^{\dagger}$	0.999 ± 0.001	$0.998^{\dagger}$	0.999 ± 0.001
LetterRecognition	$0.759^{\dagger}$	0.921 ± 0.003	$0.969^{\dagger}$	0.973 ± 0.001
Mammography	$0.940^{\dagger}$	0.943 ± 0.012	$0.953^{\dagger}$	0.957 ± 0.003
Musk	$1.000^{\dagger}$	1.000 ± 0.000	$1.000^{\dagger}$	1.000 ± 0.000
NB15	$0.899^{\dagger}$	0.931 ± 0.003	$0.929^{\dagger}$	0.938 ± 0.002
NSL-KDD	$0.895^{\dagger}$	0.947 ± 0.003	$0.926^{\dagger}$	0.951 ± 0.001
Nuclear	$0.846^{\dagger}$	0.839 ± 0.004	$0.958^{\dagger}$	0.962 ± 0.005
Optdigits	$0.995^{\dagger}$	0.997 ± 0.002	$1.000^{\dagger}$	1.000 ± 0.000
Pendigits	$0.998^{\dagger}$	0.997 ± 0.004	$0.995^{\dagger}$	0.998 ± 0.001
Satellite	$0.904^{\dagger}$	0.915 ± 0.020	$0.9206^{\dagger}$	0.958 ± 0.003
Shuttle	$0.999^{\dagger}$	0.999 ± 0.001	$0.999^{\dagger}$	0.998 ± 0.001
Thyroid	$0.991^{\dagger}$	0.996 ± 0.006	$0.995^{\dagger}$	0.991 ± 0.008
Average	$0.921^{\dagger}$	0.957	$0.969^{\dagger}$	$\mathbf{0.973}$

Tables 17 and 18 present the complete per-dataset AUPRC and AUROC results, respectively. Replacing TabPFN V2 with V3 improves the average AUPRC and AUROC of the original TabPFN by 0.107 and 0.048, respectively. TabPFN-AD similarly benefits from the stronger checkpoint, with its average AUPRC and AUROC increasing by 0.041 and 0.016, respectively. These consistent improvements show that the proposed framework remains effective across TabPFN versions and benefits from advances in the underlying backbone.