LENS-AD: ORGANIZING LLM KNOWLEDGE FOR FEW-SHOT TABULAR ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular anomaly detection becomes particularly challenging under extreme label scarcity, where limited observations provide insufficient evidence to reliably characterize abnormal patterns. Such limited evidence motivates the use of Large Language Models (LLMs) to translate task-related domain knowledge into explicit features, revealing latent concepts and feature relationships absent from raw tables. However, this expanded representation also enlarges the learning space: under scarce supervision, detectors may confuse spurious relationships with meaningful anomaly mechanisms. Our key insight is that LLM knowledge should not only expand what the detector can observe, but also constrain how the expanded information is used. Based on this insight, we propose LENS-AD, which organizes original and knowledge-expanded features into overlapping semantic subspaces, each encoding a candidate anomaly mechanism. Because target data are insufficient to reliably determine the relative importance of these subspaces, LENS-AD learns a mapping from subspace descriptors to aggregation weights across source datasets and applies it to combine their predictions without target validation labels. Extensive experiments on 20 real-world tabular anomaly detection benchmarks show that LENS-AD performs competitively under both few-labeled-anomaly and few-sample settings. Further intermediate analyses and ablation studies validate the effectiveness of the proposed framework.

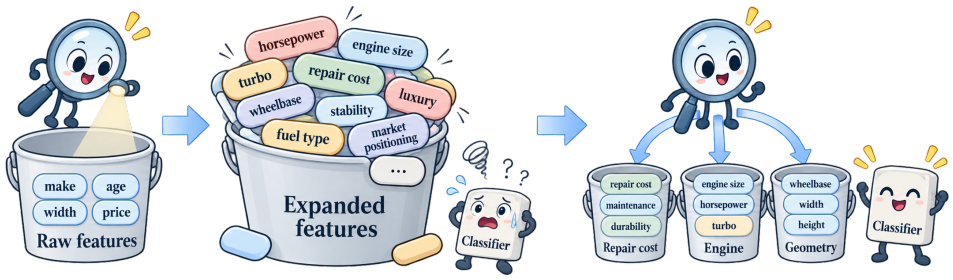


Figure 1: Motivation of LENS-AD. LLM knowledge expands the semantic evidence available beyond raw tabular features, but the resulting feature space can become difficult to navigate under scarce anomaly supervision. LENS-AD acts as a lens that focuses this expanded space into mechanism-oriented semantic subspaces, enabling more targeted anomaly identification.

1 INTRODUCTION

Tabular anomaly detection supports critical applications such as fraud detection (Fiore et al., 2019), industrial monitoring (Panza et al., 2023), and risk management (Pang et al., 2021). In these applications, collecting enough data to characterize abnormal behavior is inherently difficult: anomalies are rare and can take diverse forms (Chandola et al., 2009; Ruff et al., 2021). We focus on few-shot anomaly detection, where learning must proceed with limited training samples or few confirmed anomalies. Existing approaches typically train models from scratch on downstream anomaly detection data, using specialized objectives and representation learning within weakly or semi-supervised frameworks to mitigate limited supervision (Pang et al., 2019; Durani et al., 2025; Gao et al., 2025).

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.

Because these methods rely primarily on task-specific observations and labels, severe data scarcity makes it difficult to determine which patterns in the observed data are genuinely informative for anomaly detection. Recent LLM-based automated feature engineering (LLM-AFE) methods offer a complementary source of information by deriving semantically meaningful variables from task and column descriptions (Han et al., 2024; Hollmann et al., 2023b; Nam et al., 2024; Abhyankar et al., 2025). However, these methods often rely on held-out validation performance to select generated variables and guide subsequent generation. This assumption is poorly suited to few-shot anomaly detection, where scarce target labels cannot reliably rank a large pool of candidate features and their interactions, turning unconstrained feature expansion into an unstable model-selection problem.

Expanding the feature space introduces additional task-relevant evidence, but also creates many more possible feature interactions. With only a few target samples or anomaly labels, it is difficult to determine which of these interactions reflect meaningful anomaly mechanisms. We therefore use semantic knowledge not only to generate new variables, but also to organize original and derived variables into mechanism-oriented subspaces. This provides semantically coherent views for anomaly detection without requiring scarce target supervision to identify useful feature relations from the full expanded space. In this way, **LLM knowledge serves both to enrich the representation and to organize it for prediction.** Figure 1 illustrates this expand-and-focus intuition.

Based on this perspective, we propose LENS-AD, a semantic knowledge organization framework for few-shot tabular anomaly detection. LENS-AD implements this idea through three complementary components. **LLM Knowledge Expansion (LKE)** translates conceptual, referential, and relational knowledge into explicit variables, enriching the original table with task-relevant evidence. **Semantic Subspace Discovery (SSD)** then organizes original and derived variables into overlapping mechanism-oriented subspaces, providing focused prediction views. Because scarce target labels cannot reliably determine the relative importance of these views, **Semantic-Guided Meta-Aggregation (SGMA)** learns from source datasets how to weight them and applies the learned relation to a new task without target validation labels. Together, the three components use semantic knowledge to shape both candidate anomaly mechanisms and their relative importance. In summary, our principal contributions are as follows:

- We introduce LENS-AD, a framework for few-shot tabular anomaly detection that uses LLM knowledge to expand the evidence available from raw tables and focus the enlarged feature space around candidate anomaly mechanisms under scarce supervision.
- We develop Semantic Subspace Discovery to construct mechanism-oriented prediction views, together with Semantic-Guided Meta-Aggregation, which learns from source datasets how to weight these views on a new task without target validation labels.
- We evaluate LENS-AD on 20 real-world datasets across six scarcity settings. It shows competitive performance under both few-labeled-anomaly and few-sample regimes, with further analyses validating the roles of LKE, SSD, and SGMA and their robustness across model choices.

2 Related Work

Tabular Anomaly Detection under Limited Supervision. In supervised tabular anomaly detection, gradient-boosted trees like XGBoost and CatBoost serve as strong baselines (Chen & Guestrin, 2016; Prokhorenkova et al., 2018). To handle severe label sparsity and imbalance, XGBOD (Zhao & Hryniewicki, 2018) augments original features with unsupervised detector outputs. Under limited supervision, addressing anomaly label sparsity usually involves reference distribution separation (Pang et al., 2019; Durani et al., 2025) or sample relationship modeling (Pang et al., 2023; Gao et al., 2025). Another setting is overall sample sparsity, which has been widely explored in images (Xie et al., 2023), but remains severely understudied in tabular applications. A rare exception is MLTDD (Zhai et al., 2026), which combines a two-stage representation extractor with a tangent one-class classifier. Pretrained tabular models also show promise for anomaly detection (Hollmann et al., 2023a), with recent adaptations utilizing synthetic task pretraining (Marszałek et al., 2026; Wei & Armanfard, 2026) or lightweight dependency modules atop frozen representations (Lu et al., 2026). Nevertheless, these methods primarily extract statistical regularities from observed data and do not explicitly exploit column semantics, dataset descriptions, or external knowledge in LLMs.

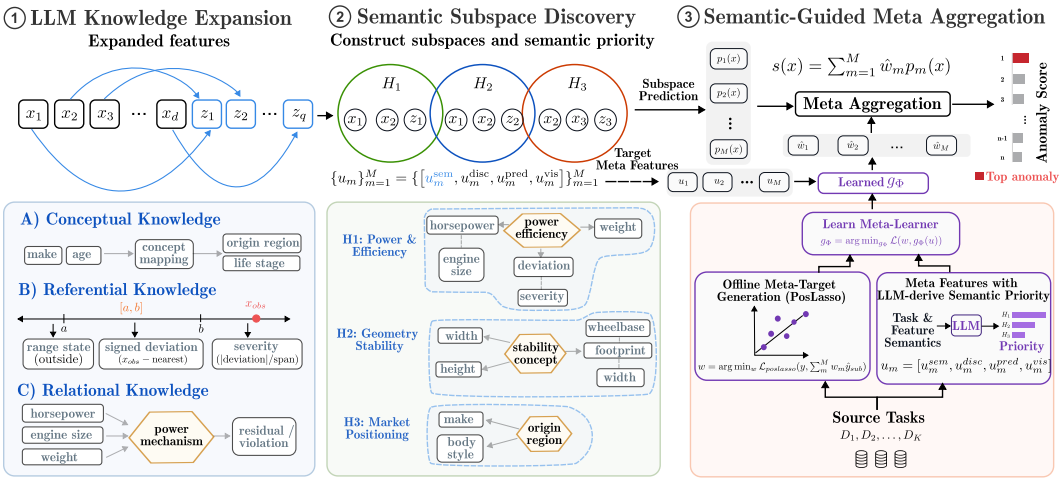


Figure 2: Overview of LENS-AD for tabular anomaly detection. The framework expands tabular representations (Expand), organizes them into mechanism-oriented semantic subspaces (Focus), and aggregates them for anomaly prediction (Aggregation).

Automatic Feature Engineering with Large Language Models. Unlike traditional automated feature engineering constrained by predefined operators, recent approaches leverage LLMs to generate semantically meaningful features from dataset metadata. While some generate features non-iteratively (Han et al., 2024), most rely on iterative refinement guided by downstream performance feedback (Hollmann et al., 2023b; Nam et al., 2024). Advanced strategies further enhance exploration through evolutionary search (Abhyankar et al., 2025), multi-agent collaboration (Dong et al., 2026), or diverse reasoning (Han et al., 2025). Despite these advances, these methods primarily target general classification and regression. Their reliance on iterative validation makes them unsuitable for AD, where extremely scarce labels render performance-based feedback unreliable.

Semantic-Aware Subspace Learning for Tabular Anomaly Detection. Feature subspace partitioning is a well-explored paradigm in anomaly detection. Early methods like Isolation Forest (Liu et al., 2008) implicitly partition spaces via random selection. Recent representation learning approaches further exploit partitions. RFGT (Tang & Li, 2025) creates complementary views for graph propagation, while PCLAD (Li et al., 2025) leverages them to capture intra- and inter-instance regularities. To optimize the partitioning strategy itself, methods like SOP (Kriegel et al., 2009) and MEXICO (Kriegel et al., 2009) rely on pure statistical correlations, whereas SemanticMask (Tao et al., 2024) groups features by semantic similarity for semantics-preserving feature augmentation. Distinct from these approaches, our method leverages LLM-derived semantic knowledge throughout both semantic subspace construction and adaptive subspace aggregation, enabling anomaly detection to be guided by coherent semantic mechanisms rather than feature correlations alone.

3 LENS-AD Framework

As illustrated in Figure 2, LENS-AD organizes task semantics into executable features, mechanism-oriented prediction views, and transferable aggregation guidance. LKE augments the original table with features computed through constrained, semantically motivated transformations. SSD organizes the expanded features into overlapping views that capture evidence for candidate anomaly mechanisms. SGMA combines the resulting predictions using a weighting rule learned across source datasets. We first define the learning problem, then describe the three stages.

3.1 Problem Formulation

We consider tabular instances $x_i \in \mathcal{X}$ with observed anomaly labels $y_i \in \{0, 1\}$, where $y_i = 1$ denotes an anomaly. The task is to learn an anomaly score $s : \mathcal{X} \to \mathbb{R}$ from the available labels: $\hat{s} = \arg\min_s \sum_{i=1}^{N} \mathcal{L}\big(s(x_i), y_i\big)$. Here N is the number of training samples with observed labels

and $\mathcal{L}$ is a binary prediction loss. Larger scores indicate stronger evidence of an anomaly. Task descriptions and feature semantics provide additional context for constructing the score.

3.2 LLM KNOWLEDGE EXPANSION

LKE enriches the original table by translating task-related domain knowledge into explicit derived variables. The LLM determines what semantic transformation to apply, while predefined validation rules and a deterministic compiler control how that transformation is executed. LKE operates only on the task description and feature schema, without access to sample values, empirical statistics, labels, model predictions, or downstream performance.

Table 1: Constrained operator library used by LKE.

Knowledge	Operator	Compiled feature	Key constraint
Conceptual	Categorical concept	Binary concept indicator	Proper subset of known categories
	Categorical mapping	Many-to-one abstraction	Reject one-to-one renaming
	Semantic binning	Named semantic intervals	Externally defined semantics
Referential	Reference interval	Range state + signed/absolute deviation	Valid bounds, units, applicability
Relational	Arithmetic relation	Difference / ratio / sum / product	Compatible numeric columns
	Aggregate relation	Sum / mean / product / equality count	Compatible nonempty feature set
	Boolean relation	Typed conjunction / disjunction	Whitelisted typed clauses

Semantic knowledge generation. LKE expands the original representation from three complementary types of semantic knowledge. **Conceptual** knowledge maps observed values or numerical regimes to domain concepts. **Referential** knowledge introduces externally defined domain anchors and represents the direction and magnitude of deviation from them. **Relational** knowledge captures interpretable cross-column quantities or semantic relationships. For example, on the Automobile dataset, LKE expands the representation of *car brand* to include a *luxury/performance-brand indicator*, uses an external reference range to derive deviation features from *compression ratio*, and combines *horsepower* with *curb weight* into a *power-to-weight ratio*.

Constrained feature generation. LKE does not generate free-form transformation code. Instead, each knowledge proposal instantiates one of the predefined operators in Table 1 and is validated against its operator-specific constraints. The compiler only evaluates the specified transformation on feature values. Valid derived columns are appended to the original features, while unusable outputs are removed. For each retained feature, LKE preserves its semantic description, source columns, knowledge type, derivation, and applicability, which are subsequently used by SSD. Detailed generation prompts, operator definitions, and validation rules are provided in Appendix C.

3.3 SEMANTIC SUBSPACE DISCOVERY

After LKE expands the feature space, SSD organizes the original and generated features around candidate semantic mechanisms of anomalous behavior. Figure 3 illustrates how mechanisms are identified, grounded in the available features, and converted into trainable prediction views.

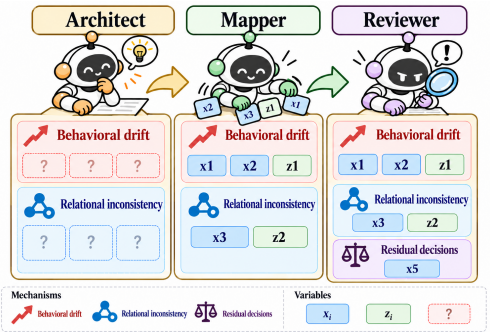


Figure 3: SSD Workflow. The architect proposes mechanisms, the mapper assigns features, and the reviewer resolves residual assignments.

Mechanism discovery and semantic assignment. SSD first identifies candidate mechanisms that are supported by the task context and available feature semantics. The **Architect** describes these mechanisms as plausible anomaly-relevant processes or conditions, rather than directly partitioning the feature space. The accepted mechanisms are then grounded in the available features by the **Mapper**, which assigns each association a semantic role: *core*, *supporting*, or *alternative proxy*. Core features provide direct evidence for a mechanism, supporting features provide complementary evidence, and alternative proxies capture related or substitutable measurements. Residual or ambiguous assignments are subsequently resolved by the **Reviewer**.

Prediction-view construction. Each accepted mechanism is materialized as a trainable prediction view using its core and supporting features, while alternative proxies are retained as semantic metadata rather than model inputs. Because a feature may support more than one mechanism, the resulting views are allowed to overlap. Features that cannot be reliably associated with a specific mechanism are placed in exploratory views, ensuring complete coverage of the expanded feature space. The resulting views are denoted by

$$\mathcal{S} = \{S_1, \ldots, S_M\}, \qquad S_m \subseteq \mathcal{F}^K.$$

Each S_m is fitted independently using the same available training data and supervision regime, producing a view-level prediction $p_m(x) \in [0, 1]$ for subsequent aggregation by SGMA. Detailed prompt schemas, validation rules, and construction procedures are provided in Appendix D.

3.4 Semantic-Guided Meta Aggregation

After SSD, each semantic subspace defines an independent prediction view. Under extreme label scarcity, however, target supervision is insufficient to reliably determine which views should receive greater weight. SGMA therefore learns this weighting rule across tasks: each view is characterized by observable semantic, structural, and predictive properties, and source datasets provide supervision for learning how these properties relate to the view's contribution to the final prediction.

View descriptors. SGMA characterizes each prediction view using four complementary descriptor families. Their information sources and roles are summarized in Table 2. **Semantic** descriptors assess the plausibility and reliability of the underlying mechanism, while **Prediction-Relation** descriptors characterize its complementarity or redundancy with other views on the unlabeled evaluation batch. Detailed definitions are provided in Appendix E.2.

Table 2: SGMA descriptor families.

Family	Source	Representative descriptors
Semantic	LLM assessment from task semantics	Prior importance, observability, task specificity
Discovery	SSD construction	View size, feature composition, roles, overlap
Prediction-Relation	Unlabeled view predictions	Cross-view correlation, rank agreement
Protocol-Visible	Training protocol	Training size, view count, supervision proxies

Source-side supervision. To learn the relationship between these descriptors and useful aggregation weights, SGMA constructs view-level supervision from source datasets under the same scarcity regime as the target task. Let Z denote the logit-transformed predictions of the candidate views on a source task. Their relative contribution is estimated with a sparse nonnegative logistic teacher:

$$(\mathbf{a}^*, b^*) = \arg\min_{\mathbf{a}\geq 0, b} \mathrm{BCE}\big(\mathbf{y}, \sigma(Z\mathbf{a} + b)\big) + \lambda\|\mathbf{a}\|_1. \tag{1}$$

Nonnegativity restricts the teacher to positive view contributions, while the ℓ_1 penalty encourages sparse allocation among redundant views. The normalized coefficients $w_m^* = a_m^* / \sum_j a_j^*$ serve as offline meta-training targets for the corresponding source views.

Cross-task transfer. SGMA pools the resulting descriptor–target pairs (u_m, w_m^*) across source datasets and learns a regression function g_ϕ that predicts view importance from its descriptors. The target dataset is excluded from this meta-training process. At inference time, descriptors are constructed for the target views and mapped to aggregation scores, which are clipped and normalized before combining their predictions:

$$\bar{w}_m = \frac{[g_\phi(u_m)]_+}{\sum_j [g_\phi(u_j)]_+}, \qquad s(x) = \sigma\left(\sum_m \bar{w}_m \operatorname{logit} p_m(x)\right). \tag{2}$$

Information boundary. Target-side aggregation relies only on information available without target evaluation labels. Because prediction-relation descriptors are computed on the batch being scored, the aggregation weights can adapt to its cross-view structure without accessing its labels. Further details are provided in Appendix E.

Table 3: Mean AUCPR over 20 datasets under three few-labeled-anomaly and three few-sample settings. Parentheses give the rank among 14 methods, Avg. Rank averages the six setting-wise ranks, and Tol. #1st counts the dataset-level wins out of 120 (6 settings × 20 datasets).

		Few-labeled-anomaly: NLA			Few-sample: train size				
Method	**Category**	**5**	**10**	**20**	**10**	**20**	**50**	**Avg. Rank**	**Tol. #1st**
DevNet	**Weak Sup.**	0.436 (10)	0.477 (11)	0.504 (11)	0.408 (9)	0.445 (10)	0.493 (11)	10.333	7
PReNet	**Weak Sup.**	0.418 (12)	0.480 (10)	0.505 (10)	0.407 (10)	0.442 (11)	0.488 (12)	10.833	5
XGBOD	**Weak Sup.**	0.481 (8)	0.546 (8)	0.616 (8)	0.346 (11)	0.476 (9)	0.582 (8)	8.667	2
CAD	**Weak Sup.**	0.257 (14)	0.289 (14)	0.311 (14)	0.222 (14)	0.223 (14)	0.241 (14)	14.000	0
WSAD-DT	**Weak Sup.**	0.348 (13)	0.351 (13)	0.357 (13)	0.334 (12)	0.334 (13)	0.356 (13)	12.833	7
XGBoost	**Sup.**	0.462 (9)	0.522 (9)	0.584 (9)	0.298 (13)	0.436 (12)	0.552 (9)	10.167	2
CatBoost	**Sup.**	0.533 (5)	0.591 (6)	0.640 (6)	0.519 (5)	0.563 (7)	0.635 (7)	6.000	5
TabPFN	**Sup.**	0.527 (7)	0.606 (5)	0.682 (4)	0.549 (2)	0.582 (2)	0.655 (3)	3.833	15
LimiX	**Sup.**	0.558 (3)	0.619 (4)	0.679 (5)	0.520 (4)	0.570 (4)	0.651 (5)	4.167	13
TabICL	**Sup.**	0.573 (2)	0.638 (2)	0.710 (2)	0.518 (6)	0.564 (6)	0.655 (4)	3.667	20
CAAFE	**LLM-AFE**	0.530 (6)	0.589 (7)	0.637 (7)	0.517 (7)	0.568 (5)	0.636 (6)	6.333	6
FeatLLM	**LLM-AFE**	0.423 (11)	0.422 (12)	0.425 (12)	0.492 (8)	0.504 (8)	0.523 (10)	10.167	6
LENS-AD (TabPFN)	**Ours**	0.556 (4)	0.633 (3)	0.698 (3)	**0.555** (1)	**0.590** (1)	**0.666** (1)	2.167	24
LENS-AD (TabICL)	**Ours**	**0.591** (1)	**0.651** (1)	**0.710** (1)	0.527 (3)	0.580 (3)	0.666 (2)	**1.833**	**30**

4 Experiments

The evaluation examines overall predictive performance, the contribution and validity of individual semantic components, component-level effects, and robustness across model choices.

4.1 Experimental Setup

Our experiments use the tabular datasets and structured textual metadata from ReTabAD (Yoon et al., 2026). Following its evaluation protocol, the benchmark covers six scarcity settings under two supervision regimes: few-labeled-anomaly, with the number of labeled anomalies (NLA) set to 5, 10, or 20, and few-sample, with training sizes of 10, 20, or 50. All methods use the same three random seeds, and AUCPR serves as the primary evaluation metric.

The comparison includes twelve representative baselines from three categories: **Weakly supervised anomaly detection** (DevNet, PReNet, XGBOD, CAD, and WSAD-DT), **Supervised tabular learning** (XGBoost, CatBoost, TabPFN, LimiX, and TabICL), and **LLM-based automated feature engineering** (CAAFE and FeatLLM).

We instantiate LENS-AD with both TabPFN and TabICL as downstream predictors, yielding two variants that share the same configuration and differ only in their base predictor. Details of the benchmark and baseline configurations are provided in Appendix A and B. Token usage and measured time for LENS-AD and the LLM-AFE baselines are analyzed by dataset in Appendix I.

4.2 Overall Performance

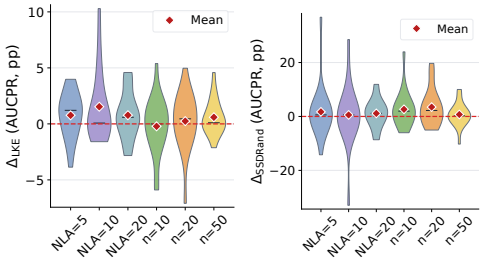
As shown in Table 3, LENS-AD variants rank among the top two methods across all six scarcity settings. The TabICL variant achieves the best performance under the three few-labeled-anomaly settings, whereas the TabPFN variant performs best under the three few-sample settings. Both variants also compare favorably with representative weakly supervised and supervised baselines. Appendix H provides complementary dataset-level W/T/L counts and paired Wilcoxon tests, showing how the aggregate improvements in Table 3 are distributed across individual datasets.

Compared with existing LLM-based automated feature engineering methods, both LENS-AD variants achieve higher aggregate AUCPR across all six settings. Because these baselines also introduce LLM-derived features, this comparison highlights the additional value of organizing the expanded representation into semantic prediction views and aggregating them accordingly, rather than feature expansion alone. The two LENS-AD variants remain competitive despite using different downstream predictors, suggesting that the observed gains are not tied to a particular prediction backbone.

4.3 Analysis of Internal Mechanisms

We next examine whether the semantic information introduced at each stage is meaningful before considering its end-to-end contribution. Specifically, we test whether LKE introduces complementary predictive signal, whether SSD provides structure beyond arbitrary feature partitioning, and whether the semantic descriptors used by SGMA are informative of view quality.

LKE adds complementary predictive signal. To isolate the information contributed by knowledge expansion, we compare each sampled subset of original features with the same subset augmented only by the LKE features derived from it. This matched comparison controls for the source features and asks whether LKE contributes predictive information beyond their original representation. Figure 4(a) shows positive mean and median AUCPR gains in most settings, suggesting that LKE adds predictive signal beyond that already captured by the original variables.



(a) LKE gain (b) SSDRand gain

Figure 4: Distributions of LKE and SSD gains over random controls. Diamonds and bars mark the mean and median, respectively.

Semantic organization matters beyond feature partitioning. We next isolate the effect of semantic organization by replacing SSD with random subspaces matched in both the number of views and their size distribution. Figure 4(b) shows that SSD consistently outperforms these shape-matched controls across all six settings. The advantage is particularly pronounced under stronger supervision scarcity, suggesting that semantic structure becomes more useful when the target data provide less evidence for discovering feature interactions directly.

Semantic descriptors provide signals of view quality. Finally, we examine whether the six semantic descriptors used by SGMA are informative before observing target performance. For each dataset, we compare descriptor rankings with the oracle AUCPR ranking of its semantic views; correlations are computed after within-dataset centering, with significance assessed by permutation tests and Holm correction. Figure 5 shows consistent directional structure across scarcity regimes. The strongest signal is the absence of strong competing explanations ($\rho = 0.32$–0.44), while prior importance, observability, and task specificity are also positively associated with view quality across all settings. The remaining descriptors are weaker and less consistent. Overall, the semantic descriptors provide informative priors for distinguishing more promising views, supporting their use in SGMA.

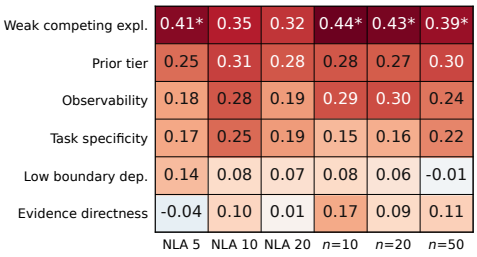


Figure 5: Rank correlations between semantic descriptors and oracle view quality. Colors encode direction and magnitude; stars mark Holm-adjusted $p < 0.05$.

4.4 End-to-End Component Analysis

Component Ablation Study We ablate LKE, SSD, and SGMA through six configurations spanning the raw and LKE-expanded feature spaces. Table 4 shows that the complete model achieves the best performance in all six scarcity settings. Beyond the individual contribution of each component, the ablation pattern suggests that their benefits depend on how they are combined. LKE alone yields only a modest gain over the raw all-feature baseline, and SSD with equal aggregation provides limited improvement. In contrast, introducing SGMA improves the corresponding SSD configuration in both feature spaces, with the largest benefit when SSD and SGMA operate on the LKE-expanded representation. Removing all three semantic components reduces the six-setting mean by 2.6%. These results suggest that knowledge expansion is most effective when the enlarged feature space is organized into semantic views and selectively aggregated.

Table 4: Component ablation study. Checkmarks denote enabled components; the last column reports the relative AUCPR change from the default LENS-AD.

				NLA			Train size n				
Condition	**LKE**	**SSD**	**SGMA**	**5**	**10**	**20**	**10**	**20**	**50**	**Mean**	**Δ (%)**
LKE SSD+SGMA (LENS-AD)	✓	✓	✓	**0.556**	**0.633**	**0.698**	**0.555**	**0.590**	**0.666**	0.616	–
LKE SSD equal ensemble	✓	✓		0.530	0.603	0.694	0.552	0.589	0.663	0.605	−1.9%
LKE all-feature	✓			0.524	0.606	0.685	0.551	0.586	0.662	0.602	−2.3%
Raw SSD+SGMA		✓	✓	0.538	0.614	0.692	0.545	0.581	0.658	0.605	−1.9%
Raw SSD equal ensemble		✓		0.531	0.603	0.685	0.543	0.580	0.656	0.600	−2.7%
Raw all-feature				0.527	0.605	0.681	0.549	0.582	0.655	0.600	−2.6%

Meta-learning versus Target-label Stacking. We compare SGMA with target-label stacking, which learns aggregation weights directly from target-task labels. Both methods use the same semantic-subspace and full-feature predictors, differing only in how their aggregation weights are learned: SGMA transfers view-quality information across datasets through the semantic descriptors in Section 3.4, without fitting on target labels. Paired significance is assessed using two-sided Wilcoxon tests over the 20 dataset-level results. Table 5 shows that SGMA outperforms target-label stacking in all six settings, with significant differences in five. The advantage is largest in the four most supervision-limited settings, reaching 3.39–3.60 AUCPR points, and gradually narrows as more target supervision becomes available, decreasing to 0.95 points at $n = 50$. SGMA also shows lower within-dataset seed variation in five settings. This pattern suggests that cross-task semantic guidance is most useful when target supervision is insufficient to reliably estimate the relative quality of candidate views. The same trend holds when excluding cases with too few anomalies for stratified cross-fitting.

Table 5: SGMA versus target-label stacking with matched ensemble members. Δ is SGMA minus stacking in AUCPR percentage points.

Setting	**Target stack**	**SGMA**	**Δ (pp)**	p
NLA 5	$0.520_{\pm 0.077}$	$\mathbf{0.556}_{\pm 0.071}$	+3.58	< .001
NLA 10	$0.597_{\pm 0.056}$	$\mathbf{0.633}_{\pm 0.047}$	+3.58	< .001
NLA 20	$0.684_{\pm 0.041}$	$\mathbf{0.698}_{\pm 0.043}$	+1.38	.040
$\boldsymbol{n = 10}$	$0.521_{\pm 0.088}$	$\mathbf{0.555}_{\pm 0.077}$	+3.39	< .001
$\boldsymbol{n = 20}$	$0.554_{\pm 0.072}$	$\mathbf{0.590}_{\pm 0.062}$	+3.60	< .001
$\boldsymbol{n = 50}$	$0.656_{\pm 0.059}$	$\mathbf{0.666}_{\pm 0.053}$	+0.95	.091

4.5 Robustness and Extensions

We examine whether the gains of LENS-AD remain stable across LLM choices, base predictors, and meta-training source pools, and then extend the framework to unsupervised anomaly detection.

