REVISITING BATCH NORMALIZATION FOR SEMI-SUPERVISED ANOMALY DETECTION ON TABULAR DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Batch Normalization (BN) is widely used to accelerate convergence and provide implicit regularization, yet its behavior remains largely unexplored in semi-supervised anomaly detection (SAD) for tabular data. We identify a previously overlooked failure mode, termed distribution collapse, in which the batch-to-batch variability of normalized normal representations progressively diminishes under anomaly-oversampled training. Because SAD repeatedly oversamples a limited set of labeled anomalies, their influence on shared BN statistics can make the standardized relation between normal and abnormal samples increasingly consistent across mini-batches, thereby reducing representation-level stochasticity and weakening a source of BN's implicit regularization. To address this issue while retaining the optimization benefits of BN, we propose Normal-Only Batch Normalization (NoBa), which computes normalization statistics exclusively from normal samples and removes the influence of oversampled anomalies from statistic estimation. Extensive experiments across eleven tabular datasets and five SAD models show that NoBa consistently improves upon conventional BN and alternative normalization methods while retaining BN's fast convergence.

1 INTRODUCTION

Anomaly detection is a critical task in many real-world applications, including fraud detection (Hilal et al., 2022; Thimonier et al., 2024), medical diagnosis (Tschuchnig & Gadermayr, 2021; Bao et al., 2024a), and industrial monitoring (Panza et al., 2023; Elía & Pagola, 2025). The goal is to identify rare and unusual patterns that deviate significantly from the majority of data. While anomaly detection is typically treated as an unsupervised learning problem due to the difficulty of obtaining labeled anomalies, in practice, one may have access to a small number of known anomalies verified by domain experts in addition to a large set of unlabeled samples. Semi-supervised Anomaly Detection (SAD) (Yoon et al., 2023; Gan et al., 2024; 2025) aims to leverage such limited labeled anomalies to achieve significant performance improvements over purely unsupervised approaches.

Several works have proposed to incorporate weak or partial supervision to leverage this insight. For instance, Deviation Networks (DevNet) (Pang et al., 2019) and Deep Semi-supervised Anomaly Detection (DSAD) (Ruff et al., 2020) formulate objective functions that encourage normal instances to lie close to a reference distribution, while forcing anomalies away. Others explore weakly supervised (Pang et al., 2023) or contamination-resilient learning frameworks, such as RoSAS (Xu et al., 2023) and NNG-Mix (Dong et al., 2024), which enhance robustness under noisy or pseudo-labeled settings. A common practice in SAD is to oversample or reweight the limited labeled anomalies to ensure they contribute sufficiently to learning (Pang et al., 2019; Xu et al., 2023).

Neural networks for tabular data lack the strong structural inductive biases inherent to architectures designed for images or sequences and therefore rely more heavily on effective regularization (Shavitt & Segal, 2018; Taghanaki et al., 2022). Nevertheless, existing tabular SAD models predominantly adopt fully connected networks with explicit regularization while typically omitting Batch Normalization (BN), despite its established benefits. BN facilitates optimization and accelerates convergence (Ioffe & Szegedy, 2015; Santurkar et al., 2018), while the stochastic variation induced by mini-batch statistics acts as a form of implicit regularization that can improve gener-

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.

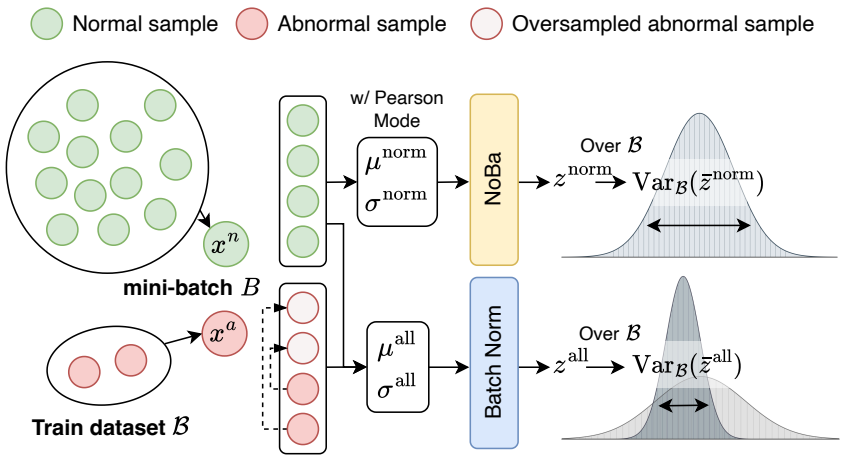


Figure 1: Overview of Normal-Only Batch Normalization (NoBa). Conventional BN estimates shared statistics from both normal and oversampled abnormal samples, allowing repeatedly reused anomalies to influence the normalization reference. NoBa instead estimates normalization statistics exclusively from normal samples, removing this direct influence while retaining batch-dependent variation in the normalized representations.

alization (Luo et al., 2019b). In tabular SAD, however, BN statistics are estimated jointly from normal samples that vary across mini-batches and a limited set of labeled anomalies that are repeatedly reused through oversampling. As training proceeds, repeated exposure to these anomalies can make the standardized relation between normal and abnormal samples increasingly consistent across mini-batches, progressively reducing the batch-to-batch variability of normalized normal representations. We refer to this attenuation of representation-level stochasticity as *distribution collapse*. Since such batch-dependent perturbations contribute to BN's implicit regularization, distribution collapse weakens this source of regularization and can limit the generalization benefit of BN. This effect is particularly consequential for tabular SAD, where effective regularization is especially important.

To address this problem while retaining the benefits of BN, we propose *Normal-Only Batch Normalization (NoBa)*, which estimates normalization statistics exclusively from normal samples while applying the resulting transformation to all samples (Figure 1). By excluding oversampled anomalies from statistic estimation, NoBa removes their direct influence on the shared normalization statistics and preserves batch-dependent variation in the normalized representations. We theoretically characterize distribution collapse through the batch-to-batch variability of normalized normal representations and show that NoBa preserves representation-level stochasticity by relying on normal-derived statistics. We further employ the Pearson mode (Kenney & Keeping, 1951), a lightweight mean–median approximation of the distributional mode, as a skewness-aware centering statistic for tabular activations. Extensive experiments across eleven tabular datasets and five representative SAD models demonstrate that NoBa consistently improves upon conventional BN and competing normalization strategies while retaining the rapid convergence afforded by BN.

Our contributions are summarized as follows:

- We identify *distribution collapse*, a previously overlooked failure mode of BN in tabular SAD, in which the batch-to-batch variability of normalized normal representations progressively diminishes under anomaly-oversampled training, weakening a source of BN's implicit regularization.
- We propose NoBa, a simple normal-only normalization scheme that estimates batch statistics exclusively from normal samples, removing the direct influence of oversampled anomalies on statistic estimation while retaining the optimization benefits of BN.
- We theoretically characterize distribution collapse through representation-level batch variability and analyze how NoBa preserves stochasticity in the normalized representations. Extensive experiments across eleven tabular datasets and five representative SAD models further validate its effectiveness.

2 Related Work

2.1 Semi-supervised Anomaly Detection

Traditional anomaly detection methods typically assume that training datasets consist exclusively of normal instances and identify anomalies as deviations from them (Batzner et al., 2024; Liu et al., 2023). However, recent studies have demonstrated improved performance by incorporating a few labeled outliers during training, a paradigm known as SAD (Yoon et al., 2023; Ruff et al., 2020; Pang et al., 2019). DevNet (Pang et al., 2019) is a representative SAD model that treats anomaly detection as a binary classification task, training on normal and a few labeled abnormal samples using a z-score-based deviation loss and oversampling to address data imbalance. DevNet suffers from overfitting to abnormal samples due to the reliance on oversampling to compensate for their scarcity. To mitigate overfitting, Pairwise Relation prediction Network (PReNet) (Pang et al., 2023) predicts the relational properties between pairs of instances, generating anomaly scores based on learned pairwise features. DSAD (Ruff et al., 2020) minimizes the distance between normal samples and a predefined vector, while maximizing the distance for abnormal samples. Feature Encoding with AutoEncoder (FeaWAD) (Zhou et al., 2022) combines hidden representations, reconstruction residuals, and reconstruction errors to enhance anomaly detection. More recently, SSRDT (Xiao et al., 2025) learns a nonlinear transformation to separate normal and abnormal samples via intermediate sample interpolation. CAD (Gao et al., 2025) adopts contrastive distance learning with denoising and a learnable bilinear distance metric.

2.2 Batch Normalization

BN accelerates convergence and stabilizes training (Santurkar et al., 2018), but tends to exacerbate overfitting on tabular datasets. Consequently, prior SAD approaches (Zhou et al., 2022; Pang et al., 2023; 2019; Xu et al., 2023) typically employ models composed solely of fully connected layers and activation functions. AnoOnly (Zhou et al., 2025) is an exception that employs a BN layer before the final FC layer, not for optimization, but to exploit the statistical properties of abundant normal data. Layer Normalization (Ba et al., 2016) can serve as an alternative, but tends to be less effective for tabular data due to high inter-feature correlation.

Several BN variants, such as GhostBN (Hoffer et al., 2017), Batch Renormalization (Ioffe, 2017), and Normalizing Batch Normalization (NBN) (Bao et al., 2024b), have been proposed to address regularization, train-inference discrepancy, and class imbalance, respectively. However, these approaches are designed primarily for supervised image classification and do not address the distribution collapse problem caused by oversampled anomalies in semi-supervised anomaly detection on tabular data. NoBa directly tackles this problem by incorporating normal-only statistics into normalization, mitigating anomaly-induced distortion and ensuring stable representation learning.

