Detection-Explanation-Improvement: A Closed-Loop Framework of Enhancing Anomaly Detection with Counterfactual Explanations

Peng Zhou[1] , **Zhiyong Huang**[1] , **Yuanting Yan**[1]
[1]School of Computer Science and Technology, Anhui University, Hefei, China
doodzhou@ahu.edu.cn, hzy@stu.ahu.edu.cn, ytyan@ahu.edu.cn

Abstract

Many state-of-the-art anomaly detection models operate as black boxes, limiting interpretability and hindering reliable deployment. While recent advances in explainable artificial intelligence have focused on explaining why individual instances are detected as anomalous, comparatively little attention has been paid to how such explanations can be systematically exploited to improve the detectors themselves. To address this gap, we propose EAD-CE (Enhancing Anomaly Detection with Counterfactual Explanations), a model-agnostic, closed-loop framework that tightly integrates detection, explanation, and improvement. Specifically, given a trained anomaly detector and its detected anomalies, EAD-CE generates minimal and semantically meaningful counterfactual explanations that reveal how targeted feature perturbations influence anomaly scores or decisions. Feature importance inferred from these counterfactual explanations is then used to guide a dynamic feature-weight optimization process, enabling detector refinement without modifying its underlying architecture. Extensive experiments on nine real-world datasets and three anomaly detection models demonstrate that EAD-CE accurately identifies anomaly-driving features, substantially enhances interpretability, and consistently improves detection performance (average AUC increases by 6.9%, up to 23%). Implementation details are provided to support reproducibility.

1 Introduction

Anomaly detection aims to identify outliers that deviate significantly from expected patterns of normal data and has been widely applied in domains such as network security, industrial monitoring, and medical diagnosis [Chandola *et al.*, 2009; Pang *et al.*, 2021]. As data dimensionality and application complexity continue to grow, stakeholders increasingly expect anomaly detectors to provide more than accurate alerts [Li *et al.*, 2023b]. Many contemporary detectors remain effectively **black boxes** due to the complexity of their internal mechanisms [Li *et al.*, 2023b]. In practice, users often ask not only (1) **why a particular instance was flagged as anomalous**, but also (2) **how insights derived from anomalies can be leveraged to improve the detector itself**. Consequently, modern anomaly detection systems are expected to go beyond outlier identification by explaining the underlying causes of anomalous decisions, thereby offering reliable and actionable evidence for human decision-makers. Equally important is the question of how such explanations can be systematically exploited to enhance detection performance.

Explainable Artificial Intelligence (XAI) seeks to provide transparent, comprehensible, and human-aligned explanations for model decisions, while satisfying core requirements such as fidelity, usability, and reliability [Gunning and Aha, 2019]. Among XAI techniques, **counterfactual explanations** have recently attracted increasing attention [Wachter *et al.*, 2017]. By identifying minimal and feasible feature modifications that would alter a model's decision under realistic constraints, counterfactuals offer intuitive, actionable explanations that align well with human causal reasoning [Byrne, 2019; Akula *et al.*, 2020; Kanamori *et al.*, 2020; Alfano *et al.*, 2025]. These properties make counterfactual explanations particularly suitable for interpreting anomalous predictions and suggesting potential remediation strategies. Recent studies have explored counterfactual explanations tailored specifically to anomalous instances (e.g., [Xiao *et al.*, 2023; Tang *et al.*, 2023; Han *et al.*, 2023; Zhou *et al.*, 2025]). However, most existing work remains **purely post-hoc**, focusing on explanation generation without feeding the extracted knowledge back into model learning. In particular, the outlier-explanation relationships revealed by counterfactuals are rarely translated into a principled mechanism for detector optimization. As a result, the practical value of interpretability is only partially realized. How to **leverage explanations to systematically improve anomaly detection models** therefore remains an open and important research problem.

Given the diversity of anomaly detection models, we need to find a way to improve model performance that is independent of the underlying structure. Fundamental anomaly detection approaches are categorized as proximity-based methods and projection-based methods [Boukerche *et al.*, 2020]. A shared limitation across many of these techniques is the implicit assumption that all input features contribute equally to the anomaly score. This uniform treatment obscures the

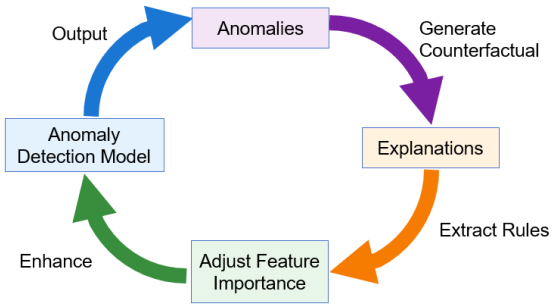

Figure 1: Detection-Explanation-Improvement: A model-agnostic, closed-loop framework for enhancing anomaly detection models with counterfactual explanations.