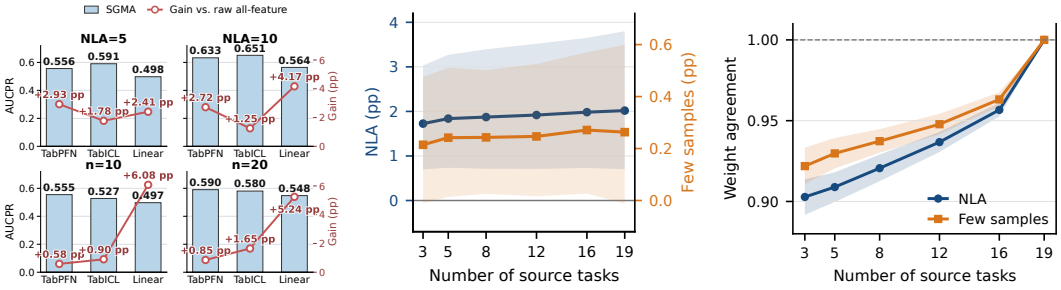
Stability across LLMs. We evaluate whether the effectiveness of LENS-AD depends on the choice of LLM. Each LLM independently generates both the LKE features and semantic mechanisms, while the downstream predictor (TabPFN), data splits, and random seeds remain fixed. To isolate the effect of semantic organization rather than differences in the generated feature representations, each complete pipeline is compared with the corresponding all-feature TabPFN using the same LKE features. Performance is reported as the AUCPR improvement of the complete pipeline over its corresponding all-feature baseline. We further compare the dataset-level performance profiles of different LLMs using Spearman correlation, with DeepSeek-Flash serving as the reference.

Table 6: SGMA gains across LLM variants over matched all-feature baselines. Pos. counts improved settings; ρ denotes agreement with DeepSeek-Flash.

LLM	**NLA=5**	**NLA=10**	$\boldsymbol{n = 10}$	$\boldsymbol{n = 20}$	**Pos.**	ρ
DeepSeek-Flash	+3.22	+2.71	+0.42	+0.45	4/4	1.000
GPT-Sol	+2.06	+3.01	+1.28	+0.75	4/4	0.997
DeepSeek-Pro	+0.34	+1.24	+0.86	+0.42	4/4	0.986
GPT-Luna	+3.06	+2.35	+0.14	−0.19	3/4	0.986

Table 6 shows positive gains in 15 of the 16 LLM–setting comparisons. DeepSeek-Flash and GPT-Sol improve over their matched all-feature baselines in all four settings, with complementary strengths across the two scarcity regimes; DeepSeek-Pro also improves in all four settings. All three alternative variants have correlations above 0.98 with DeepSeek-Flash, showing highly consistent performance profiles. These results demonstrate that the benefit is reproduced across multiple independently generated semantic representations rather than being confined to one LLM.

Stability across Base Predictors. We next hold the LLM model and semantic contexts fixed and replace the predictor used to evaluate. For each predictor, the matched reference applies the same



(a) Base predictors. (b) Gain over equal aggregation. (c) Weight agreement.

Figure 6: SGMA stability. In (a), red markers show gains over matched raw-feature predictors. In (b,c), shading shows 95% target-bootstrap intervals; the gain curves use separate vertical scales.

predictor directly to all original features. This comparison measures the gain of the complete semantic pipeline over a predictor-matched raw all-feature baseline.

Figure 6a shows that the complete pipeline improves the matched raw all-feature predictor in all 12 predictor–setting comparisons. TabICL provides the strongest absolute performance in both few-labeled-anomaly settings, while the Linear predictor achieves the largest gains under few-sample training. The complete semantic pipeline therefore complements both high-capacity tabular foundation models and a substantially simpler predictor, showing that its utility is not specific to TabPFN.

Stability of Meta-Learning. SGMA learns its aggregation rule from the remaining source datasets, raising the question of how strongly it depends on source-pool size and composition. Figure 6b shows positive mean SGMA gains over equal aggregation even with substantially fewer source tasks. Meanwhile, Figure 6c shows that the learned target weights become increasingly consistent with those from the complete source pool. With 16 sources, weight agreement already exceeds approximately 0.95 under both supervision regimes, with only minor changes toward the full 19-source setting. Additional composition and leave-one-source analyses in Appendix E.6 further show that the result is not driven by any single source dataset.

Extension to Unsupervised Anomaly Detection. We further examine whether the benefit of semantic knowledge organization persists without target-task labels. We retain LKE and SSD, replace the subspace predictors with unsupervised anomaly scorers, and use label-free descriptors for SGMA; target labels are used only for evaluation. Across 20 datasets, the Isolation Forest instantiation ranks first under all four training-set sizes, while the COPOD instantiation ranks second overall. The TabPFN-Unsupervised variant also improves over its matched full-feature detector. These results suggest that the benefit of LENS-AD arises from semantic organization rather than its reliance on supervised prediction. Full settings and results are provided in Appendix F.

5 Conclusion

This paper studies how LLM knowledge can be effectively used for tabular anomaly detection under extreme supervision scarcity. Our key finding is that expanding the feature space alone is insufficient: the introduced knowledge becomes more useful when it is organized around candidate anomaly mechanisms and selectively aggregated. Based on this insight, LENS-AD combines knowledge-based feature expansion, mechanism-oriented semantic subspaces, and cross-task meta-learning to guide prediction without relying on target validation labels. Experiments on 20 real-world datasets show competitive performance across both scarcity regimes, while controlled analyses further demonstrate that LKE introduces complementary predictive signal, semantic subspaces outperform matched random partitions, and semantic descriptors provide informative signals of view quality. The resulting gains remain stable across different LLMs, base predictors, and source-task configurations, suggesting that the benefit stems from semantic knowledge organization rather than a particular model choice. Future work will explore richer hierarchical and multi-granularity semantic structures and incorporate additional external knowledge sources.

AI USE STATEMENT

In this work, we used large language models as part of the proposed method to generate structured semantic knowledge units, organize features into semantic subspaces, and assess the resulting semantic mechanisms for aggregation. These model-generated research artifacts were subject to the predefined schema and feasibility checks described in the paper and appendix. Additionally, we used generative AI tools to edit manuscript text for clarity and readability and to identify potentially relevant literature. The authors reviewed and revised AI-assisted text and manually checked the existence, relevance, and bibliographic details of suggested references against the original publications before citing them. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The appendix specifies the benchmark tasks, data splits, and label-visibility protocols in Appendix A; implementation settings, baseline configurations, and the computing environment in Appendix B; and the procedures for LLM knowledge expansion, semantic subspace discovery, and semantic-guided meta-aggregation in Appendices C–E. The supplementary package includes source code, run instructions, checkpoint acquisition guidance, dependency specifications, the benchmark data and task metadata, split manifests and experiment configurations, as well as the recorded LLM prompts and outputs, generated feature definitions, semantic subspace assignments, and dataset-by-seed results. It also includes scripts for rerunning the evaluation stages and exporting aggregate tables from the archived results.

REFERENCES

Nikhil Abhyankar, Parshin Shojaee, and Chandan K Reddy. Llm-fe: Automated feature engineering for tabular data with llms as evolutionary optimizers, 2025.

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

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.

Fengxian Dong, Zhi Zheng, Xiao Han, Wei Chen, Jingqing Ruan, Tong Xu, Yong Chen, and Enhong Chen. Memory-augmented llm-based multi-agent system for automated feature generation on tabular data. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10988–11007, 2026.

Walid Durani, Tobias Nitzl, Claudia Plant, and Christian Böhm. Weakly supervised anomaly detection via dual-tailed kernel. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hxUGmRusz5.

Ugo Fiore, Alfredo De Santis, Francesca Perla, Paolo Zanetti, and Francesco Palmieri. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479:448–455, 2019. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2017.12.030. URL https://www.sciencedirect.com/science/article/pii/S0020025517311519.

Jianling Gao, Chongyang Tao, Zhenchao Sun, Xiya Jiang, and Shuai Ma. Semi-supervised anomaly detection through denoising-aware contrastive distance learning. In *Proceedings of the ACM on Web Conference 2025*, pp. 2111–2119, 2025.

Sungwon Han, Jinsung Yoon, Sercan O Arik, and Tomas Pfister. Large language models can automatically engineer features for few-shot tabular learning. In *International Conference on Machine Learning*, pp. 17454–17479. PMLR, 2024.

Sungwon Han, Sungkyu Park, and Seungeon Lee. Tabular feature discovery with reasoning type exploration, 2025.

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*, 2023a. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, and Frank Hutter. Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. *Advances in Neural Information Processing Systems*, 36:44753–44775, 2023b.

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Outlier detection in axis-parallel subspaces of high dimensional data. In *Advances in Knowledge Discovery and Data Mining*, pp. 831–838. Springer Berlin Heidelberg, 2009.

Yizhou Li, Yijie Wang, Hongzuo Xu, Bin Li, and Xiaohui Zhou. Deep anomaly detection with partition contrastive learning for tabular data: Deep anomaly detection with partition contrastive learning... *Data Min. Knowl. Discov.*, 39(4), May 2025. ISSN 1384-5810. doi: 10.1007/s10618-025-01102-w. URL https://doi.org/10.1007/s10618-025-01102-w.

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

Sha Lu, Jixue Liu, Stefan Peters, Thuc Duy Le, Craig Xie, Lin Liu, and Jiuyong Li. ulead-tabpfn: Uncertainty-aware dependency-based anomaly detection with tabpfn, 2026.

Patryk Marszałek, Tomasz Kuśmierczyk, and Marek Śmieja. Tactic for navigating the unknown: Tabular anomaly detection via in-context inference, 2026.

Jaehyun Nam, Kyuyoung Kim, Seunghyuk Oh, Jihoon Tack, Jaehyung Kim, and Jinwoo Shin. Optimized feature generation for tabular data via llms with decision tree reasoning. *Advances in neural information processing systems*, 37:92352–92380, 2024.

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, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 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.

Maria Antonietta Panza, Marco Pota, and Massimo Esposito. Anomaly detection methods for industrial applications: A comparative study. *Electronics*, 12(18):3971, 2023.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Wen-Wei Tang and Cheng-Te Li. Random-feature graph neural networks with representation tokenized transformer for robust one-class anomaly detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 5279–5283, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720406. doi: 10.1145/3746252.3760938. URL https://doi.org/10.1145/3746252.3760938.

Shuting Tao, Peng Zhu, Jinyu Wang, and Jian Wang. Semanticmask: A contrastive view design for anomaly detection in tabular data. In Kate Saenko (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 2370–2378. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/262. URL https://doi.org. Main Track.

Jack Yi Wei and Narges Armanfard. Iclad: In-context learning for unified tabular anomaly detection across supervision regimes, 2026.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of few-shot anomaly detection in industry vision: Graphcore. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xzmqxHdZAwO.

Sanghyu Yoon, Dongmin Kim, Suhee Yoon, Ye Seul Sim, Seungdong Yoa, Hye-Seung Cho, Soonyoung Lee, Hankook Lee, and Woohyung Lim. RetabAD: A benchmark for restoring semantic context in tabular anomaly detection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UFwgg44VZq.

Yuejing Zhai, Yiping Li, and Wuman Luo. A tangent few-shot anomaly detection framework based on two-stage embedding representation for tabular data. *Complex & Intelligent Systems*, 2026.

Yue Zhao and Maciej K Hryniewicki. Xgbod: improving supervised outlier detection with unsupervised representation learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2018.

A Benchmark and Evaluation Protocols

A.1 Benchmark Tasks and Semantic Metadata

We evaluate all 20 binary anomaly-detection tasks from ReTabAD (Yoon et al., 2026). Each task provides a dataset description, anomaly definition, column schema, and column-level semantic descriptions. Dataset statistics and concise task definitions are reported in Appendix A.3, while the complete metadata supplied to the semantic modules is included in the reproducibility artifacts.

A.2 Main Evaluation Protocols

Common split and randomization. For each dataset and root seed, we construct a stratified 70/30 train–test split using root seeds 42, 2024, and 999. Within each dataset–seed pair, all methods and supervision settings share the same test split. Methods evaluated under the same setting additionally receive the same training rows and visible labels. Test labels are withheld during model construction and used only for final evaluation. The resulting splits and visible-label sets are recorded to support exact reconstruction.

Few-labeled-anomaly supervision. All training rows are available, but only $k \in \{5, 10, 20\}$ true anomalies are revealed as positive labels. All remaining training rows are treated as unlabeled negatives and may therefore contain unobserved anomalies. A seeded permutation of the training anomalies defines nested visible sets, so the anomalies visible at $k = 5$ remain visible at $k = 10$ and $k = 20$. All compared methods receive the same visible set for a given dataset–seed pair.

Few-sample supervision. Only $n \in \{10, 20, 50\}$ labeled training rows are available. Class counts follow the training-set class proportions, subject to at least one example from each class. Seeded within-class permutations produce nested subsets across the three values of n. All labels in the selected subset are visible.

Table A1: Main evaluation protocol.

Item	Main evaluation setting
Benchmark	20 ReTabAD tasks
Few-labeled-anomaly	NLA $\in \{5, 10, 20\}$
Few-sample	$n \in \{10, 20, 50\}$ labeled rows
Root seeds	42, 2024, and 999
LENS-AD variants	TabPFN and TabICL base predictors
Metrics	AUCPR
Pairing	Same split, training context, visible labels, and root seed

Table A2: Statistics of the 20 ReTabAD datasets before model-specific preprocessing. Num., Cat., and Bin. denote numerical, categorical, and binary feature columns, respectively.

Dataset	Domain	Samples	Features	Num.	Cat.	Bin.	Normal	Anomaly	Anom. (%)
Automobile	Insurance risk	159	25	15	7	3	117	42	26.42
Backdoor	Network security	29,223	42	37	2	3	29,113	110	0.38
Campaign	Bank marketing	7,842	16	7	5	4	6,056	1,786	22.77
Cardiotocography	Fetal health	2,126	21	20	1	0	1,655	471	22.15
Census	Income	50,000	41	9	29	3	47,121	2,879	5.76
Churn	Telecommunications	7,032	19	3	10	6	5,163	1,869	26.58
Cirrhosis	Clinical survival	247	17	10	2	5	165	82	33.20
Covertype	Forest ecology	50,000	12	10	2	0	49,520	480	0.96
Credit	Credit risk	30,000	23	14	8	1	23,364	6,636	22.12
Equip	Industrial monitoring	7,672	6	4	2	0	6,905	767	10.00
Gallstone	Clinical diagnosis	241	38	31	2	5	161	80	33.20
Glass	Forensic materials	214	9	9	0	0	163	51	23.83
Glioma	Cancer grading	730	23	1	1	21	487	243	33.29
Quasar	Astronomy	50,000	8	8	0	0	40,520	9,480	18.96
Seismic	Mining safety	2,584	18	14	2	2	2,414	170	6.58
Stroke	Clinical outcome	4,909	10	3	3	4	4,700	209	4.26
Vertebral	Biomechanics	310	6	6	0	0	210	100	32.26
WBC	Cancer diagnosis	535	30	30	0	0	357	178	33.27
Wine	Food chemistry	178	13	13	0	0	130	48	26.97
Yeast	Protein localization	1,484	8	7	0	1	1,389	95	6.40

A.3 Dataset Statistics and Task Definitions

Table A2 provides the complete dataset-level statistics used throughout the experiments. Full dataset descriptions and column-level semantics are retained in the reproducibility artifacts.

Task Definitions and Anomaly Classes. The following catalog gives the benchmark-defined detection objective for each dataset. Each entry states the task context and the positive anomaly class; complete dataset and column descriptions are provided in the reproducibility artifacts.

- **Automobile.** Automobile specifications are used for insurance-risk assessment; anomalous cases are cars with symboling $+2$ or $+3$.
- **Backdoor.** Network flows are monitored for intrusion detection; backdoor attacks form the anomaly class.
- **Campaign.** Client and campaign records are used to model response to bank direct marketing; clients who subscribe to a term deposit form the anomaly class.
- **Cardiotocography.** Cardiotocograms are used for fetal-state assessment; suspect or pathologic fetal states form the anomaly class.
- **Census.** Demographic and employment records are used for income classification; annual income above USD 50K forms the anomaly class.
- **Churn.** Telecom account and service records are used for customer retention analysis; customers who churn form the anomaly class.
- **Cirrhosis.** Clinical records are used for survival assessment in primary biliary cirrhosis; patient death, rather than censoring or transplantation censoring, forms the anomaly class.
- **Covertype.** Cartographic measurements are used for forest-cover identification; cottonwood/willow cover, rather than lodgepole pine, forms the anomaly class.
- **Credit.** Account and payment histories are used for credit-card risk assessment; customers who default form the anomaly class.
- **Equip.** Industrial sensor readings are used for equipment monitoring; faulty operation forms the anomaly class.
- **Gallstone.** Demographic, laboratory, and body-composition measurements are used for clinical diagnosis; the presence of gallstones forms the anomaly class.
- **Glass.** Refractive-index and oxide-content measurements are used for glass-type identification; container, tableware, and headlamp glass form the anomaly class.

- **Glioma.** Clinical and mutation features are used for glioma grading; glioblastoma multiforme, rather than lower-grade glioma, forms the anomaly class.
- **Quasar.** Spectral and photometric measurements are used for astronomical-object identification; quasars, rather than stars or galaxies, form the anomaly class.
- **Seismic.** Mining seismic and operational measurements are used for hazard forecasting; a high-energy seismic bump in the next shift forms the anomaly class.
- **Stroke.** Clinical and demographic factors are used for stroke-risk assessment; patients with a recorded stroke form the anomaly class.
- **Vertebral.** Biomechanical measurements are used for vertebral-condition identification; the benchmark's sampled original "Normal" class is treated as anomalous.
- **WBC.** Cell-nucleus image features are used for breast-mass diagnosis; malignant masses form the anomaly class.
- **Wine.** Chemical measurements are used for wine-cultivar identification; cultivar 3 forms the anomaly class.
- **Yeast.** Biochemical sequence-derived scores are used for protein localization; ME1/ME2 endoplasmic-reticulum membrane localization forms the anomaly class.

B Hyperparameter Configuration

This section reports the implementation settings, baseline configurations, hyperparameter-selection protocol, and computing environment used in the experiments. Further implementation details of LKE, SSD, and SGMA are provided in Appendices C–E.

B.1 LLM Configuration

The main semantic-generation pipeline uses the preview version of `deepseek-v4-flash`. The preview version of `deepseek-v4-pro`, together with `gpt-5.6-luna` and `gpt-5.6-sol`, is evaluated only in the LLM-stability analysis. Each alternative model independently regenerates the corresponding semantic artifacts from the same fixed prompt templates, structured-output schemas, and information boundaries.

All LLM calls use temperature 0, top-p equal to 1, and structured JSON output. The DeepSeek preview models use the maximum available reasoning setting, whereas the two GPT variants use the high reasoning setting. The token limit is fixed to 393,216 for all evaluated models. Prompt templates, schemas, parsing rules, and validation procedures are held constant across datasets and model variants. No LLM output is selected or edited manually.

B.2 Base Predictors and Baselines

Both variants of TabPFN and TabICL use eight estimators for every member predictor. Automatic estimator scaling is disabled for TabPFN so that the ensemble size remains fixed across datasets, semantic views, and scarcity settings. The linear base-predictor variant uses scikit-learn logistic regression with `max_iter=1000` and default ℓ_2 regularization.

All preprocessing statistics are estimated exclusively from the training rows available under the corresponding evaluation setting. Numerical missing values are replaced by training-set medians. Categorical vocabularies are constructed from the selected training rows, and categories not observed during fitting are represented by a fixed unknown-category code. The same preprocessing procedure is used for the full-feature predictor and all semantic-view predictors.

Table B1 summarizes the implementations, checkpoints, and fixed configurations of the comparison methods. Parameters not explicitly listed follow the corresponding official implementation defaults. Every method is evaluated using the same dataset split, protocol-visible training rows and labels, and root seed as its paired competitors. No baseline configuration is selected using target evaluation performance. For CAAFE (Hollmann et al., 2023b) and FeatLLM (Han et al., 2024), the generation budgets are fixed for each dataset and setting, and generated features are not filtered using final evaluation results.

Table B1: Baseline configurations. All methods use the same data split and root seeds for a given dataset and setting.

Method	Implementation/checkpoint	Main parameters	Selection rule
DevNet	Official implementation	50 epochs; batch 512; 20 batches; depth 2	Fixed configuration
PReNet	Official implementation	50 epochs; batch 512; lr 10^{-3}; wd 10^{-2}	Fixed configuration
XGBOD	PyOD 3.6.0	Tuning disabled	Fixed configuration
CAD	Official implementation	Batch 128; lr 10^{-3}; adaptive epochs	Fixed implementation rule
WSAD-DT	Official implementation	5 splits; 100 epochs; latent dim. 128	Fixed configuration
XGBoost	XGBoost 3.2.0	Official defaults; one CPU worker	No validation search
CatBoost	CatBoost 1.2.10	Official defaults; one CPU worker	No validation search
TabPFN	`tabpfn-v2.5-classifier-v2.5_default`	batch 10,000; max 50,000 train rows	Fixed checkpoint
LimiX	`LimiX-16M.ckpt`	Softmax temperature 0.9	Fixed checkpoint
TabICL	`tabicl-classifier-v2-20260212`	Official inference defaults	Fixed checkpoint
CAAFE	Official implementation	10 iterations; 4 folds; temperature 0.5	Same query budget per task
FeatLLM	Official implementation	5 queries; 4 shots; temperatures 0.5/0	Same query budget per task

B.3 Hyperparameter Selection Protocol

We distinguish globally fixed method-design choices from model-internal selection performed using permitted source-side information. The LLM prompt templates, structured-output schemas, LKE operator definitions, SSD construction constraints, SGMA descriptor definitions, model checkpoints, ensemble sizes, and baseline configurations are fixed globally before the reported benchmark comparisons. They are shared across datasets and scarcity settings and are not selected using target validation or evaluation performance.

No dataset-specific search is performed over LLM prompts, generated features, semantic subspaces, descriptor subsets, member-predictor checkpoints, or baseline hyperparameters. The only recurring model-internal selection in the main pipeline is the source-side regularization selection used to construct the sparse SGMA teacher. Its candidate values, inner-validation objective, fold construction, tie-breaking rule, and fallback procedure are specified in Appendix E.3 together with the teacher optimization. This selection is performed independently within each source task and does not use target validation or evaluation labels.

All globally fixed configurations were frozen before the final benchmark tables were produced. In particular, downstream AUCPR was not used to select or revise LKE knowledge units, generated columns, SSD mechanisms, feature memberships, semantic descriptors, LLM outputs, or model configurations.

B.4 Hardware, Software, and Reproducibility Artifacts

Experiments are executed on a Linux server equipped with two Intel Xeon Gold 6530 processors (64 physical CPU cores in total), 1 TiB of system memory, and eight NVIDIA L20 GPUs with 46,068 MiB of memory per GPU. The recorded software environment uses Python 3.11.15, PyTorch 2.5.1 with CUDA 12.4, NumPy 1.26.4, pandas 2.3.3, scikit-learn 1.6.1, SciPy 1.17.1, TabPFN 8.0.8, CatBoost 1.2.10, XGBoost 3.2.0, and PyOD 3.6.0.

The supplementary package provides the exact environment specification, experiment configurations, split identifiers, prompts, structured outputs, generated feature definitions, semantic-subspace assignments, and unrounded result files used in the appendix.

C LLM Knowledge Expansion

The main paper introduces LKE at the level of structured knowledge units. This appendix provides the generation, validation, and compilation details omitted from the main text. LKE is executed once for each dataset, and the resulting feature catalog is reused unchanged across all scarcity settings. Therefore, neither the LKE prompt nor the generated features depend on the number of available samples or labels.

Algorithm 1 LLM Knowledge Expansion

Require: Task metadata $\mathcal{D}$ and feature table $\mathbf{X}$
Ensure: Expanded table $\mathbf{X}^K$ and feature metadata $\mathcal{M}^K$

Construct a schema-level task profile from $\mathcal{D}$, excluding sample rows and the label column.
Construct the original-feature metadata $\mathcal{M}^X$ from the input schema.
for $\tau \in \{\text{CONCEPTUAL}, \text{REFERENTIAL}, \text{RELATIONAL}\}$ **do**
    Construct the fixed family-specific prompt using the task profile, output schema, and admissible operators.
    Request a structured set of knowledge proposals $\mathcal{K}_\tau$.
    Validate required fields, column references, argument types, and operator-family compatibility.
end for
Merge the accepted proposals and remove invalid or duplicate knowledge units to obtain $\mathcal{K}$.
Initialize the generated feature set $\mathbf{Z} \leftarrow \varnothing$ and metadata set $\mathcal{M}^Z \leftarrow \varnothing$.
for each accepted knowledge unit $\kappa_j \in \mathcal{K}$ **do**
    Compile its operator o_j on the referenced columns $\mathbf{X}[S_j]$, producing candidate outputs $\{z_{j,q}\}_{q=1}^{Q_j}$ and their operator-specific output roles.
    for each candidate output $z_{j,q}$ **do**
        Construct feature-level metadata $m_{j,q}$ from κ_j, its output role, and its inferred logical type.
        if $z_{j,q}$ is nonempty, finite, nonconstant, and nonduplicate **then**
            Add $z_{j,q}$ to $\mathbf{Z}$ and add $m_{j,q}$ to $\mathcal{M}^Z$.
        end if
    end for
end for
Construct $\mathbf{X}^K = [\mathbf{X}, \mathbf{Z}]$ and $\mathcal{M}^K = \mathcal{M}^X \cup \mathcal{M}^Z$.
return $(\mathbf{X}^K, \mathcal{M}^K)$

C.1 Generation Inputs and Workflow

LKE receives a schema-level description of the anomaly-detection task. The input contains the dataset description, the definition of the anomaly class, column identifiers and logical types, column-level descriptions, and the schema-defined vocabularies of categorical and binary columns. The categorical vocabularies specify admissible values but do not include their empirical frequencies.

Knowledge generation does not access sample rows, individual feature values, empirical distributions, category frequencies, labels, anomaly prevalence, model predictions, or downstream performance. After a knowledge unit has been generated and validated, a deterministic compiler accesses the referenced feature columns only to evaluate the already-fixed transformation. The compiler does not fit thresholds, estimate reference values, or use labels.

Algorithm 1 summarizes the complete workflow. For each dataset, the LLM is queried separately for conceptual, referential, and relational knowledge. The resulting structured proposals are validated against the corresponding operator specification and the dataset schema. Accepted operators are then applied to the feature table, after which unusable or duplicated outputs are removed. The remaining generated columns are appended to the original table, and their source columns and semantic definitions are retained as feature metadata.

C.2 Structured Generation and Executable Operators

For each dataset, the three knowledge families are generated through separate calls under the same information boundary. Conceptual knowledge maps an individual column to an explicit domain concept, many-to-one abstraction, or externally defined semantic band. Referential knowledge introduces an external reference interval and represents both the state and magnitude of deviation from that interval. Relational knowledge expresses interpretable cross-column arithmetic, aggregation, or Boolean relationships.

Role Definition and Generation Principles

You are generating conceptual LLM Knowledge Expansion units for a tabular anomaly task. Inspect every compatible schema column. Generate every materially distinct executable direction supported by this family boundary and operator catalog. …

Knowledge Family Constraint

Use one source column per knowledge unit. Map categorical values to concepts or map a single numeric column to real-world semantic regimes. Geographic origin, product class, technology class, and market identity belong here. …

Operator Specification

Allowed structured operators for this family:
- categorical_concept: column, concept, positive_values
- categorical_mapping: column, mapping, default
- semantic_binning: column, definition_context, bins [{label, lower, upper}]

Information Boundary

Information boundary: no sample rows, observed numerical distributions, value frequencies, label statistics, per-value label rates, validation or test results, downstream results, or empirical rarity assumptions are available …

Dataset Context

<<<DATASET_TASK>>> {"dataset": "automobile", "dataset_description": "…", "task_definition": "…"}…
<<<FEATURE SCHEMA>>> {"name": "drive-wheels", "kind": "categorical", …}

Output Schema

Return ONLY a JSON array. Every object must contain exactly this complete envelope:
[{semantic_name, derivation, semantic_meaning, task_relevance, operator, applicability, confidence, mechanism_tags, semantic_tags}]

Figure C1: Structure of the LKE generation prompt (using conceptual knowledge as an example).

The prompt does not prescribe a fixed number of proposals. Instead, it requests materially distinct and executable transformations and instructs the model to stop rather than add synonymous or non-executable variations. Each prompt specifies the selected knowledge family, its admissible operators, the information boundary, the task definition, the feature schema, and a JSON-only output contract.

Each LLM-generated knowledge unit contains a semantic name, a description of the derivation, the semantic meaning of the resulting feature, its relevance to the anomaly-detection task, one structured operator, applicability conditions, an ordinal confidence level, and semantic tags. The confidence value belongs to $\{\text{LOW}, \text{MEDIUM}, \text{HIGH}\}$ and represents a semantic prior rather than an empirical probability. Operator type, source-column scope, and internal identifiers are derived programmatically and are not requested from the LLM. Figure C1 shows the LLM prompt for generating features driven by conceptual knowledge.

LKE accepts only the seven operator types summarized in Table C1. Free-form code or transformations outside this library are rejected. One accepted knowledge unit may compile into multiple derived columns. In particular, a reference interval produces three complementary outputs: a categorical relation to the reference range, a signed deviation, and an absolute deviation. Validation is performed before compilation and verifies schema references, feature-type compatibility, categorical vocabularies, and operator-specific semantic constraints.