3 Methodology

In this section, we present Normal-Only Batch Normalization (NoBa) for tabular SAD. We first review conventional BN and its role in mini-batch training. We then introduce distribution collapse and characterize how anomaly-oversampled training can reduce the batch-to-batch variability of normalized normal representations, weakening a source of BN's implicit regularization. Based on this analysis, we formulate NoBa, which estimates normalization statistics exclusively from normal samples to remove the direct influence of oversampled anomalies while preserving representation-level stochasticity and the optimization benefits of BN. Finally, we consider the Pearson mode as a skewness-aware centering statistic for tabular activations.

3.1 Preliminary

BN is a widely used technique to accelerate training and improve generalization in deep neural networks. It normalizes the input to each layer across the batch, reducing internal covariate shift. Given a layer with d-dimensional input $\mathbf{x} = \{x^1, x^2, \ldots, x^d\}$, BN normalizes each dimension using the mean and variance computed across a mini-batch $B = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+m}\}$, where n and m represent the numbers of normal and abnormal samples in a batch, respectively. Focusing on a

particular activation x^k (omitting k for simplicity), BN performs the transformation as follows:

$$BN(x_i) = \gamma \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta, \quad \mu = \frac{1}{n+m} \sum_{i=1}^{n+m} x_i, \quad \sigma^2 = \frac{1}{n+m} \sum_{i=1}^{n+m} (x_i - \mu)^2, \tag{1}$$

where μ and σ^2 denote the batch mean and variance. Here, ε is a small constant for numerical stability, and γ and β are learnable scale and shift parameters.

3.2 Distribution Collapse in Tabular SAD

Existing tabular SAD methods predominantly rely on fully connected networks with explicit regularization and typically omit BN. This is notable because BN provides two important benefits: it facilitates optimization and accelerates convergence (Ioffe, 2017; Santurkar et al., 2018), while the stochastic perturbations induced by mini-batch-dependent normalization can improve generalization through implicit regularization (Luo et al., 2019b). This regularization effect is particularly relevant to neural networks for tabular data, which have relatively weak structural inductive biases and therefore rely heavily on effective regularization (Grinsztajn et al., 2022; Shavitt & Segal, 2018).

We identify a previously overlooked failure mode of BN in tabular SAD. Because labeled anomalies are scarce, SAD commonly oversamples them to provide sufficient supervision, causing a limited set of abnormal samples to repeatedly constitute a substantial fraction of each mini-batch. Conventional BN therefore estimates its statistics jointly from varying normal samples and repeatedly reused abnormal samples. As the same abnormal samples recur across mini-batches, the abnormal-dependent component of the normalization reference becomes increasingly similar across training steps, reducing the batch-to-batch variation expressed in the normalized normal representation. Consequently, normal activations become progressively less sensitive to changes in mini-batch composition. We refer to this attenuation of batch-dependent representation variability as *distribution collapse*. Since such batch-dependent perturbations contribute to BN's implicit regularization, distribution collapse reduces a source of stochastic regularization during training.

Consider a mini-batch $\mathcal{B}$ containing n normal and m abnormal samples, with abnormal ratio $\pi = m/(n+m)$. Let μ^{norm} and μ^{abnorm} denote the means of the normal and abnormal samples, respectively. Conventional BN normalizes with the shared mean $\mu^{\text{all}} = (1-\pi)\mu^{\text{norm}} + \pi\mu^{\text{abnorm}}$ and the corresponding standard deviation σ^{all}, producing normalized activations $z_i^{\text{all}} = (x_i - \mu^{\text{all}})/\sigma^{\text{all}}$. We denote the mean normalized activation over the normal subset by $\bar{z}_N^{\text{all}} = \frac{1}{n}\sum_{i=1}^{n} z_i^{\text{all}}$.

Proposition 1 (Distribution collapse of normalized normal representations) *Let the standardized normal–abnormal gap be*

$$q_{\mathcal{B}} = \frac{\mu^{\text{abnorm}} - \mu^{\text{norm}}}{\sigma^{\text{all}}}. \tag{2}$$

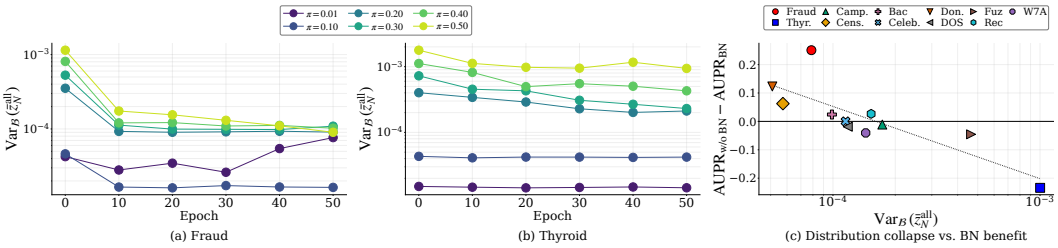
For a fixed abnormal ratio π, the mean normalized activation of the normal subset satisfies

$$\bar{z}_N^{\text{all}} = -\pi\, q_{\mathcal{B}}, \tag{3}$$

so that its batch-to-batch variance is

$$\mathrm{Var}_{\mathcal{B}}\big(\bar{z}_N^{\text{all}}\big) = \pi^2\, \mathrm{Var}_{\mathcal{B}}(q_{\mathcal{B}})\,. \tag{4}$$

Proposition 1 shows that the batch-to-batch variability of the normalized normal representation is determined by the variability of the standardized normal–abnormal gap $q_{\mathcal{B}}$ (See Appendix B.1 for the proof). Two characteristics of tabular SAD training contribute to the attenuation of this variability. First, anomaly oversampling increases the abnormal ratio π, giving abnormal samples a substantial influence on the shared statistics μ^{all} and σ^{all}. Second, because only a limited set of labeled anomalies is available, the same abnormal samples are repeatedly reused across mini-batches. As training proceeds, this repeated exposure makes the abnormal-dependent component of the normalization increasingly consistent across batches, reducing the variability of $q_{\mathcal{B}}$. Through Eq. 4, this progressively reduces $\mathrm{Var}_{\mathcal{B}}(\bar{z}_N^{\text{all}})$. We refer to this attenuation of representation-level batch variability under anomaly-oversampled training as *distribution collapse*. Since batch-dependent perturbations of normalized activations contribute to BN's implicit regularization (Luo et al., 2019b), distribution collapse weakens this source of regularization.



(a) Fraud (b) Thyroid (c) Distribution collapse vs. BN benefit

Figure 2: Empirical analysis of distribution collapse under anomaly-oversampled SAD training. (a–b) Batch-to-batch variability of the normalized normal representation, $\mathrm{Var}_{\mathcal{B}}(\bar{z}_N^{\mathrm{all}})$, under different abnormal batch ratios π on fraud and thyroid. Decreasing variability over training indicates distribution collapse. (c) Across eleven datasets, lower representation variability is associated with a larger performance advantage of removing BN in terms of AUPR.

Proposition 1 shows that the batch-to-batch variability of the normalized normal representation is determined by the variability of the standardized normal–abnormal gap $q_{\mathcal{B}}$ (see Appendix B.1 for the proof). Two characteristics of tabular SAD training make this variability susceptible to attenuation. First, anomaly oversampling increases the abnormal ratio π, giving the limited labeled anomalies greater influence on the shared statistics μ^{all} and σ^{all}. Second, the same anomalies are repeatedly reused across mini-batches. As training proceeds, this repeated exposure can make their standardized relation to the normal samples increasingly consistent across batches, reducing the variability of $q_{\mathcal{B}}$ and, through Eq. 4, progressively attenuating $\mathrm{Var}_{\mathcal{B}}(\bar{z}_N^{\mathrm{all}})$. We refer to this attenuation of representation-level batch variability as *distribution collapse*. Since batch-dependent perturbations of normalized activations contribute to BN's implicit regularization (Luo et al., 2019b), distribution collapse reduces these stochastic perturbations and thereby weakens this source of regularization.

3.3 Normal-Only Batch Normalization

To address distribution collapse while retaining the optimization benefits of BN, we propose *Normal-Only Batch Normalization (NoBa)*, which estimates normalization statistics exclusively from normal samples. In contrast to conventional BN, whose statistics depend jointly on normal and oversampled abnormal samples, NoBa excludes the abnormal group from statistic estimation. For a mini-batch containing n normal samples, NoBa is defined as

$$\mathrm{NoBa}(x_j) = \gamma \frac{x_j - \mu^{\mathrm{norm}}}{\sqrt{(\sigma^{\mathrm{norm}})^2 + \varepsilon}} + \beta, \quad \mu^{\mathrm{norm}} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad (\sigma^{\mathrm{norm}})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu^{\mathrm{norm}})^2. \tag{5}$$

The resulting normal-derived statistics are applied to both normal and abnormal samples. Because oversampled anomalies do not contribute to their estimation, NoBa removes their direct influence on the shared normalization statistics and decouples the normalization reference from repeated abnormal samples.

We next examine how this normal-only reference affects variability in the resulting representations. Let $(\mu^{\mathrm{all}}, \sigma^{\mathrm{all}})$ denote the statistics computed over the entire mini-batch and $(\mu^{\mathrm{norm}}, \sigma^{\mathrm{norm}})$ those computed only from normal samples. The corresponding normalized activations are $z_i^{\mathrm{all}} = (x_i - \mu^{\mathrm{all}})/\sigma^{\mathrm{all}}$ and $z_i^{\mathrm{norm}} = (x_i - \mu^{\mathrm{norm}})/\sigma^{\mathrm{norm}}$, with batch averages $\bar{z}^{\mathrm{all}} = \frac{1}{n+m}\sum_{i=1}^{n+m} z_i^{\mathrm{all}}$ and $\bar{z}^{\mathrm{norm}} = \frac{1}{n+m}\sum_{i=1}^{n+m} z_i^{\mathrm{norm}}$.