identification of core anomaly-driving features and makes detectors susceptible to redundant dimensions and noise, particularly in high-dimensional settings, ultimately degrading detection accuracy. Moreover, such methods provide limited interpretability: they rarely indicate which features dominate a decision or how individual features influence the anomaly score. This lack of transparency hampers result verification and undermines trust, significantly constraining real-world deployment. Meanwhile, some approaches attempt to estimate feature importance using marginal statistical properties. For example, CELOF [Chen *et al.*, 2021] assigns feature weights based on information entropy. However, these methods evaluate feature importance independently of the detector's decision function, ignoring feature interactions and the model's internal scoring logic. Consequently, features that appear informative in isolation may be irrelevant or even misleading, once correlations and model-specific behavior are taken into account. Therefore, we propose to **dynamically infer and adjust feature importance using counterfactual explanations**, which are directly grounded in the detector's decision boundary.

Figure 1 presents our proposed closed-loop framework for enhancing anomaly detection via counterfactual explanations. Given a trained detector and a set of identified anomalous instances, we first generate counterfactual explanations that specify the minimal, feasible feature perturbations required to flip a prediction from anomalous to normal. These counterfactuals expose actionable relationships between feature changes and model outputs, offering insight into the features that most strongly influence anomaly decisions and enabling the formulation of causal hypotheses. We then mine the counterfactual set to extract recurring patterns and rules, from which feature importance is quantified at both the global and instance levels. These importance signals are subsequently used to guide targeted detector refinement, such as feature reweighting, constrained retraining with feature-aware regularization, attention mechanisms, or focused data augmentation in influential feature subspaces. This **reverse-optimization** strategy aims to mitigate what we term *undifferentiated processing*, wherein all features are treated uniformly regardless of their relevance. By emphasizing consistently informative features and suppressing noisy or redundant dimensions, the proposed framework enables measurable improvements in both detection accuracy and interpretability, as validated through empirical evaluation.

We summarize our main contributions as follows:

- We propose **one of the first** closed-loop framework that tightly integrates counterfactual explanation, systematic feature importance quantification, and anomaly detection model improvement. By using counterfactual knowledge to directly guide model refinement, the framework addresses key limitations of prior work, including weak coupling between explanation and optimization.
- We develop a model-agnostic, sparsity-constrained counterfactual explanation method tailored to anomaly detection. The method identifies minimal feature edits that reverse anomaly predictions while enforcing feasibility and factuality constraints. We then quantify each feature's contribution to decision reversal by leveraging the perturbation magnitudes required for generating counterfactuals. Finally, we update the feature weights with the historical data to drive an iterative optimization of the anomaly detection model.
- We conduct extensive experiments on multiple public datasets with diverse anomaly detection models. The evaluation includes quantitative analysis of (i) detection performance gains measured by AUC, (ii)the accuracy of identified anomaly-driving features, and (iii) interpretability assessed through qualitative case studies. The results demonstrate the practical effectiveness of the proposed framework in improving both interpretability and detection performance for a wide class of distance-based anomaly detectors.

2 Related Work

Anomaly detection has been studied for decades [Boukerche *et al.*, 2020], yet most advanced methods lack intrinsic interpretability [Li *et al.*, 2023b]. To explain detection results, some approaches rely on pre-model feature engineering or post-hoc explanation tools such as SHAP [Lundberg and Lee, 2017]. While effective for retrospective analysis, these techniques are decoupled from the detector's internal decision process and do not support systematic model refinement. A small number of studies incorporate interpretability directly into the detection model via architectural design, producing explanations alongside anomaly decisions. Representative examples include CDT [Kraiem *et al.*, 2021], which extracts temporal anomaly rules using a composition-based decision tree, and GPs [Berns *et al.*, 2020], which offer probabilistic explanations via Gaussian processes. Although these methods improve transparency, they often sacrifice expressive power or generality and do not scale well to diverse, high-dimensional anomaly-detection scenarios.

Counterfactual explanations provide intuitive, actionable insights into model decisions [Verma *et al.*, 2024] and have recently attracted widespread attention, including Transformer-VAE approaches for mixed-type features [Panagiotou *et al.*, 2024] and mixed-integer programming formulations with optimality guarantees [Tsiourvas *et al.*, 2024].

Directly adapting these techniques to anomaly detection remains challenging because of asymmetric decision boundaries, limited supervision, and domain-specific feasibility constraints. Recently, studies have begun exploring counterfactual explanations for anomalous instances [Ernst, 2024; Xiao *et al.*, 2023; Tang *et al.*, 2023; Han *et al.*, 2023; Zhou *et al.*, 2025]; however, these approaches are still post-hoc and do not use counterfactual information to guide detector optimization, leaving explanation–anomaly relationships untapped for model improvement.