During compilation, missing values, unseen categories, and undefined arithmetic are handled by fixed rules. Generated outputs that are empty, non-finite, constant, or exact duplicates are removed. These checks can prevent an unusable proposed feature from entering the final catalog, but the observed values cannot change the semantic rule originally proposed by the LLM.

C.3 End-to-End Example

We use the Automobile dataset to illustrate how the three knowledge families are converted into executable features. The examples document the generation and compilation process rather than the individual predictive contribution of each derived feature.

Conceptual knowledge. Given the categorical manufacturer feature, LKE introduces a luxury/performance-brand concept covering Audi, BMW, Jaguar, Mercedes-Benz, and Porsche. The

Table C1: Executable operators supported by LKE. Proposals outside this library are not executed.

Family	Operator	Input / specification	Compiled output(s)	Main validation constraints
Conceptual	Categorical concept	One categorical feature; a named concept; categories associated with the concept	Binary concept indicator	Referenced column must be categorical; selected values must be known categories and form a proper subset
	Categorical mapping	One categorical feature; a mapping from observed categories to higher-level semantic groups	Many-to-one categorical abstraction	Mapping keys must belong to the known vocabulary; complete one-to-one renaming is rejected
	Semantic binning	One numerical feature; a named external definition; labeled numerical intervals	Semantic interval representation	Source must be numerical; interval semantics must come from an external definition rather than thresholds inferred from the observed data
Referential	Reference interval	One numerical feature; an externally defined reference range; optional applicability conditions	Range state (below / within / above), signed deviation, and absolute deviation	Bounds must be ordered and unit-compatible; the reference must be applicable under the specified context
Relational	Arithmetic relation	Two compatible numerical features; difference, ratio, sum, or product	One derived numerical relation	Both operands must be valid numerical columns with compatible semantics / units for the requested operation
	Aggregate relation	A nonempty set of compatible features; sum, mean, product, or equality count	One aggregate feature	Referenced feature set must be nonempty and compatible with the requested aggregation
	Boolean relation	Typed comparison clauses over schema features, combined by conjunction or disjunction	Binary relation indicator	Comparisons and operands must belong to the predefined whitelist and satisfy their type constraints

resulting binary feature makes this overlapping market-segment concept explicit, rather than requiring the predictor to infer it independently from the original manufacturer identifier.

Referential knowledge. For turbocharged gasoline vehicles, LKE associates compression ratio with an external reference interval of $[8, 9]$. The resulting features encode whether an observed value lies below, within, or above the interval, together with its signed and absolute deviations. They therefore represent both the direction and magnitude of departure from the stated reference.

Relational knowledge. LKE combines horsepower and curb weight to construct a power-to-weight-ratio feature. This variable represents engine power relative to vehicle mass, a domain-relevant quantity that would otherwise need to be inferred from two separately represented measurements.

C.4 Dataset-Level Expansion Statistics

Table C2 reports the number of accepted knowledge units and generated columns for each dataset. Across the 20 datasets, the LLM produced 611 proposals, of which 519 passed structured validation. The compiler retained 604 generated columns after removing 141 invalid, constant, non-finite, or

Dataset	Original	Accepted	Generated	Final catalog
Automobile	25	39	49	74
Backdoor	42	48	43	85
Campaign	16	43	34	50
Cardiotocography	21	24	32	53
Census	41	39	37	78
Churn	19	13	22	41
Cirrhosis	17	29	40	57
Covertype	12	10	14	26
Credit	23	52	51	74
Equip	6	15	21	27
Gallstone	38	61	93	131
Glass	9	13	16	25
Glioma	23	11	17	40
Quasar	8	9	12	20
Seismic	18	22	32	50
Stroke	10	25	28	38
Vertebral	6	10	9	15
WBC	30	27	17	47
Wine	13	12	19	32
Yeast	8	17	18	26
Total	385	519	604	989
Mean	19.25	25.95	30.20	49.45

Table C2: Dataset-level LKE statistics using *deepseek-v4-flash*. "Accepted" denotes validated knowledge units, while "Generated" denotes retained derived columns.

duplicated candidate outputs. The number of accepted units differs from the number of generated columns because a referential unit can produce multiple outputs.

The number of retained generated columns ranges from 9 to 93, with a mean of 30.2 per dataset. No fixed feature quota is imposed; the final catalog size is determined by the number of distinct executable semantic transformations that pass the common validation procedure. All reported feature catalogs were generated and compiled without dataset-specific manual editing or selection based on downstream performance.

C.5 Controlled LKE Augmentation

To isolate the contribution of semantic knowledge expansion from that of the original features, we construct controlled feature groups for each dataset. Each group contains four original parent columns, and an LKE feature is included only when all of its declared source columns belong to the same group. This strict closure ensures that any performance difference between conditions is attributable to the introduced semantic features rather than information from unselected original columns. We retain up to eight valid groups per dataset, yielding 152 dataset–group pairs across the 20 datasets.

Each group is evaluated using the parent columns alone and their concatenation with the corresponding LKE features. The main analysis reports the paired AUCPR gain of the augmented condition over the parent-only condition across the six scarcity settings. Results are first averaged over the valid groups within each dataset and then averaged across datasets with equal weight. The feature groups are fixed before model fitting, and no group selection is performed.

D Semantic Subspace Discovery

The main paper explains how SSD organizes expanded features around reasonable anomaly mechanisms. This appendix specifies how these symbolic mechanisms and feature descriptions are converted into a complete set of trainable prediction views.

Algorithm 2 Semantic Subspace Discovery

Require: Task metadata D, symbolic feature catalog $\mathcal{F}$, maximum repair retries R
Ensure: Prediction views $\mathcal{S}$ and view metadata $\mathcal{M}^S$

Construct the closed feature identifier set from $\mathcal{F}$ and load the fixed assignment vocabularies.
Initialize the frozen stage context $\mathcal{C} \leftarrow (D, \mathcal{F})$.
for $g \in \{\text{ARCHITECT}, \text{MAPPER}, \text{REVIEWER}\}$ **do**
  Construct the stage-specific prompt from $\mathcal{C}$.
  Initialize $r \leftarrow 0$, $v \leftarrow \text{false}$, and diagnostics $\Delta \leftarrow \varnothing$.
  while $\neg v \wedge r \leq R$ **do**
    Request structured output $\mathcal{O}_g$ using the prompt and current diagnostics Δ.
    Validate required fields, closed identifiers, and the stage-specific constraints, obtaining v and Δ.
    if $\neg v$ **then**
      Set $r \leftarrow r + 1$.
    end if
  end while
  if $\neg v$ **then**
    abort with validation diagnostics Δ.
  end if
  Freeze $\mathcal{O}_g$ and append it to $\mathcal{C}$.
end for
Construct confirmed, singleton, and exploratory views from the frozen stage outputs.
Include core and supporting features and retain alternative proxies as metadata.
Merge views with identical trainable feature sets.
Verify that every catalog feature is covered by a mechanism or exploratory view.
Construct view metadata $\mathcal{M}^S$ in catalog order.
return $(\mathcal{S}, \mathcal{M}^S)$

D.1 SSD CONSTRUCTION

Inputs and information boundary. SSD receives the task description, anomaly definition, and symbolic feature catalog. Each catalog entry records its identifier, description, logical type, source columns, and LKE provenance when applicable. SSD does not access sample values, labels, empirical statistics, prediction outputs, feature importance, or downstream performance. The raw-feature control uses the same interface after removing the LKE-generated entries.

Discovery workflow. Algorithm 2 summarizes how SSD converts the symbolic feature catalog into mechanism-based and exploratory prediction views. The architect first proposes an accepted mechanism set, which is frozen after validation. The mapper then assigns features to these mechanisms and records residual features, while the reviewer resolves only eligible residual decisions without modifying the accepted mechanisms or primary assignments. If validation fails, only the current stage is repeated using the returned diagnostics; previously accepted outputs remain fixed.

The architect receives the task metadata and feature catalog, the mapper additionally receives the accepted mechanisms, and the reviewer receives the validated assignments and unresolved residual records. Each stage instantiates a fixed prompt template with the required symbolic context and structured-output schema. Figure D1 illustrates this construction for the semantic architect.

Feature roles and view construction. For each accepted mechanism, SSD classifies the associated features as core, supporting, or alternative proxy variables. Core and supporting features form the corresponding trainable prediction view, whereas alternative proxies are retained as auxiliary semantic information but are not included in that view.

Each accepted mechanism produces one mechanism-oriented view. A mechanism supported by only one trainable feature produces a single-feature view, while features that cannot be reliably associated with an accepted mechanism are organized into exploratory views. Mechanism views with identical feature sets are merged, and every catalog feature is included in either a mechanism-oriented or exploratory view.

Role Definition and Generation Principles

You are semantic architect proposer. You may write only proposals and consensus inside the exact root keys shown in the schema. Produce grounded proposals first, then adjudicate only those proposals into frozen mechanisms. …

Information Boundary and Semantic Context Policy

Do not use rows, feature values, label values, statistics, correlations, predictions, feature importance, or model performance. Treat descriptions as exact measurement contracts and preserve uncertainty, boundary conditions, competing explanations, and observability limitations.

Output Format Constraint and Schema

Return strict JSON only. Do not add root keys or use identifiers outside the closed constraints.
Schema: {"mechanisms": [{"id": "", "name": "short name",
"latent_factor": "one latent real-world state or process",
"tendency": "positive | negative | non_monotonic | conditional | uncertain",
…}, …

Dataset Context and Feature List

<<<DATASET_TASK>>> {“dataset”: “automobile”, “dataset_description”: “…”, “task_definition”: “…”}…
<<<FEATURE SCHEMA>>> {“name”: “drive-wheels”, “kind”: “categorical”, …}

Figure D1: Example prompt structure for the SSD semantic architect. Fixed role instructions, information-boundary constraints, and an output schema are combined with dataset-specific task context and feature metadata.

D.2 Subspace Construction Example

On the Automobile dataset, SSD identifies two confirmed mechanisms: luxury-make risk and high-performance-engine risk. The generated luxury-make indicator is assigned as a core feature of the first view, while the power-to-weight ratio is assigned as a core feature of the second. Vehicle volume is retained as supporting evidence rather than as an independent mechanism indicator.

The remaining features that cannot be grounded in either mechanism are placed in one exploratory view. The final construction therefore contains two mechanism views and one exploratory view, yielding three unique trainable prediction views. LKE provenance records how a generated feature is constructed, whereas SSD membership records how that feature is used in a prediction view.

D.3 Dataset-Level Construction Statistics

Table D1 summarizes the SSD constructions across the 20 datasets. Feature Count includes both original and LKE-generated features. Final Views counts the unique trainable views after merging mechanism subspaces with identical feature sets. No such merging occurs in these constructions, so Final Views equals the sum of Mechanism Subspaces and Exploratory Groups. Complete feature coverage holds by construction and is not reported separately.

D.4 Additional Details for the SSD Control Analysis

To determine whether the benefits of SSD arise from semantic organization rather than feature partitioning alone, we compare the semantic subspaces with shape-matched random subspaces. For each dataset, a random collection is constructed to match the semantic subspaces exactly in the number of views and their ordered view-size profile, while assigning features uniformly at random from the complete LKE feature catalog. Consequently, any performance difference reflects semantic organization rather than differences in subspace size or view count.

The random collections are generated once and kept fixed throughout the evaluation. Only the view shapes are matched; feature overlap, feature occurrence frequency, and the proportion of original and LKE features are left to the random assignment. Semantic and random subspaces are evaluated under identical experimental settings, and the reported results are based on the paired AUCPR gain of the semantic organization over its random counterpart.

Table D1: SSD construction statistics across the 20 datasets in the LKE feature space.

Dataset	# Feature	Mechanism	Exploratory	Views
Automobile	74	2	1	3
Backdoor	85	3	0	3
Campaign	50	2	0	2
Cardiotocography	53	3	0	3
Census	78	4	0	4
Churn	41	3	0	3
Cirrhosis	57	4	1	5
Covertype	26	3	1	4
Credit	74	3	1	4
Equip	27	4	1	5
Gallstone	131	5	1	6
Glass	25	2	1	3
Glioma	40	3	2	5
Quasar	20	2	1	3
Seismic	50	2	0	2
Stroke	38	4	1	5
Vertebral	15	3	0	3
WBC	47	2	1	3
Wine	32	5	1	6
Yeast	26	2	1	3

E SEMANTIC-GUIDED META-AGGREGATION

The key challenge addressed by SGMA is how to estimate the relative importance of prediction views when target supervision is insufficient for learning reliable aggregation weights. SGMA addresses this problem by learning from source datasets how view descriptors relate to useful aggregation weights, and then applying this relationship to a held-out target. This appendix makes this transfer process explicit, clarifies the information boundary between source-side learning and target-side inference, and examines whether the learned weighting rule is meaningful and stable.

E.1 TRAINING AND INFERENCE WORKFLOW

For each scarcity setting, the source tasks and the held-out target task are instantiated under the same training budget and label-visibility rule. Each semantic subspace defines one prediction member, and the full-feature predictor is included as an additional member. When dataset D_t is treated as the target, all of its splits and root seeds are excluded from meta-regressor training. The complete SGMA procedure is:

1. fit the semantic-view member predictors for each source task under the corresponding scarcity setting;
2. construct the descriptor vector for every source member;
3. fit a sparse nonnegative logistic teacher on each source dataset and normalize its coefficients to obtain the offline meta-training targets w_m^*;
4. train the meta-regressor on the resulting descriptor–target pairs, excluding the held-out target dataset;
5. construct the corresponding target-member descriptors, predict nonnegative aggregation weights, and combine the target member predictions.

Target evaluation labels are used only to compute the final evaluation metrics and do not enter meta-regressor training or target weight prediction. SGMA therefore transfers the relationship between member descriptors and aggregation weights, rather than source samples, feature columns, or fitted source predictors.

E.2 DESCRIPTOR DESIGN

For prediction view m and evaluation batch $\mathcal{B}_{\mathrm{ev}}$, SGMA constructs

$$u_m(\mathcal{B}_{\mathrm{ev}}) = \left[u_m^{\mathrm{sem}}, u_m^{\mathrm{disc}}, u_m^{\mathrm{pred}}(\mathcal{B}_{\mathrm{ev}}), u_m^{\mathrm{vis}}\right]. \tag{3}$$

All descriptor definitions and numerical mappings are fixed across datasets and scarcity settings.

Role Definition and Generation Principles

You are a semantic prior adjudicator. Score every frozen subspace below on the six closed ordinal vocabularies so downstream modeling can weigh subspaces by the strength of their semantic prior. Judge only from the frozen metadata artifacts supplied here: …

Information Boundary and Judge rules

The frozen metadata below is the only evidence you may use. …
Rate prior_tier = exploratory for an exploratory_group whose features are plausible to observe jointly for this task …

Output Format Constraint and Schema

Return strict JSON only under the exact root key "subspaces". …
Schema: {"subspaces": [{ "subspace_id": "exact frozen subspace id",
"subspace_type": "exact frozen subspace type",
"prior_tier": "weak | exploratory | plausible | primary",
"task_specificity": "low | medium | high",
"observability": "poor | partial | full",
"evidence_directness": "proxy_dominated | mixed | direct",
"boundary_dependence": "high | medium | low",
"competing_explanation_strength": " high | medium | low", …}]

Dataset Context and Semantic Views

<<<DATASET_TASK>>> {"dataset": "automobile", "dataset_description": "…", "task_definition": "…"}…
<<<VIEW SCHEMA>>> {"id": "view_id", "type": "confirmed_mechanism | exploratory_group", "feature_names": [Feature list under this view], "features": [Feature schemas], …}

Discovery Stability Provenance

- Mechanism stability
- Mapping agreement
- Feature Coverage
- …

Figure E1: The prompt for generating semantic evaluations.

Semantic descriptors. This family contains six LLM assessments: prior tier, observability, task specificity, competing explanations, boundary dependence, and evidence directness. Ordinal assessments are mapped to fixed numerical levels; free-form rationales are not used. The prompt for generating these semantic evaluations across views is shown in Figure E1.

Discovery descriptors. This family summarizes the corresponding SSD view, including its size, original/LKE feature composition, feature-role proportions, overlap with other views, and confirmed, exploratory, or singleton type.

Prediction-relation descriptors. This family describes the relationship between a view and the other views on the unlabeled evaluation batch. It includes the mean and maximum absolute Pearson correlation with other views, correlation with the leave-one-view-out mean, and Spearman agreement with the cross-view rank consensus.

Protocol-observable descriptors. This family contains quantities available under the active evaluation protocol, including training-set size, number of prediction views, and visible-supervision proxies. The proxies are derived from the percentile ranks of the visible anomalies under fixed kNN-distance and ECOD scores. True anomaly prevalence, hidden-label counts, oracle performance, and the reference weights themselves are excluded.

Because the prediction-relation descriptors are computed jointly on $\mathcal{B}_{\mathrm{ev}}$, the aggregation is transductive: the predicted weights may depend on the unlabeled batch being scored.

E.3 Meta-Training Targets and Meta-Regressor

For each source dataset, SGMA fits a sparse nonnegative logistic teacher to estimate the relative importance of its prediction views. The normalized coefficients define the offline meta-training targets w_m^*, which the meta-regressor learns to predict from view descriptors across datasets.

Meta-training targets. Let E denote the source rows used to fit the teacher, and let $p_m(x_i)$ be the prediction of view m for sample x_i. Probabilities are clipped to $[\epsilon, 1-\epsilon]$, where $\epsilon = 10^{-6}$, and transformed as $\ell_{im} = \mathrm{logit}(p_m^\epsilon(x_i))$. The teacher is defined by

$$q_i = \sigma\left(b + \sum_{m=1}^{M} a_m \ell_{im}\right),$$
$$\mathrm{BCE}(y, q) = -y \log q - (1-y)\log(1-q), \tag{4}$$
$$(a^*, b^*) = \arg\min_{a \geq 0,\, b} \left[\frac{1}{|E|}\sum_{i \in E} \mathrm{BCE}(y_i, q_i) + \lambda \|a\|_1\right].$$

Here, M includes the semantic-view predictors and the additional full-feature predictor. Nonnegativity restricts the coefficients to positive contributions, while the ℓ_1 penalty encourages sparse allocation across redundant views. The intercept is discarded because SGMA transfers relative view importance rather than source-task probability calibration. The constrained problem is solved using SLSQP with explicit nonnegative bounds on a, while the intercept b remains unconstrained.

Let $A = \sum_{j=1}^{M} [a_j^*]_+$. The offline meta-training targets are defined as

$$w_m^* = \begin{cases} [a_m^*]_+ / A, & A > 0, \\ 0, & A = 0. \end{cases} \tag{5}$$

They represent the relative importance of the prediction views on the corresponding source task. If the nonnegative coefficients sum to zero, all targets for that task remain zero.

The regularization coefficient is selected independently for each source dataset from $\Lambda = \{0, 10^{-4}, 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$ using stratified cross-validation and mean held-out log loss, with at most five folds. When stratification is infeasible, $\lambda = 10^{-3}$ is used.

Meta-regressor. Each source view contributes one training pair (u_m, w_m^*). For each scarcity setting, a separate meta-regressor is fitted using all source datasets except the held-out target and predicts $\widehat{w}_m = g_\phi(u_m)$. We use TabPFN with the *tabpfn-v2.5-regression-v2.5_default* checkpoint.

Because the full-feature predictor does not correspond to a single semantic mechanism, its input descriptor is the component-wise mean of the active descriptor columns across the semantic views from the same dataset, scarcity setting, and root seed.

E.4 Target Aggregation

For a target evaluation batch $\mathcal{B}_{\mathrm{ev}}$, the meta-regressor predicts an importance score $\widehat{w}_m = g_\phi(u_m(\mathcal{B}_{\mathrm{ev}}))$ for each prediction view. Let $Z = \sum_{j=1}^{M} [\widehat{w}_j]_+$. The aggregation weights are

$$\bar{w}_m(\mathcal{B}_{\mathrm{ev}}) = \begin{cases} [\widehat{w}_m]_+ / Z, & Z > 0, \\ 1/M, & Z = 0. \end{cases} \tag{6}$$

Thus, negative predictions are clipped to zero before normalization, and uniform aggregation is used when all clipped scores sum to zero.

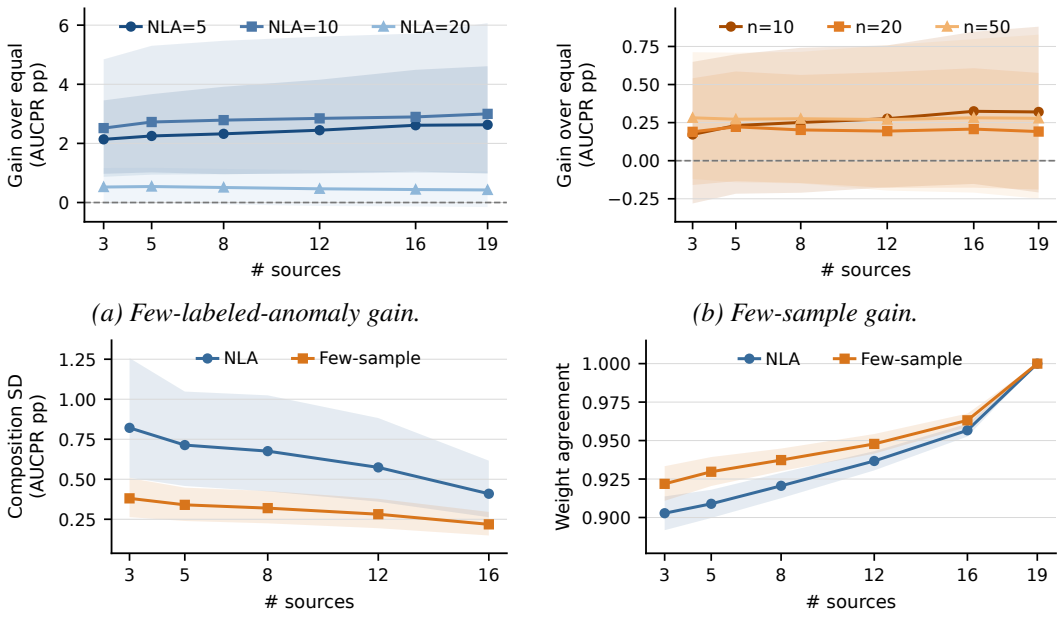
For probability-valued view predictors, probabilities are clipped to $[\epsilon, 1-\epsilon]$, with $\epsilon = 10^{-6}$, before applying the logit transform. The final anomaly score is

$$s(x; \mathcal{B}_{\mathrm{ev}}) = \sigma\left(\sum_{m=1}^{M} \bar{w}_m(\mathcal{B}_{\mathrm{ev}}) \,\mathrm{logit}(p_m^\epsilon(x))\right). \tag{7}$$

The resulting value is used for anomaly ranking rather than interpreted as a calibrated probability.

E.5 Additional Details for SGMA Analyses

The main paper reports two analyses of SGMA. We provide their evaluation protocols and implementation details below.



(a) Few-labeled-anomaly gain.

(b) Few-sample gain.

(c) Composition variability.

(d) Weight agreement.

Figure E2: Source-pool scaling of SGMA. Panels (a,b) show AUCPR gain over equal aggregation, (c) composition variability, and (d) agreement with the complete-pool weights. Shading denotes 95% target-dataset bootstrap intervals.

Semantic-descriptor audit. For each dataset and scarcity setting, semantic views are ranked separately by oracle AUCPR and by each of the six semantic descriptors. Average-tie percentile ranks are centered within datasets and then pooled to measure whether each descriptor preserves the relative quality ordering of views. The full-feature view is excluded because it does not correspond to a single semantic mechanism. Oracle AUCPR is used only in this analysis and never enters SGMA training or inference.

Two-sided p-values are estimated from 100,000 Monte Carlo permutations of the oracle ranks within each dataset. The six semantic descriptors form one prespecified hypothesis family within each scarcity setting, and their p-values are adjusted using Holm's procedure.

Target-label stacking comparison. The target-label stacker uses the same semantic-view predictors, full-feature predictor, training rows, and evaluation predictions as SGMA. It differs only in how the aggregation weights are obtained: SGMA predicts them through cross-dataset meta-learning, whereas the stacker fits them directly from the protocol-visible target labels.

The stacker constructs out-of-fold member predictions using $K = \min(5, n_{\min})$ stratified folds, where $n_{\min}$ is the minority-class count, and fits the sparse nonnegative logistic objective defined in Appendix E.3. When stratified cross-fitting is infeasible, the reported in-sample fallback is used.

E.6 Source-Pool Size and Stability

The main experiments use the complete leave-one-dataset-out source pool, consisting of the other 19 datasets for each held-out target. We examine whether SGMA depends strongly on the number or composition of these source tasks. We first vary the source-pool size to characterize performance, composition sensitivity, and weight convergence, and then remove individual sources to evaluate stability near the complete-pool configuration.

Scaling with Source-pool Size. For each held-out target dataset, we generate 50 random permutations of the remaining 19 datasets and use nested prefixes of size $m \in \{3, 5, 8, 12, 16, 19\}$ as source pools. The same permutations are shared across scarcity settings and root seeds, enabling

paired comparisons as the source pool grows. The complete $m = 19$ pool has only one possible composition and is evaluated once.

All upstream components are kept fixed, including the LKE features, SSD views, member predictions, SGMA descriptors, and source-side teacher weights. Only the descriptor-to-weight meta-regressor is refitted for each reduced source pool.

Figure E2 evaluates source-pool behavior from three complementary perspectives. In panels (a) and (b), values above zero indicate that SGMA improves over equal aggregation. In panel (c), lower values indicate less sensitivity to which source datasets are selected. In panel (d), values closer to one indicate that the reduced-pool meta-regressor assigns aggregation mass across the target prediction views similarly to the meta-regressor trained on all 19 source datasets. This panel measures stability of the target-side aggregation rule, not agreement with oracle view importance.

Panels (a) and (b) show that SGMA achieves positive average gains over equal aggregation even with small source pools, although the magnitude differs substantially between the two scarcity regimes. Under few-labeled-anomaly supervision, the gains are considerably larger, particularly for NLA $= 5$ and NLA $= 10$. Increasing the number of sources yields moderate additional improvements for these two settings, whereas the NLA $= 20$ curve remains approximately flat. Under few-sample supervision, the average gains are smaller and change only modestly with the pool size. Thus, increasing the source pool does not uniformly produce large improvements in mean AUCPR.

The clearer effect of additional sources is improved stability. Panel (c) shows that variation across different source compositions decreases steadily as the source pool grows. This trend is especially pronounced under few-labeled-anomaly supervision, where small pools are more sensitive to which datasets are included. At the same time, panel (d) shows a monotonic increase in agreement with the complete-pool weight vector. With 16 sources, the agreement exceeds approximately 0.95 in both scarcity regimes, while the value at 19 sources equals one by construction.

These results indicate that the main benefit of increasing source-task availability is not a large or strictly monotonic increase in mean performance. Instead, larger source pools reduce dependence on the particular source composition and produce more stable aggregation weights. The relatively small changes between 16 and 19 sources suggest that the complete configuration lies near the stable end of the observed source-pool scaling trend.

Stability Near the Complete Source Pool. To test whether the complete SGMA result depends on any particular source task, we remove each source dataset in turn and refit the meta-regressor using the remaining 18 sources, while keeping all target-side data, semantic views, predictions, descriptors, teacher weights, and experimental settings fixed. In Figure E3, proximity between the open and filled markers indicates limited sensitivity to the removed source, while filled markers above zero indicate that SGMA retains advantage over equal aggregation.

The gain remains positive after every deletion, ranging from 0.77 to 1.29 AUCPR percentage points. The largest average decrease is only 0.07 points, and the mean absolute change is 0.19 points. Combined with the decreasing composition variability and increasing weight agreement observed as the source pool grows, these results indicate that the complete 19-source configuration is stable and is not driven by any single source dataset.

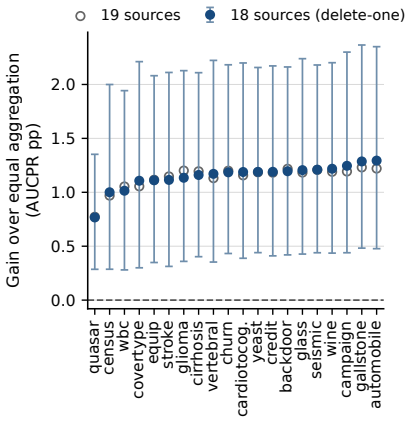


Figure E3: SGMA gain after deleting individual source datasets. Bars are 95% target-bootstrap intervals.

F Extension to Unsupervised Anomaly Detection

F.1 Unsupervised Setting and Method Adaptation

We further evaluate whether LENS-AD can be extended to unsupervised anomaly detection, where no labels from the evaluated target dataset are used for model fitting or weight prediction. The same

Table F1: Unsupervised anomaly detection on 20 datasets. Full-feature baselines operate directly on the original feature space, whereas all LENS-AD variants use the complete LKE, SSD, and SGMA pipeline. Avg. Rank reports the average rank across the four training-set sizes.