Proposition 2 (Aggregate representation variability under NoBa) *If the aggregate abnormal activation normalized by the normal-derived statistics varies across mini-batches, NoBa preserves nonzero batch-to-batch variability in the aggregate normalized representation:*

$$\mathrm{Var}_{\mathcal{B}}(\bar{z}^{\mathrm{norm}}) > \mathrm{Var}_{\mathcal{B}}\big(\bar{z}^{\mathrm{all}}\big) = 0. \tag{6}$$

Under conventional BN, centering by the shared batch mean ensures that $\sum_{i=1}^{n+m} z_i^{\mathrm{all}} = 0$ for every mini-batch and therefore $\mathrm{Var}_{\mathcal{B}}(\bar{z}^{\mathrm{all}}) = 0$. Under NoBa, only the normalized normal samples sum to

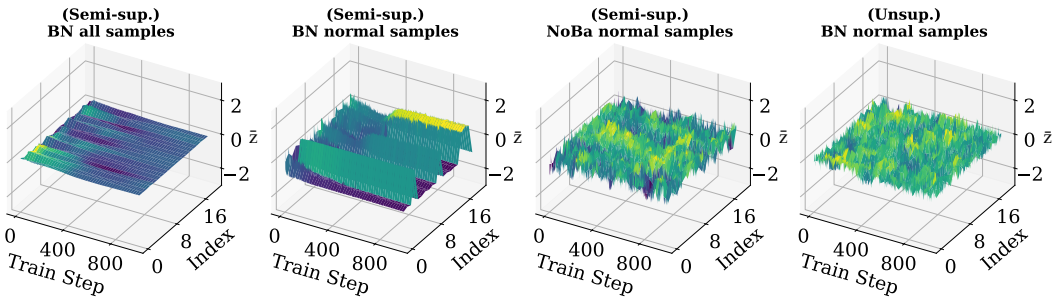


Figure 3: Training dynamics of the aggregate normalized representation under different normalization settings. NoBa preserves substantial batch-to-batch variation, with dynamics similar to the unsupervised normal-only BN reference, whereas conventional semi-supervised BN remains nearly invariant.

zero. Let $\mathcal{A} = \sum_{i=1}^{m} z_{n+i}^{\text{norm}}$ denote the aggregate abnormal activation. Then,

$$\bar{z}^{\text{norm}} = \frac{\mathcal{A}}{n+m}, \qquad \text{Var}_{\mathcal{B}}(\bar{z}^{\text{norm}}) = \frac{1}{(n+m)^2}\,\text{Var}_{\mathcal{B}}(\mathcal{A}) > 0, \tag{7}$$

whenever $\text{Var}_{\mathcal{B}}(\mathcal{A}) > 0$. Thus, normal-only normalization allows batch-dependent variation to remain reflected in the aggregate representation rather than being canceled by full-batch centering. Since such mini-batch-dependent perturbations contribute to the implicit regularization of BN (Luo et al., 2019b), this provides a representation-level explanation for how NoBa retains a source of stochastic regularization while removing the influence of oversampled anomalies from statistic estimation. See Appendix B.2 for the proof. Figure 3 empirically illustrates this distinction. Conventional BN yields an approximately invariant aggregate normalized representation, whereas NoBa retains substantial batch-to-batch variation with dynamics similar to the unsupervised normal-only reference.

3.4 Mode-based Normalization

The key contribution of NoBa lies in computing normalization statistics exclusively from normal data. Building on this, we further address the skewness commonly observed in deep feature distributions. This is particularly relevant for tabular data, where heterogeneous feature types can lead to activations that deviate from symmetric, Gaussian-like distributions (Grinsztajn et al., 2022; Shavitt & Segal, 2018). We therefore adopt the Pearson mode (Kenney & Keeping, 1951) as a practical skewness-aware estimate of central tendency, computed from normal samples only:

$$\mu^{\text{mode}} = 3\text{median}_{\tilde{B}}(x) - 2\mathbb{E}_{\tilde{B}}[x], \tag{8}$$

where $\tilde{B}$ denotes the normal samples within a mini-batch. Eq. 8 follows Pearson's empirical relation for moderately skewed unimodal distributions, providing an inexpensive approximation to the mode using only the mean and median, without explicit density estimation.

Under skew, mean-based centering shifts toward the longer tail and away from the region where most samples lie. Pearson-mode centering accounts for this asymmetry by combining the mean and median while retaining the computational simplicity of standard batch statistics. Since nonlinearities such as ReLU are sensitive to the pre-activation distribution, mis-centered inputs can leave neurons predominantly active or inactive, limiting representational capacity. The Pearson mode thus provides a practical, skewness-aware alternative to mean-based centering for normal-feature distributions. See Appendix D for further analysis.

4 Experiments

This section presents a performance comparison between NoBa and several baseline approaches across eleven tabular datasets, following the experimental protocol of the widely recognized SAD method PReNet (Pang et al., 2023). The benchmark spans diverse application domains, including

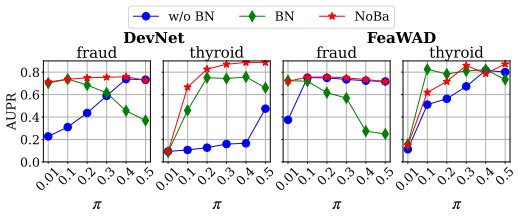


Figure 4: AUPR on fraud and thyroid datasets (DevNet, FeaWAD), where increasing the abnormal batch ratio π improves NoBa and w/o BN but may degrade BN.

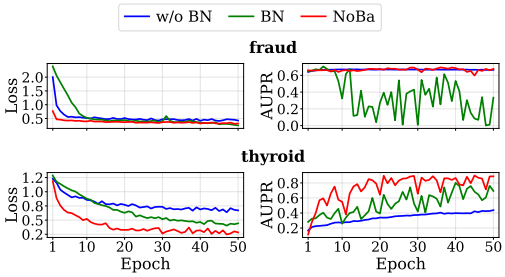


Figure 5: Training dynamics (loss and AUPR per epoch) of w/o BN, BN, and NoBa.

fundraising and socio-economic records, financial fraud detection, attribute-based image metadata, network intrusion categories from UNSW-NB15, webpage classification, marketing response prediction, and medical diagnosis. Across datasets, anomalies are defined as rare or abnormal cases (e.g., exceptionally exciting projects, high-income individuals, fraudulent transactions, the "bald" attribute, specific attack types, minority classes, rare positive responses, and hypothyroidism). See Appendix E for details. We evaluate our method on five representative SAD models spanning diverse learning paradigms: DevNet (Pang et al., 2019), DSAD (Ruff et al., 2020), PReNet (Pang et al., 2023), FeaWAD (Zhou et al., 2022), and CAD (Gao et al., 2025). This evaluation demonstrates the effectiveness of our method across different SAD formulations.

For comparative evaluation, we include a baseline model consisting of FC layers without BN, alongside four representative normalization techniques, and one BN-utilizing variant. To assess the effectiveness and generalization capability of the proposed NoBa, we evaluate the following normalization techniques: conventional BN (Ioffe & Szegedy, 2015), Layer Normalization (LN) (Ba et al., 2016), Switchable Normalization (SN) (Luo et al., 2019a), and Normalizing Batch Normalization (NBN) (Bao et al., 2024b). LN performs normalization across the feature dimension, thereby reducing dependence on batch statistics. SN adaptively learns to combine BN and LN via trainable weighting coefficients, enabling dynamic adjustment based on data characteristics. NBN regularizes the BN parameters γ and β by decomposing them into scale and direction components, improving robustness under class imbalance. We also include AnoOnly, which applies BN before the final linear layer with only an anomaly loss. Here, BN primarily captures normal statistics, while optimization is driven solely by abnormal samples. AnoOnly is not evaluated with CAD because CAD does not employ an anomaly-only loss.

4.1 Experimental Setup

Our experiments were conducted following the experimental settings of DevNet and PReNet. Each dataset was randomly partitioned into training and test sets with an 80:20 ratio. In all experiments, the number of abnormal instances was fixed at 60, except for the analysis of varying anomaly counts. To address the rarity of anomalies, we applied oversampling to maintain a 1:1 ratio between normal and abnormal instances in each mini-batch with a batch size of 512, following the standard SAD setup. Given the rarity of anomalies, we adopted the Area Under the Precision-Recall Curve (AUPR) as our primary evaluation metric, as it offers a more informative and reliable assessment than the Area Under the Receiver Operating Characteristic in highly imbalanced settings. Reported results are averaged across 10 runs with distinct random seeds to reduce performance variability caused by randomness. All experiments were carried out on a workstation equipped with an Intel Core i7-7820X CPU @ 3.60 GHz and an NVIDIA GeForce RTX 2080 Ti GPU (11 GB). See Appendix F for implementation details.

4.2 Interaction between Anomaly Oversampling and Batch Normalization

Figure 4 reports the effect of anomaly oversampling under different normalization schemes. On thyroid, increasing the number of anomalies per mini-batch improves performance both with and without BN. On fraud, however, stronger oversampling improves the w/o BN baseline while substantially degrading conventional BN. These contrasting responses show that the effect of anomaly

Table 1: Comparison of normalization methods under different label contamination ratios. Each entry reports the average AUPR, with the average rank shown in the row below.