3 Preliminaries

3.1 Anomaly Detection

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ denote a dataset of n samples, where each $x_i \in \mathbb{R}^d$ is a d-dimensional instance. Assume $\mathcal{X}$ is partitioned into normal and anomalous subsets: $\mathcal{X}_n$ and $\mathcal{X}_a$, respectively. So, $\mathcal{X} = \mathcal{X}_n \cup \mathcal{X}_a$ and $\mathcal{X}_n \cap \mathcal{X}_a = \emptyset$. The goal of anomaly detection is to learn a decision function $h(x_i) \to \{0, 1\}$, where $h(x_i) = 0$ indicates that the sample x_i belongs to the normal class and $h(x) = 1$ for anomalous. More commonly, we learn an anomaly scoring function $f(x_i) \to \mathbb{R}$, where larger values indicate greater anomaly. With a predefined threshold τ, the model identifies a sample as an anomaly if $f(x_i) > \tau$.

The weight of features quantifies the contribution of each dimension to the decision on the anomaly. These weights play three practical roles: (1) Dimensionality control: by upweighting high weights and removing or down-weighting low ones, the effective input dimensionality and model complexity are reduced, which helps mitigate overfitting and lower computational cost in high-dimensional settings; (2) Targeted modeling: importance signals can steer the detector away from uniformly processing all features toward focusing on core feature patterns that drive anomalies, often reducing false positives and false negatives; and (3) Interpretability: numeric weights provide a compact, quantitative basis for explaining anomaly decisions. Dynamically adjusting feature weights creates a principled bridge between explanation and model optimization, enabling simultaneous gains in detection performance and transparency.

To measure whether the anomaly detection model has improved, we define the unsupervised metric **Separability Score** as follows:

$$SS = \frac{\mu_{\mathrm{a}} - \mu_{\mathrm{n}}}{\sigma_{\mathrm{n}} + \epsilon_0}, \tag{1}$$

where μ_{a} and μ_{n} are the mean anomaly scores of anomaly samples and normal samples, respectively. σ_{n} is the score standard deviation of normal samples, while ϵ_0 is a fixed positive smoothing term to avoid division by zero. The Separability Score quantifies the difference between the score distributions of anomalous and normal samples, reflecting the clarity of the model's decision boundary. A larger Separability Score indicates a clearer class separation and improved anomaly detection accuracy.

3.2 Counterfactual Explanation

Counterfactual explanation aims to answer the question of "which conditions need to be changed to reverse the decision result". In the scenario of anomaly detection, counterfactual explanations establish the causal relationship between feature adjustments and decision results by constructing minimally modified samples as:

$$\arg\min_{x_{cf}} \left\{ \ell\left(h(x_{cf}), y^0\right) + \mathrm{d}(x_{cf}, x_a) \right\}, \tag{2}$$

where x_a is the target sample (anomaly) to be explained, x_{cf} is the generated counterfactual. $h()$ is the anomaly detector, and $h(x_a) = y^1$. $\ell()$ is the loss function between the desired target y^0 and the current prediction $h(x_{cf})$. The generated counterfactual explanation x_{cf} must satisfy the following three properties:

1. **Decision Reversibility**: $h(x_{cf}) = y^0$ (anomaly $\to$ normal);
2. **Minimal Modification**: $d(x_a, x_{cf}) = ||x_{cf} - x_a||_p \leq \epsilon$ (where $p \in \{0, 1, 2\}$ denotes the regularization term and ϵ is the preset maximum modification threshold);
3. **Reasonableness**: x_{cf} must conform to the practical meaning of data distribution (e.g., feature values fall within reasonable intervals).

By comparing the feature differences between x_a and its counterfactual x_{cf}, we can directly identify the core features that drive the anomaly decision, thereby providing a causal-level basis for feature importance quantification. Unlike traditional statistical measures (e.g., variance- or distance-based criteria), which are independent of the detector's decision process, this comparison links feature importance explicitly to the model's decision boundary. As a result, the quantified importance reflects not merely statistical salience but the actual features whose minimal changes are sufficient to reverse the anomaly judgment.

4 Methodology

4.1 Overview of the EAD-CE Framework

Algorithm 1 summarizes the complete execution pipeline of the proposed EAD-CE framework. EAD-CE is organized as a three-stage progressive closed loop: (1) sparse counterfactual generation for detected anomalies, (2) feature importance quantification based on counterfactual feature perturbations, and (3) iterative model optimization and validation driven by the derived importance weights. Through repeated iterations of these stages, the framework continuously refines both detector performance and interpretability.

4.2 Detailed Methodology

Step1: Sparse Counterfactual Generation

Input data in anomaly detection tasks are typically high-dimensional. Without sparsity constraints on feature modifications in counterfactual samples, the resulting feature importance signals can become diluted, thereby weakening the interpretability of the counterfactual explanations. To address this issue, we enforce sparsity during counterfactual generation, requiring that only a small number of key features be modified to reverse the detector's decision from "anomaly" to "normal." This constraint ensures that the identified feature changes are concise, focused, and causally meaningful.