Method	$n = 10$	$n = 20$	$n = 50$	Full	Avg. Rank
PCA	$0.313_{\pm 0.066}$ (10)	$0.317_{\pm 0.060}$ (10)	$0.295_{\pm 0.036}$ (11)	$0.287_{\pm 0.028}$ (10)	10.250
IForest	$0.337_{\pm 0.067}$ (5)	$0.360_{\pm 0.061}$ (4)	$0.359_{\pm 0.037}$ (3)	$0.358_{\pm 0.033}$ (3)	3.750
LOF	$0.289_{\pm 0.048}$ (11)	$0.279_{\pm 0.051}$ (11)	$0.339_{\pm 0.034}$ (6)	$0.264_{\pm 0.025}$ (11)	9.750
OCSVM	$0.329_{\pm 0.056}$ (7)	$0.327_{\pm 0.035}$ (8)	$0.325_{\pm 0.031}$ (9)	$0.325_{\pm 0.022}$ (8)	8.000
ECOD	$0.324_{\pm 0.023}$ (8)	$0.328_{\pm 0.022}$ (7)	$0.328_{\pm 0.021}$ (8)	$0.330_{\pm 0.025}$ (6)	7.250
COPOD	$0.348_{\pm 0.036}$ (4)	$0.361_{\pm 0.028}$ (3)	$0.359_{\pm 0.026}$ (2)	$0.352_{\pm 0.028}$ (4)	3.250
KNN	$0.352_{\pm 0.037}$ (3)	$0.342_{\pm 0.031}$ (5)	$0.336_{\pm 0.031}$ (7)	$0.330_{\pm 0.022}$ (7)	5.500
TabPFN-Unsupervised	$0.314_{\pm 0.069}$ (9)	$0.329_{\pm 0.051}$ (6)	$0.325_{\pm 0.040}$ (10)	$0.323_{\pm 0.025}$ (9)	8.500
LENS-AD (TabPFN)	$0.337_{\pm 0.080}$ (6)	$0.321_{\pm 0.048}$ (9)	$0.344_{\pm 0.040}$ (5)	$0.333_{\pm 0.032}$ (5)	6.250
LENS-AD (COPOD)	$0.359_{\pm 0.039}$ (2)	$0.363_{\pm 0.026}$ (2)	$0.358_{\pm 0.031}$ (4)	$0.374_{\pm 0.030}$ (2)	2.500
LENS-AD (IForest)	**$0.360_{\pm 0.072}$** (1)	**$0.365_{\pm 0.063}$** (1)	**$0.369_{\pm 0.044}$** (1)	**$0.404_{\pm 0.031}$** (1)	**1.000**

benchmark splits as in the main experiments are adopted, except that all target training labels are discarded before fitting anomaly scorers. Evaluation labels are used only for computing AUCPR.

LKE and SSD are reused without modification. Each semantic subspace and the all-feature view are scored independently using an unsupervised anomaly detector. The supervised target-side descriptors used by SGMA are replaced by four deterministic label-free distributional descriptors derived from KNN and ECOD scores, while the semantic, discovery, and cross-view agreement descriptors remain unchanged. The resulting anomaly scores are aggregated directly on their native scale rather than through probability-logit transformation.

This experiment evaluates whether the semantic pipeline remains useful when the target-task training rows are supplied without annotations. LKE and SSD are reused unchanged, view predictors are replaced by unsupervised anomaly scorers, and SGMA uses a label-free descriptor set. Target annotations are introduced only when computing the final AUCPR.

F.2 Scorers and Baselines

We extend LENS-AD with three representative unsupervised member scorers: Isolation Forest, COPOD, and TabPFN-Unsupervised. Each scorer is applied independently to every semantic subspace and to the all-feature view.

For comparison, we evaluate the corresponding full-feature detectors together with representative unsupervised baselines including PCA, LOF, OCSVM, ECOD, KNN, Isolation Forest, COPOD, and TabPFN-Unsupervised.

F.3 Results and Discussion

Table F1 shows that the semantic organization pipeline remains effective in the absence of target supervision. The Isolation-Forest instantiation consistently achieves the best average performance across all four training-set sizes, while the COPOD instantiation ranks second overall. The TabPFN-Unsupervised instantiation also consistently improves upon its corresponding full-feature detector.

These results indicate that the benefits of LENS-AD are not tied to supervised base predictors. Although the member scorers become entirely unsupervised, semantic knowledge expansion, semantic subspace discovery, and cross-dataset semantic aggregation continue to improve anomaly detection. Complete dataset-level results are provided in Appendix G.

G Complete Dataset-Level Results

This section reports the complete dataset-level AUCPR results underlying the aggregate comparisons presented in the main paper and Appendix F. Each entry reports the mean and sample standard deviation over three seeds in the form mean $\pm$ std. The parenthesized number denotes the within-

dataset rank across all methods in the table, with the best and second-best results highlighted in bold and underlined, respectively. Aggregate means and mean ranks are computed from the unrounded dataset-level means. No dataset-level significance tests are reported.

G.1 Few-Labeled-Anomaly Results

Table G1: Dataset-level AUCPR for NLA = 5.

(a) Conventional and supervised baselines

Dataset	DevNet	PReNet	XGBOD	CAD	WSAD-DT	XGBoost	CatBoost
automobile	0.353$_{\pm 0.064}$ (12)	0.412$_{\pm 0.065}$ (10)	0.473$_{\pm 0.114}$ (8)	0.274$_{\pm 0.030}$ (14)	0.333$_{\pm 0.078}$ (13)	0.454$_{\pm 0.150}$ (9)	0.635$_{\pm 0.083}$ (3)
backdoor	0.012$_{\pm 0.009}$ (11)	0.010$_{\pm 0.010}$ (12)	0.245$_{\pm 0.063}$ (7)	0.006$_{\pm 0.000}$ (14)	0.007$_{\pm 0.001}$ (13)	0.216$_{\pm 0.087}$ (9)	0.263$_{\pm 0.151}$ (6)
campaign	0.322$_{\pm 0.079}$ (10)	0.282$_{\pm 0.026}$ (13)	0.329$_{\pm 0.032}$ (9)	0.285$_{\pm 0.026}$ (12)	0.310$_{\pm 0.014}$ (11)	0.362$_{\pm 0.028}$ (7)	0.374$_{\pm 0.091}$ (5)
cardiotocography	0.584$_{\pm 0.087}$ (7)	0.527$_{\pm 0.065}$ (11)	0.538$_{\pm 0.141}$ (9)	0.420$_{\pm 0.087}$ (12)	0.336$_{\pm 0.008}$ (13)	0.557$_{\pm 0.051}$ (8)	0.643$_{\pm 0.046}$ (4)
census	0.313$_{\pm 0.030}$ (3)	0.217$_{\pm 0.077}$ (9)	0.209$_{\pm 0.063}$ (11)	0.059$_{\pm 0.002}$ (14)	0.146$_{\pm 0.050}$ (13)	0.230$_{\pm 0.032}$ (8)	0.266$_{\pm 0.040}$ (6)
churn	**0.422**$_{\pm 0.198}$ (1)	0.387$_{\pm 0.160}$ (4)	0.331$_{\pm 0.049}$ (10)	0.272$_{\pm 0.045}$ (13)	0.202$_{\pm 0.004}$ (14)	0.344$_{\pm 0.072}$ (6)	0.335$_{\pm 0.078}$ (9)
cirrhosis	0.555$_{\pm 0.179}$ (10)	0.471$_{\pm 0.048}$ (13)	0.595$_{\pm 0.204}$ (9)	0.457$_{\pm 0.023}$ (14)	0.544$_{\pm 0.040}$ (12)	0.551$_{\pm 0.133}$ (11)	0.700$_{\pm 0.105}$ (5)
covertype	0.319$_{\pm 0.158}$ (11)	0.163$_{\pm 0.134}$ (12)	0.810$_{\pm 0.045}$ (6)	0.147$_{\pm 0.045}$ (13)	0.005$_{\pm 0.000}$ (14)	0.718$_{\pm 0.039}$ (9)	0.813$_{\pm 0.016}$ (5)
credit	0.249$_{\pm 0.043}$ (9)	0.242$_{\pm 0.023}$ (10)	0.274$_{\pm 0.046}$ (4)	0.228$_{\pm 0.012}$ (11)	0.184$_{\pm 0.003}$ (12)	0.260$_{\pm 0.039}$ (6)	0.255$_{\pm 0.008}$ (7)
equip	0.406$_{\pm 0.053}$ (12)	0.326$_{\pm 0.091}$ (14)	0.821$_{\pm 0.024}$ (7)	0.376$_{\pm 0.037}$ (13)	0.494$_{\pm 0.019}$ (11)	0.674$_{\pm 0.031}$ (9)	0.773$_{\pm 0.032}$ (8)
gallstone	0.437$_{\pm 0.088}$ (11)	0.415$_{\pm 0.046}$ (12)	0.508$_{\pm 0.112}$ (7)	0.345$_{\pm 0.006}$ (14)	0.370$_{\pm 0.046}$ (13)	0.550$_{\pm 0.094}$ (5)	0.528$_{\pm 0.074}$ (6)
glass	0.750$_{\pm 0.148}$ (8)	0.720$_{\pm 0.107}$ (10)	0.729$_{\pm 0.142}$ (9)	0.223$_{\pm 0.014}$ (13)	**0.882**$_{\pm 0.101}$ (1)	0.707$_{\pm 0.121}$ (11)	0.826$_{\pm 0.129}$ (4)
glioma	**0.698**$_{\pm 0.027}$ (1)	0.569$_{\pm 0.041}$ (10)	0.529$_{\pm 0.085}$ (11)	0.419$_{\pm 0.291}$ (13)	0.657$_{\pm 0.015}$ (2)	0.601$_{\pm 0.043}$ (7)	0.615$_{\pm 0.060}$ (5)
quasar	0.376$_{\pm 0.113}$ (12)	0.511$_{\pm 0.140}$ (8)	0.574$_{\pm 0.203}$ (6)	0.184$_{\pm 0.012}$ (14)	0.207$_{\pm 0.001}$ (13)	0.515$_{\pm 0.099}$ (7)	0.736$_{\pm 0.098}$ (3)
seismic	0.082$_{\pm 0.044}$ (12)	0.117$_{\pm 0.046}$ (11)	0.127$_{\pm 0.022}$ (10)	0.069$_{\pm 0.003}$ (13)	0.161$_{\pm 0.020}$ (6)	0.131$_{\pm 0.037}$ (9)	0.149$_{\pm 0.056}$ (7)
stroke	0.085$_{\pm 0.021}$ (8)	0.073$_{\pm 0.034}$ (11)	0.072$_{\pm 0.023}$ (12)	0.107$_{\pm 0.059}$ (4)	**0.143**$_{\pm 0.014}$ (1)	0.096$_{\pm 0.023}$ (5)	0.074$_{\pm 0.029}$ (10)
vertebral	0.625$_{\pm 0.161}$ (3)	0.587$_{\pm 0.128}$ (5)	0.564$_{\pm 0.182}$ (10)	0.422$_{\pm 0.152}$ (13)	0.235$_{\pm 0.010}$ (14)	0.510$_{\pm 0.178}$ (11)	0.574$_{\pm 0.094}$ (8)
wbc	0.830$_{\pm 0.086}$ (2)	0.827$_{\pm 0.066}$ (3)	0.634$_{\pm 0.151}$ (12)	0.473$_{\pm 0.202}$ (14)	**0.957**$_{\pm 0.010}$ (1)	0.627$_{\pm 0.102}$ (13)	0.757$_{\pm 0.170}$ (8)
wine	0.693$_{\pm 0.242}$ (12)	0.915$_{\pm 0.075}$ (4)	0.832$_{\pm 0.114}$ (10)	0.268$_{\pm 0.016}$ (13)	0.237$_{\pm 0.022}$ (14)	0.767$_{\pm 0.174}$ (11)	0.864$_{\pm 0.084}$ (7)
yeast	**0.615**$_{\pm 0.095}$ (1)	0.595$_{\pm 0.099}$ (2)	0.420$_{\pm 0.128}$ (11)	0.101$_{\pm 0.078}$ (14)	0.552$_{\pm 0.072}$ (4)	0.362$_{\pm 0.195}$ (12)	0.491$_{\pm 0.082}$ (9)
Mean	0.436	0.418	0.481	0.257	0.348	0.462	0.533
Mean Rank	8.150	9.650	9.200	13.100	9.950	8.850	6.500
#1st	3	0	0	0	3	0	0

(b) Pretrained, LLM-based, and LENS-AD methods

Dataset	TabPFN	LimiX	TabICL	CAAFE	FeatLLM	LENS-AD (TabPFN)	LENS-AD (TabICL)
automobile	0.570$_{\pm 0.261}$ (7)	0.612$_{\pm 0.171}$ (5)	0.587$_{\pm 0.204}$ (6)	0.629$_{\pm 0.046}$ (4)	0.355$_{\pm 0.104}$ (11)	**0.641**$_{\pm 0.116}$ (1)	0.640$_{\pm 0.207}$ (2)
backdoor	**0.509**$_{\pm 0.124}$ (1)	0.476$_{\pm 0.057}$ (2)	0.466$_{\pm 0.099}$ (3)	0.231$_{\pm 0.153}$ (8)	0.142$_{\pm 0.100}$ (10)	0.401$_{\pm 0.092}$ (5)	0.454$_{\pm 0.032}$ (4)
campaign	0.367$_{\pm 0.021}$ (6)	**0.426**$_{\pm 0.045}$ (1)	0.393$_{\pm 0.016}$ (4)	0.405$_{\pm 0.074}$ (3)	0.310$_{\pm 0.091}$ (11)	0.351$_{\pm 0.025}$ (8)	0.406$_{\pm 0.020}$ (2)
cardiotocography	0.619$_{\pm 0.145}$ (6)	0.645$_{\pm 0.079}$ (3)	**0.683**$_{\pm 0.122}$ (1)	0.631$_{\pm 0.039}$ (5)	0.530$_{\pm 0.159}$ (10)	0.643$_{\pm 0.129}$ (4)	0.673$_{\pm 0.174}$ (2)
census	0.163$_{\pm 0.062}$ (12)	0.300$_{\pm 0.014}$ (4)	0.342$_{\pm 0.012}$ (2)	0.252$_{\pm 0.032}$ (7)	0.215$_{\pm 0.060}$ (10)	0.269$_{\pm 0.020}$ (5)	**0.364**$_{\pm 0.094}$ (1)
churn	0.314$_{\pm 0.056}$ (12)	0.341$_{\pm 0.072}$ (8)	0.379$_{\pm 0.095}$ (5)	0.343$_{\pm 0.071}$ (7)	0.419$_{\pm 0.057}$ (2)	0.328$_{\pm 0.081}$ (11)	0.395$_{\pm 0.140}$ (3)
cirrhosis	0.658$_{\pm 0.193}$ (8)	0.664$_{\pm 0.215}$ (7)	0.694$_{\pm 0.164}$ (6)	0.734$_{\pm 0.070}$ (2)	**0.768**$_{\pm 0.047}$ (1)	0.716$_{\pm 0.082}$ (3)	0.715$_{\pm 0.112}$ (4)
covertype	0.770$_{\pm 0.054}$ (8)	0.893$_{\pm 0.008}$ (3)	0.854$_{\pm 0.031}$ (4)	0.771$_{\pm 0.026}$ (7)	0.323$_{\pm 0.141}$ (10)	**0.948**$_{\pm 0.007}$ (1)	0.944$_{\pm 0.025}$ (2)
credit	0.253$_{\pm 0.029}$ (8)	**0.291**$_{\pm 0.074}$ (1)	0.253$_{\pm 0.034}$ (8)	0.261$_{\pm 0.027}$ (5)	0.255$_{\pm 0.035}$ (7)	0.281$_{\pm 0.037}$ (3)	0.289$_{\pm 0.064}$ (2)
equip	0.867$_{\pm 0.057}$ (6)	0.911$_{\pm 0.016}$ (2)	**0.915**$_{\pm 0.012}$ (1)	0.869$_{\pm 0.029}$ (5)	0.618$_{\pm 0.193}$ (10)	0.887$_{\pm 0.022}$ (4)	0.902$_{\pm 0.013}$ (3)
gallstone	0.654$_{\pm 0.068}$ (2)	0.560$_{\pm 0.074}$ (3)	**0.687**$_{\pm 0.094}$ (1)	0.480$_{\pm 0.017}$ (9)	0.455$_{\pm 0.077}$ (10)	0.507$_{\pm 0.070}$ (8)	0.553$_{\pm 0.085}$ (4)
glass	0.820$_{\pm 0.100}$ (5)	0.808$_{\pm 0.105}$ (7)	0.809$_{\pm 0.096}$ (6)	0.826$_{\pm 0.146}$ (4)	0.520$_{\pm 0.179}$ (12)	0.845$_{\pm 0.089}$ (2)	0.843$_{\pm 0.093}$ (3)
glioma	0.579$_{\pm 0.072}$ (9)	0.595$_{\pm 0.036}$ (8)	0.608$_{\pm 0.100}$ (6)	0.619$_{\pm 0.011}$ (4)	0.492$_{\pm 0.137}$ (12)	0.638$_{\pm 0.101}$ (3)	0.615$_{\pm 0.111}$ (5)
quasar	0.441$_{\pm 0.069}$ (10)	0.749$_{\pm 0.013}$ (2)	0.717$_{\pm 0.068}$ (4)	0.650$_{\pm 0.180}$ (5)	0.461$_{\pm 0.291}$ (9)	0.428$_{\pm 0.021}$ (11)	**0.789**$_{\pm 0.104}$ (1)
seismic	**0.172**$_{\pm 0.017}$ (1)	0.171$_{\pm 0.022}$ (2)	0.171$_{\pm 0.018}$ (2)	0.164$_{\pm 0.053}$ (4)	0.162$_{\pm 0.017}$ (5)	0.167$_{\pm 0.030}$ (3)	0.141$_{\pm 0.053}$ (8)
stroke	0.088$_{\pm 0.008}$ (6)	0.096$_{\pm 0.027}$ (5)	0.088$_{\pm 0.031}$ (6)	0.083$_{\pm 0.026}$ (9)	0.119$_{\pm 0.027}$ (3)	0.131$_{\pm 0.056}$ (2)	0.087$_{\pm 0.061}$ (7)
vertebral	0.581$_{\pm 0.083}$ (6)	0.580$_{\pm 0.114}$ (7)	0.629$_{\pm 0.123}$ (2)	0.569$_{\pm 0.129}$ (9)	0.434$_{\pm 0.103}$ (12)	0.593$_{\pm 0.104}$ (4)	**0.663**$_{\pm 0.138}$ (1)
wbc	0.741$_{\pm 0.202}$ (10)	0.668$_{\pm 0.147}$ (11)	0.759$_{\pm 0.219}$ (7)	0.754$_{\pm 0.212}$ (9)	0.793$_{\pm 0.245}$ (5)	0.813$_{\pm 0.218}$ (4)	0.780$_{\pm 0.232}$ (6)
wine	0.857$_{\pm 0.083}$ (8)	0.856$_{\pm 0.015}$ (9)	0.905$_{\pm 0.077}$ (5)	0.875$_{\pm 0.057}$ (6)	0.979$_{\pm 0.005}$ (3)	0.991$_{\pm 0.008}$ (2)	**0.994**$_{\pm 0.007}$ (1)
yeast	0.515$_{\pm 0.126}$ (7)	0.508$_{\pm 0.173}$ (8)	0.516$_{\pm 0.147}$ (6)	0.448$_{\pm 0.043}$ (10)	0.109$_{\pm 0.082}$ (13)	0.547$_{\pm 0.108}$ (5)	0.564$_{\pm 0.118}$ (3)
Mean	0.527	0.558	0.573	0.530	0.423	0.556	0.591
Mean Rank	7.200	5.050	4.550	6.300	8.600	4.550	3.350
#1st	2	2	3	0	1	2	4

Table G2: Dataset-level AUCPR for NLA = 10.

(a) Conventional and supervised baselines

Dataset	DevNet	PReNet	XGBOD	CAD	WSAD-DT	XGBoost	CatBoost
automobile	0.296$_{\pm 0.060}$ (13)	0.419$_{\pm 0.048}$ (10)	0.470$_{\pm 0.140}$ (9)	0.267$_{\pm 0.010}$ (14)	0.333$_{\pm 0.078}$ (12)	0.595$_{\pm 0.202}$ (8)	0.653$_{\pm 0.102}$ (3)
backdoor	0.009$_{\pm 0.001}$ (11)	0.006$_{\pm 0.006}$ (13)	0.419$_{\pm 0.069}$ (8)	0.005$_{\pm 0.002}$ (14)	0.007$_{\pm 0.002}$ (12)	0.388$_{\pm 0.064}$ (9)	0.507$_{\pm 0.061}$ (7)
campaign	0.335$_{\pm 0.101}$ (9)	0.297$_{\pm 0.036}$ (13)	0.417$_{\pm 0.047}$ (8)	0.308$_{\pm 0.059}$ (12)	0.309$_{\pm 0.013}$ (11)	0.420$_{\pm 0.024}$ (7)	0.485$_{\pm 0.040}$ (5)
cardiotocography	0.684$_{\pm 0.052}$ (7)	0.670$_{\pm 0.023}$ (8)	0.647$_{\pm 0.029}$ (9)	0.523$_{\pm 0.169}$ (12)	0.338$_{\pm 0.009}$ (13)	0.577$_{\pm 0.082}$ (10)	0.722$_{\pm 0.033}$ (5)
census	0.239$_{\pm 0.097}$ (10)	0.271$_{\pm 0.073}$ (8)	0.290$_{\pm 0.098}$ (5)	0.066$_{\pm 0.008}$ (14)	0.146$_{\pm 0.031}$ (13)	0.259$_{\pm 0.068}$ (9)	0.285$_{\pm 0.077}$ (7)
churn	0.274$_{\pm 0.090}$ (13)	**0.542**$_{\pm 0.067}$ (1)	0.361$_{\pm 0.046}$ (11)	0.342$_{\pm 0.089}$ (12)	0.202$_{\pm 0.004}$ (14)	0.370$_{\pm 0.056}$ (9)	0.392$_{\pm 0.072}$ (6)
cirrhosis	0.589$_{\pm 0.026}$ (10)	0.476$_{\pm 0.042}$ (12)	0.647$_{\pm 0.026}$ (9)	0.406$_{\pm 0.049}$ (13)	0.544$_{\pm 0.040}$ (11)	0.647$_{\pm 0.081}$ (9)	0.731$_{\pm 0.024}$ (6)
covertype	0.544$_{\pm 0.167}$ (9)	0.152$_{\pm 0.160}$ (12)	0.833$_{\pm 0.057}$ (7)	0.173$_{\pm 0.018}$ (11)	0.005$_{\pm 0.000}$ (13)	0.749$_{\pm 0.097}$ (8)	0.878$_{\pm 0.056}$ (5)
credit	0.234$_{\pm 0.048}$ (11)	0.232$_{\pm 0.037}$ (12)	0.280$_{\pm 0.032}$ (4)	0.227$_{\pm 0.009}$ (13)	0.184$_{\pm 0.003}$ (14)	0.268$_{\pm 0.012}$ (7)	0.260$_{\pm 0.023}$ (9)
equip	0.480$_{\pm 0.141}$ (12)	0.478$_{\pm 0.146}$ (13)	0.848$_{\pm 0.059}$ (6)	0.422$_{\pm 0.153}$ (14)	0.492$_{\pm 0.017}$ (11)	0.720$_{\pm 0.082}$ (9)	0.799$_{\pm 0.071}$ (8)
gallstone	0.523$_{\pm 0.036}$ (10)	0.513$_{\pm 0.053}$ (11)	0.551$_{\pm 0.067}$ (9)	0.334$_{\pm 0.028}$ (14)	0.370$_{\pm 0.046}$ (13)	0.591$_{\pm 0.016}$ (8)	0.614$_{\pm 0.051}$ (6)
glass	0.766$_{\pm 0.127}$ (10)	0.803$_{\pm 0.156}$ (9)	0.849$_{\pm 0.085}$ (5)	0.231$_{\pm 0.000}$ (12)	0.846$_{\pm 0.159}$ (6)	0.849$_{\pm 0.066}$ (5)	0.881$_{\pm 0.083}$ (2)
glioma	**0.728**$_{\pm 0.095}$ (1)	0.654$_{\pm 0.136}$ (7)	0.586$_{\pm 0.022}$ (12)	0.720$_{\pm 0.027}$ (2)	0.692$_{\pm 0.112}$ (4)	0.610$_{\pm 0.014}$ (11)	0.626$_{\pm 0.055}$ (10)
quasar	0.629$_{\pm 0.137}$ (10)	0.670$_{\pm 0.109}$ (8)	0.757$_{\pm 0.075}$ (6)	0.171$_{\pm 0.010}$ (14)	0.207$_{\pm 0.001}$ (13)	0.577$_{\pm 0.091}$ (11)	0.763$_{\pm 0.120}$ (5)
seismic	0.057$_{\pm 0.012}$ (13)	0.141$_{\pm 0.023}$ (8)	0.126$_{\pm 0.054}$ (10)	0.078$_{\pm 0.013}$ (12)	0.161$_{\pm 0.019}$ (5)	0.110$_{\pm 0.021}$ (11)	0.135$_{\pm 0.036}$ (9)
stroke	0.150$_{\pm 0.028}$ (5)	0.135$_{\pm 0.009}$ (8)	0.109$_{\pm 0.025}$ (10)	0.105$_{\pm 0.064}$ (11)	0.143$_{\pm 0.014}$ (6)	0.104$_{\pm 0.013}$ (12)	0.101$_{\pm 0.014}$ (13)
vertebral	0.705$_{\pm 0.107}$ (2)	0.643$_{\pm 0.101}$ (7)	0.566$_{\pm 0.063}$ (10)	0.442$_{\pm 0.123}$ (12)	0.235$_{\pm 0.009}$ (14)	0.553$_{\pm 0.028}$ (11)	0.635$_{\pm 0.033}$ (8)
wbc	0.867$_{\pm 0.073}$ (3)	0.861$_{\pm 0.079}$ (4)	0.786$_{\pm 0.106}$ (10)	0.552$_{\pm 0.160}$ (13)	**0.957**$_{\pm 0.010}$ (1)	0.731$_{\pm 0.113}$ (12)	0.851$_{\pm 0.067}$ (6)
wine	0.778$_{\pm 0.101}$ (11)	0.994$_{\pm 0.010}$ (2)	0.908$_{\pm 0.094}$ (9)	0.307$_{\pm 0.050}$ (12)	0.237$_{\pm 0.022}$ (13)	0.893$_{\pm 0.145}$ (10)	0.955$_{\pm 0.062}$ (7)
yeast	0.655$_{\pm 0.084}$ (2)	0.649$_{\pm 0.086}$ (3)	0.478$_{\pm 0.139}$ (11)	0.108$_{\pm 0.082}$ (14)	0.608$_{\pm 0.012}$ (5)	0.435$_{\pm 0.041}$ (12)	0.546$_{\pm 0.035}$ (9)
Mean	0.477	0.480	0.546	0.289	0.351	0.522	0.591
Mean Rank	9.000	8.850	8.900	12.700	10.600	9.800	7.100
#1st	1	1	0	0	1	0	0

(b) Pretrained, LLM-based, and LENS-AD methods

Dataset	TabPFN	LimiX	TabICL	CAAFE	FeatLLM	LENS-AD (TabPFN)	LENS-AD (TabICL)
automobile	0.606$_{\pm 0.076}$ (7)	0.616$_{\pm 0.114}$ (5)	0.609$_{\pm 0.137}$ (6)	0.652$_{\pm 0.160}$ (4)	0.354$_{\pm 0.104}$ (11)	**0.675**$_{\pm 0.038}$ (1)	0.666$_{\pm 0.100}$ (2)
backdoor	0.662$_{\pm 0.054}$ (5)	**0.689**$_{\pm 0.089}$ (1)	0.663$_{\pm 0.065}$ (4)	0.527$_{\pm 0.061}$ (6)	0.133$_{\pm 0.086}$ (10)	0.664$_{\pm 0.070}$ (3)	0.679$_{\pm 0.057}$ (2)
campaign	0.501$_{\pm 0.014}$ (4)	0.515$_{\pm 0.068}$ (2)	0.515$_{\pm 0.067}$ (2)	0.482$_{\pm 0.033}$ (6)	0.310$_{\pm 0.091}$ (10)	0.511$_{\pm 0.065}$ (3)	**0.517**$_{\pm 0.057}$ (1)
cardiotocography	0.741$_{\pm 0.040}$ (3)	0.762$_{\pm 0.020}$ (2)	**0.785**$_{\pm 0.013}$ (1)	0.699$_{\pm 0.012}$ (6)	0.529$_{\pm 0.160}$ (11)	0.728$_{\pm 0.019}$ (4)	0.722$_{\pm 0.061}$ (5)
census	0.231$_{\pm 0.059}$ (11)	0.368$_{\pm 0.037}$ (3)	0.422$_{\pm 0.037}$ (2)	0.289$_{\pm 0.086}$ (6)	0.215$_{\pm 0.060}$ (12)	0.343$_{\pm 0.033}$ (4)	**0.425**$_{\pm 0.044}$ (1)
churn	0.369$_{\pm 0.085}$ (10)	0.385$_{\pm 0.080}$ (8)	0.428$_{\pm 0.099}$ (3)	0.396$_{\pm 0.060}$ (5)	0.419$_{\pm 0.057}$ (4)	0.387$_{\pm 0.085}$ (7)	0.434$_{\pm 0.114}$ (2)
cirrhosis	0.692$_{\pm 0.078}$ (8)	0.712$_{\pm 0.115}$ (7)	0.752$_{\pm 0.062}$ (4)	0.740$_{\pm 0.016}$ (5)	0.767$_{\pm 0.044}$ (2)	0.762$_{\pm 0.043}$ (3)	**0.776**$_{\pm 0.049}$ (1)
covertype	0.941$_{\pm 0.029}$ (2)	0.894$_{\pm 0.042}$ (4)	0.911$_{\pm 0.040}$ (3)	0.837$_{\pm 0.017}$ (6)	0.323$_{\pm 0.140}$ (10)	**0.964**$_{\pm 0.022}$ (1)	**0.964**$_{\pm 0.021}$ (1)
credit	0.263$_{\pm 0.037}$ (8)	0.300$_{\pm 0.054}$ (2)	0.272$_{\pm 0.031}$ (6)	0.273$_{\pm 0.030}$ (5)	0.255$_{\pm 0.035}$ (10)	**0.302**$_{\pm 0.030}$ (1)	0.298$_{\pm 0.058}$ (3)
equip	0.870$_{\pm 0.052}$ (5)	0.907$_{\pm 0.024}$ (2)	**0.921**$_{\pm 0.003}$ (1)	0.828$_{\pm 0.038}$ (7)	0.618$_{\pm 0.193}$ (10)	0.887$_{\pm 0.011}$ (4)	0.904$_{\pm 0.005}$ (3)
gallstone	**0.724**$_{\pm 0.126}$ (1)	0.666$_{\pm 0.110}$ (4)	0.721$_{\pm 0.139}$ (2)	0.596$_{\pm 0.122}$ (7)	0.455$_{\pm 0.081}$ (12)	0.647$_{\pm 0.122}$ (5)	0.681$_{\pm 0.175}$ (3)
glass	**0.885**$_{\pm 0.074}$ (1)	0.844$_{\pm 0.089}$ (7)	0.836$_{\pm 0.035}$ (8)	0.873$_{\pm 0.102}$ (3)	0.516$_{\pm 0.176}$ (11)	0.867$_{\pm 0.089}$ (4)	0.867$_{\pm 0.076}$ (4)
glioma	0.626$_{\pm 0.093}$ (10)	0.650$_{\pm 0.078}$ (8)	0.661$_{\pm 0.110}$ (6)	0.645$_{\pm 0.036}$ (9)	0.493$_{\pm 0.137}$ (13)	0.688$_{\pm 0.063}$ (5)	0.694$_{\pm 0.038}$ (3)
quasar	0.635$_{\pm 0.091}$ (9)	0.808$_{\pm 0.033}$ (3)	0.866$_{\pm 0.054}$ (2)	0.766$_{\pm 0.057}$ (4)	0.462$_{\pm 0.292}$ (12)	0.676$_{\pm 0.070}$ (7)	**0.906**$_{\pm 0.038}$ (1)
seismic	0.173$_{\pm 0.007}$ (2)	0.173$_{\pm 0.028}$ (2)	**0.176**$_{\pm 0.015}$ (1)	0.151$_{\pm 0.008}$ (7)	0.162$_{\pm 0.017}$ (4)	0.171$_{\pm 0.011}$ (3)	0.160$_{\pm 0.016}$ (6)
stroke	0.159$_{\pm 0.039}$ (4)	0.136$_{\pm 0.027}$ (7)	0.161$_{\pm 0.034}$ (3)	0.088$_{\pm 0.031}$ (14)	0.119$_{\pm 0.028}$ (9)	0.166$_{\pm 0.044}$ (2)	**0.172**$_{\pm 0.055}$ (1)
vertebral	0.659$_{\pm 0.075}$ (6)	0.678$_{\pm 0.088}$ (4)	**0.713**$_{\pm 0.088}$ (1)	0.592$_{\pm 0.044}$ (9)	0.434$_{\pm 0.103}$ (13)	0.675$_{\pm 0.075}$ (5)	0.691$_{\pm 0.077}$ (3)
wbc	0.802$_{\pm 0.106}$ (8)	0.769$_{\pm 0.096}$ (11)	0.802$_{\pm 0.110}$ (8)	0.853$_{\pm 0.074}$ (5)	0.794$_{\pm 0.244}$ (9)	0.892$_{\pm 0.011}$ (2)	0.842$_{\pm 0.070}$ (7)
wine	0.992$_{\pm 0.014}$ (3)	0.963$_{\pm 0.061}$ (6)	0.985$_{\pm 0.026}$ (4)	0.943$_{\pm 0.091}$ (8)	0.976$_{\pm 0.023}$ (5)	**0.999**$_{\pm 0.002}$ (1)	**0.999**$_{\pm 0.002}$ (1)
yeast	0.582$_{\pm 0.112}$ (6)	0.551$_{\pm 0.068}$ (8)	0.565$_{\pm 0.038}$ (7)	0.540$_{\pm 0.039}$ (10)	0.110$_{\pm 0.083}$ (13)	**0.656**$_{\pm 0.034}$ (1)	0.617$_{\pm 0.025}$ (4)
Mean	0.606	0.619	0.638	0.589	0.422	0.633	0.651
Mean Rank	5.850	5.050	4.000	6.850	10.000	3.425	2.875
#1st	2	1	4	0	0	5	7

Table G3: Dataset-level AUCPR for NLA = 20.