Type	Contamination ratio 0%						Contamination ratio 2%					
	DevNet	DSAD	PReNet	FeaWAD	CAD	Avg.	DevNet	DSAD	PReNet	FeaWAD	CAD	Avg.
w/o BN	0.591 (2.450)	0.615 (3.820)	0.571 (3.180)	0.603 (3.360)	0.613 (4.364)	0.599 (3.600)	0.566 (2.360)	**0.578** (**2.730**)	0.556 (**2.090**)	0.504 (4.360)	**0.590** (3.273)	0.559 (2.000)
BN	0.571 (3.360)	0.649 (2.910)	0.580 (3.180)	0.548 (4.820)	0.624 (3.818)	0.594 (3.600)	0.518 (4.090)	0.532 (3.910)	0.557 (3.270)	0.477 (4.730)	0.532 (5.091)	0.523 (4.800)
LN	0.466 (5.910)	0.603 (5.000)	0.464 (6.360)	0.551 (4.270)	0.633 (3.818)	0.543 (5.400)	0.470 (5.270)	0.520 (4.820)	0.427 (6.270)	0.488 (4.270)	0.565 (2.818)	0.494 (5.200)
SN	0.493 (4.730)	0.626 (4.360)	0.525 (4.550)	0.567 (4.180)	0.630 (3.455)	0.568 (4.400)	0.464 (4.820)	0.502 (5.270)	0.439 (5.360)	0.504 (4.820)	0.552 (4.091)	0.492 (5.200)
NBN	0.561 (3.730)	0.633 (4.000)	0.562 (3.270)	0.571 (4.270)	0.614 (4.091)	0.588 (4.000)	0.538 (3.450)	0.532 (4.180)	0.544 (3.180)	0.504 (3.640)	0.543 (4.273)	0.532 (3.600)
AnoOnly	0.410 (6.000)	0.580 (5.550)	0.465 (5.450)	0.578 (5.000)	–	0.508 (5.750)	0.406 (5.820)	0.567 (4.000)	0.427 (5.450)	0.495 (5.000)	–	0.473 (5.250)
NoBa (ours)	**0.644** (**1.820**)	**0.670** (**2.360**)	**0.624** (**2.000**)	**0.633** (**2.090**)	**0.653** (**1.455**)	**0.645** (**1.000**)	**0.589** (**2.180**)	0.559 (3.090)	**0.598** (2.360)	**0.559** (**1.180**)	0.589 (**1.455**)	**0.579** (**1.600**)

oversampling depends critically on the normalization scheme. In comparison, NoBa remains robust across oversampling levels and consistently benefits from increased anomaly supervision. This behavior is consistent with its exclusion of abnormal samples from normalization-statistic estimation, which prevents repeatedly oversampled anomalies from directly altering the normalization reference and thereby mitigates distribution collapse.

4.3 Preserving BN's Optimization Benefits

Although conventional BN is susceptible to distribution collapse under anomaly-oversampled training, it remains effective at accelerating optimization. Figure 5 compares the training dynamics of w/o BN, conventional BN, and NoBa. Conventional BN converges faster than w/o BN but exhibits pronounced AUPR fluctuations despite smooth loss convergence. NoBa retains BN's rapid convergence while achieving smoother loss reduction and more stable AUPR throughout training. These results show that NoBa preserves the optimization benefits of BN while removing the direct influence of oversampled anomalies from normalization-statistic estimation.

4.4 Normalization Comparison

Table 1 compares normalization methods across eleven tabular datasets. The effectiveness of conventional normalization varies substantially across SAD architectures. The w/o BN baseline performs best among the baselines for DevNet and FeaWAD, whereas BN performs best for DSAD and PReNet, and LN for CAD. LN and SN provide no consistent improvement over BN, while NBN exhibits model-dependent performance. AnoOnly yields the weakest overall results, suggesting that its normalization strategy does not generalize effectively across SAD architectures. In contrast, NoBa achieves the highest average AUPR for all five backbones, outperforming the strongest competing normalization methods by 0.053, 0.021, 0.044, 0.030, and 0.020 on DevNet, DSAD, PReNet, FeaWAD, and CAD, respectively. NoBa also achieves the best average rank across datasets, indicating that its improvements are consistent rather than concentrated on a few datasets. Notably, NoBa further improves DSAD despite the regularization already imposed by its one-class objective, demonstrating its effectiveness across models with different training objectives.

To evaluate robustness to noisy supervision, we randomly corrupt 2% of the normal-labeled training samples following the PReNet protocol. NoBa achieves the highest average AUPR on DevNet, PReNet, and FeaWAD, while remaining competitive with the strongest variants on DSAD and CAD and substantially outperforming BN, LN, SN, and NBN. These results show that NoBa remains robust to moderate contamination of normal-labeled data and largely preserves its performance gains observed under clean supervision.

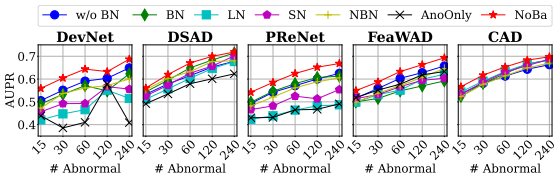


Figure 6: Average AUPR over the eleven tabular datasets following the PReNet protocol as the number of labeled abnormal instances increases, reported for five SAD models.

Table 2: Ablation study of normalization statistics and reference sets. Med. denotes the median statistic.

Method	Normal&Abnormal			Normal-Only		
	Mean	Med.	Mode	Mean	Med.	Mode
DevNet	0.566	0.605	0.610	**0.652**	0.639	0.644
DSAD	0.648	0.650	0.668	0.655	0.661	**0.670**
PReNet	0.585	0.611	0.598	0.616	0.621	**0.624**
FeaWAD	0.555	0.611	0.614	0.617	0.632	**0.633**
CAD	0.624	0.633	0.639	0.649	0.651	**0.653**
Avg.	0.596	0.622	0.626	0.638	0.641	**0.645**

Figure 6 reports performance as the number of labeled abnormal instances varies. As the anomaly count increases, all methods except AnoOnly generally improve, while NoBa maintains a consistent advantage over both w/o BN and conventional BN. Overall, NoBa remains effective across different levels of anomaly supervision.

4.5 Ablation Study

Table 2 examines two complementary design factors in normalization: the reference set and the centering statistic. First, restricting statistic estimation to normal samples consistently improves average AUPR across all centering statistics, from 0.596 to 0.638 for the mean, 0.622 to 0.641 for the median, and 0.626 to 0.645 for the Pearson mode. This consistent improvement indicates that excluding repeatedly oversampled anomalies from statistic estimation is the primary source of the gain. Second, the choice of centering statistic provides an additional benefit. Even when both normal and abnormal samples are used, replacing the mean with the median or Pearson mode improves average AUPR from 0.596 to 0.622 and 0.626, respectively. Mean-based centering can be shifted by extreme or asymmetric activations, whereas the median provides a robust alternative and Pearson mode explicitly accounts for skewness. A similar trend appears under normal-only estimation, where the median and Pearson mode further improve average AUPR from 0.638 to 0.641 and 0.645, respectively, suggesting that normal feature distributions can themselves be asymmetric. Overall, the ablation reveals two complementary effects: normal-only reference selection removes the influence of oversampled anomalies from statistic estimation, while robust or skewness-aware centering better accommodates asymmetric feature distributions within the selected reference set. Pearson-mode centering achieves the highest average AUPR and the best performance on four of the five backbones.

5 Conclusion

We propose NoBa, a simple normalization scheme for tabular semi-supervised anomaly detection (SAD). We identify *distribution collapse*, a failure mode in which the batch-to-batch variability of normalized normal representations progressively diminishes under anomaly-oversampled training, weakening a source of BN's implicit regularization. NoBa addresses this issue by estimating normalization statistics exclusively from normal samples, removing the direct influence of oversampled anomalies while preserving representation-level stochasticity and the optimization benefits of BN. Experiments across eleven datasets and five representative SAD models demonstrate consistent improvements over conventional BN and alternative normalization methods. A limitation of NoBa is its reliance on sufficiently reliable normal samples for statistic estimation, which may reduce its effectiveness under severe contamination or when only a small number of normal samples are available per mini-batch. Nevertheless, NoBa can be readily incorporated into existing tabular SAD architectures without modifying their training objectives, establishing normal-only statistic estimation as a simple and broadly applicable normalization strategy for this setting.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript refinement, including improving the clarity and readability of the writing and organizing and formatting tables. We reviewed all AI-assisted content and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our results. The complete NoBa procedure is described in Section 3, with the full algorithm provided in Algorithm 1 (Appendix A). All datasets are publicly available, and their original sources, access URLs, and statistics are provided in Appendix E (Tables 4 and 5); we follow the same data-processing protocol as PReNet (Pang et al., 2023). Model architectures, hyperparameters, optimizer settings, oversampling protocols, and evaluation metrics are specified in Section 4 and Appendix F. All results are averaged over 10 runs with distinct random seeds to account for run-to-run variability. We will release the source code upon publication.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4042–4053, 2024a. doi: 10.1109/CVPRW63382.2024.00408.

Yuxiang Bao, Guoliang Kang, Linlin Yang, Xiaoyue Duan, Bo Zhao, and Baochang Zhang. Normalizing batch normalization for long-tailed recognition. *IEEE Transactions on Image Processing*, 2024b. doi: 10.1109/TIP.2024.3518099.

Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 127–137, 2024. doi: 10.1109/WACV57701.2024.00020.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 4(5):11, 2015.

Hao Dong, Gaëtan Frusque, Yue Zhao, Eleni Chatzi, and Olga Fink. Nng-mix: Improving semi-supervised anomaly detection with pseudo-anomaly generation. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024. doi: 10.1109/TNNLS.2024.3497801.

Iñaki Elía and Miguel Pagola. Anomaly detection in smart-manufacturing era: A review. *Engineering Applications of Artificial Intelligence*, 139:109578, 2025. doi: 10.1016/j.engappai.2024.109578.

Hongping Gan, Hejie Zheng, Zhangfa Wu, Chunyan Ma, and Jie Liu. Tfd-net: Transformer deviation network for weakly supervised anomaly detection. *IEEE Transactions on Network and Service Management*, 22(1):941–954, 2024.

Hongping Gan, Hejie Zheng, J Nescolarde Selva, Yunong Chen, Zhangfa Wu, and Farhan Ullah. Tcad-net: Integrating transformer and convolution for weakly supervised anomaly data discovery. *IEEE Transactions on Consumer Electronics*, 2025.

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.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35: 507–520, 2022.