Algorithm 1 Overall Procedure of EAD-CE

Input: $\mathcal{X}$:dataset, f:anomaly detection model, θ:tolerance threshold, $T_{\max}$:maximum number of iterations
Output: $\mathcal{X}_a$:outliers, w:feature importance

Initialize $t = 0$, $w = \mathbf{1}$;
do
  $\mathcal{X}_a \leftarrow f(\mathcal{X}, w)$;
  Calculate Separability Score SS_t by Eq. 1;
  $t \leftarrow t + 1$;
/* 1. Spare Counterfactual Generation */
  $\mathcal{X}_{cf} \leftarrow \text{SCG}(\mathcal{X}_a)$;
/* 2. Feature Importance Quantification */
  $w \leftarrow \text{FIQ}(\mathcal{X}_a, \mathcal{X}_{cf})$;
/* 3. Iterative Optimization and Validation */
  $SS_{(t)}, \mathcal{X}_a, w \leftarrow \text{IQV}(w, f)$;
  If $(SS_{(t)} - SS_{(t-1)} \leq \theta)$ **break;**
while $t < T_{\max}$
return $\mathcal{X}_a$**,** w

To balance the dual goals of enforcing sparsity and minimizing adjustment magnitude, we formulate the objective function for counterfactual generation as:

$$\begin{aligned} \arg\min_{x_{cf}} \quad & ||x_{cf} - x_a||_1 + ||x_a - x_{nn}||_2 + ||x_{cf} - x_a||_0 \\ \text{s.t.} \quad & f(x_a + \alpha \cdot \boldsymbol{p}_{\text{masked}}) \leq \tau, \end{aligned} \tag{3}$$

where the first term is the L_1 distance between counterfactual sample x_{cf} and anomalous sample x_a, constraining the overall magnitude of feature adjustments. The second term is the L_2 distance between x_{cf} and nearest normal sample x_{nn}, ensuring the counterfactual conforms to normal data distribution. The third term is the L_0 norm, counting modified feature dimensions to enforce sparsity and ensure only a few features are altered. α is the step size parameter in $[0, 1]$, controlling the degree of feature adjustment. $\boldsymbol{p}_{\text{masked}}$ is the masked perturbation direction vector, non-zero only in modified feature dimensions and zero elsewhere. τ is the anomaly score threshold for normal samples.

Based on the loss function defined above, the generation of sparse counterfactual explanations is described in Algorithm 2. The implementation consists of three core steps: nearest-neighbor guidance, greedy feature subspace selection, and bisection-based numerical optimization. The detailed procedure is as follows:

Nearest Neighbor Guidance. The search direction for counterfactual samples in high-dimensional space is highly uncertain. Blind random search often leads to low computational efficiency and counterfactuals with limited plausibility. To address this issue, we adopt the manifold hypothesis—namely, that valid counterfactual samples should lie on or near the distribution manifold of normal data—and use a nearest-neighbor strategy to provide high-confidence guidance for perturbation directions.

Specifically, the nearest normal neighbor sample $\boldsymbol{x}_{nn}$ of $\boldsymbol{x_a}$ is found as:

$$\boldsymbol{x}_{nn} = \arg\min_{\boldsymbol{z} \in \mathcal{X}_n} \|\boldsymbol{x_a} - \boldsymbol{z}\|_2. \tag{4}$$

Algorithm 2 SCG (Sparse Counterfactual Generation)

Input: $\mathcal{X}_a$:outliers
Output: $\mathcal{X}_{cf}$:counterfactual explanations

for $\boldsymbol{x_a} \in \mathcal{X}_a$ **do**
  Find the nearest normal sample x_{nn} by Eq. 4;
  Satisfy the L_0 sparsity constraint of counterfactuals via a greedy strategy;
  Generate counterfactual samples x_{cf} satisfying Eq. 5 via the bisection method;
end for
return $\mathcal{X}_{cf}$

The initial perturbation direction vector is defined as the difference between the nearest-neighbor sample and the original sample, i.e., $\boldsymbol{p} = \boldsymbol{x}_{\text{nn}} - \boldsymbol{x_a}$. This direction vector naturally points from the anomaly region toward the normal region, thereby providing a clear and well-motivated initialization for the subsequent sparse search.