(a) Conventional and supervised baselines

Dataset	DevNet	PReNet	XGBOD	CAD	WSAD-DT	XGBoost	CatBoost
automobile	$0.429_{\pm 0.120}$ (10)	$0.406_{\pm 0.050}$ (11)	$0.752_{\pm 0.067}$ (9)	$0.246_{\pm 0.031}$ (14)	$0.332_{\pm 0.077}$ (13)	$0.777_{\pm 0.024}$ (7)	$0.797_{\pm 0.051}$ (5)
backdoor	$0.022_{\pm 0.023}$ (11)	$0.005_{\pm 0.003}$ (13)	$0.629_{\pm 0.063}$ (7)	$0.005_{\pm 0.001}$ (13)	$0.007_{\pm 0.001}$ (12)	$0.530_{\pm 0.050}$ (9)	$0.627_{\pm 0.059}$ (8)
campaign	$0.337_{\pm 0.070}$ (12)	$0.343_{\pm 0.034}$ (11)	$0.453_{\pm 0.038}$ (9)	$0.346_{\pm 0.075}$ (10)	$0.310_{\pm 0.011}$ (13)	$0.467_{\pm 0.016}$ (8)	$0.523_{\pm 0.014}$ (6)
cardiotocography	$0.754_{\pm 0.024}$ (8)	$0.730_{\pm 0.051}$ (9)	$0.764_{\pm 0.052}$ (7)	$0.622_{\pm 0.176}$ (11)	$0.339_{\pm 0.007}$ (13)	$0.669_{\pm 0.075}$ (10)	$0.832_{\pm 0.009}$ (3)
census	$0.222_{\pm 0.049}$ (10)	$0.215_{\pm 0.111}$ (11)	$0.327_{\pm 0.066}$ (6)	$0.084_{\pm 0.017}$ (13)	$0.143_{\pm 0.025}$ (12)	$0.292_{\pm 0.052}$ (9)	$0.307_{\pm 0.045}$ (7)
churn	$0.332_{\pm 0.153}$ (12)	$0.471_{\pm 0.188}$ (4)	$0.409_{\pm 0.030}$ (9)	$0.390_{\pm 0.059}$ (10)	$0.202_{\pm 0.003}$ (13)	$0.368_{\pm 0.042}$ (11)	$0.426_{\pm 0.048}$ (6)
cirrhosis	$0.598_{\pm 0.018}$ (9)	$0.478_{\pm 0.039}$ (11)	$0.658_{\pm 0.025}$ (7)	$0.354_{\pm 0.029}$ (12)	$0.542_{\pm 0.031}$ (10)	$0.634_{\pm 0.086}$ (8)	$0.732_{\pm 0.075}$ (5)
covertype	$0.324_{\pm 0.232}$ (12)	$0.186_{\pm 0.128}$ (13)	$0.898_{\pm 0.046}$ (7)	$0.529_{\pm 0.317}$ (10)	$0.005_{\pm 0.000}$ (14)	$0.860_{\pm 0.055}$ (9)	$0.915_{\pm 0.013}$ (6)
credit	$0.253_{\pm 0.050}$ (11)	$0.208_{\pm 0.033}$ (13)	$0.322_{\pm 0.027}$ (3)	$0.226_{\pm 0.008}$ (12)	$0.184_{\pm 0.001}$ (14)	$0.282_{\pm 0.013}$ (9)	$0.301_{\pm 0.027}$ (6)
equip	$0.664_{\pm 0.064}$ (10)	$0.629_{\pm 0.033}$ (11)	$0.854_{\pm 0.058}$ (6)	$0.481_{\pm 0.080}$ (14)	$0.492_{\pm 0.018}$ (13)	$0.708_{\pm 0.085}$ (9)	$0.739_{\pm 0.097}$ (8)
gallstone	$0.531_{\pm 0.001}$ (10)	$0.581_{\pm 0.062}$ (9)	$0.643_{\pm 0.090}$ (8)	$0.312_{\pm 0.015}$ (13)	$0.369_{\pm 0.047}$ (12)	$0.648_{\pm 0.049}$ (7)	$0.672_{\pm 0.083}$ (5)
glass	$0.807_{\pm 0.113}$ (11)	$0.857_{\pm 0.081}$ (10)	$0.790_{\pm 0.111}$ (12)	$0.231_{\pm 0.000}$ (14)	$0.883_{\pm 0.166}$ (9)	$0.910_{\pm 0.065}$ (8)	**$0.958_{\pm 0.014}$** (1)
glioma	$0.733_{\pm 0.041}$ (2)	$0.675_{\pm 0.072}$ (8)	$0.596_{\pm 0.038}$ (12)	$0.562_{\pm 0.304}$ (13)	**$0.755_{\pm 0.049}$** (1)	$0.641_{\pm 0.047}$ (11)	$0.658_{\pm 0.056}$ (10)
quasar	$0.661_{\pm 0.152}$ (10)	$0.772_{\pm 0.078}$ (8)	$0.838_{\pm 0.011}$ (6)	$0.180_{\pm 0.015}$ (14)	$0.207_{\pm 0.001}$ (13)	$0.646_{\pm 0.011}$ (11)	$0.799_{\pm 0.040}$ (7)
seismic	$0.088_{\pm 0.080}$ (10)	$0.077_{\pm 0.048}$ (11)	$0.131_{\pm 0.022}$ (8)	$0.070_{\pm 0.006}$ (12)	$0.162_{\pm 0.020}$ (6)	$0.130_{\pm 0.007}$ (9)	$0.158_{\pm 0.021}$ (7)
stroke	$0.182_{\pm 0.030}$ (2)	$0.168_{\pm 0.021}$ (6)	$0.116_{\pm 0.013}$ (11)	$0.138_{\pm 0.055}$ (9)	$0.143_{\pm 0.014}$ (8)	$0.109_{\pm 0.018}$ (12)	$0.108_{\pm 0.018}$ (13)
vertebral	$0.735_{\pm 0.111}$ (7)	$0.740_{\pm 0.019}$ (4)	$0.672_{\pm 0.070}$ (8)	$0.383_{\pm 0.099}$ (13)	$0.236_{\pm 0.008}$ (14)	$0.612_{\pm 0.055}$ (11)	$0.663_{\pm 0.040}$ (9)
wbc	$0.914_{\pm 0.055}$ (7)	$0.906_{\pm 0.060}$ (10)	$0.907_{\pm 0.037}$ (9)	$0.619_{\pm 0.246}$ (13)	$0.957_{\pm 0.010}$ (2)	$0.887_{\pm 0.011}$ (11)	$0.951_{\pm 0.027}$ (3)
wine	$0.858_{\pm 0.082}$ (8)	**$1.000_{\pm 0.000}$** (1)	$0.969_{\pm 0.019}$ (7)	$0.336_{\pm 0.091}$ (9)	$0.237_{\pm 0.022}$ (10)	$0.987_{\pm 0.005}$ (6)	$0.997_{\pm 0.002}$ (3)
yeast	$0.649_{\pm 0.104}$ (8)	$0.657_{\pm 0.107}$ (6)	$0.602_{\pm 0.087}$ (11)	$0.111_{\pm 0.078}$ (13)	$0.644_{\pm 0.067}$ (9)	$0.521_{\pm 0.056}$ (12)	$0.632_{\pm 0.002}$ (10)
Mean	0.504	0.505	0.616	0.311	0.357	0.584	0.640
Mean Rank	9.600	9.525	8.700	12.750	11.150	9.950	6.900
#1st	0	1	0	0	1	0	1

(b) Pretrained, LLM-based, and LENS-AD methods

Dataset	TabPFN	LimiX	TabICL	CAAFE	FeatLLM	LENS-AD (TabPFN)	LENS-AD (TabICL)
automobile	$0.766_{\pm 0.037}$ (8)	$0.835_{\pm 0.061}$ (4)	$0.849_{\pm 0.047}$ (3)	$0.793_{\pm 0.019}$ (6)	$0.351_{\pm 0.088}$ (12)	**$0.884_{\pm 0.039}$** (1)	$0.883_{\pm 0.042}$ (2)
backdoor	$0.764_{\pm 0.051}$ (4)	$0.792_{\pm 0.032}$ (3)	**$0.829_{\pm 0.026}$** (1)	$0.649_{\pm 0.025}$ (6)	$0.133_{\pm 0.086}$ (10)	$0.746_{\pm 0.049}$ (5)	$0.806_{\pm 0.020}$ (2)
campaign	$0.576_{\pm 0.033}$ (3)	$0.551_{\pm 0.042}$ (5)	$0.578_{\pm 0.050}$ (2)	$0.477_{\pm 0.029}$ (7)	$0.310_{\pm 0.091}$ (13)	**$0.591_{\pm 0.041}$** (1)	$0.575_{\pm 0.042}$ (4)
cardiotocography	$0.875_{\pm 0.019}$ (2)	$0.805_{\pm 0.009}$ (6)	**$0.877_{\pm 0.031}$** (1)	$0.816_{\pm 0.032}$ (5)	$0.530_{\pm 0.160}$ (12)	$0.832_{\pm 0.026}$ (3)	$0.825_{\pm 0.042}$ (4)
census	$0.342_{\pm 0.042}$ (5)	$0.380_{\pm 0.015}$ (4)	$0.439_{\pm 0.045}$ (2)	$0.303_{\pm 0.067}$ (8)	$0.215_{\pm 0.060}$ (11)	$0.428_{\pm 0.051}$ (3)	**$0.469_{\pm 0.048}$** (1)
churn	$0.452_{\pm 0.071}$ (5)	$0.452_{\pm 0.047}$ (5)	$0.539_{\pm 0.051}$ (2)	$0.420_{\pm 0.050}$ (7)	$0.419_{\pm 0.057}$ (8)	$0.512_{\pm 0.042}$ (3)	**$0.553_{\pm 0.024}$** (1)
cirrhosis	$0.732_{\pm 0.049}$ (5)	$0.744_{\pm 0.063}$ (4)	$0.751_{\pm 0.046}$ (3)	$0.728_{\pm 0.073}$ (6)	$0.758_{\pm 0.043}$ (2)	$0.744_{\pm 0.054}$ (4)	**$0.772_{\pm 0.065}$** (1)
covertype	$0.967_{\pm 0.013}$ (4)	$0.950_{\pm 0.010}$ (5)	$0.975_{\pm 0.008}$ (3)	$0.896_{\pm 0.021}$ (8)	$0.327_{\pm 0.138}$ (11)	$0.981_{\pm 0.009}$ (2)	**$0.983_{\pm 0.009}$** (1)
credit	$0.291_{\pm 0.038}$ (8)	$0.319_{\pm 0.039}$ (4)	$0.307_{\pm 0.033}$ (5)	$0.300_{\pm 0.031}$ (7)	$0.255_{\pm 0.035}$ (10)	$0.329_{\pm 0.030}$ (2)	**$0.350_{\pm 0.041}$** (1)
equip	$0.889_{\pm 0.038}$ (5)	$0.895_{\pm 0.026}$ (4)	**$0.931_{\pm 0.012}$** (1)	$0.828_{\pm 0.054}$ (7)	$0.620_{\pm 0.196}$ (12)	$0.897_{\pm 0.017}$ (3)	$0.918_{\pm 0.011}$ (2)
gallstone	$0.747_{\pm 0.115}$ (2)	$0.721_{\pm 0.058}$ (3)	**$0.755_{\pm 0.112}$** (1)	$0.662_{\pm 0.037}$ (6)	$0.454_{\pm 0.079}$ (11)	$0.694_{\pm 0.160}$ (4)	$0.721_{\pm 0.170}$ (3)
glass	$0.934_{\pm 0.055}$ (4)	$0.913_{\pm 0.024}$ (7)	$0.931_{\pm 0.049}$ (5)	$0.922_{\pm 0.033}$ (6)	$0.565_{\pm 0.200}$ (13)	$0.950_{\pm 0.025}$ (2)	$0.942_{\pm 0.037}$ (3)
glioma	$0.701_{\pm 0.066}$ (5)	$0.685_{\pm 0.085}$ (7)	$0.727_{\pm 0.089}$ (3)	$0.668_{\pm 0.063}$ (9)	$0.493_{\pm 0.137}$ (14)	$0.693_{\pm 0.070}$ (6)	$0.721_{\pm 0.053}$ (4)
quasar	$0.860_{\pm 0.011}$ (5)	$0.878_{\pm 0.019}$ (4)	$0.940_{\pm 0.006}$ (2)	$0.758_{\pm 0.008}$ (9)	$0.461_{\pm 0.291}$ (12)	$0.903_{\pm 0.017}$ (3)	**$0.947_{\pm 0.010}$** (1)
seismic	$0.185_{\pm 0.021}$ (3)	$0.185_{\pm 0.025}$ (3)	**$0.195_{\pm 0.029}$** (1)	$0.164_{\pm 0.022}$ (5)	$0.162_{\pm 0.017}$ (6)	$0.178_{\pm 0.016}$ (4)	$0.188_{\pm 0.033}$ (2)
stroke	$0.178_{\pm 0.041}$ (4)	$0.156_{\pm 0.021}$ (7)	$0.175_{\pm 0.039}$ (5)	$0.105_{\pm 0.011}$ (14)	$0.118_{\pm 0.025}$ (10)	**$0.217_{\pm 0.039}$** (1)	$0.179_{\pm 0.057}$ (3)
vertebral	$0.775_{\pm 0.032}$ (2)	$0.765_{\pm 0.051}$ (3)	**$0.784_{\pm 0.034}$** (1)	$0.654_{\pm 0.059}$ (10)	$0.435_{\pm 0.098}$ (12)	$0.739_{\pm 0.049}$ (5)	$0.738_{\pm 0.024}$ (6)
wbc	$0.936_{\pm 0.063}$ (6)	$0.909_{\pm 0.069}$ (8)	$0.945_{\pm 0.039}$ (5)	$0.946_{\pm 0.024}$ (4)	$0.796_{\pm 0.244}$ (12)	**$0.961_{\pm 0.038}$** (1)	$0.957_{\pm 0.030}$ (2)
wine	$0.999_{\pm 0.002}$ (2)	**$1.000_{\pm 0.000}$** (1)	**$1.000_{\pm 0.000}$** (1)	$0.996_{\pm 0.004}$ (4)	$0.992_{\pm 0.008}$ (5)	$0.999_{\pm 0.002}$ (2)	**$1.000_{\pm 0.000}$** (1)
yeast	$0.667_{\pm 0.046}$ (4)	$0.652_{\pm 0.080}$ (7)	$0.670_{\pm 0.040}$ (3)	$0.662_{\pm 0.034}$ (5)	$0.110_{\pm 0.083}$ (14)	**$0.684_{\pm 0.076}$** (1)	$0.680_{\pm 0.090}$ (2)
Mean	0.682	0.679	0.710	0.637	0.425	0.698	0.710
Mean Rank	4.650	5.025	2.625	7.500	11.000	3.100	2.525
#1st	0	1	7	0	0	5	7

G.2 Few-Sample Results

Table G4: Dataset-level AUCPR for n = 10.

(a) Conventional and supervised baselines

Dataset	DevNet	PReNet	XGBOD	CAD	WSAD-DT	XGBoost	CatBoost
automobile	0.333$_{\pm 0.089}$ (12)	0.319$_{\pm 0.010}$ (13)	0.444$_{\pm 0.257}$ (9)	0.278$_{\pm 0.026}$ (14)	0.334$_{\pm 0.078}$ (11)	0.546$_{\pm 0.225}$ (2)	0.504$_{\pm 0.160}$ (5)
backdoor	0.007$_{\pm 0.004}$ (9)	0.011$_{\pm 0.005}$ (8)	0.004$_{\pm 0.000}$ (11)	0.006$_{\pm 0.002}$ (10)	0.007$_{\pm 0.001}$ (9)	0.004$_{\pm 0.000}$ (11)	0.073$_{\pm 0.022}$ (7)
campaign	0.305$_{\pm 0.100}$ (10)	0.309$_{\pm 0.099}$ (9)	0.290$_{\pm 0.020}$ (12)	0.236$_{\pm 0.021}$ (13)	0.294$_{\pm 0.011}$ (11)	0.228$_{\pm 0.000}$ (14)	0.446$_{\pm 0.015}$ (2)
cardiotocography	0.497$_{\pm 0.221}$ (7)	0.467$_{\pm 0.236}$ (10)	0.378$_{\pm 0.134}$ (12)	0.403$_{\pm 0.169}$ (11)	0.321$_{\pm 0.044}$ (13)	0.221$_{\pm 0.000}$ (14)	0.483$_{\pm 0.203}$ (8)
census	0.171$_{\pm 0.083}$ (7)	0.180$_{\pm 0.103}$ (6)	0.058$_{\pm 0.000}$ (13)	0.120$_{\pm 0.048}$ (10)	0.106$_{\pm 0.055}$ (12)	0.058$_{\pm 0.000}$ (13)	0.134$_{\pm 0.038}$ (9)
churn	0.539$_{\pm 0.057}$ (2)	0.468$_{\pm 0.137}$ (8)	0.360$_{\pm 0.043}$ (10)	0.256$_{\pm 0.043}$ (11)	0.202$_{\pm 0.004}$ (12)	0.420$_{\pm 0.063}$ (9)	0.493$_{\pm 0.044}$ (4)
cirrhosis	0.600$_{\pm 0.068}$ (10)	0.496$_{\pm 0.120}$ (12)	0.643$_{\pm 0.091}$ (9)	0.377$_{\pm 0.074}$ (13)	0.541$_{\pm 0.047}$ (11)	0.653$_{\pm 0.133}$ (8)	0.777$_{\pm 0.045}$ (3)
covertype	0.019$_{\pm 0.007}$ (9)	0.011$_{\pm 0.002}$ (11)	0.010$_{\pm 0.000}$ (12)	0.013$_{\pm 0.002}$ (10)	0.005$_{\pm 0.000}$ (13)	0.010$_{\pm 0.000}$ (12)	0.629$_{\pm 0.121}$ (4)
credit	0.211$_{\pm 0.016}$ (12)	0.218$_{\pm 0.022}$ (11)	0.232$_{\pm 0.004}$ (9)	0.218$_{\pm 0.019}$ (11)	0.185$_{\pm 0.002}$ (13)	0.221$_{\pm 0.000}$ (10)	0.319$_{\pm 0.066}$ (6)
equip	0.349$_{\pm 0.067}$ (11)	0.360$_{\pm 0.033}$ (10)	0.100$_{\pm 0.000}$ (13)	0.242$_{\pm 0.027}$ (12)	0.485$_{\pm 0.077}$ (5)	0.100$_{\pm 0.000}$ (13)	0.428$_{\pm 0.050}$ (6)
gallstone	0.351$_{\pm 0.078}$ (12)	0.382$_{\pm 0.056}$ (10)	0.485$_{\pm 0.207}$ (3)	0.337$_{\pm 0.052}$ (14)	0.373$_{\pm 0.045}$ (11)	0.442$_{\pm 0.090}$ (8)	0.446$_{\pm 0.145}$ (6)
glass	0.950$_{\pm 0.040}$ (3)	0.909$_{\pm 0.053}$ (8)	0.745$_{\pm 0.076}$ (9)	0.231$_{\pm 0.000}$ (11)	0.943$_{\pm 0.014}$ (4)	0.231$_{\pm 0.000}$ (11)	0.957$_{\pm 0.021}$ (2)
glioma	0.701$_{\pm 0.026}$ (2)	0.698$_{\pm 0.023}$ (3)	0.577$_{\pm 0.087}$ (10)	0.229$_{\pm 0.027}$ (14)	0.559$_{\pm 0.076}$ (12)	0.511$_{\pm 0.033}$ (13)	0.683$_{\pm 0.051}$ (4)
quasar	0.361$_{\pm 0.091}$ (10)	0.431$_{\pm 0.215}$ (8)	0.392$_{\pm 0.118}$ (9)	0.195$_{\pm 0.011}$ (12)	0.208$_{\pm 0.002}$ (11)	0.190$_{\pm 0.000}$ (13)	0.812$_{\pm 0.115}$ (4)
seismic	0.121$_{\pm 0.070}$ (6)	0.046$_{\pm 0.003}$ (11)	0.066$_{\pm 0.000}$ (10)	0.074$_{\pm 0.012}$ (9)	**0.162**$_{\pm 0.020}$ (1)	0.066$_{\pm 0.000}$ (10)	0.137$_{\pm 0.040}$ (5)
stroke	0.123$_{\pm 0.029}$ (8)	0.090$_{\pm 0.033}$ (10)	0.043$_{\pm 0.000}$ (11)	0.043$_{\pm 0.009}$ (11)	0.139$_{\pm 0.011}$ (4)	0.043$_{\pm 0.000}$ (11)	0.136$_{\pm 0.042}$ (7)
vertebral	0.521$_{\pm 0.023}$ (9)	0.473$_{\pm 0.034}$ (10)	0.565$_{\pm 0.032}$ (7)	0.397$_{\pm 0.028}$ (12)	0.245$_{\pm 0.019}$ (13)	0.473$_{\pm 0.033}$ (10)	0.583$_{\pm 0.091}$ (3)
wbc	0.912$_{\pm 0.032}$ (9)	0.877$_{\pm 0.064}$ (10)	0.622$_{\pm 0.221}$ (13)	0.359$_{\pm 0.064}$ (14)	0.956$_{\pm 0.010}$ (2)	0.765$_{\pm 0.123}$ (12)	0.927$_{\pm 0.084}$ (7)
wine	0.591$_{\pm 0.308}$ (9)	0.863$_{\pm 0.151}$ (6)	0.835$_{\pm 0.152}$ (7)	0.299$_{\pm 0.037}$ (10)	0.236$_{\pm 0.021}$ (11)	0.711$_{\pm 0.256}$ (8)	0.990$_{\pm 0.018}$ (5)
yeast	0.503$_{\pm 0.244}$ (2)	**0.527**$_{\pm 0.186}$ (1)	0.065$_{\pm 0.000}$ (12)	0.134$_{\pm 0.148}$ (11)	0.378$_{\pm 0.127}$ (6)	0.065$_{\pm 0.000}$ (12)	0.422$_{\pm 0.198}$ (3)
Mean	0.408	0.407	0.346	0.222	0.334	0.298	0.519
Mean Rank	8.450	9.500	11.025	12.475	9.900	11.750	5.450
#1st	0	1	0	0	1	0	0

(b) Pretrained, LLM-based, and LENS-AD methods

Dataset	TabPFN	LimiX	TabICL	CAAFE	FeatLLM	LENS-AD (TabPFN)	LENS-AD (TabICL)
automobile	0.493$_{\pm 0.197}$ (7)	0.453$_{\pm 0.077}$ (8)	0.540$_{\pm 0.153}$ (3)	0.499$_{\pm 0.153}$ (6)	0.358$_{\pm 0.111}$ (10)	**0.626**$_{\pm 0.120}$ (1)	0.522$_{\pm 0.180}$ (4)
backdoor	0.076$_{\pm 0.041}$ (6)	0.079$_{\pm 0.009}$ (4)	**0.140**$_{\pm 0.095}$ (1)	0.077$_{\pm 0.060}$ (5)	0.137$_{\pm 0.080}$ (2)	0.082$_{\pm 0.035}$ (3)	**0.140**$_{\pm 0.042}$ (1)
campaign	0.364$_{\pm 0.013}$ (7)	0.414$_{\pm 0.057}$ (3)	0.369$_{\pm 0.045}$ (6)	**0.454**$_{\pm 0.011}$ (1)	0.332$_{\pm 0.098}$ (8)	0.389$_{\pm 0.071}$ (4)	0.372$_{\pm 0.065}$ (5)
cardiotocography	0.563$_{\pm 0.246}$ (3)	0.544$_{\pm 0.176}$ (4)	0.508$_{\pm 0.254}$ (6)	0.480$_{\pm 0.202}$ (9)	**0.620**$_{\pm 0.139}$ (1)	0.575$_{\pm 0.168}$ (2)	0.536$_{\pm 0.170}$ (5)
census	0.116$_{\pm 0.041}$ (11)	0.184$_{\pm 0.059}$ (5)	0.186$_{\pm 0.075}$ (4)	0.144$_{\pm 0.055}$ (8)	0.201$_{\pm 0.053}$ (2)	0.191$_{\pm 0.081}$ (3)	**0.209**$_{\pm 0.073}$ (1)
churn	0.490$_{\pm 0.047}$ (5)	0.487$_{\pm 0.049}$ (6)	0.504$_{\pm 0.065}$ (3)	0.493$_{\pm 0.038}$ (4)	**0.572**$_{\pm 0.053}$ (1)	0.469$_{\pm 0.061}$ (7)	0.493$_{\pm 0.033}$ (4)
cirrhosis	0.775$_{\pm 0.082}$ (4)	0.784$_{\pm 0.056}$ (2)	0.738$_{\pm 0.118}$ (7)	0.777$_{\pm 0.047}$ (3)	**0.785**$_{\pm 0.026}$ (1)	0.742$_{\pm 0.049}$ (6)	0.744$_{\pm 0.053}$ (5)
covertype	0.787$_{\pm 0.274}$ (2)	0.579$_{\pm 0.177}$ (7)	0.553$_{\pm 0.294}$ (8)	0.617$_{\pm 0.144}$ (5)	0.581$_{\pm 0.199}$ (6)	**0.911**$_{\pm 0.051}$ (1)	0.728$_{\pm 0.076}$ (3)
credit	0.330$_{\pm 0.085}$ (3)	0.265$_{\pm 0.044}$ (8)	0.333$_{\pm 0.066}$ (2)	0.325$_{\pm 0.060}$ (4)	0.276$_{\pm 0.047}$ (7)	0.323$_{\pm 0.067}$ (5)	**0.334**$_{\pm 0.055}$ (1)
equip	0.675$_{\pm 0.177}$ (3)	**0.722**$_{\pm 0.144}$ (1)	0.404$_{\pm 0.158}$ (7)	0.403$_{\pm 0.084}$ (8)	0.720$_{\pm 0.079}$ (2)	0.610$_{\pm 0.201}$ (4)	0.393$_{\pm 0.181}$ (9)
gallstone	**0.597**$_{\pm 0.171}$ (1)	0.340$_{\pm 0.030}$ (13)	0.544$_{\pm 0.139}$ (2)	0.441$_{\pm 0.131}$ (9)	0.463$_{\pm 0.073}$ (5)	0.443$_{\pm 0.048}$ (7)	0.465$_{\pm 0.020}$ (4)
glass	0.934$_{\pm 0.033}$ (5)	0.928$_{\pm 0.050}$ (6)	0.950$_{\pm 0.033}$ (3)	**0.964**$_{\pm 0.024}$ (1)	0.714$_{\pm 0.127}$ (10)	0.934$_{\pm 0.018}$ (5)	0.926$_{\pm 0.047}$ (7)
glioma	0.669$_{\pm 0.068}$ (7)	**0.708**$_{\pm 0.071}$ (1)	0.649$_{\pm 0.069}$ (9)	0.673$_{\pm 0.056}$ (5)	0.571$_{\pm 0.198}$ (11)	0.670$_{\pm 0.066}$ (6)	0.651$_{\pm 0.061}$ (8)
quasar	**0.940**$_{\pm 0.007}$ (1)	0.744$_{\pm 0.085}$ (6)	0.911$_{\pm 0.013}$ (3)	0.806$_{\pm 0.109}$ (5)	0.634$_{\pm 0.342}$ (7)	0.924$_{\pm 0.007}$ (2)	0.911$_{\pm 0.014}$ (3)
seismic	0.137$_{\pm 0.021}$ (5)	0.121$_{\pm 0.047}$ (6)	0.115$_{\pm 0.046}$ (8)	0.142$_{\pm 0.032}$ (3)	0.154$_{\pm 0.006}$ (2)	0.139$_{\pm 0.044}$ (4)	0.118$_{\pm 0.045}$ (7)
stroke	**0.142**$_{\pm 0.049}$ (1)	0.123$_{\pm 0.050}$ (8)	0.097$_{\pm 0.026}$ (9)	0.137$_{\pm 0.051}$ (6)	0.141$_{\pm 0.032}$ (2)	0.140$_{\pm 0.036}$ (3)	0.138$_{\pm 0.037}$ (5)
vertebral	0.586$_{\pm 0.098}$ (2)	0.564$_{\pm 0.107}$ (8)	0.567$_{\pm 0.097}$ (6)	**0.592**$_{\pm 0.115}$ (1)	0.451$_{\pm 0.055}$ (11)	0.570$_{\pm 0.056}$ (5)	0.576$_{\pm 0.101}$ (4)
wbc	0.947$_{\pm 0.064}$ (3)	**0.974**$_{\pm 0.010}$ (1)	0.938$_{\pm 0.086}$ (5)	0.937$_{\pm 0.071}$ (6)	0.828$_{\pm 0.200}$ (11)	0.939$_{\pm 0.064}$ (4)	0.923$_{\pm 0.049}$ (8)
wine	**0.999**$_{\pm 0.002}$ (1)	0.991$_{\pm 0.015}$ (4)	0.993$_{\pm 0.006}$ (3)	0.990$_{\pm 0.014}$ (5)	0.994$_{\pm 0.011}$ (2)	**0.999**$_{\pm 0.002}$ (1)	**0.999**$_{\pm 0.002}$ (1)
yeast	0.361$_{\pm 0.237}$ (8)	0.398$_{\pm 0.178}$ (4)	0.329$_{\pm 0.211}$ (9)	0.389$_{\pm 0.185}$ (5)	0.304$_{\pm 0.164}$ (10)	0.422$_{\pm 0.293}$ (3)	0.369$_{\pm 0.239}$ (7)
Mean	0.549	0.520	0.518	0.517	0.492	0.555	0.527
Mean Rank	4.700	5.800	5.600	5.250	5.950	4.075	5.075
#1st	4	3	1	3	3	3	4

Table G5: Dataset-level AUCPR for n = 20.