Waleed Hilal, S Andrew Gadsden, and John Yawney. Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems With applications*, 193:116429, 2022. doi: 10.1016/j.eswa.2021.116429.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1729–1739, 2017.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. *Advances in neural information processing systems*, 30, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pp. 448–456, 2015.

John F Kenney and ES Keeping. Mathematics of statistics, vol. ii, 1951.

Ron Kohavi. Census Income. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5GP7S.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023. doi: 10.1109/CVPR52729.2023.01954.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, and Jingyu Li. Differentiable learning-to-normalize via switchable normalization. *International Conference on Learning Representation (ICLR)*, 2019a.

Ping Luo, Xinjiang Wang, Wenqi Shao, and Zhanglin Peng. Towards understanding regularization in batch normalization. In *International Conference on Learning Representations*, 2019b. URL `https://openreview.net/forum?id=HJlLKjR9FQ`.

Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. Identifying suspicious urls: an application of large-scale online learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 681–688, 2009.

Tohgoroh Matsui. Credit card fraud detection. `https://kaggle.com/competitions/1056lab-credit-card-fraud-detection`, 2019. Kaggle.

S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5K306.

Nour Moustafa and Jill Slay. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *2015 military communications and information systems conference (MilCIS)*, pp. 1–6. IEEE, 2015. doi: 10.1109/MilCIS.2015.7348942.

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. Association for Computing Machinery, 2019. doi: 10.1145/3292500.3330871.

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. Association for Computing Machinery, 2023. doi: 10.1145/3580305.3599302.

Maria Antonietta Panza, Marco Pota, and Massimo Esposito. Anomaly detection methods for industrial applications: A comparative study. *Electronics*, 12(18):3971, 2023. doi: 10.3390/electronics12183971.

Claudia Perlich, Vladimir Dubovskiy, and Will Cukierski. Kdd cup 2014 - predicting excitement at donorschoose.org. https://kaggle.com/competitions/kdd-cup-2014-predicting-excitement-at-donors-choose, 2014. Kaggle.

Ross Quinlan. Thyroid Disease. UCI Machine Learning Repository, 1986. DOI: https://doi.org/10.24432/C5D010.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

Ira Shavitt and Eran Segal. Regularization learning networks: deep learning for tabular datasets. *Advances in neural information processing systems*, 31, 2018.

Saeid Asgari Taghanaki, Ali Gholami, Fereshte Khani, Kristy Choi, Linh Tran, Ran Zhang, and Aliasghar Khani. Counterbalancing teacher: Regularizing batch normalized models for robustness. *arXiv preprint arXiv:2207.01548*, 2022.

Hugo Thimonier, Fabrice Popineau, Arpad Rimmel, Bich-Liên Doan, and Fabrice Daniel. Comparative evaluation of anomaly detection methods for fraud detection in online credit card payments. In *International Congress on Information and Communication Technology*, pp. 37–50, 2024. doi: 10.1007/978-981-97-4581-4_4.

Maximilian E Tschuchnig and Michael Gadermayr. Anomaly detection in medical imaging-a mini review. In *International Data Science Conference*, pp. 33–38, 2021. doi: 10.1007/978-3-658-36295-9_5.

Feng Xiao, Youqing Wang, S Joe Qin, and Jicong Fan. Semi-supervised anomaly detection using restricted distribution transformation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Hongzuo Xu, Yijie Wang, Guansong Pang, Songlei Jian, Ning Liu, and Yongjun Wang. Rosas: Deep semi-supervised anomaly detection with contamination-resilient continuous supervision. *Information Processing & Management*, 60(5), 2023. doi: 10.1016/j.ipm.2023.103459.

Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan O Arik, and Tomas Pfister. SPADE: Semi-supervised anomaly detection under distribution mismatch. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JwDpZSv3yz. Featured Certification.

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

Yixuan Zhou, Peiyu Yang, Yi Qu, Xing Xu, Zhe Sun, and Cichocki Andrzej. Anoonly: Semi-supervised anomaly detection with the only loss on anomalies. *Expert Systems with Applications*, 262:125597, 2025. doi: 10.1016/j.eswa.2024.125597.

A ALGORITHM

Algorithm 1 outlines the proposed NoBa procedure. In contrast to the conventional BN, NoBa computes normalization statistics solely from normal data and substitutes the mean with the Pearson mode as defined in Eq. 8. Aside from its modified statistical computation, NoBa follows the standard BN framework. It employs a moving average with a decay factor of $\alpha = 0.1$ for inference-time normalization and includes the learnable scale (γ) and shift (β) parameters to facilitate adaptable feature transformations. By preserving the architectural structure of BN, NoBa inherits key advantages such as compatibility with large learning rates and a smoother optimization landscape.

Algorithm 1 Normal-Only Batch Normalization

Input: Mini-batch layer activations $B = \{x_i\}_{\forall i}$ with labels y_i, parameters γ and β, current moving mean μ and variance σ^2, cumulative statistic rate α
Set $\tilde{B} = \{x_i : y_i = 0\}_{\forall i}$
for $k = 1 \dots K$ **do**
  //$x \equiv x^k$ For clarity, omitting the dimension.
  if Training **then**
    $\mu^{\text{mode}} = 3\text{median}_{\tilde{x}\in\tilde{B}}(\tilde{x}) - 2\mathbb{E}_{\tilde{x}\in\tilde{B}}(\tilde{x})$
    $(\sigma^{\text{mode}})^2 = \frac{1}{|\tilde{B}|}\sum_{\tilde{x}\in\tilde{B}}(\tilde{x} - \mu^{\text{mode}})^2$
    $\hat{x_i} = (x_i - \mu^{\text{mode}})/\sqrt{(\sigma^{\text{mode}})^2 + \epsilon}$
    $\mu = (1-\alpha)\mu + \alpha\mu^{\text{mode}}$
    $\sigma^2 = (1-\alpha)\sigma^2 + \alpha\frac{|\tilde{B}|}{|\tilde{B}|-1}(\sigma^{\text{mode}})^2$
  else
    $\hat{x_i} = (x_i - \mu)/\sqrt{\sigma^2 + \epsilon}$
  end if
  $\hat{x_i} = \gamma\hat{x_i} + \beta$
end for

B Theoretical Analysis of Distribution Collapse and NoBa

B.1 Proof of Proposition 1

Recall that conventional BN uses the shared batch mean

$$\mu^{\text{all}} = (1-\pi)\mu^{\text{norm}} + \pi\mu^{\text{abnorm}}. \tag{9}$$

The mean normalized activation of the normal subset is

$$\begin{aligned}\bar{z}_N^{\text{all}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \mu^{\text{all}}}{\sigma^{\text{all}}} \\ &= \frac{\mu^{\text{norm}} - \mu^{\text{all}}}{\sigma^{\text{all}}} \\ &= -\pi\frac{\mu^{\text{abnorm}} - \mu^{\text{norm}}}{\sigma^{\text{all}}} = -\pi q_{\mathcal{B}}.\end{aligned} \tag{10}$$

Since the abnormal batch ratio π is fixed across mini-batches, taking the variance over $\mathcal{B}$ gives

$$\text{Var}_{\mathcal{B}}\left(\bar{z}_N^{\text{all}}\right) = \pi^2\,\text{Var}_{\mathcal{B}}\left(q_{\mathcal{B}}\right), \tag{11}$$

which proves Proposition 1.

B.2 Proof of Proposition 2

We show that the batch-dependent variability retained by NoBa can remain reflected in the normalized representations. Under conventional BN, the normalized activation before the affine transformation is

$$z_i^{\text{all}} = \frac{x_i - \mu^{\text{all}}}{\sigma^{\text{all}}}. \tag{12}$$

Since $\mu^{\text{all}} = \frac{1}{n+m}\sum_{i=1}^{n+m} x_i$, the aggregate normalized activation satisfies

$$\begin{aligned}\bar{z}^{\text{all}} &= \frac{1}{n+m}\sum_{i=1}^{n+m} z_i^{\text{all}} \\ &= \frac{1}{(n+m)\sigma^{\text{all}}}\sum_{i=1}^{n+m}\left(x_i - \mu^{\text{all}}\right) = 0.\end{aligned} \tag{13}$$

This holds for every mini-batch, and therefore

$$\mathrm{Var}_{\mathcal{B}}\big(\bar{z}^{\mathrm{all}}\big) = 0. \tag{14}$$

Under NoBa, all samples are normalized using statistics estimated exclusively from the n normal samples:

$$z_i^{\mathrm{norm}} = \frac{x_i - \mu^{\mathrm{norm}}}{\sigma^{\mathrm{norm}}}. \tag{15}$$

The aggregate normalized activation is

$$\bar{z}^{\mathrm{norm}} = \frac{1}{n+m}\left(\sum_{i=1}^{n} z_i^{\mathrm{norm}} + \sum_{i=1}^{m} z_{n+i}^{\mathrm{norm}}\right). \tag{16}$$

By the definition of μ^{norm},

$$\sum_{i=1}^{n} z_i^{\mathrm{norm}} = \frac{1}{\sigma^{\mathrm{norm}}} \sum_{i=1}^{n} (x_i - \mu^{\mathrm{norm}}) = 0. \tag{17}$$

Defining the aggregate normalized abnormal activation as

$$\mathcal{A} := \sum_{i=1}^{m} z_{n+i}^{\mathrm{norm}}, \tag{18}$$

we obtain

$$\bar{z}^{\mathrm{norm}} = \frac{\mathcal{A}}{n+m}. \tag{19}$$

Taking the variance across mini-batches gives

$$\mathrm{Var}_{\mathcal{B}}(\bar{z}^{\mathrm{norm}}) = \frac{1}{(n+m)^2}\,\mathrm{Var}_{\mathcal{B}}(\mathcal{A}). \tag{20}$$

Whenever $\mathrm{Var}_{\mathcal{B}}(\mathcal{A}) > 0$, Eqs. 14 and 20 imply

$$\mathrm{Var}_{\mathcal{B}}(\bar{z}^{\mathrm{norm}}) > \mathrm{Var}_{\mathcal{B}}\big(\bar{z}^{\mathrm{all}}\big) = 0, \tag{21}$$

which proves Proposition 2.