Greedy Feature Subspace Selection. The initial perturbation direction vector $\boldsymbol{p}$ encodes adjustments across all feature dimensions; applying it directly typically violates the L_0 sparsity constraint. Consequently, we adopt a greedy strategy to identify an active feature subspace, reducing the original high-dimensional optimization to a tractable low-dimensional problem. The specific procedure is as follows:

1. Calculate the absolute values of each component of the initial perturbation direction vector $\boldsymbol{p}$, sort the feature dimensions in descending order according to the absolute values, and obtain the feature index sequence $\pi = [\pi_1, \pi_2, ..., \pi_d]$, where π_1 corresponds to the feature dimension with the largest adjustment magnitude.
2. Define the active feature set of the k-th iteration as $S_k = \{\pi_1, \pi_2, ..., \pi_k\}$, where $k \in \{1, 2, ..., d\}$ and d is the total number of feature dimensions of the sample.
3. Construct the masked perturbation direction vector $\boldsymbol{p}_{\text{masked}}$, which only retains the components corresponding to the active feature set S_k and sets the components of other inactive dimensions to 0.
4. Iteratively increase the value of k until a sparse counterfactual sample satisfying the constraint conditions is found.

We employ a greedy strategy that prioritizes adjustments to the most influential features, modifying only a small subset of active features in each iteration. This approach enforces the L_0 sparsity constraint while substantially reducing the search complexity in high-dimensional spaces.

Numerical Optimization via Bisection. After selecting the active feature subspace and the perturbation direction, the counterfactual generation problem reduces to finding the minimal step size α that drives the anomaly score below the normal threshold as follows:

$$\min \alpha \quad \text{s.t.} \quad f(\boldsymbol{x_a} + \alpha \cdot \boldsymbol{p}_{\text{masked}}) \leq \tau. \tag{5}$$

Because the framework must apply to general anomaly detection models without analytic gradients, and the score transition from "anomaly" to "normal" is monotonic, we employ

Algorithm 3 FIQ (Feature Importance Quantification)

Input: $\mathcal{X}_a$:outliers, $\mathcal{X}_{cf}$:counterfactual explanations
Output: $\boldsymbol{w}$:feature importance vector

1: **for** $i \in 1, 2, ..., d$ **do**
2: Calculate the average perturbation of feature f_i by Eq. 6;
3: Calculate the weight w^i of feature f_i by Eq. 7;
4: **end for**
5: **return** $\boldsymbol{w}$

a **bisection search** to determine the optimal step size α. Bisection requires only evaluations of the anomaly score and efficiently locates the smallest α that satisfies Eq.5 up to a prescribed tolerance. Details can be seen in the **Appendix**.

Step2: Feature Importance Quantification

Updating feature weights provides the foundation for subsequent iterative optimization. The core idea is to quantify each feature's contribution to decision reversal indirectly by using the perturbation magnitudes required for counterfactual samples. If inducing a transition from "anomaly" to "normal" requires a larger perturbation for a given feature, that feature plays a more critical role in the anomaly decision and should therefore receive a higher weight. Given perturbation vectors for multiple anomalous samples, the feature-weight vector $\boldsymbol{w} \in \mathbb{R}^d$ is obtained by statistically aggregating the per-feature perturbation magnitudes. The detailed procedure for feature-importance quantification is presented in Algorithm 3.

Specifically, the contribution of the i-th feature f_i to decision reversal is characterized by the average perturbation magnitude across valid counterfactual samples. This metric effectively excludes the influence of invalid samples while capturing the average adjustment intensity required for the feature to induce a decision reversal. The corresponding calculation is given by:

$$\bar{\Delta}_i = \frac{1}{M} \sum_{m=1}^{M} \Delta_{m,i}, \tag{6}$$

where $\Delta_{m,i}$ represents the perturbation value of feature f_i in the m-th suspected anomaly sample, and $M = |\mathcal{X}_{cf}|$ is the total number of valid counterfactual samples.

Considering that different features vary in dimensions and value ranges, directly using the average perturbation magnitude to calculate weights will lead to unbalanced weight distribution. Therefore, we normalize $\bar{\Delta}_i$ to derive each feature weight as:

$$w^i = \frac{\bar{\Delta}_i}{\sum_{j=1}^{d} \bar{\Delta}_j + \epsilon_0} \cdot \boldsymbol{p}, \tag{7}$$

where $\epsilon_0 = 10^{-12}$ is a smoothing term to prevent numerical instability caused by a zero denominator. Multiplying by $\boldsymbol{p}$ is to make the mean value of the initial weights approach 1, facilitating weight adjustment and interpretation in subsequent iterations.

For the characteristics of high-dimensional data, features that contribute nothing or extremely little to decision reversal will have their average perturbation magnitudes $\bar{\Delta}_i$ naturally approach 0, and the corresponding initial weights will be directly reduced to 0. In this way, automatic sparsification of feature weights in the high-dimensional space is achieved.