(a) Conventional and supervised baselines

Dataset	DevNet	PReNet	XGBOD	CAD	WSAD-DT	XGBoost	CatBoost
automobile	$0.386_{\pm 0.029}$ (10)	$0.338_{\pm 0.009}$ (12)	$0.599_{\pm 0.142}$ (9)	$0.243_{\pm 0.026}$ (14)	$0.334_{\pm 0.078}$ (13)	$0.723_{\pm 0.055}$ (3)	$0.705_{\pm 0.048}$ (4)
backdoor	$0.011_{\pm 0.009}$ (10)	$0.005_{\pm 0.003}$ (12)	$0.029_{\pm 0.017}$ (9)	$0.005_{\pm 0.001}$ (12)	$0.007_{\pm 0.001}$ (11)	$0.004_{\pm 0.000}$ (13)	$0.085_{\pm 0.026}$ (7)
campaign	$0.369_{\pm 0.053}$ (8)	$0.331_{\pm 0.033}$ (9)	$0.313_{\pm 0.029}$ (11)	$0.236_{\pm 0.024}$ (13)	$0.300_{\pm 0.010}$ (12)	$0.383_{\pm 0.024}$ (7)	$0.491_{\pm 0.084}$ (3)
cardiotocography	$0.634_{\pm 0.060}$ (9)	$0.641_{\pm 0.029}$ (8)	$0.550_{\pm 0.195}$ (11)	$0.278_{\pm 0.064}$ (13)	$0.318_{\pm 0.013}$ (12)	$0.622_{\pm 0.035}$ (10)	$0.687_{\pm 0.084}$ (6)
census	$0.188_{\pm 0.088}$ (6)	$0.180_{\pm 0.102}$ (7)	$0.141_{\pm 0.043}$ (9)	$0.114_{\pm 0.067}$ (12)	$0.109_{\pm 0.065}$ (13)	$0.058_{\pm 0.000}$ (14)	$0.131_{\pm 0.027}$ (10)
churn	$0.463_{\pm 0.178}$ (9)	$0.276_{\pm 0.022}$ (12)	$0.399_{\pm 0.066}$ (10)	$0.295_{\pm 0.073}$ (11)	$0.200_{\pm 0.003}$ (13)	$0.499_{\pm 0.029}$ (5)	$0.464_{\pm 0.015}$ (8)
cirrhosis	$0.574_{\pm 0.055}$ (10)	$0.500_{\pm 0.093}$ (12)	$0.721_{\pm 0.084}$ (9)	$0.387_{\pm 0.073}$ (13)	$0.539_{\pm 0.045}$ (11)	$0.789_{\pm 0.054}$ (4)	$0.763_{\pm 0.038}$ (5)
covertype	$0.037_{\pm 0.036}$ (10)	$0.032_{\pm 0.014}$ (11)	$0.158_{\pm 0.091}$ (9)	$0.011_{\pm 0.002}$ (12)	$0.005_{\pm 0.000}$ (14)	$0.010_{\pm 0.000}$ (13)	$0.718_{\pm 0.189}$ (4)
credit	$0.230_{\pm 0.021}$ (11)	$0.210_{\pm 0.036}$ (13)	$0.270_{\pm 0.063}$ (9)	$0.218_{\pm 0.023}$ (12)	$0.187_{\pm 0.003}$ (14)	$0.276_{\pm 0.058}$ (8)	**$0.323_{\pm 0.075}$** (1)
equip	$0.468_{\pm 0.113}$ (12)	$0.490_{\pm 0.037}$ (11)	$0.653_{\pm 0.168}$ (7)	$0.270_{\pm 0.013}$ (13)	$0.502_{\pm 0.022}$ (10)	$0.100_{\pm 0.000}$ (14)	$0.614_{\pm 0.124}$ (8)
gallstone	$0.380_{\pm 0.014}$ (11)	$0.394_{\pm 0.082}$ (10)	**$0.575_{\pm 0.195}$** (1)	$0.334_{\pm 0.029}$ (13)	$0.373_{\pm 0.046}$ (12)	$0.538_{\pm 0.220}$ (4)	$0.529_{\pm 0.199}$ (5)
glass	$0.929_{\pm 0.031}$ (7)	$0.904_{\pm 0.067}$ (9)	$0.885_{\pm 0.054}$ (10)	$0.231_{\pm 0.000}$ (13)	$0.845_{\pm 0.155}$ (11)	$0.927_{\pm 0.036}$ (8)	$0.958_{\pm 0.017}$ (4)
glioma	**$0.743_{\pm 0.015}$** (1)	$0.732_{\pm 0.061}$ (2)	$0.692_{\pm 0.020}$ (7)	$0.300_{\pm 0.114}$ (14)	$0.628_{\pm 0.083}$ (12)	$0.694_{\pm 0.063}$ (6)	$0.713_{\pm 0.058}$ (4)
quasar	$0.515_{\pm 0.228}$ (11)	$0.634_{\pm 0.206}$ (9)	$0.642_{\pm 0.186}$ (8)	$0.189_{\pm 0.008}$ (13)	$0.208_{\pm 0.001}$ (12)	$0.598_{\pm 0.186}$ (10)	$0.863_{\pm 0.038}$ (4)
seismic	$0.108_{\pm 0.051}$ (10)	$0.102_{\pm 0.052}$ (11)	$0.101_{\pm 0.039}$ (12)	$0.068_{\pm 0.002}$ (13)	**$0.162_{\pm 0.020}$** (1)	$0.066_{\pm 0.000}$ (14)	$0.144_{\pm 0.026}$ (3)
stroke	$0.110_{\pm 0.024}$ (8)	$0.103_{\pm 0.032}$ (9)	$0.066_{\pm 0.023}$ (11)	$0.044_{\pm 0.004}$ (12)	$0.139_{\pm 0.015}$ (4)	$0.043_{\pm 0.000}$ (13)	$0.135_{\pm 0.054}$ (6)
vertebral	**$0.644_{\pm 0.055}$** (1)	$0.586_{\pm 0.058}$ (6)	$0.573_{\pm 0.154}$ (7)	$0.384_{\pm 0.060}$ (13)	$0.242_{\pm 0.009}$ (14)	$0.529_{\pm 0.102}$ (11)	$0.564_{\pm 0.043}$ (9)
wbc	$0.936_{\pm 0.037}$ (7)	$0.892_{\pm 0.078}$ (9)	$0.869_{\pm 0.086}$ (11)	$0.437_{\pm 0.175}$ (12)	$0.957_{\pm 0.010}$ (6)	$0.893_{\pm 0.068}$ (8)	$0.971_{\pm 0.012}$ (4)
wine	$0.640_{\pm 0.263}$ (6)	$0.961_{\pm 0.039}$ (3)	$0.946_{\pm 0.045}$ (4)	$0.284_{\pm 0.014}$ (7)	$0.236_{\pm 0.021}$ (8)	$0.899_{\pm 0.026}$ (5)	**$1.000_{\pm 0.000}$** (1)
yeast	$0.536_{\pm 0.228}$ (2)	**$0.538_{\pm 0.218}$** (1)	$0.335_{\pm 0.077}$ (9)	$0.137_{\pm 0.153}$ (11)	$0.391_{\pm 0.015}$ (6)	$0.065_{\pm 0.000}$ (12)	$0.409_{\pm 0.219}$ (3)
Mean	0.445	0.442	0.476	0.223	0.334	0.436	0.563
Mean Rank	8.750	9.600	9.450	13.250	11.250	9.950	5.500
#1st	2	1	1	0	1	0	2

(b) Pretrained, LLM-based, and LENS-AD methods

Dataset	TabPFN	LimiX	TabICL	CAAFE	FeatLLM	LENS-AD (TabPFN)	LENS-AD (TabICL)
automobile	$0.654_{\pm 0.044}$ (8)	$0.672_{\pm 0.030}$ (7)	$0.697_{\pm 0.074}$ (5)	$0.685_{\pm 0.060}$ (6)	$0.342_{\pm 0.099}$ (11)	$0.728_{\pm 0.037}$ (2)	**$0.764_{\pm 0.066}$** (1)
backdoor	$0.115_{\pm 0.074}$ (4)	$0.117_{\pm 0.006}$ (3)	$0.112_{\pm 0.026}$ (5)	$0.081_{\pm 0.023}$ (8)	**$0.165_{\pm 0.102}$** (1)	$0.103_{\pm 0.010}$ (6)	$0.153_{\pm 0.098}$ (2)
campaign	$0.503_{\pm 0.143}$ (2)	$0.491_{\pm 0.039}$ (3)	$0.459_{\pm 0.091}$ (6)	$0.481_{\pm 0.105}$ (4)	$0.327_{\pm 0.096}$ (10)	**$0.512_{\pm 0.116}$** (1)	$0.472_{\pm 0.093}$ (5)
cardiotocography	$0.731_{\pm 0.049}$ (2)	$0.712_{\pm 0.078}$ (3)	$0.660_{\pm 0.043}$ (7)	$0.695_{\pm 0.066}$ (4)	$0.691_{\pm 0.120}$ (5)	**$0.739_{\pm 0.062}$** (1)	$0.712_{\pm 0.098}$ (3)
census	$0.152_{\pm 0.054}$ (8)	$0.193_{\pm 0.044}$ (5)	**$0.234_{\pm 0.101}$** (1)	$0.129_{\pm 0.026}$ (11)	$0.201_{\pm 0.061}$ (4)	$0.215_{\pm 0.040}$ (3)	$0.216_{\pm 0.060}$ (2)
churn	$0.506_{\pm 0.013}$ (4)	$0.511_{\pm 0.006}$ (2)	$0.481_{\pm 0.022}$ (7)	$0.485_{\pm 0.009}$ (6)	**$0.554_{\pm 0.046}$** (1)	$0.511_{\pm 0.015}$ (2)	$0.507_{\pm 0.003}$ (3)
cirrhosis	$0.747_{\pm 0.024}$ (6)	**$0.811_{\pm 0.033}$** (1)	$0.721_{\pm 0.059}$ (9)	$0.790_{\pm 0.035}$ (3)	$0.795_{\pm 0.035}$ (2)	$0.734_{\pm 0.052}$ (8)	$0.737_{\pm 0.043}$ (7)
covertype	$0.726_{\pm 0.387}$ (3)	$0.657_{\pm 0.331}$ (7)	$0.715_{\pm 0.291}$ (5)	$0.692_{\pm 0.231}$ (6)	$0.605_{\pm 0.192}$ (8)	**$0.927_{\pm 0.041}$** (1)	$0.917_{\pm 0.052}$ (2)
credit	$0.314_{\pm 0.125}$ (2)	$0.266_{\pm 0.078}$ (10)	$0.291_{\pm 0.097}$ (6)	$0.311_{\pm 0.098}$ (3)	$0.277_{\pm 0.043}$ (7)	$0.299_{\pm 0.107}$ (5)	$0.300_{\pm 0.086}$ (4)
equip	$0.786_{\pm 0.129}$ (2)	**$0.880_{\pm 0.040}$** (1)	$0.721_{\pm 0.166}$ (4)	$0.703_{\pm 0.172}$ (6)	$0.711_{\pm 0.151}$ (5)	$0.732_{\pm 0.133}$ (3)	$0.596_{\pm 0.208}$ (9)
gallstone	$0.570_{\pm 0.185}$ (2)	$0.441_{\pm 0.109}$ (9)	$0.539_{\pm 0.120}$ (3)	$0.529_{\pm 0.190}$ (5)	$0.517_{\pm 0.127}$ (6)	$0.475_{\pm 0.126}$ (8)	$0.501_{\pm 0.132}$ (7)
glass	$0.957_{\pm 0.010}$ (5)	$0.961_{\pm 0.026}$ (3)	$0.956_{\pm 0.016}$ (6)	$0.961_{\pm 0.020}$ (3)	$0.653_{\pm 0.237}$ (12)	**$0.974_{\pm 0.010}$** (1)	$0.971_{\pm 0.018}$ (2)
glioma	$0.700_{\pm 0.071}$ (5)	$0.681_{\pm 0.094}$ (8)	$0.673_{\pm 0.084}$ (10)	$0.720_{\pm 0.080}$ (3)	$0.560_{\pm 0.189}$ (13)	$0.679_{\pm 0.078}$ (9)	$0.664_{\pm 0.079}$ (11)
quasar	$0.922_{\pm 0.058}$ (2)	$0.814_{\pm 0.020}$ (6)	$0.916_{\pm 0.068}$ (3)	$0.843_{\pm 0.045}$ (5)	$0.760_{\pm 0.160}$ (7)	**$0.923_{\pm 0.034}$** (1)	$0.922_{\pm 0.038}$ (2)
seismic	$0.125_{\pm 0.029}$ (7)	$0.128_{\pm 0.042}$ (6)	$0.118_{\pm 0.035}$ (8)	$0.143_{\pm 0.022}$ (4)	$0.157_{\pm 0.025}$ (2)	$0.132_{\pm 0.034}$ (5)	$0.116_{\pm 0.041}$ (9)
stroke	$0.144_{\pm 0.057}$ (3)	$0.138_{\pm 0.053}$ (5)	$0.100_{\pm 0.030}$ (10)	$0.134_{\pm 0.063}$ (7)	$0.138_{\pm 0.037}$ (5)	$0.146_{\pm 0.042}$ (2)	**$0.157_{\pm 0.041}$** (1)
vertebral	$0.625_{\pm 0.043}$ (2)	$0.562_{\pm 0.051}$ (10)	$0.572_{\pm 0.067}$ (8)	$0.587_{\pm 0.065}$ (5)	$0.468_{\pm 0.040}$ (12)	$0.606_{\pm 0.052}$ (3)	$0.591_{\pm 0.064}$ (4)
wbc	$0.967_{\pm 0.031}$ (5)	**$0.982_{\pm 0.007}$** (1)	$0.973_{\pm 0.025}$ (3)	$0.971_{\pm 0.015}$ (4)	$0.878_{\pm 0.126}$ (10)	$0.973_{\pm 0.017}$ (3)	$0.976_{\pm 0.010}$ (2)
wine	**$1.000_{\pm 0.000}$** (1)	**$1.000_{\pm 0.000}$** (1)	**$1.000_{\pm 0.000}$** (1)	**$1.000_{\pm 0.000}$** (1)	$0.992_{\pm 0.014}$ (2)	**$1.000_{\pm 0.000}$** (1)	**$1.000_{\pm 0.000}$** (1)
yeast	$0.394_{\pm 0.254}$ (5)	$0.378_{\pm 0.169}$ (7)	$0.338_{\pm 0.221}$ (8)	$0.409_{\pm 0.184}$ (3)	$0.286_{\pm 0.158}$ (10)	$0.402_{\pm 0.229}$ (4)	$0.335_{\pm 0.185}$ (9)
Mean	0.582	0.570	0.564	0.568	0.504	0.590	0.580
Mean Rank	4.300	5.300	6.250	5.350	7.450	3.800	4.800
#1st	1	4	2	1	2	6	3

Table G6: Dataset-level AUCPR for n = 50.

(a) Conventional and supervised baselines

Dataset	DevNet	PReNet	XGBOD	CAD	WSAD-DT	XGBoost	CatBoost
automobile	0.352$_{\pm 0.109}$ (12)	0.367$_{\pm 0.039}$ (10)	0.673$_{\pm 0.094}$ (9)	0.294$_{\pm 0.025}$ (14)	0.334$_{\pm 0.078}$ (13)	0.755$_{\pm 0.130}$ (8)	0.798$_{\pm 0.113}$ (6)
backdoor	0.019$_{\pm 0.020}$ (9)	0.010$_{\pm 0.002}$ (10)	0.023$_{\pm 0.007}$ (8)	0.005$_{\pm 0.003}$ (12)	0.007$_{\pm 0.001}$ (11)	0.004$_{\pm 0.000}$ (13)	0.071$_{\pm 0.016}$ (7)
campaign	0.389$_{\pm 0.069}$ (10)	0.326$_{\pm 0.022}$ (12)	0.452$_{\pm 0.093}$ (9)	0.254$_{\pm 0.037}$ (14)	0.300$_{\pm 0.010}$ (13)	0.519$_{\pm 0.036}$ (7)	0.565$_{\pm 0.049}$ (3)
cardiotocography	0.718$_{\pm 0.030}$ (10)	0.718$_{\pm 0.038}$ (10)	0.833$_{\pm 0.037}$ (4)	0.315$_{\pm 0.151}$ (12)	0.320$_{\pm 0.017}$ (11)	0.747$_{\pm 0.041}$ (8)	0.827$_{\pm 0.013}$ (6)
census	0.221$_{\pm 0.062}$ (8)	0.199$_{\pm 0.079}$ (11)	0.141$_{\pm 0.012}$ (13)	0.101$_{\pm 0.028}$ (14)	0.220$_{\pm 0.055}$ (9)	0.155$_{\pm 0.023}$ (12)	0.206$_{\pm 0.087}$ (10)
churn	**0.573**$_{\pm 0.039}$ (1)	0.357$_{\pm 0.164}$ (11)	0.444$_{\pm 0.021}$ (10)	0.320$_{\pm 0.085}$ (12)	0.200$_{\pm 0.003}$ (13)	0.477$_{\pm 0.082}$ (9)	0.511$_{\pm 0.040}$ (8)
cirrhosis	0.619$_{\pm 0.021}$ (11)	0.510$_{\pm 0.078}$ (13)	0.776$_{\pm 0.047}$ (10)	0.424$_{\pm 0.036}$ (14)	0.542$_{\pm 0.047}$ (12)	0.789$_{\pm 0.060}$ (9)	0.809$_{\pm 0.035}$ (6)
covertype	0.215$_{\pm 0.241}$ (10)	0.033$_{\pm 0.008}$ (11)	0.507$_{\pm 0.188}$ (9)	0.011$_{\pm 0.001}$ (12)	0.005$_{\pm 0.000}$ (14)	0.010$_{\pm 0.000}$ (13)	0.764$_{\pm 0.194}$ (5)
credit	0.215$_{\pm 0.039}$ (11)	0.203$_{\pm 0.018}$ (13)	0.303$_{\pm 0.058}$ (9)	0.213$_{\pm 0.019}$ (12)	0.187$_{\pm 0.003}$ (14)	0.309$_{\pm 0.051}$ (8)	**0.358**$_{\pm 0.066}$ (1)
equip	0.558$_{\pm 0.085}$ (10)	0.574$_{\pm 0.029}$ (9)	**0.914**$_{\pm 0.031}$ (1)	0.303$_{\pm 0.054}$ (13)	0.508$_{\pm 0.016}$ (11)	0.451$_{\pm 0.082}$ (12)	0.746$_{\pm 0.053}$ (8)
gallstone	0.510$_{\pm 0.189}$ (12)	0.512$_{\pm 0.174}$ (11)	0.690$_{\pm 0.175}$ (9)	0.318$_{\pm 0.030}$ (14)	0.376$_{\pm 0.049}$ (13)	0.728$_{\pm 0.142}$ (7)	0.759$_{\pm 0.116}$ (2)
glass	0.896$_{\pm 0.107}$ (9)	0.938$_{\pm 0.049}$ (7)	0.892$_{\pm 0.042}$ (10)	0.231$_{\pm 0.000}$ (12)	0.936$_{\pm 0.081}$ (8)	0.955$_{\pm 0.025}$ (6)	0.972$_{\pm 0.015}$ (4)
glioma	0.748$_{\pm 0.015}$ (2)	0.743$_{\pm 0.038}$ (3)	0.669$_{\pm 0.059}$ (9)	0.382$_{\pm 0.232}$ (11)	0.730$_{\pm 0.035}$ (5)	0.702$_{\pm 0.079}$ (8)	0.725$_{\pm 0.056}$ (6)
quasar	0.641$_{\pm 0.189}$ (11)	0.686$_{\pm 0.167}$ (10)	0.876$_{\pm 0.074}$ (8)	0.191$_{\pm 0.004}$ (13)	0.210$_{\pm 0.002}$ (12)	0.887$_{\pm 0.057}$ (7)	0.935$_{\pm 0.021}$ (6)
seismic	0.094$_{\pm 0.019}$ (12)	0.129$_{\pm 0.015}$ (10)	0.123$_{\pm 0.039}$ (11)	0.069$_{\pm 0.005}$ (13)	0.161$_{\pm 0.020}$ (6)	**0.184**$_{\pm 0.036}$ (1)	0.181$_{\pm 0.055}$ (2)
stroke	0.120$_{\pm 0.036}$ (10)	0.115$_{\pm 0.053}$ (11)	0.112$_{\pm 0.023}$ (12)	0.046$_{\pm 0.004}$ (13)	0.140$_{\pm 0.041}$ (5)	0.043$_{\pm 0.000}$ (14)	0.136$_{\pm 0.069}$ (6)
vertebral	0.668$_{\pm 0.025}$ (10)	0.733$_{\pm 0.050}$ (5)	0.681$_{\pm 0.037}$ (9)	0.387$_{\pm 0.076}$ (12)	0.239$_{\pm 0.009}$ (13)	0.764$_{\pm 0.051}$ (3)	0.727$_{\pm 0.051}$ (6)
wbc	0.946$_{\pm 0.030}$ (8)	0.941$_{\pm 0.028}$ (9)	0.978$_{\pm 0.011}$ (6)	0.516$_{\pm 0.270}$ (11)	0.957$_{\pm 0.010}$ (7)	0.979$_{\pm 0.009}$ (5)	0.980$_{\pm 0.009}$ (4)
wine	0.679$_{\pm 0.146}$ (5)	0.973$_{\pm 0.047}$ (4)	0.999$_{\pm 0.002}$ (2)	0.297$_{\pm 0.028}$ (6)	0.236$_{\pm 0.021}$ (7)	**1.000**$_{\pm 0.000}$ (1)	**1.000**$_{\pm 0.000}$ (1)
yeast	0.684$_{\pm 0.056}$ (2)	**0.688**$_{\pm 0.082}$ (1)	0.550$_{\pm 0.076}$ (10)	0.140$_{\pm 0.153}$ (13)	0.524$_{\pm 0.093}$ (11)	0.581$_{\pm 0.101}$ (9)	0.635$_{\pm 0.089}$ (7)
Mean	0.493	0.488	0.582	0.241	0.356	0.552	0.635
Mean Rank	9.600	9.950	9.450	13.550	11.450	8.825	5.925
#1st	1	1	1	0	0	2	2

(b) Pretrained, LLM-based, and LENS-AD methods

Dataset	TabPFN	LimiX	TabICL	CAAFE	FeatLLM	LENS-AD (TabPFN)	LENS-AD (TabICL)
automobile	0.806$_{\pm 0.115}$ (5)	0.839$_{\pm 0.093}$ (3)	0.822$_{\pm 0.117}$ (4)	0.796$_{\pm 0.125}$ (7)	0.365$_{\pm 0.117}$ (11)	0.876$_{\pm 0.124}$ (2)	**0.881**$_{\pm 0.104}$ (1)
backdoor	0.120$_{\pm 0.071}$ (5)	0.135$_{\pm 0.052}$ (3)	0.120$_{\pm 0.038}$ (5)	0.073$_{\pm 0.007}$ (6)	0.148$_{\pm 0.091}$ (2)	0.128$_{\pm 0.035}$ (4)	**0.161**$_{\pm 0.035}$ (1)
campaign	**0.580**$_{\pm 0.044}$ (1)	0.569$_{\pm 0.020}$ (2)	0.515$_{\pm 0.106}$ (8)	0.530$_{\pm 0.012}$ (5)	0.327$_{\pm 0.094}$ (11)	0.554$_{\pm 0.083}$ (4)	0.526$_{\pm 0.089}$ (6)
cardiotocography	**0.844**$_{\pm 0.019}$ (1)	0.832$_{\pm 0.009}$ (5)	0.820$_{\pm 0.031}$ (7)	0.820$_{\pm 0.009}$ (7)	0.746$_{\pm 0.048}$ (9)	0.843$_{\pm 0.018}$ (2)	0.838$_{\pm 0.024}$ (3)
census	0.238$_{\pm 0.091}$ (6)	0.278$_{\pm 0.089}$ (4)	0.302$_{\pm 0.038}$ (3)	0.224$_{\pm 0.097}$ (7)	0.240$_{\pm 0.081}$ (5)	0.318$_{\pm 0.096}$ (2)	**0.350**$_{\pm 0.058}$ (1)
churn	0.539$_{\pm 0.034}$ (6)	0.556$_{\pm 0.026}$ (3)	0.541$_{\pm 0.057}$ (5)	0.554$_{\pm 0.032}$ (4)	0.527$_{\pm 0.059}$ (7)	0.556$_{\pm 0.038}$ (3)	0.567$_{\pm 0.037}$ (2)
cirrhosis	0.819$_{\pm 0.028}$ (4)	**0.833**$_{\pm 0.042}$ (1)	0.805$_{\pm 0.036}$ (7)	0.818$_{\pm 0.038}$ (5)	0.793$_{\pm 0.009}$ (8)	0.824$_{\pm 0.043}$ (3)	0.832$_{\pm 0.055}$ (2)
covertype	0.824$_{\pm 0.254}$ (4)	0.701$_{\pm 0.396}$ (7)	0.837$_{\pm 0.201}$ (3)	0.750$_{\pm 0.229}$ (6)	0.568$_{\pm 0.136}$ (8)	**0.921**$_{\pm 0.050}$ (1)	0.920$_{\pm 0.071}$ (2)
credit	0.336$_{\pm 0.084}$ (5)	0.340$_{\pm 0.073}$ (4)	0.333$_{\pm 0.083}$ (7)	0.345$_{\pm 0.093}$ (3)	0.275$_{\pm 0.037}$ (10)	0.355$_{\pm 0.087}$ (2)	0.335$_{\pm 0.074}$ (6)
equip	0.829$_{\pm 0.056}$ (6)	0.838$_{\pm 0.071}$ (5)	0.869$_{\pm 0.053}$ (2)	0.805$_{\pm 0.109}$ (7)	0.746$_{\pm 0.158}$ (8)	0.843$_{\pm 0.032}$ (4)	0.847$_{\pm 0.016}$ (3)
gallstone	**0.778**$_{\pm 0.119}$ (1)	0.740$_{\pm 0.145}$ (5)	0.753$_{\pm 0.139}$ (3)	0.743$_{\pm 0.109}$ (4)	0.546$_{\pm 0.089}$ (10)	0.735$_{\pm 0.176}$ (6)	0.715$_{\pm 0.148}$ (8)
glass	**0.984**$_{\pm 0.011}$ (1)	0.980$_{\pm 0.017}$ (2)	**0.984**$_{\pm 0.011}$ (1)	0.966$_{\pm 0.031}$ (5)	0.791$_{\pm 0.072}$ (11)	0.979$_{\pm 0.012}$ (3)	0.979$_{\pm 0.015}$ (3)
glioma	0.722$_{\pm 0.097}$ (7)	0.725$_{\pm 0.074}$ (6)	0.722$_{\pm 0.069}$ (7)	**0.753**$_{\pm 0.077}$ (1)	0.561$_{\pm 0.190}$ (10)	0.734$_{\pm 0.082}$ (4)	0.725$_{\pm 0.078}$ (6)
quasar	0.954$_{\pm 0.016}$ (3)	0.949$_{\pm 0.021}$ (4)	**0.960**$_{\pm 0.011}$ (1)	0.942$_{\pm 0.017}$ (5)	0.819$_{\pm 0.161}$ (9)	0.956$_{\pm 0.009}$ (2)	0.956$_{\pm 0.010}$ (2)
seismic	0.179$_{\pm 0.064}$ (3)	0.176$_{\pm 0.048}$ (4)	0.159$_{\pm 0.074}$ (7)	0.179$_{\pm 0.037}$ (3)	0.164$_{\pm 0.019}$ (5)	0.153$_{\pm 0.037}$ (9)	0.156$_{\pm 0.055}$ (8)
stroke	0.127$_{\pm 0.058}$ (9)	0.134$_{\pm 0.062}$ (7)	0.149$_{\pm 0.058}$ (3)	0.145$_{\pm 0.052}$ (4)	0.133$_{\pm 0.027}$ (8)	0.164$_{\pm 0.058}$ (2)	**0.165**$_{\pm 0.043}$ (1)
vertebral	**0.784**$_{\pm 0.046}$ (1)	0.757$_{\pm 0.046}$ (4)	0.765$_{\pm 0.070}$ (2)	0.725$_{\pm 0.022}$ (7)	0.539$_{\pm 0.044}$ (11)	0.717$_{\pm 0.024}$ (8)	0.727$_{\pm 0.012}$ (6)
wbc	0.984$_{\pm 0.009}$ (3)	0.985$_{\pm 0.006}$ (2)	0.980$_{\pm 0.009}$ (4)	0.978$_{\pm 0.010}$ (6)	0.891$_{\pm 0.119}$ (10)	**0.986**$_{\pm 0.004}$ (1)	0.984$_{\pm 0.011}$ (3)
wine	**1.000**$_{\pm 0.000}$ (1)	**1.000**$_{\pm 0.000}$ (1)	**1.000**$_{\pm 0.000}$ (1)	**1.000**$_{\pm 0.000}$ (1)	0.997$_{\pm 0.002}$ (3)	**1.000**$_{\pm 0.000}$ (1)	**1.000**$_{\pm 0.000}$ (1)
yeast	0.660$_{\pm 0.040}$ (4)	0.642$_{\pm 0.070}$ (6)	0.660$_{\pm 0.026}$ (4)	0.583$_{\pm 0.172}$ (8)	0.293$_{\pm 0.126}$ (12)	0.672$_{\pm 0.044}$ (3)	0.650$_{\pm 0.050}$ (5)
Mean	0.655	0.651	0.655	0.636	0.523	0.666	0.666
Mean Rank	4.325	4.375	4.675	5.775	9.450	3.675	3.975
#1st	6	2	3	2	0	3	5

G.3 Unsupervised Results

The following tables expand Table F1. The eight full-feature scorers operate on the original feature space; the three LENS-AD rows use the complete LKE, SSD, and SGMA pipeline with the indicated member scorer.