The source of this variability is the use of normal-derived statistics for all samples. Under NoBa, abnormal samples are normalized relative to μ^{norm} and σ^{norm}, both of which vary with the composition of the normal mini-batch. Consequently, even repeatedly sampled abnormal observations can exhibit different normalized activations as the normal reference changes. Proposition 2 therefore establishes that normal-only normalization permits batch-dependent variability to remain in the aggregate normalized representation. This result identifies a representation-level source of stochasticity that can contribute to the implicit regularization associated with mini-batch normalization (Luo et al., 2019b).

C Further Analysis of NoBa

Distribution Visualization Figure 7 visualizes the evolution of feature statistics under BN and NoBa on Census with DevNet. As training progresses, conventional BN exhibits progressively reduced batch-to-batch variation in the normalized representations, consistent with distribution collapse. In contrast, NoBa estimates normalization statistics exclusively from normal samples, thereby removing the direct influence of repeatedly oversampled anomalies on the normalization reference. Consequently, NoBa retains greater batch-dependent variation throughout training, consistent with the representation-level stochasticity characterized in Proposition 2.

Learning Rate BN is known to support relatively large learning rates, contributing to efficient optimization and fast convergence. To assess whether NoBa retains this property, we evaluate each normalization method across a range of learning rates. As shown in Figure 8, conventional BN maintains relatively stable performance across learning rates, whereas AnoOnly (Zhou et al., 2025), Layer Normalization (LN) (Ba et al., 2016), Switchable Normalization (SN) (Luo et al., 2019a), and Normalizing Batch Normalization (NBN) (Bao et al., 2024b) exhibit greater sensitivity. NoBa shows robustness comparable to conventional BN while consistently achieving higher performance. These results provide further evidence that NoBa retains BN's favorable optimization behavior, including tolerance to large learning rates, while avoiding the direct influence of oversampled anomalies on normalization-statistic estimation.

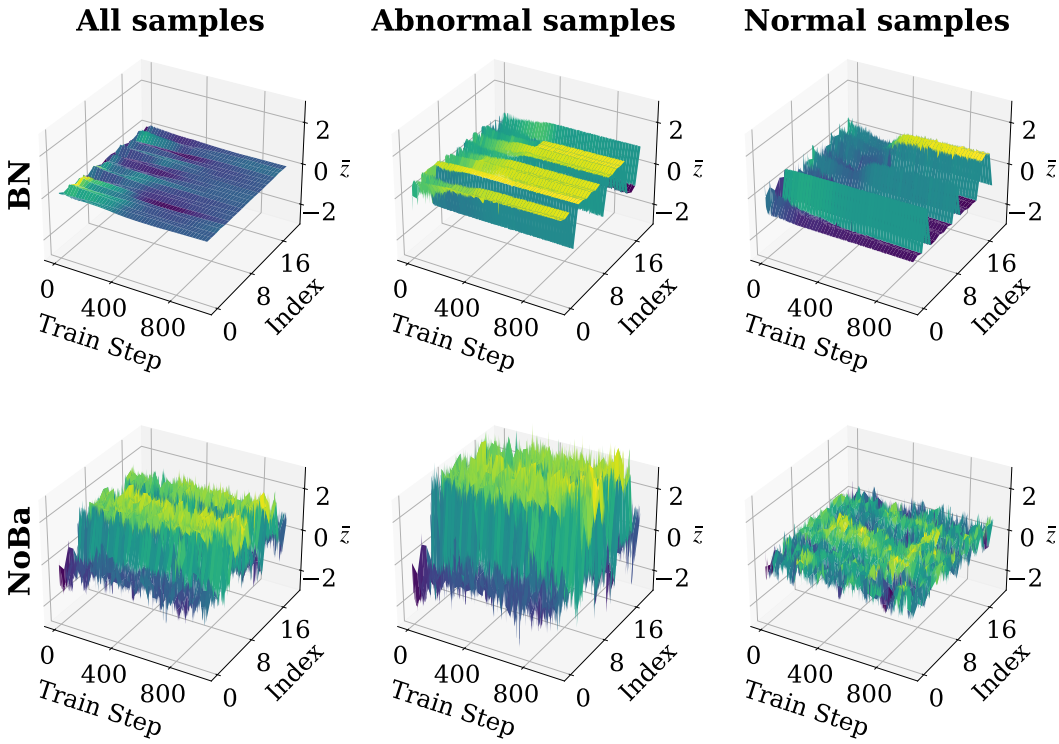


Figure 7: Progression of batch normalized features after BN and NoBa under different training setups on the Census dataset. The x-, y-, and z-axes denote the feature index, training step, and feature value $\bar{z}$, respectively. Here, $\bar{z}$ denotes the mean of each normalized feature within a mini-batch. While BN exhibits reduced variance and statistical conflation caused by oversampling, NoBa preserves statistical variability that closely mirrors the natural variability of normal data.

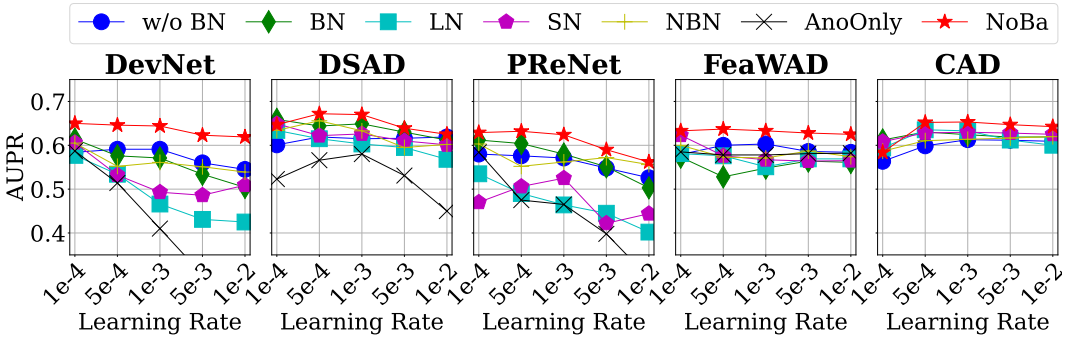


Figure 8: Average AUPR across different learning rates for five SAD backbones (DevNet, DSAD, PReNet, FeaWAD, and CAD), averaged over the eleven tabular datasets following the PReNet protocol. Higher is better.

Batch Size Normalization methods based on batch statistics can be sensitive to batch size. Figure 9 reports AUPR across different batch sizes. Performance generally decreases as the batch size increases, which is consistent with reduced mini-batch stochasticity and weaker implicit regularization from batch-dependent normalization. Batch-independent methods such as LN exhibit comparatively smaller changes, although some degradation remains. Among BN-based methods, NoBa shows substantially lower sensitivity to batch size and maintains stronger performance across

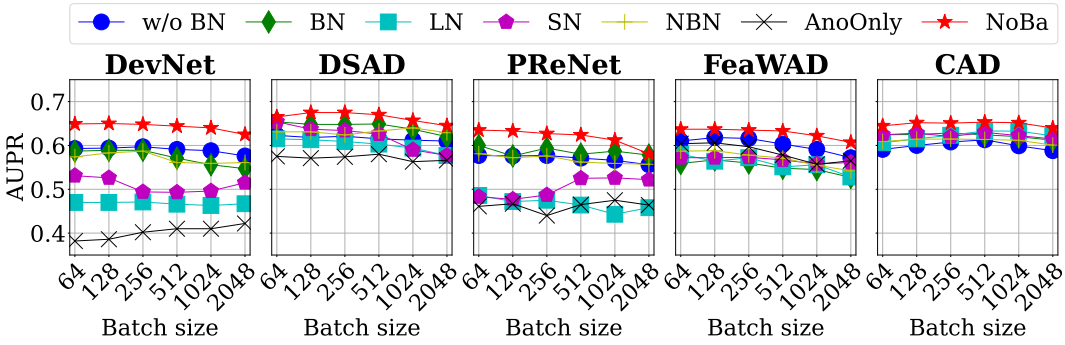


Figure 9: Batch-size sensitivity of normalization methods. Each point reports the average AUPR across the eleven tabular datasets for five SAD backbones (DevNet, DSAD, PReNet, FeaWAD, and CAD).

the evaluated range. These results further support the robustness of NoBa to changes in mini-batch composition.

Resource Usage Table 3 shows the computation resources. NoBa incurs minimal computational overhead compared to BN, as it replaces batch statistic estimation with a median-based computation.

Table 3: Resource usage comparison.

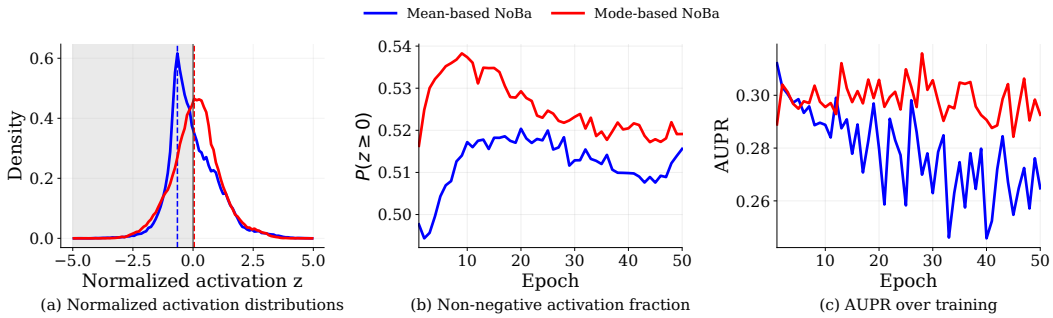
Method	# Params	Memory (MB)	MACs (M)	Train (s)	Test (s)
w/o BN	266K	23.23	136.34	13.61	0.13
BN	269K	23.29	137.64	14.51	0.16
SN	269K	31.26	140.24	18.69	0.26
NBN	270K	27.20	136.99	18.19	0.25
NoBa	269K	27.19	137.64	19.63	0.19

D Analysis of Skewness-Aware Centering

We further examine the effect of skewness-aware centering using DevNet in Figure 10. Under right-skewed activations, mean centering places the density peak in the negative normalized region, where ELU exhibits reduced gradients. Pearson-mode centering shifts the high-density region closer to the origin and increases the proportion of activations in ELU's non-negative, unit-gradient region. Figure 10 illustrates this behavior over training. Mean centering places the density peak near -0.65, whereas mode centering shifts it to approximately $+0.05$. This shift increases the fraction of activations in the non-negative ELU region and is accompanied by improved AUPR.