Algorithm 4 IOV (Iterative Optimization and Validation)

Input: $\mathcal{X}$:dataset, $\boldsymbol{w}$:feature importance vector, f:anomaly detection model
Output: $\mathcal{X}_a$:outliers, SS:separability score , $\boldsymbol{w'}$:updated feature importance vector

1: The iterative weight $\boldsymbol{w'}$ of the current round is updated by Eq. 8;
2: $\mathcal{X}_a \leftarrow f(\boldsymbol{w'}, \mathcal{X})$;
3: Calculate separability score SS by Eq. 1;
4: **return** SS , $\mathcal{X}_a$, $\boldsymbol{w'}$

Step3: Iterative Optimization and Validation

During optimization of the feature-importance-based anomaly detection model, the historical weights capture optimization information accumulated across iterations and thus mitigate the randomness introduced by counterfactual perturbations in any single round. Accordingly, the feature importance obtained in the t-th iteration is combined with the historical weights as follows:

$$\boldsymbol{w}_{(t)} = \lambda \cdot \boldsymbol{w}_{(t-1)} + (1 - \lambda) \cdot \boldsymbol{w}, \tag{8}$$

where λ is designed to balance the stability of the historical weight $\boldsymbol{w}_{(t-1)}$ and the adaptability of the current weight $\boldsymbol{w}$. We set $\lambda = 0.4$ as an empirical value. Subsequently, the updated iterative weight $\boldsymbol{w}_{(t)}$ is used as the core guidance to perform targeted optimization of the base anomaly detection model through weighted feature inputs. This mechanism encourages the model to focus more on the key feature dimensions relevant to anomaly judgment, ensuring that the optimization direction remains highly consistent with the quantified feature importance. The specific process of model iterative optimization based on feature importance quantification is shown in Algorithm 4.

Meanwhile, we compute the unsupervised Separability Score to quantitatively evaluate the performance of the optimized model. Dual convergence criteria: indicator stability and a maximum iteration limit, are employed. Specifically, if the variation in the Separability Score between two consecutive iterations satisfies $\Delta SS = SS_{(t)} - SS_{(t-1)} \leq \theta$ (where θ denotes a predefined tolerance threshold). The model is considered to have reached an optimal state, and the iteration is terminated since further updates are unlikely to yield significant improvement. Additionally, if the number of iterations reaches the preset maximum, the optimization process is forcibly stopped to prevent infinite looping.

Theorem 1. *The Separability Score converges under continuous optimization and iterative updates of the feature weights.*

The proof is provided in the **Appendix**.

Time Complexity Analysis

The overall time complexity of the proposed framework is $\mathcal{O}(N_a N_n d + d \log d + d + N)$, and the detailed complexity analysis is provided in the **Appendix.**

Dataset	Instances	Attributes	Outliers(Ratio)
Yeast	1483	8	35(2%)
Wine-Red	1599	11	10(0.6%)
Wine-White	4898	11	20(0.4%)
Http	567479	3	2211 (0.4%)
Mnist	7603	100	700 (9.2%)
Musk	3062	166	97 (3.2%)
Arrhythmia	452	274	66 (15%)
Glass	214	9	9 (4.21%)
Cardio	1831	21	176(9.61%)

Table 1: Datasets

5 Experiments and Results

This paper presents comprehensive experiments to evaluate the effectiveness of the proposed method by addressing the following research questions:

- **Q1:** Can EAD-CE consistently improve the detection performance of diverse baseline anomaly-detection models and produce stable gains across multiple datasets?
- **Q2:** Is the improvement in anomaly-detection performance positively correlated with changes in feature importance?
- **Q3:** Can EAD-CE provide intuitive and reliable interpretability for the decision outcomes of anomaly-detection models?

5.1 Experiment Setups

Datasets. We conducted experiments on nine real-world datasets[1] spanning multiple domains, including Bioinformatics, Food Science, Cybersecurity, Computer Vision, Cheminformatics, Healthcare, and Materials Science, as summarized in 1.

Baselines. This paper selects three representative anomaly detection models as baselines: the density-based Local Outlier Factor (LOF [Breunig *et al.*, 2000]) algorithm, the Empirical Cumulative Distribution-based Outlier Detection (ECOD [Li *et al.*, 2023a]) algorithm, and the Granular-ball Computing-based Mean-shift Outlier Detection (GBMOD [Cheng *et al.*, 2024]) method.

Evaluation metrics. We select Area Under the Curve (AUC) and the Separability Score (Eq.1) as the primary evaluation metrics to comprehensively assess anomaly detection performance.

5.2 Analysis of the Experimental Results

RQ1: Performance Gains Across Different Datasets and Anomaly Detection Models

Figure 2 compares AUC performance between the original baseline models (Base, blue dashed line) and models optimized by our framework (CE-Enhanced, red solid line) across three baseline types (ECOD, LOF, GBMOD) and eight datasets. Across every model–dataset pair, our new framework yields substantially higher AUC values, demonstrating consistent performance gains. For example, ECOD on dataset Musk increases from 0.94 to 1.00, LOF on Cardio from 0.78 to 0.83, and GBMOD on Arrhythmia from 0.75 to 0.82. These results indicate that EAD-CE framework reliably improves detection performance regardless of the underlying algorithmic paradigm (density-, probability-, or kNN-based) across different datasets, underscoring the broad applicability of our framework.