Table G7: Dataset-level unsupervised AUCPR for $n = 10$. Entries are mean $\pm$ standard deviation over three root seeds.

(a) Conventional full-feature scorers

Dataset	PCA	IForest	LOF	OCSVM	ECOD	COPOD	KNN
automobile	$0.351_{\pm 0.117}$ (2)	$0.349_{\pm 0.123}$ (3)	$0.277_{\pm 0.013}$ (10)	$0.270_{\pm 0.018}$ (11)	$0.330_{\pm 0.039}$ (6)	$0.324_{\pm 0.041}$ (7)	$0.320_{\pm 0.073}$ (8)
backdoor	$0.004_{\pm 0.000}$ (8)	$0.012_{\pm 0.010}$ (3)	$0.007_{\pm 0.003}$ (6)	$0.009_{\pm 0.002}$ (5)	$0.015_{\pm 0.002}$ (2)	$0.012_{\pm 0.002}$ (3)	$0.012_{\pm 0.009}$ (3)
campaign	$0.306_{\pm 0.020}$ (5)	$0.297_{\pm 0.036}$ (7)	$0.231_{\pm 0.008}$ (11)	$0.277_{\pm 0.009}$ (9)	$0.301_{\pm 0.014}$ (6)	$0.337_{\pm 0.027}$ (3)	$0.287_{\pm 0.006}$ (8)
cardiotocography	$0.286_{\pm 0.019}$ (11)	$0.408_{\pm 0.103}$ (8)	$0.373_{\pm 0.109}$ (10)	$0.497_{\pm 0.154}$ (3)	**$0.512_{\pm 0.022}$** (1)	$0.423_{\pm 0.048}$ (7)	$0.498_{\pm 0.101}$ (2)
census	$0.245_{\pm 0.126}$ (3)	$0.049_{\pm 0.011}$ (11)	$0.227_{\pm 0.059}$ (4)	$0.272_{\pm 0.016}$ (2)	$0.071_{\pm 0.002}$ (9)	$0.083_{\pm 0.004}$ (7)	**$0.291_{\pm 0.005}$** (1)
churn	$0.271_{\pm 0.097}$ (3)	$0.194_{\pm 0.008}$ (9)	$0.278_{\pm 0.050}$ (2)	**$0.282_{\pm 0.053}$** (1)	$0.213_{\pm 0.016}$ (7)	$0.176_{\pm 0.002}$ (11)	$0.244_{\pm 0.057}$ (4)
cirrhosis	$0.666_{\pm 0.107}$ (6)	$0.705_{\pm 0.090}$ (3)	$0.418_{\pm 0.029}$ (10)	$0.459_{\pm 0.086}$ (9)	$0.616_{\pm 0.036}$ (7)	**$0.776_{\pm 0.057}$** (1)	$0.509_{\pm 0.069}$ (8)
covertype	$0.015_{\pm 0.008}$ (8)	$0.080_{\pm 0.045}$ (5)	$0.011_{\pm 0.004}$ (9)	$0.009_{\pm 0.002}$ (10)	$0.193_{\pm 0.045}$ (2)	$0.135_{\pm 0.046}$ (4)	$0.008_{\pm 0.002}$ (11)
credit	$0.181_{\pm 0.010}$ (10)	$0.244_{\pm 0.025}$ (5)	$0.209_{\pm 0.009}$ (8)	$0.190_{\pm 0.011}$ (9)	$0.247_{\pm 0.007}$ (4)	$0.231_{\pm 0.006}$ (6)	$0.190_{\pm 0.004}$ (9)
equip	$0.593_{\pm 0.122}$ (7)	$0.590_{\pm 0.103}$ (8)	$0.792_{\pm 0.154}$ (6)	$0.886_{\pm 0.023}$ (4)	**$0.904_{\pm 0.012}$** (1)	$0.892_{\pm 0.016}$ (3)	$0.895_{\pm 0.012}$ (2)
gallstone	$0.374_{\pm 0.036}$ (4)	$0.369_{\pm 0.093}$ (6)	$0.365_{\pm 0.036}$ (7)	$0.376_{\pm 0.053}$ (3)	$0.342_{\pm 0.034}$ (10)	$0.345_{\pm 0.038}$ (9)	$0.389_{\pm 0.057}$ (2)
glass	$0.659_{\pm 0.177}$ (6)	$0.724_{\pm 0.109}$ (2)	$0.379_{\pm 0.071}$ (11)	$0.526_{\pm 0.155}$ (9)	$0.548_{\pm 0.051}$ (8)	$0.513_{\pm 0.086}$ (10)	$0.646_{\pm 0.032}$ (7)
glioma	$0.411_{\pm 0.019}$ (8)	$0.369_{\pm 0.065}$ (11)	$0.560_{\pm 0.145}$ (2)	$0.557_{\pm 0.139}$ (3)	$0.410_{\pm 0.033}$ (9)	$0.449_{\pm 0.032}$ (7)	**$0.643_{\pm 0.119}$** (1)
quasar	**$0.214_{\pm 0.036}$** (1)	$0.180_{\pm 0.018}$ (9)	$0.198_{\pm 0.004}$ (3)	$0.195_{\pm 0.002}$ (4)	$0.188_{\pm 0.002}$ (6)	$0.194_{\pm 0.002}$ (5)	$0.195_{\pm 0.003}$ (4)
seismic	$0.160_{\pm 0.022}$ (4)	$0.171_{\pm 0.035}$ (2)	$0.103_{\pm 0.039}$ (10)	$0.142_{\pm 0.022}$ (7)	$0.160_{\pm 0.010}$ (4)	**$0.177_{\pm 0.010}$** (1)	$0.154_{\pm 0.022}$ (6)
stroke	$0.067_{\pm 0.001}$ (11)	$0.086_{\pm 0.039}$ (6)	$0.075_{\pm 0.015}$ (9)	$0.113_{\pm 0.008}$ (5)	$0.081_{\pm 0.017}$ (8)	$0.074_{\pm 0.017}$ (10)	$0.124_{\pm 0.027}$ (4)
vertebral	$0.280_{\pm 0.022}$ (3)	$0.225_{\pm 0.008}$ (8)	**$0.327_{\pm 0.025}$** (1)	$0.296_{\pm 0.068}$ (2)	$0.256_{\pm 0.008}$ (5)	$0.216_{\pm 0.003}$ (10)	$0.224_{\pm 0.008}$ (9)
wbc	$0.746_{\pm 0.153}$ (8)	$0.866_{\pm 0.085}$ (3)	$0.519_{\pm 0.055}$ (10)	$0.783_{\pm 0.192}$ (6)	$0.493_{\pm 0.052}$ (11)	$0.791_{\pm 0.075}$ (5)	**$0.941_{\pm 0.010}$** (1)
wine	$0.288_{\pm 0.102}$ (8)	$0.541_{\pm 0.185}$ (4)	$0.292_{\pm 0.004}$ (7)	$0.231_{\pm 0.039}$ (10)	$0.391_{\pm 0.016}$ (6)	$0.528_{\pm 0.129}$ (5)	$0.208_{\pm 0.051}$ (11)
yeast	$0.151_{\pm 0.122}$ (6)	**$0.286_{\pm 0.159}$** (1)	$0.148_{\pm 0.128}$ (7)	$0.199_{\pm 0.068}$ (5)	$0.199_{\pm 0.051}$ (5)	$0.283_{\pm 0.088}$ (2)	$0.254_{\pm 0.079}$ (3)
Mean	0.313	0.337	0.289	0.329	0.324	0.348	0.352
Mean Rank	6.450	5.850	7.450	6.200	5.950	6.000	5.400
#1st	1	1	1	1	2	2	3

(b) TabPFN-Unsupervised and LENS-AD variants

Dataset	TabPFN-Unsup.	LENS-AD (TabPFN)	LENS-AD (COPOD)	LENS-AD (IForest)
automobile	$0.281_{\pm 0.068}$ (9)	**$0.421_{\pm 0.124}$** (1)	$0.331_{\pm 0.042}$ (5)	$0.333_{\pm 0.088}$ (4)
backdoor	$0.005_{\pm 0.002}$ (7)	$0.010_{\pm 0.010}$ (4)	$0.010_{\pm 0.001}$ (4)	**$0.021_{\pm 0.020}$** (1)
campaign	$0.271_{\pm 0.081}$ (10)	$0.313_{\pm 0.066}$ (4)	$0.377_{\pm 0.058}$ (2)	**$0.396_{\pm 0.085}$** (1)
cardiotocography	$0.427_{\pm 0.189}$ (5)	$0.450_{\pm 0.176}$ (4)	$0.406_{\pm 0.029}$ (9)	$0.424_{\pm 0.091}$ (6)
census	$0.058_{\pm 0.014}$ (10)	$0.123_{\pm 0.036}$ (5)	$0.093_{\pm 0.017}$ (6)	$0.074_{\pm 0.032}$ (8)
churn	$0.233_{\pm 0.032}$ (5)	$0.217_{\pm 0.033}$ (6)	$0.186_{\pm 0.004}$ (10)	$0.210_{\pm 0.021}$ (8)
cirrhosis	$0.666_{\pm 0.166}$ (6)	$0.685_{\pm 0.046}$ (4)	$0.752_{\pm 0.029}$ (2)	$0.683_{\pm 0.090}$ (5)
covertype	$0.017_{\pm 0.005}$ (7)	$0.037_{\pm 0.021}$ (6)	$0.153_{\pm 0.020}$ (3)	**$0.510_{\pm 0.160}$** (1)
credit	$0.251_{\pm 0.084}$ (3)	$0.261_{\pm 0.017}$ (2)	$0.224_{\pm 0.003}$ (7)	**$0.262_{\pm 0.017}$** (1)
equip	$0.855_{\pm 0.027}$ (5)	$0.579_{\pm 0.198}$ (9)	$0.546_{\pm 0.059}$ (10)	$0.233_{\pm 0.064}$ (11)
gallstone	**$0.421_{\pm 0.094}$** (1)	$0.364_{\pm 0.156}$ (8)	$0.372_{\pm 0.064}$ (5)	$0.300_{\pm 0.031}$ (11)
glass	$0.672_{\pm 0.158}$ (5)	$0.707_{\pm 0.144}$ (3)	$0.692_{\pm 0.062}$ (4)	**$0.747_{\pm 0.078}$** (1)
glioma	$0.393_{\pm 0.137}$ (10)	$0.508_{\pm 0.127}$ (4)	$0.495_{\pm 0.043}$ (5)	$0.490_{\pm 0.072}$ (6)
quasar	$0.187_{\pm 0.024}$ (7)	$0.174_{\pm 0.039}$ (10)	$0.185_{\pm 0.034}$ (8)	$0.206_{\pm 0.019}$ (2)
seismic	$0.165_{\pm 0.066}$ (3)	$0.121_{\pm 0.018}$ (8)	$0.157_{\pm 0.010}$ (5)	$0.113_{\pm 0.053}$ (9)
stroke	$0.083_{\pm 0.042}$ (7)	$0.130_{\pm 0.020}$ (3)	$0.155_{\pm 0.020}$ (2)	**$0.160_{\pm 0.012}$** (1)
vertebral	$0.225_{\pm 0.011}$ (8)	$0.257_{\pm 0.037}$ (4)	$0.226_{\pm 0.010}$ (7)	$0.251_{\pm 0.042}$ (6)
wbc	$0.670_{\pm 0.088}$ (9)	$0.775_{\pm 0.061}$ (7)	$0.844_{\pm 0.072}$ (4)	$0.925_{\pm 0.077}$ (2)
wine	$0.276_{\pm 0.044}$ (9)	$0.548_{\pm 0.275}$ (3)	$0.731_{\pm 0.169}$ (2)	**$0.748_{\pm 0.325}$** (1)
yeast	$0.130_{\pm 0.043}$ (8)	$0.056_{\pm 0.004}$ (10)	$0.238_{\pm 0.042}$ (4)	$0.113_{\pm 0.072}$ (9)
Mean	0.314	0.337	0.359	0.360
Mean Rank	6.950	5.550	5.400	4.800
#1st	1	1	0	7

Table G8: Dataset-level unsupervised AUCPR for $n = 20$. Entries are mean $\pm$ standard deviation over three root seeds.

(a) Conventional full-feature scorers

Dataset	PCA	IForest	LOF	OCSVM	ECOD	COPOD	KNN
automobile	$0.353_{\pm 0.190}$ (4)	$0.344_{\pm 0.085}$ (5)	$0.320_{\pm 0.040}$ (8)	$0.264_{\pm 0.024}$ (9)	$0.336_{\pm 0.043}$ (6)	$0.362_{\pm 0.047}$ (2)	$0.320_{\pm 0.057}$ (8)
backdoor	$0.005_{\pm 0.002}$ (7)	$0.007_{\pm 0.004}$ (5)	$0.007_{\pm 0.003}$ (5)	$0.008_{\pm 0.003}$ (4)	**$0.015_{\pm 0.002}$** (1)	$0.011_{\pm 0.002}$ (2)	$0.008_{\pm 0.002}$ (4)
campaign	$0.271_{\pm 0.024}$ (8)	$0.276_{\pm 0.050}$ (7)	$0.231_{\pm 0.007}$ (11)	$0.253_{\pm 0.015}$ (9)	$0.311_{\pm 0.006}$ (4)	$0.352_{\pm 0.008}$ (3)	$0.296_{\pm 0.011}$ (6)
cardiotocography	$0.309_{\pm 0.050}$ (11)	$0.474_{\pm 0.064}$ (5)	$0.381_{\pm 0.096}$ (9)	**$0.546_{\pm 0.100}$** (1)	$0.513_{\pm 0.021}$ (2)	$0.424_{\pm 0.046}$ (7)	$0.503_{\pm 0.046}$ (3)
census	$0.251_{\pm 0.103}$ (3)	$0.060_{\pm 0.007}$ (10)	$0.208_{\pm 0.076}$ (4)	$0.275_{\pm 0.018}$ (2)	$0.071_{\pm 0.003}$ (9)	$0.083_{\pm 0.004}$ (8)	**$0.300_{\pm 0.003}$** (1)
churn	$0.280_{\pm 0.025}$ (3)	$0.219_{\pm 0.013}$ (8)	$0.283_{\pm 0.044}$ (2)	**$0.298_{\pm 0.053}$** (1)	$0.219_{\pm 0.004}$ (8)	$0.177_{\pm 0.003}$ (10)	$0.246_{\pm 0.072}$ (4)
cirrhosis	$0.643_{\pm 0.041}$ (7)	**$0.781_{\pm 0.071}$** (1)	$0.347_{\pm 0.078}$ (11)	$0.413_{\pm 0.092}$ (10)	$0.619_{\pm 0.036}$ (8)	$0.779_{\pm 0.056}$ (2)	$0.465_{\pm 0.047}$ (9)
covertype	$0.015_{\pm 0.010}$ (8)	$0.179_{\pm 0.195}$ (3)	$0.009_{\pm 0.001}$ (10)	$0.011_{\pm 0.002}$ (9)	$0.227_{\pm 0.015}$ (2)	$0.154_{\pm 0.022}$ (4)	$0.008_{\pm 0.001}$ (11)
credit	$0.180_{\pm 0.005}$ (11)	$0.243_{\pm 0.009}$ (4)	$0.209_{\pm 0.007}$ (8)	$0.200_{\pm 0.009}$ (9)	$0.253_{\pm 0.003}$ (2)	$0.238_{\pm 0.002}$ (5)	$0.191_{\pm 0.010}$ (10)
equip	$0.581_{\pm 0.072}$ (7)	$0.815_{\pm 0.043}$ (5)	$0.728_{\pm 0.141}$ (6)	$0.892_{\pm 0.019}$ (3)	**$0.903_{\pm 0.010}$** (1)	$0.892_{\pm 0.015}$ (3)	$0.895_{\pm 0.006}$ (2)
gallstone	**$0.468_{\pm 0.046}$** (1)	$0.345_{\pm 0.072}$ (8)	$0.353_{\pm 0.029}$ (6)	$0.367_{\pm 0.046}$ (5)	$0.341_{\pm 0.035}$ (10)	$0.348_{\pm 0.046}$ (7)	$0.382_{\pm 0.039}$ (3)
glass	$0.649_{\pm 0.121}$ (7)	$0.711_{\pm 0.146}$ (3)	$0.334_{\pm 0.125}$ (11)	$0.571_{\pm 0.027}$ (8)	$0.551_{\pm 0.057}$ (10)	$0.557_{\pm 0.023}$ (9)	$0.699_{\pm 0.074}$ (5)
glioma	$0.387_{\pm 0.047}$ (9)	$0.342_{\pm 0.018}$ (11)	**$0.580_{\pm 0.032}$** (1)	$0.504_{\pm 0.093}$ (3)	$0.407_{\pm 0.036}$ (8)	$0.449_{\pm 0.035}$ (5)	$0.503_{\pm 0.108}$ (4)
quasar	$0.223_{\pm 0.038}$ (3)	$0.225_{\pm 0.043}$ (2)	$0.196_{\pm 0.003}$ (4)	$0.194_{\pm 0.001}$ (5)	$0.188_{\pm 0.002}$ (7)	$0.194_{\pm 0.002}$ (5)	$0.193_{\pm 0.002}$ (6)
seismic	$0.153_{\pm 0.032}$ (6)	**$0.187_{\pm 0.030}$** (1)	$0.091_{\pm 0.026}$ (11)	$0.135_{\pm 0.012}$ (8)	$0.160_{\pm 0.010}$ (3)	$0.177_{\pm 0.010}$ (2)	$0.156_{\pm 0.020}$ (5)
stroke	$0.086_{\pm 0.008}$ (8)	$0.117_{\pm 0.022}$ (4)	$0.070_{\pm 0.003}$ (11)	$0.114_{\pm 0.012}$ (5)	$0.081_{\pm 0.016}$ (9)	$0.074_{\pm 0.018}$ (10)	$0.109_{\pm 0.015}$ (6)
vertebral	$0.280_{\pm 0.030}$ (2)	$0.223_{\pm 0.012}$ (9)	**$0.327_{\pm 0.056}$** (1)	$0.272_{\pm 0.017}$ (3)	$0.254_{\pm 0.005}$ (5)	$0.217_{\pm 0.003}$ (11)	$0.222_{\pm 0.003}$ (10)
wbc	$0.683_{\pm 0.156}$ (8)	$0.799_{\pm 0.109}$ (4)	$0.491_{\pm 0.079}$ (9)	$0.788_{\pm 0.105}$ (5)	$0.487_{\pm 0.046}$ (10)	$0.788_{\pm 0.076}$ (5)	**$0.935_{\pm 0.014}$** (1)
wine	$0.333_{\pm 0.082}$ (7)	$0.562_{\pm 0.071}$ (4)	$0.277_{\pm 0.038}$ (9)	$0.233_{\pm 0.013}$ (10)	$0.416_{\pm 0.031}$ (6)	$0.650_{\pm 0.060}$ (3)	$0.226_{\pm 0.048}$ (11)
yeast	$0.190_{\pm 0.110}$ (6)	$0.282_{\pm 0.166}$ (2)	$0.135_{\pm 0.142}$ (8)	$0.202_{\pm 0.030}$ (4)	$0.201_{\pm 0.052}$ (5)	**$0.286_{\pm 0.089}$** (1)	$0.189_{\pm 0.045}$ (7)
Mean	0.317	0.360	0.279	0.327	0.328	0.361	0.342
Mean Rank	6.650	5.350	7.600	5.900	6.050	5.400	6.000
#1st	1	2	2	2	2	1	2

(b) TabPFN-Unsupervised and LENS-AD variants

Dataset	TabPFN-Unsup.	LENS-AD (TabPFN)	LENS-AD (COPOD)	LENS-AD (IForest)
automobile	$0.359_{\pm 0.108}$ (3)	$0.359_{\pm 0.077}$ (3)	$0.329_{\pm 0.029}$ (7)	**$0.365_{\pm 0.079}$** (1)
backdoor	$0.006_{\pm 0.002}$ (6)	$0.008_{\pm 0.002}$ (4)	$0.009_{\pm 0.001}$ (3)	$0.011_{\pm 0.003}$ (2)
campaign	$0.236_{\pm 0.037}$ (10)	$0.302_{\pm 0.075}$ (5)	**$0.394_{\pm 0.008}$** (1)	$0.363_{\pm 0.072}$ (2)
cardiotocography	$0.315_{\pm 0.083}$ (10)	$0.390_{\pm 0.004}$ (8)	$0.427_{\pm 0.012}$ (6)	$0.475_{\pm 0.082}$ (4)
census	$0.100_{\pm 0.025}$ (6)	$0.153_{\pm 0.033}$ (5)	$0.100_{\pm 0.010}$ (6)	$0.091_{\pm 0.012}$ (7)
churn	$0.240_{\pm 0.039}$ (5)	$0.230_{\pm 0.058}$ (7)	$0.184_{\pm 0.005}$ (9)	$0.237_{\pm 0.021}$ (6)
cirrhosis	$0.773_{\pm 0.082}$ (3)	$0.729_{\pm 0.022}$ (6)	$0.763_{\pm 0.040}$ (4)	$0.733_{\pm 0.061}$ (5)
covertype	$0.062_{\pm 0.069}$ (6)	$0.048_{\pm 0.016}$ (7)	$0.139_{\pm 0.003}$ (5)	**$0.454_{\pm 0.104}$** (1)
credit	$0.222_{\pm 0.064}$ (7)	$0.244_{\pm 0.039}$ (3)	$0.223_{\pm 0.000}$ (6)	**$0.262_{\pm 0.044}$** (1)
equip	$0.857_{\pm 0.030}$ (4)	$0.448_{\pm 0.026}$ (9)	$0.510_{\pm 0.042}$ (8)	$0.348_{\pm 0.167}$ (10)
gallstone	$0.464_{\pm 0.029}$ (2)	$0.297_{\pm 0.013}$ (11)	$0.374_{\pm 0.079}$ (4)	$0.343_{\pm 0.074}$ (9)
glass	$0.667_{\pm 0.132}$ (6)	**$0.738_{\pm 0.076}$** (1)	$0.701_{\pm 0.060}$ (4)	$0.735_{\pm 0.109}$ (2)
glioma	$0.377_{\pm 0.064}$ (10)	$0.423_{\pm 0.190}$ (7)	$0.505_{\pm 0.028}$ (2)	$0.435_{\pm 0.041}$ (6)
quasar	**$0.269_{\pm 0.053}$** (1)	$0.169_{\pm 0.020}$ (9)	$0.169_{\pm 0.013}$ (9)	$0.187_{\pm 0.003}$ (8)
seismic	$0.139_{\pm 0.053}$ (7)	$0.130_{\pm 0.028}$ (9)	$0.157_{\pm 0.010}$ (4)	$0.103_{\pm 0.035}$ (10)
stroke	$0.106_{\pm 0.017}$ (7)	$0.161_{\pm 0.022}$ (2)	**$0.170_{\pm 0.030}$** (1)	$0.160_{\pm 0.028}$ (3)
vertebral	$0.248_{\pm 0.023}$ (7)	$0.266_{\pm 0.048}$ (4)	$0.229_{\pm 0.011}$ (8)	$0.250_{\pm 0.045}$ (6)
wbc	$0.686_{\pm 0.063}$ (7)	$0.779_{\pm 0.045}$ (6)	$0.855_{\pm 0.066}$ (3)	$0.881_{\pm 0.045}$ (2)
wine	$0.330_{\pm 0.016}$ (8)	$0.460_{\pm 0.153}$ (5)	**$0.780_{\pm 0.038}$** (1)	$0.748_{\pm 0.194}$ (2)
yeast	$0.124_{\pm 0.026}$ (10)	$0.083_{\pm 0.016}$ (11)	$0.236_{\pm 0.041}$ (3)	$0.127_{\pm 0.035}$ (9)
Mean	0.329	0.321	0.363	0.365
Mean Rank	6.650	6.450	4.950	5.000
#1st	1	1	3	3

Table G9: Dataset-level unsupervised AUCPR for $n = 50$. Entries are mean $\pm$ standard deviation over three root seeds.