E Dataset Details

Following the experimental protocol established by the widely recognized semi-supervised anomaly detection (SAD) method PReNet (Pang et al., 2023), we evaluate our approach on eleven tabular datasets: donors, census, fraud, celeba, dos, reconnaissance (rec), fuzzers (fuz), backdoor (bac), w7a, campaign, and thyroid. The donors dataset, derived from the KDD Cup 2014, involves predicting the excitement level of donation projects, with exceptionally exciting projects labeled as anomalies. The census dataset is sourced from the U.S. Census Bureau, where high-income individuals, due to their rarity, are treated as anomalies. The fraud dataset is designed for detecting anomalous credit card transactions indicative of fraudulent behavior. The celeba dataset comprises 200K celebrity images annotated with 40 binary attributes, of which images labeled with the "bald" attribute are considered anomalous due to their rarity. The dos, rec, fuz, and bac datasets are drawn from the UNSW-NB15 benchmark (Moustafa & Slay, 2015), each representing distinct categories of cyberattacks, which are designated as anomalies. The w7a dataset pertains to webpage classification, where instances from minority classes are treated as anomalous. The campaign dataset



(a) Normalized activation distributions (b) Non-negative activation fraction (c) AUPR over training

Figure 10: Effect of skewness-aware centering with DevNet. (a) Pearson-mode centering shifts the normalized activation density peak from approximately -0.65 to $+0.05$. (b) The shift increases the fraction of activations in ELU's non-negative, unit-gradient region throughout training. (c) Corresponding AUPR trajectories.

Table 4: The datasets exhibit considerable variation in anomaly prevalence, with anomaly ratios spanning from 0.17% to 14.95%. The columns labeled # Normal and # Abnormal correspond to the counts of normal and abnormal samples, respectively, and D denotes the feature dimensionality.

Data	# Normal	# Abnormal	Ratio	D
donors	582,616	36,710	5.93%	10
census	280,717	18,568	6.20%	500
fraud	284,315	492	0.17%	29
celeba	198,052	4,547	2.24%	39
dos	93,000	16,353	14.95%	196
rec	93,000	13,987	13.07%	196
fuz	93,000	3,000	3.13%	196
bac	93,000	2,329	2.44%	196
w7a	48,270	1,479	2.97%	300
campaign	36,548	4,640	11.27%	62
thyroid	6,666	534	7.42%	21

contains data from a bank marketing campaign, with rare positive responses identified as anomalies. Lastly, the thyroid dataset focuses on medical diagnosis, where patients with hypothyroidism are classified as anomalies. Table 4 provides a summary of the datasets, including the number of normal and abnormal instances, the anomaly ratio, and the feature dimensionality. For reproducibility, the original sources and publicly accessible URLs of all datasets are listed in Table 5.

F Implementation Details

Each SAD model adhered to the official architecture provided in its respective codebase. DevNet (Pang et al., 2019) and PReNet (Pang et al., 2023) were implemented using three linear layers with hidden dimensions of 1000, 250, and 20. FeaWAD employed an autoencoder comprising two fully connected (FC) layers with hidden sizes of 128 and 64, followed by a score generator consisting of two FC layers with dimensions 256 and 32. DSAD (Ruff et al., 2020) utilized a three-layer FC architecture with hidden sizes of 128, 64, and 32. CAD (Gao et al., 2025) utilized a three layer FC architecture with hidden sizes of 128, 64, and 8 for feature extraction. All SAD models employed ReLU activation after each FC layer, except for the output layer. However, in Batch Normalization (BN)-based architectures (excluding AnoOnly and w/o BN), ReLU was replaced with ELU (Clevert et al., 2015). This modification was motivated by the tendency of BN to induce strong zero-centering, which can suppress activations and impede learning. ELU retains small non-zero values near zero, preserving gradient flow while maintaining ReLU-like behavior. Training was performed over 50 epochs, with 20 mini-batches per epoch and a batch size of 512. To address the rarity of anomalies, we employed oversampling to achieve a balanced dataset, maintaining a 1:1 ratio between normal and abnormal instances per mini-batch, following the standard SAD setup.

Table 5: Original sources and public URLs of the datasets used in our experiments. The dos, rec, fuz, and bac datasets correspond to the DoS, reconnaissance, fuzzers, and backdoor attack categories drawn from the UNSW-NB15 benchmark.

Data	Original source and URL
donors	KDD Cup 2014: Predicting Excitement at DonorsChoose.org (Perlich et al., 2014) https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose
census	Census Income, UCI ML Repository (Kohavi, 1996) https://archive.ics.uci.edu/ml/datasets/census+income
fraud	Credit Card Fraud Detection (Matsui, 2019) https://www.kaggle.com/c/1056lab-credit-card-fraud-detection
celeba	CelebA (Liu et al., 2015) http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
dos / rec / fuz / bac	UNSW-NB15 (Moustafa & Slay, 2015) https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
w7a	web-page classification (URL features) (Ma et al., 2009) https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
campaign	Bank Marketing dataset (Moro et al., 2014) https://archive.ics.uci.edu/ml/datasets/bank+marketing
thyroid	Thyroid Disease (Quinlan, 1986) https://www.openml.org/d/40497

We employed the RMSprop optimizer (Graves, 2013) with a learning rate of 0.001 and a weight decay of 0.01 to implement L^2 regularization. For FeaWAD (Zhou et al., 2022), the autoencoder was pretrained for 100 epochs, followed by an additional 30 epochs of joint training with the score generator.

G Detailed Experimental Results

The main paper reports average AUPR and average rank across datasets. This section provides dataset-level results for a more detailed comparison of the normalization strategies. Tables 6 and 7 report the performance of all variants (w/o BN, BN, LN, SN, NBN, AnoOnly, and NoBa) on the eleven benchmark datasets under the standard and 2% label contamination settings, respectively. Within each model block, the best and second-best results are highlighted in bold and underlined, respectively.

Conventional BN substantially degrades performance on several datasets, whereas NoBa largely mitigates this degradation. The clearest example is fraud, where DevNet decreases from 0.733 to 0.369 with BN, while NoBa largely recovers performance to 0.726. Similar patterns are observed for PReNet ($0.690 \rightarrow 0.431 \rightarrow 0.616$) and FeaWAD ($0.705 \rightarrow 0.249 \rightarrow 0.727$). On bac, NoBa further outperforms both w/o BN (0.866) and BN (0.853), achieving 0.912 with DevNet. These results further support the benefit of excluding oversampled anomalies from normalization-statistic estimation.

BN is beneficial on other datasets. On thyroid, BN improves DevNet, PReNet, FeaWAD, and CAD over their w/o BN counterparts, while NoBa yields further gains, reaching AUPRs of 0.846, 0.858, 0.862, and 0.872, respectively. On campaign, NBN achieves the best performance for several models, whereas NoBa remains competitive. On donors, several backbones already attain near-perfect performance without BN, leaving limited room for improvement.

Overall, NoBa achieves the highest average AUPR and average rank across all five backbones under the standard setting. Under 2% label contamination, it retains the highest average AUPR for DevNet, PReNet, and FeaWAD and remains competitive for DSAD and CAD, where w/o BN performs slightly better on average. Nevertheless, NoBa consistently outperforms conventional BN in average performance under contamination, indicating greater robustness to noisy supervision. These dataset-level results are consistent with the main findings: NoBa retains the benefits of normalization while reducing the adverse influence of oversampled anomalies on normalization-statistic estimation.

Table 6: AUPR comparison across normalization types for each base model. For each dataset, the best result within each model block is in **bold** and the second best is underlined.