RQ2: Relationship Between Feature Importance Adjustment and Detection Performance

To verify the relationship between feature importance adjustment and detection performance, we conduct experiments on the HTTP network security detection dataset, as shown in Table 1. The core task is to identify outlier network attack traffic, which includes three key traffic features: src_bytes (the number of bytes transmitted from the source end to the destination end), dst_bytes (the number of bytes transmitted from the destination end to the source end), and duration (network connection duration).

Figure 3 demonstrates the cascading impact of counterfactual-derived feature importance on iterative model optimization and performance. The left panel shows feature weights computed from counterfactuals: core feature *src_bytes* (identified as decision-relevant) rises from 1.0 to over 2.5, while noncritical features (*duration*, *dst_bytes*) decline, indicating increased emphasis on high-impact features. The middle panel shows score separation increasing in parallel (from 3.5907 to 4.1419 in the first two iterations and stabilizing around 4.37), reflecting improved distinction between anomalous and normal samples. The right panel confirms that these changes translate into better detection: AUC climbs from 0.9641 to 0.995 in the early iterations, mirroring feature weighting and separability. Overall, these results indicate that counterfactual-driven feature importance learning produces consistent, positive optimization of separability score and overall detection performance.

RQ3: Interpretability Analysis On Real-world Dataset

We use the low-dimensional HTTP dataset to intuitively visualize optimization mechanisms. This dual analysis of counterfactuals and feature importance demonstrates the interpretability of the framework. For counterfactuals, we select samples the model labels as attacks and generate counterfactual variants, observing prediction changes revealing each feature's causal influence on attack classification. For feature-importance quantification, we compare the model's feature-weight distributions before and after optimization and present differences highlighting shifts in core feature contributions. These complementary analyses connect causal evidence from counterfactuals with quantitative changes in feature importance, demonstrating the framework's interpretability in the network-security detection settings.

Table 2 shows that the primary cause for the model labeling traffic as an attack (Outlier) in the original samples is a deviation of *src_bytes* from its normal range. *src_bytes* denotes the number of bytes sent from the source (e.g., an attack-initiating node) to the destination and a key descriptor of traf-

[1]http://archive.ics.uci.edu/

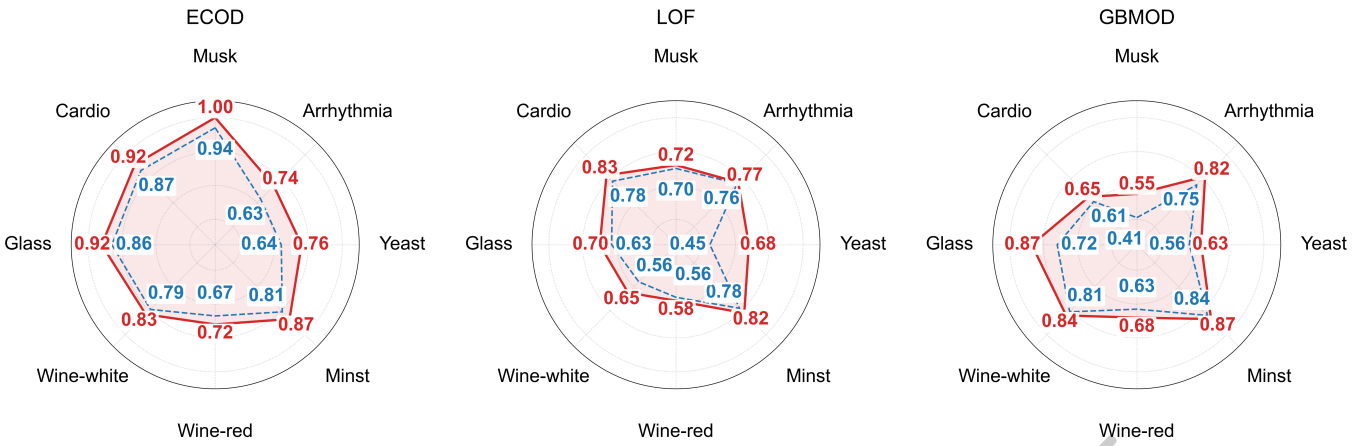

Figure 2: Performance across different datasets and anomaly detection models between the original baseline models (Base, blue dashed line) and models optimized by our framework (CE-Enhanced, red solid line).