(a) Conventional full-feature scorers

Dataset	PCA	IForest	LOF	OCSVM	ECOD	COPOD	KNN
automobile	$0.374_{\pm 0.063}$ (3)	$0.346_{\pm 0.075}$ (6)	$0.341_{\pm 0.063}$ (8)	$0.250_{\pm 0.037}$ (11)	$0.336_{\pm 0.045}$ (9)	$0.342_{\pm 0.031}$ (7)	$0.321_{\pm 0.055}$ (10)
backdoor	$0.004_{\pm 0.001}$ (10)	$0.011_{\pm 0.002}$ (5)	$0.008_{\pm 0.002}$ (7)	$0.012_{\pm 0.004}$ (4)	$0.015_{\pm 0.002}$ (2)	$0.011_{\pm 0.002}$ (5)	$0.013_{\pm 0.005}$ (3)
campaign	$0.264_{\pm 0.028}$ (8)	$0.311_{\pm 0.039}$ (5)	$0.285_{\pm 0.014}$ (7)	$0.254_{\pm 0.017}$ (9)	$0.319_{\pm 0.007}$ (4)	$0.356_{\pm 0.005}$ (3)	$0.301_{\pm 0.014}$ (6)
cardiotocography	$0.282_{\pm 0.050}$ (10)	$0.488_{\pm 0.031}$ (5)	**$0.567_{\pm 0.037}$** (1)	$0.565_{\pm 0.047}$ (2)	$0.515_{\pm 0.021}$ (3)	$0.426_{\pm 0.046}$ (8)	$0.499_{\pm 0.029}$ (4)
census	$0.132_{\pm 0.011}$ (4)	$0.061_{\pm 0.004}$ (11)	$0.300_{\pm 0.005}$ (2)	$0.275_{\pm 0.016}$ (3)	$0.071_{\pm 0.002}$ (10)	$0.085_{\pm 0.005}$ (8)	**$0.309_{\pm 0.003}$** (1)
churn	$0.249_{\pm 0.014}$ (2)	$0.204_{\pm 0.012}$ (8)	$0.227_{\pm 0.011}$ (3)	**$0.275_{\pm 0.024}$** (1)	$0.220_{\pm 0.003}$ (4)	$0.177_{\pm 0.003}$ (10)	$0.207_{\pm 0.007}$ (7)
cirrhosis	$0.573_{\pm 0.059}$ (8)	$0.779_{\pm 0.031}$ (2)	$0.465_{\pm 0.085}$ (10)	$0.426_{\pm 0.087}$ (11)	$0.620_{\pm 0.041}$ (7)	**$0.783_{\pm 0.059}$** (1)	$0.528_{\pm 0.120}$ (9)
covertype	$0.024_{\pm 0.012}$ (8)	$0.177_{\pm 0.044}$ (3)	$0.008_{\pm 0.001}$ (11)	$0.015_{\pm 0.004}$ (9)	$0.255_{\pm 0.028}$ (2)	$0.164_{\pm 0.021}$ (4)	$0.011_{\pm 0.001}$ (10)
credit	$0.184_{\pm 0.003}$ (9)	$0.239_{\pm 0.012}$ (4)	$0.194_{\pm 0.005}$ (7)	$0.197_{\pm 0.004}$ (6)	$0.255_{\pm 0.004}$ (2)	$0.240_{\pm 0.002}$ (3)	$0.186_{\pm 0.006}$ (8)
equip	$0.519_{\pm 0.009}$ (10)	$0.893_{\pm 0.013}$ (5)	$0.902_{\pm 0.006}$ (3)	$0.901_{\pm 0.012}$ (4)	$0.903_{\pm 0.010}$ (2)	$0.892_{\pm 0.015}$ (6)	$0.901_{\pm 0.006}$ (4)
gallstone	$0.483_{\pm 0.058}$ (2)	$0.392_{\pm 0.041}$ (4)	$0.390_{\pm 0.046}$ (6)	$0.366_{\pm 0.050}$ (9)	$0.344_{\pm 0.039}$ (10)	$0.367_{\pm 0.045}$ (8)	$0.391_{\pm 0.038}$ (5)
glass	$0.682_{\pm 0.031}$ (3)	$0.657_{\pm 0.042}$ (5)	$0.702_{\pm 0.048}$ (2)	$0.582_{\pm 0.056}$ (9)	$0.508_{\pm 0.026}$ (11)	$0.559_{\pm 0.034}$ (10)	$0.649_{\pm 0.087}$ (7)
glioma	$0.399_{\pm 0.067}$ (9)	$0.391_{\pm 0.052}$ (10)	$0.481_{\pm 0.145}$ (3)	$0.470_{\pm 0.113}$ (4)	$0.408_{\pm 0.043}$ (7)	$0.450_{\pm 0.040}$ (6)	$0.496_{\pm 0.116}$ (2)
quasar	$0.269_{\pm 0.035}$ (2)	$0.243_{\pm 0.033}$ (4)	$0.194_{\pm 0.001}$ (6)	$0.195_{\pm 0.002}$ (5)	$0.188_{\pm 0.002}$ (7)	$0.194_{\pm 0.002}$ (6)	$0.169_{\pm 0.011}$ (9)
seismic	$0.095_{\pm 0.025}$ (10)	$0.164_{\pm 0.026}$ (2)	$0.153_{\pm 0.021}$ (6)	$0.142_{\pm 0.015}$ (8)	$0.159_{\pm 0.012}$ (4)	**$0.177_{\pm 0.010}$** (1)	$0.160_{\pm 0.025}$ (3)
stroke	$0.089_{\pm 0.027}$ (9)	$0.121_{\pm 0.042}$ (4)	$0.115_{\pm 0.022}$ (6)	$0.117_{\pm 0.024}$ (5)	$0.081_{\pm 0.016}$ (10)	$0.074_{\pm 0.017}$ (11)	$0.097_{\pm 0.026}$ (8)
vertebral	$0.268_{\pm 0.008}$ (2)	$0.227_{\pm 0.008}$ (7)	$0.218_{\pm 0.011}$ (8)	**$0.271_{\pm 0.018}$** (1)	$0.257_{\pm 0.002}$ (3)	$0.216_{\pm 0.005}$ (9)	$0.211_{\pm 0.005}$ (10)
wbc	$0.586_{\pm 0.127}$ (10)	$0.745_{\pm 0.090}$ (8)	$0.861_{\pm 0.096}$ (2)	$0.794_{\pm 0.040}$ (5)	$0.491_{\pm 0.054}$ (11)	$0.789_{\pm 0.078}$ (6)	**$0.908_{\pm 0.031}$** (1)
wine	$0.314_{\pm 0.059}$ (7)	$0.484_{\pm 0.076}$ (5)	$0.190_{\pm 0.011}$ (11)	$0.212_{\pm 0.012}$ (9)	$0.406_{\pm 0.018}$ (6)	$0.555_{\pm 0.034}$ (4)	$0.199_{\pm 0.007}$ (10)
yeast	$0.103_{\pm 0.026}$ (10)	$0.247_{\pm 0.072}$ (3)	$0.183_{\pm 0.046}$ (6)	$0.190_{\pm 0.044}$ (5)	$0.200_{\pm 0.048}$ (4)	**$0.327_{\pm 0.064}$** (1)	$0.156_{\pm 0.029}$ (7)
Mean	0.295	0.359	0.339	0.325	0.328	0.359	0.336
Mean Rank	7.000	5.450	5.850	6.000	6.000	6.050	6.450
#1st	0	0	1	2	0	3	2

(b) TabPFN-Unsupervised and LENS-AD variants

Dataset	TabPFN-Unsup.	LENS-AD (TabPFN)	LENS-AD (COPOD)	LENS-AD (IForest)
automobile	$0.365_{\pm 0.103}$ (4)	**$0.448_{\pm 0.171}$** (1)	$0.350_{\pm 0.052}$ (5)	$0.408_{\pm 0.046}$ (2)
backdoor	$0.006_{\pm 0.001}$ (9)	$0.007_{\pm 0.001}$ (8)	$0.010_{\pm 0.001}$ (6)	**$0.030_{\pm 0.005}$** (1)
campaign	$0.251_{\pm 0.008}$ (10)	$0.247_{\pm 0.020}$ (11)	**$0.393_{\pm 0.007}$** (1)	$0.392_{\pm 0.046}$ (2)
cardiotocography	$0.272_{\pm 0.056}$ (11)	$0.344_{\pm 0.012}$ (9)	$0.443_{\pm 0.036}$ (7)	$0.477_{\pm 0.005}$ (6)
census	$0.107_{\pm 0.035}$ (5)	$0.082_{\pm 0.006}$ (9)	$0.104_{\pm 0.001}$ (6)	$0.099_{\pm 0.023}$ (7)
churn	$0.220_{\pm 0.020}$ (4)	$0.211_{\pm 0.011}$ (6)	$0.186_{\pm 0.002}$ (9)	$0.216_{\pm 0.018}$ (5)
cirrhosis	$0.771_{\pm 0.087}$ (4)	$0.772_{\pm 0.024}$ (3)	$0.764_{\pm 0.034}$ (5)	$0.693_{\pm 0.056}$ (6)
covertype	$0.043_{\pm 0.009}$ (7)	$0.077_{\pm 0.015}$ (6)	$0.120_{\pm 0.010}$ (5)	**$0.352_{\pm 0.083}$** (1)
credit	$0.181_{\pm 0.004}$ (10)	$0.194_{\pm 0.006}$ (7)	$0.217_{\pm 0.001}$ (5)	**$0.258_{\pm 0.023}$** (1)
equip	**$0.921_{\pm 0.011}$** (1)	$0.731_{\pm 0.078}$ (7)	$0.527_{\pm 0.060}$ (9)	$0.626_{\pm 0.018}$ (8)
gallstone	**$0.516_{\pm 0.011}$** (1)	$0.467_{\pm 0.087}$ (3)	$0.380_{\pm 0.075}$ (7)	$0.339_{\pm 0.047}$ (11)
glass	$0.588_{\pm 0.123}$ (8)	$0.656_{\pm 0.036}$ (6)	**$0.715_{\pm 0.066}$** (1)	$0.676_{\pm 0.047}$ (4)
glioma	$0.405_{\pm 0.105}$ (8)	$0.380_{\pm 0.018}$ (11)	**$0.510_{\pm 0.039}$** (1)	$0.465_{\pm 0.080}$ (5)
quasar	**$0.333_{\pm 0.005}$** (1)	$0.260_{\pm 0.063}$ (3)	$0.163_{\pm 0.015}$ (10)	$0.183_{\pm 0.006}$ (8)
seismic	$0.156_{\pm 0.014}$ (5)	$0.142_{\pm 0.006}$ (8)	$0.152_{\pm 0.009}$ (7)	$0.103_{\pm 0.027}$ (9)
stroke	$0.098_{\pm 0.043}$ (7)	$0.180_{\pm 0.047}$ (3)	$0.181_{\pm 0.035}$ (2)	**$0.194_{\pm 0.035}$** (1)
vertebral	$0.218_{\pm 0.004}$ (8)	$0.229_{\pm 0.007}$ (6)	$0.231_{\pm 0.011}$ (5)	$0.255_{\pm 0.015}$ (4)
wbc	$0.628_{\pm 0.113}$ (9)	$0.769_{\pm 0.085}$ (7)	$0.856_{\pm 0.064}$ (3)	$0.829_{\pm 0.078}$ (4)
wine	$0.313_{\pm 0.040}$ (8)	$0.599_{\pm 0.082}$ (3)	$0.609_{\pm 0.058}$ (2)	**$0.640_{\pm 0.154}$** (1)
yeast	$0.107_{\pm 0.011}$ (9)	$0.080_{\pm 0.020}$ (11)	$0.249_{\pm 0.039}$ (2)	$0.155_{\pm 0.066}$ (8)
Mean	0.325	0.344	0.358	0.369
Mean Rank	6.550	6.650	5.100	4.900
#1st	3	1	3	5

Table G10: Dataset-level unsupervised AUCPR for Full. Entries are mean ± standard deviation over three root seeds.

(a) Conventional full-feature scorers

Dataset	PCA	IForest	LOF	OCSVM	ECOD	COPOD	KNN
automobile	$0.420_{\pm 0.120}$ (4)	$0.350_{\pm 0.065}$ (5)	$0.292_{\pm 0.009}$ (9)	$0.258_{\pm 0.028}$ (11)	$0.342_{\pm 0.053}$ (7)	$0.348_{\pm 0.044}$ (6)	$0.283_{\pm 0.021}$ (10)
backdoor	$0.009_{\pm 0.002}$ (8)	$0.016_{\pm 0.002}$ (4)	$0.006_{\pm 0.002}$ (9)	$0.010_{\pm 0.003}$ (7)	$0.015_{\pm 0.002}$ (5)	$0.011_{\pm 0.002}$ (6)	$0.021_{\pm 0.003}$ (3)
campaign	$0.254_{\pm 0.005}$ (11)	$0.321_{\pm 0.011}$ (5)	$0.297_{\pm 0.015}$ (7)	$0.259_{\pm 0.011}$ (9)	$0.308_{\pm 0.006}$ (6)	$0.345_{\pm 0.012}$ (4)	$0.348_{\pm 0.016}$ (3)
cardiotocography	$0.241_{\pm 0.017}$ (10)	$0.461_{\pm 0.015}$ (3)	$0.366_{\pm 0.037}$ (8)	**$0.514_{\pm 0.019}$** (1)	$0.507_{\pm 0.019}$ (2)	$0.401_{\pm 0.020}$ (6)	$0.375_{\pm 0.034}$ (7)
census	$0.054_{\pm 0.001}$ (11)	$0.063_{\pm 0.002}$ (9)	$0.062_{\pm 0.002}$ (10)	**$0.290_{\pm 0.005}$** (1)	$0.070_{\pm 0.002}$ (8)	$0.081_{\pm 0.002}$ (7)	$0.246_{\pm 0.002}$ (2)
churn	$0.241_{\pm 0.003}$ (3)	$0.202_{\pm 0.008}$ (9)	**$0.323_{\pm 0.015}$** (1)	$0.265_{\pm 0.005}$ (2)	$0.218_{\pm 0.004}$ (5)	$0.178_{\pm 0.004}$ (11)	$0.222_{\pm 0.010}$ (4)
cirrhosis	$0.544_{\pm 0.030}$ (9)	$0.777_{\pm 0.062}$ (3)	$0.571_{\pm 0.086}$ (8)	$0.419_{\pm 0.055}$ (11)	$0.648_{\pm 0.084}$ (7)	**$0.789_{\pm 0.075}$** (1)	$0.537_{\pm 0.096}$ (10)
covertype	$0.035_{\pm 0.010}$ (8)	$0.161_{\pm 0.052}$ (5)	$0.032_{\pm 0.003}$ (9)	$0.013_{\pm 0.000}$ (10)	$0.235_{\pm 0.023}$ (2)	$0.190_{\pm 0.029}$ (4)	$0.063_{\pm 0.020}$ (6)
credit	$0.182_{\pm 0.002}$ (8)	**$0.277_{\pm 0.003}$** (1)	$0.206_{\pm 0.005}$ (6)	$0.206_{\pm 0.001}$ (6)	$0.254_{\pm 0.003}$ (3)	$0.247_{\pm 0.001}$ (4)	$0.172_{\pm 0.002}$ (10)
equip	$0.524_{\pm 0.030}$ (10)	**$0.930_{\pm 0.013}$** (1)	$0.472_{\pm 0.018}$ (11)	$0.909_{\pm 0.004}$ (3)	$0.901_{\pm 0.010}$ (5)	$0.886_{\pm 0.005}$ (7)	$0.905_{\pm 0.008}$ (4)
gallstone	$0.478_{\pm 0.009}$ (2)	$0.430_{\pm 0.060}$ (3)	$0.404_{\pm 0.031}$ (5)	$0.365_{\pm 0.049}$ (10)	$0.357_{\pm 0.032}$ (11)	$0.373_{\pm 0.049}$ (9)	$0.390_{\pm 0.037}$ (7)
glass	**$0.735_{\pm 0.092}$** (1)	$0.598_{\pm 0.042}$ (5)	$0.391_{\pm 0.026}$ (11)	$0.592_{\pm 0.076}$ (6)	$0.540_{\pm 0.024}$ (10)	$0.573_{\pm 0.026}$ (8)	$0.581_{\pm 0.028}$ (7)
glioma	$0.348_{\pm 0.029}$ (11)	$0.395_{\pm 0.041}$ (8)	$0.386_{\pm 0.068}$ (9)	$0.488_{\pm 0.041}$ (2)	$0.424_{\pm 0.052}$ (6)	$0.465_{\pm 0.053}$ (4)	$0.476_{\pm 0.055}$ (3)
quasar	$0.345_{\pm 0.023}$ (2)	$0.299_{\pm 0.010}$ (3)	$0.234_{\pm 0.003}$ (5)	$0.199_{\pm 0.002}$ (8)	$0.188_{\pm 0.002}$ (11)	$0.194_{\pm 0.001}$ (9)	$0.229_{\pm 0.004}$ (7)
seismic	$0.069_{\pm 0.002}$ (11)	$0.153_{\pm 0.017}$ (5)	$0.081_{\pm 0.018}$ (10)	$0.135_{\pm 0.028}$ (8)	$0.163_{\pm 0.019}$ (2)	**$0.175_{\pm 0.016}$** (1)	$0.158_{\pm 0.028}$ (4)
stroke	$0.118_{\pm 0.018}$ (6)	$0.129_{\pm 0.024}$ (4)	$0.053_{\pm 0.005}$ (11)	$0.123_{\pm 0.021}$ (5)	$0.087_{\pm 0.023}$ (9)	$0.093_{\pm 0.028}$ (7)	$0.074_{\pm 0.005}$ (10)
vertebral	$0.257_{\pm 0.032}$ (4)	$0.228_{\pm 0.008}$ (7)	**$0.279_{\pm 0.015}$** (1)	$0.263_{\pm 0.005}$ (2)	$0.259_{\pm 0.004}$ (3)	$0.216_{\pm 0.004}$ (10)	$0.217_{\pm 0.005}$ (9)
wbc	$0.493_{\pm 0.075}$ (11)	$0.677_{\pm 0.117}$ (7)	$0.524_{\pm 0.114}$ (8)	$0.794_{\pm 0.022}$ (3)	$0.500_{\pm 0.062}$ (10)	$0.790_{\pm 0.096}$ (4)	**$0.947_{\pm 0.027}$** (1)
wine	$0.302_{\pm 0.050}$ (8)	$0.486_{\pm 0.060}$ (4)	$0.225_{\pm 0.005}$ (10)	$0.198_{\pm 0.018}$ (11)	$0.397_{\pm 0.018}$ (5)	$0.369_{\pm 0.017}$ (6)	$0.229_{\pm 0.017}$ (9)
yeast	$0.089_{\pm 0.013}$ (8)	$0.214_{\pm 0.052}$ (3)	$0.080_{\pm 0.019}$ (10)	$0.203_{\pm 0.054}$ (4)	$0.196_{\pm 0.062}$ (5)	**$0.323_{\pm 0.068}$** (1)	$0.117_{\pm 0.020}$ (7)
Mean	0.287	0.358	0.264	0.325	0.330	0.352	0.330
Mean Rank	7.450	4.750	8.050	6.150	6.100	5.800	6.250
#1st	1	2	2	2	0	3	1

(b) TabPFN-Unsupervised and LENS-AD variants

Dataset	TabPFN-Unsup.	LENS-AD (TabPFN)	LENS-AD (COPOD)	LENS-AD (IForest)
automobile	$0.424_{\pm 0.109}$ (3)	**$0.434_{\pm 0.143}$** (1)	$0.338_{\pm 0.038}$ (8)	$0.433_{\pm 0.083}$ (2)
backdoor	**$0.033_{\pm 0.007}$** (1)	$0.011_{\pm 0.001}$ (6)	$0.010_{\pm 0.001}$ (7)	$0.024_{\pm 0.001}$ (2)
campaign	$0.260_{\pm 0.014}$ (8)	$0.257_{\pm 0.007}$ (10)	**$0.391_{\pm 0.008}$** (1)	$0.389_{\pm 0.007}$ (2)
cardiotocography	$0.230_{\pm 0.018}$ (11)	$0.254_{\pm 0.026}$ (9)	$0.430_{\pm 0.025}$ (5)	$0.441_{\pm 0.025}$ (4)
census	$0.182_{\pm 0.002}$ (4)	$0.167_{\pm 0.003}$ (5)	$0.128_{\pm 0.005}$ (6)	$0.223_{\pm 0.007}$ (3)
churn	$0.213_{\pm 0.006}$ (6)	$0.203_{\pm 0.001}$ (8)	$0.186_{\pm 0.003}$ (10)	$0.209_{\pm 0.006}$ (7)
cirrhosis	$0.768_{\pm 0.058}$ (4)	$0.723_{\pm 0.051}$ (6)	$0.788_{\pm 0.045}$ (2)	$0.737_{\pm 0.052}$ (5)
covertype	$0.032_{\pm 0.006}$ (9)	$0.042_{\pm 0.006}$ (7)	$0.207_{\pm 0.016}$ (3)	**$0.440_{\pm 0.007}$** (1)
credit	$0.180_{\pm 0.000}$ (9)	$0.187_{\pm 0.002}$ (7)	$0.225_{\pm 0.000}$ (5)	$0.263_{\pm 0.004}$ (2)
equip	$0.929_{\pm 0.007}$ (2)	$0.741_{\pm 0.033}$ (8)	$0.644_{\pm 0.010}$ (9)	$0.898_{\pm 0.017}$ (6)
gallstone	**$0.532_{\pm 0.017}$** (1)	$0.381_{\pm 0.042}$ (8)	$0.393_{\pm 0.080}$ (6)	$0.407_{\pm 0.070}$ (4)
glass	$0.541_{\pm 0.040}$ (9)	$0.649_{\pm 0.017}$ (4)	$0.729_{\pm 0.052}$ (2)	$0.676_{\pm 0.015}$ (3)
glioma	$0.396_{\pm 0.060}$ (7)	$0.349_{\pm 0.029}$ (10)	**$0.523_{\pm 0.050}$** (1)	$0.443_{\pm 0.051}$ (5)
quasar	**$0.347_{\pm 0.012}$** (1)	$0.278_{\pm 0.004}$ (4)	$0.192_{\pm 0.003}$ (10)	$0.232_{\pm 0.003}$ (6)
seismic	$0.162_{\pm 0.019}$ (3)	$0.136_{\pm 0.004}$ (7)	$0.147_{\pm 0.014}$ (6)	$0.107_{\pm 0.018}$ (9)
stroke	$0.092_{\pm 0.005}$ (8)	**$0.197_{\pm 0.034}$** (1)	$0.189_{\pm 0.034}$ (3)	$0.191_{\pm 0.024}$ (2)
vertebral	$0.222_{\pm 0.010}$ (8)	$0.228_{\pm 0.009}$ (7)	$0.230_{\pm 0.010}$ (6)	$0.255_{\pm 0.011}$ (5)
wbc	$0.507_{\pm 0.059}$ (9)	$0.731_{\pm 0.108}$ (6)	$0.861_{\pm 0.069}$ (2)	$0.771_{\pm 0.106}$ (5)
wine	$0.321_{\pm 0.037}$ (7)	$0.613_{\pm 0.104}$ (3)	$0.632_{\pm 0.071}$ (2)	**$0.789_{\pm 0.051}$** (1)
yeast	$0.085_{\pm 0.008}$ (9)	$0.071_{\pm 0.010}$ (11)	$0.241_{\pm 0.060}$ (2)	$0.153_{\pm 0.053}$ (6)
Mean	0.323	0.333	0.374	0.404
Mean Rank	6.050	6.500	4.900	4.000
#1st	3	2	2	2

H PAIRED DATASET-LEVEL COMPARISONS

Table H1 reports the complete paired comparisons available for NLA 5, NLA 10, and training sizes 10 and 20. For each method and dataset, AUCPR is first averaged over the same three random seeds used in the main evaluation. The resulting 20 dataset-level means form each pairwise comparison. Wins, ties, and losses compare LENS-AD against the named baseline, with differences of at most 10^{-12} counted as ties. The one-sided Wilcoxon signed-rank test uses the alternative that LENS-AD has higher AUCPR; reported p-values are unadjusted across comparisons.

Table H1: Dataset-level wins/ties/losses (W/T/L) and one-sided Wilcoxon p-values for LENS-AD against selected baselines. Each cell gives W/T/L followed by p in parentheses; all comparisons use the same 20 datasets.

Comparison	NLA 5	NLA 10	$n = 10$	$n = 20$
LENS-AD (TabICL) vs. TabICL	$14/0/6$ ($p = .011$)	$15/0/5$ ($p = .038$)	$12/0/8$ ($p = .215$)	$13/1/6$ ($p = .035$)
LENS-AD (TabICL) vs. CAAFE	$17/0/3$ ($p < .001$)	$18/0/2$ ($p < .001$)	$10/0/10$ ($p = .337$)	$11/1/8$ ($p = .300$)
LENS-AD (TabICL) vs. FeatLLM	$15/0/5$ ($p < .001$)	$19/0/1$ ($p < .001$)	$14/0/6$ ($p = .053$)	$14/0/6$ ($p = .016$)
LENS-AD (TabICL) vs. XGBOD	$20/0/0$ ($p < .001$)	$20/0/0$ ($p < .001$)	$19/0/1$ ($p < .001$)	$16/0/4$ ($p < .001$)
LENS-AD (TabPFN) vs. TabPFN	$15/0/5$ ($p = .012$)	$16/0/4$ ($p = .003$)	$10/1/9$ ($p = .484$)	$12/1/7$ ($p = .420$)
LENS-AD (TabPFN) vs. CAAFE	$16/0/4$ ($p = .013$)	$17/0/3$ ($p < .001$)	$12/0/8$ ($p = .101$)	$13/1/6$ ($p = .049$)
LENS-AD (TabPFN) vs. FeatLLM	$17/0/3$ ($p < .001$)	$18/0/2$ ($p < .001$)	$11/0/9$ ($p = .049$)	$15/0/5$ ($p = .009$)
LENS-AD (TabPFN) vs. XGBOD	$17/0/3$ ($p < .001$)	$19/0/1$ ($p < .001$)	$19/0/1$ ($p < .001$)	$18/0/2$ ($p < .001$)

I LLM COST ANALYSIS

This section quantifies the LLM usage of LENS-AD and the two LLM-AFE baselines, CAAFE (Hollmann et al., 2023b) and FeatLLM (Han et al., 2024), on all 20 benchmark datasets. All three methods are driven by the same `deepseek-v4-flash` backbone. Token counts are net prompt plus completion tokens without double-counting cached prefixes, and cover the logged successful responses of each pipeline. Table I1 reports the complete per-dataset accounting.

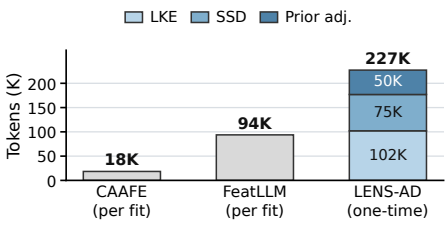


Figure I1: LLM token consumption averaged over the 20 benchmark datasets. CAAFE and FeatLLM regenerate features at every fit (single-fit usage shown), whereas LENS-AD performs a one-time dataset-level construction, decomposed here into the LKE, SSD, and semantic-prior adjudication (Prior adj.) stages.

The two sides differ in how the LLM cost is incurred. LENS-AD concentrates all LLM usage into a one-time, dataset-level construction: LKE, SSD, and semantic-prior adjudication operate exclusively on dataset-level metadata (task descriptions, column names, types, and schema vocabularies; Appendices C and D) and never observe training rows or labels, so the resulting expanded features, semantic subspaces, and prior descriptors are constructed once per dataset. In contrast, CAAFE and FeatLLM condition feature generation on the visible training rows and labels — sample values for CAAFE and serialized labeled few-shot examples for FeatLLM — so their LLM usage is paid per fit. Their entries in Table I1 measure a single fit under NLA $= 5$ with root seed 42. The LENS-AD time is the critical path of its staged pipeline, in which independent calls within a stage run concurrently, and excludes downstream predictor fitting, whereas the baseline times are end-to-end fits that additionally cover feature evaluation and downstream model fitting.

A single construction of LENS-AD costs 117K–393K tokens (mean 227K) and completes within 5–15 minutes of LLM time. The LKE stage accounts for the largest share of tokens (45%), followed by SSD (33%) and semantic-prior adjudication (22%). Per individual fit, CAAFE uses a mean of 18K tokens and completes in about 70 seconds, while FeatLLM uses a mean of 94K tokens and takes about 5.4 minutes. Figure I1 summarizes these averages.

Table I1: Dataset-level LLM usage of LENS-AD and the LLM-AFE baselines on all 20 benchmark datasets. LENS-AD performs a one-time, dataset-level construction; we report the net tokens of its three stages (LKE, SSD, and semantic-prior adjudication), their total, and the critical-path LLM time, which excludes downstream predictor fitting. CAAFE and FeatLLM regenerate features at every fit; their columns report the token usage and end-to-end fit time (LLM calls, feature evaluation, and downstream fitting) of a single fit measured under NLA $= 5$ with root seed 42. All three methods use the same `deepseek-v4-flash` backbone.

	LENS-AD (one-time per dataset)					**CAAFE (per fit)**		**FeatLLM (per fit)**	
Dataset	**LKE**	**SSD**	**Prior**	**Total**	**Time (s)**	**Tokens**	**Time (s)**	**Tokens**	**Time (s)**
automobile	108,563	102,045	65,069	275,677	561	13,009	50	94,437	341
backdoor	123,955	98,209	65,734	287,898	622	16,342	60	79,992	281
campaign	114,959	62,891	43,813	221,663	555	22,008	86	109,050	382
cardiotocography	86,843	76,693	47,682	211,218	568	13,727	59	70,572	259
census	150,861	86,824	63,752	301,437	657	32,069	105	106,565	384
churn	105,707	63,498	43,744	212,949	616	20,854	72	104,975	350
cirrhosis	88,107	81,405	59,038	228,550	554	20,343	78	96,495	330
covertype	106,165	48,967	37,441	192,573	575	11,895	49	89,567	310
credit	107,409	91,287	71,226	269,922	558	16,614	54	104,357	346
equip	82,534	69,467	40,512	192,513	575	18,452	68	52,568	183
gallstone	112,768	165,082	115,496	393,346	900	18,232	72	109,864	377
glass	79,100	48,309	30,578	157,987	430	31,229	129	109,338	374
glioma	109,896	70,610	48,995	229,501	480	27,950	100	97,348	343
quasar	82,475	42,049	27,287	151,811	343	20,080	77	92,034	320
seismic	105,511	101,034	47,763	254,308	612	19,717	72	93,383	316
stroke	142,956	72,344	51,874	267,174	713	10,701	44	89,029	322
vertebral	49,813	37,934	29,080	116,827	305	15,221	66	80,493	277
wbc	128,027	63,424	41,337	232,788	450	16,072	53	102,618	358
wine	69,071	55,577	40,216	164,864	404	12,743	52	96,370	321
yeast	88,213	58,502	34,010	180,725	452	11,600	43	96,522	336
Mean	102,147	74,808	50,232	227,187	546.5	18,443	69.5	93,779	325.5