Model	BN type	donors	census	fraud	celeba	dos	rec	fuz	bac	w7a	camp-aign	thy-roid	Avg. AUPR	Avg. Rank
DevNet	w/o BN	**1.000**	**0.360**	**0.733**	0.281	0.923	**0.880**	0.167	0.866	**0.507**	0.308	0.475	0.591	2.450
	BN	0.943	0.275	0.369	0.275	0.923	0.857	0.197	0.853	0.504	0.423	0.661	0.571	3.360
	LN	0.976	0.270	0.317	0.178	0.839	0.640	0.118	0.758	0.424	0.290	0.313	0.466	5.910
	SN	0.982	0.224	0.329	0.213	0.879	0.772	0.168	0.721	0.480	0.333	0.321	0.493	4.730
	NBN	0.835	0.322	0.621	0.283	0.865	0.787	0.231	0.817	0.373	**0.526**	0.512	0.561	3.730
	AnoOnly	0.909	0.327	0.226	0.226	0.606	0.672	0.131	0.589	0.334	0.329	0.163	0.410	6.000
	NoBa (ours)	0.953	0.349	0.726	**0.286**	**0.937**	0.869	**0.234**	**0.912**	0.504	0.465	**0.846**	**0.644**	**1.820**
DSAD	w/o BN	**1.000**	0.317	0.715	0.314	0.943	**0.905**	0.284	0.915	0.549	0.304	0.515	0.615	3.820
	BN	0.999	0.317	0.672	0.234	**0.960**	0.894	0.292	**0.940**	**0.619**	0.401	**0.810**	0.649	2.910
	LN	0.999	0.306	**0.744**	0.266	0.936	0.892	0.224	0.889	0.483	0.338	0.556	0.603	5.000
	SN	0.997	0.269	0.688	0.254	0.952	0.894	0.242	0.934	0.553	0.344	0.760	0.626	4.360
	NBN	0.997	0.330	0.668	0.312	0.952	0.886	0.283	0.929	0.596	0.409	0.597	0.633	4.000
	AnoOnly	0.995	**0.362**	0.692	0.295	0.923	0.811	0.182	0.868	0.464	0.377	0.405	0.580	5.550
	NoBa (ours)	0.992	0.359	0.743	**0.319**	0.949	0.896	**0.308**	0.935	0.614	**0.450**	0.809	**0.670**	**2.360**
PReNet	w/o BN	**1.000**	0.320	**0.690**	0.267	0.887	**0.867**	0.159	0.863	0.482	0.298	0.449	0.571	3.180
	BN	0.944	0.301	0.431	0.261	**0.938**	0.845	0.178	0.877	**0.530**	0.464	0.610	0.580	3.180
	LN	0.978	0.235	0.372	0.183	0.826	0.689	0.111	0.749	0.405	0.247	0.306	0.464	6.360
	SN	0.998	0.245	0.331	0.245	0.908	0.719	0.135	0.866	0.457	0.358	0.514	0.525	4.550
	NBN	0.913	0.310	0.464	**0.287**	0.892	0.815	**0.197**	0.874	0.415	**0.513**	0.498	0.562	3.270
	AnoOnly	0.998	0.305	0.254	0.271	0.820	0.770	0.114	0.825	0.363	0.264	0.129	0.465	5.450
	NoBa (ours)	0.983	**0.338**	0.616	0.278	0.928	0.856	0.185	**0.893**	0.469	0.464	**0.858**	**0.624**	**2.000**
FeaWAD	w/o BN	0.998	0.228	0.705	0.242	**0.927**	0.864	0.212	**0.833**	0.499	0.426	0.701	0.603	3.360
	BN	0.999	0.269	0.249	0.227	0.901	0.861	0.151	0.740	0.532	0.367	0.735	0.548	4.820
	LN	**1.000**	0.201	0.363	0.226	0.921	0.804	0.179	0.798	0.523	0.346	0.704	0.551	4.270
	SN	**1.000**	0.223	0.420	0.209	0.924	0.870	0.138	0.779	0.520	0.317	0.834	0.567	4.180
	NBN	0.999	0.299	0.353	0.256	0.861	0.860	0.193	0.767	0.490	0.432	0.774	0.571	4.270
	AnoOnly	0.996	0.201	0.657	0.242	0.914	0.867	0.175	0.788	0.490	0.337	0.687	0.578	5.000
	NoBa (ours)	**1.000**	**0.322**	**0.727**	**0.287**	0.893	**0.877**	**0.220**	0.756	**0.582**	**0.435**	**0.862**	**0.633**	**2.090**
CAD	w/o BN	**1.000**	0.327	0.646	0.264	0.846	**0.896**	0.192	0.857	0.501	0.394	0.817	0.613	4.364
	BN	**1.000**	0.273	0.626	0.276	0.923	0.883	0.194	0.892	0.517	0.437	0.844	0.624	3.818
	LN	**1.000**	0.341	0.697	0.284	0.891	0.874	0.192	0.883	0.495	0.458	0.847	0.633	3.818
	SN	**1.000**	0.309	0.667	0.280	0.911	0.881	**0.205**	0.873	0.503	0.465	0.833	0.630	3.455
	NBN	0.980	0.318	0.489	0.277	0.925	0.885	0.185	0.903	0.497	0.478	0.813	0.614	4.091
	AnoOnly	-	-	-	-	-	-	-	-	-	-	-	-	-
	NoBa (ours)	**1.000**	**0.357**	**0.698**	**0.299**	**0.928**	0.878	0.194	**0.916**	**0.552**	**0.492**	**0.872**	**0.653**	**1.455**

Table 7: AUPR comparison under 2% label contamination in the normal-labeled training set. For each dataset, we corrupt 2% of the training instances labeled as normal, following the contamination ratio used in PReNet, to simulate noisy normal supervision in semi-supervised anomaly detection. Within each model block, the best result per dataset is in **bold** and the second best is underlined.

Model	BN type	donors	census	fraud	celeba	dos	rec	fuz	bac	w7a	camp-aign	thy-roid	Avg. AUPR	Avg. Rank
DevNet	w/o BN	**1.000**	0.327	**0.715**	0.266	**0.896**	**0.873**	0.161	**0.862**	0.403	0.285	0.435	0.566	2.360
	BN	0.931	0.217	0.336	0.210	0.803	0.839	0.160	0.703	0.415	0.394	0.688	0.518	4.090
	LN	0.936	0.242	0.379	0.184	0.867	0.737	0.124	0.805	0.340	0.268	0.292	0.470	5.270
	SN	0.973	0.190	0.346	0.193	0.844	0.770	0.148	0.672	0.353	0.302	0.317	0.464	4.820
	NBN	0.907	0.297	0.511	0.274	0.770	0.756	0.205	0.825	0.365	**0.523**	0.491	0.538	3.450
	AnoOnly	0.853	**0.329**	0.258	0.224	0.705	0.612	0.139	0.457	0.353	0.294	0.241	0.406	5.820
	NoBa (ours)	0.959	0.318	0.567	**0.282**	0.846	0.838	**0.219**	0.780	**0.432**	0.433	**0.808**	**0.589**	**2.180**
DSAD	w/o BN	**1.000**	0.270	0.711	**0.289**	**0.937**	**0.883**	**0.250**	**0.864**	0.367	0.282	0.507	**0.578**	**2.730**
	BN	0.997	0.248	0.321	0.215	0.815	0.861	0.216	0.632	0.398	0.376	**0.769**	0.532	3.910
	LN	0.996	0.275	**0.737**	0.236	0.793	0.847	0.168	0.581	0.229	0.331	0.529	0.520	4.820
	SN	0.930	0.236	0.581	0.203	0.813	0.871	0.171	0.628	0.285	0.314	0.490	0.502	5.270
	NBN	0.995	0.280	0.435	0.270	0.763	0.853	0.228	0.587	0.415	0.397	0.635	0.532	4.180
	AnoOnly	0.994	**0.356**	0.684	0.284	0.922	0.804	0.162	0.857	0.404	0.367	0.397	0.567	4.000
	NoBa (ours)	0.995	0.293	0.484	0.271	0.777	0.867	0.234	0.625	**0.474**	**0.416**	0.713	0.559	3.090
PReNet	w/o BN	**0.999**	**0.312**	**0.669**	0.258	0.874	**0.856**	**0.164**	0.862	0.401	0.275	0.446	0.556	**2.090**
	BN	0.981	0.269	0.515	0.227	0.851	0.827	0.163	0.764	**0.438**	0.410	0.687	0.557	3.270
	LN	0.931	0.223	0.355	0.185	0.729	0.684	0.112	0.702	0.239	0.213	0.325	0.427	6.270
	SN	0.994	0.205	0.270	0.188	0.787	0.773	0.126	0.541	0.312	0.310	0.323	0.439	5.360
	NBN	0.863	0.308	0.534	**0.275**	0.830	0.796	0.159	0.855	0.378	**0.511**	0.477	0.544	3.180
	AnoOnly	**0.999**	0.279	0.231	0.254	0.764	0.752	0.119	0.614	0.295	0.240	0.146	0.427	5.450
	NoBa (ours)	0.979	0.302	0.595	0.254	**0.912**	0.838	0.158	**0.890**	**0.438**	0.408	**0.803**	**0.598**	2.360
FeaWAD	w/o BN	0.994	0.181	0.381	0.236	0.783	0.838	0.164	0.626	0.277	0.398	0.668	0.504	4.360
	BN	0.959	0.208	0.131	0.187	0.829	0.840	0.146	0.588	0.366	0.325	0.671	0.477	4.730
	LN	**0.999**	0.165	0.298	0.196	**0.854**	0.755	0.147	0.609	0.313	0.344	0.685	0.488	4.270
	SN	0.996	0.171	**0.502**	0.186	0.802	0.824	0.115	0.521	0.334	0.270	0.822	0.504	4.820
	NBN	0.981	0.260	0.228	0.232	0.733	0.843	0.166	0.556	0.377	0.401	0.773	0.504	3.640
	AnoOnly	0.995	0.152	0.430	0.220	0.825	0.838	0.140	0.519	0.282	0.301	0.740	0.495	5.000
	NoBa (ours)	**0.999**	**0.287**	0.440	**0.250**	0.831	**0.850**	**0.171**	**0.642**	**0.419**	**0.412**	**0.842**	**0.559**	**1.180**
CAD	w/o BN	0.999	0.298	0.648	0.250	**0.825**	**0.883**	**0.177**	**0.851**	0.372	0.394	0.796	**0.590**	3.273
	BN	0.999	0.215	0.599	0.243	0.691	0.841	0.161	0.561	0.310	0.397	0.836	0.532	5.091
	LN	0.999	0.300	**0.695**	0.265	0.737	0.849	0.163	0.602	0.331	0.429	0.843	0.565	2.818
	SN	**1.000**	0.259	0.675	0.252	0.741	0.849	0.160	0.590	0.303	0.427	0.817	0.552	4.091
	NBN	0.975	0.290	0.479	0.258	0.703	0.842	0.157	0.579	0.390	0.464	0.838	0.543	4.273
	AnoOnly	-	-	-	-	-	-	-	-	-	-	-	-	-
	NoBa (ours)	**1.000**	**0.330**	0.688	**0.282**	0.769	0.857	0.176	0.648	**0.402**	**0.476**	**0.855**	0.589	**1.455**