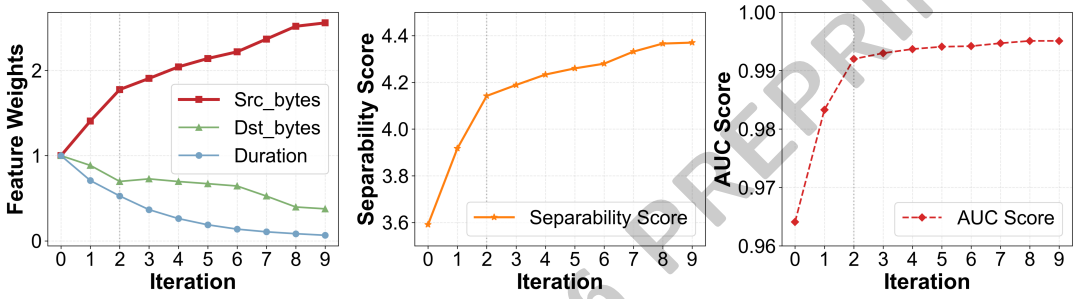

Figure 3: The left panel tracks changes in feature importance, quantified via counterfactuals, across the iterative process; the middle panel shows the corresponding evolution of separability score throughout the optimization; and the right panel reports the AUC trajectory over iterations, illustrating how detection performance improves as the model is iteratively refined.

ID	Type	src_bytes	dst_bytes	duration	Prediction
316098	Original	-18.0693	-7.435	-0.0732	Outlier
	Counterfactual	-0.3347	-7.435	-0.0732	Normal
201669	Original	-18.0693	-7.435	8.7608	Outlier
	Counterfactual	-0.1437	-7.435	8.7608	Normal
433927	Original	-8.6192	-1.9125	-0.0732	Outlier
	Counterfactual	-0.7867	-1.0034	-0.0732	Normal

Table 2: The ECOD anomaly detector flagged three outliers, and our framework generated three corresponding counterfactual explanations (the features changed to flip the prediction are highlighted in green).

fic behavior. Legitimate traffic exhibits stable *src_bytes* values, whereas attack traffic produces outlier fluctuations (sudden surges or sharp drops). These three counterfactual explanations adjust *src_bytes* back into a reasonable range that change the model's prediction to normal. This result establishes a direct, business-relevant causal link between outlier *src_bytes* and attack classification, addressing the shortcoming of traditional anomaly detectors that provide labels without explaining the operational rationale behind them.

Combined with the feature-importance visualization in Figure 3, analysis reveals clear post-optimization attention reallocation: *src_bytes* importance increases from 1.0 to 2.56, strengthening its role as the primary attack-classification driver; *dst_bytes* remains a secondary, supportive signal; and *duration* (connection time), having weak attack correlation, declines in influence with no notable effect on performance. This pattern aligns with counterfactual evidence: minimal adjustments to *src_bytes* can flip predictions. Thus, our interpretability analysis confirms outlier *src_bytes* drive detection, and the optimized model concentrates on network-security features consistent with real attack–defense dynamics.

6 Conclusions

This paper tackles two user concerns: why samples are flagged as anomalous and how such explanations can improve the anomaly detection model. We propose a model-agnostic, counterfactual-enhanced anomaly detection framework that closes the loop between detection, explanation, and improvement. It uses counterfactuals as decision-aligned references and extracts feature importance from model-response differences, guiding detectors to emphasize core anomaly signals while suppressing noise. Extensive experiments show consistent improvements in accuracy and interpretability across multiple models and datasets. Future work will extend it to multimodal settings and deep-learning-based detectors.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under grants 62376001, the Natural Science Foundation of Anhui Province of China under grants 2308085MF215, and the Open Project of the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China) grant BigKEOpen2025-05.

Data availability

The datasets in this article are all publicly available datasets. The source code for EAD-CE is available at https://github.com/doodzhou/XAI.

References

[Akula *et al.*, 2020] Arjun Akula, Shuai Wang, and Song-Chun Zhu. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 2594–2601, 2020.

[Alfano *et al.*, 2025] Gianvincenzo Alfano, Adam Gould, Francesco Leofante, Antonio Rago, and Francesca Toni. Counterfactual explanations under model multiplicity and their use in computational argumentation. In *Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 4321–4329, 2025.

[Berns *et al.*, 2020] Fabian Berns, Markus Lange-Hegermann, and Christian Beecks. Towards gaussian processes for automatic and interpretable anomaly detection in industry 4.0. In *In Proceedings of the International Conference on Innovative Intelligent Industrial Production and Logistic*, pages 87–92, 2020.

[Boukerche *et al.*, 2020] Azzedine Boukerche, Lining Zheng, and Omar Alfandi. Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, 53(3):1–37, 2020.

[Breunig *et al.*, 2000] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104, New York, NY, USA, 2000. ACM.

[Byrne, 2019] Ruth MJ Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 6276–6282. California, CA, 2019.

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

[Chen *et al.*, 2021] Liang Chen, Wei Wang, and Yun Yang. Celof: Effective and fast memory efficient local outlier detection in high-dimensional data streams. *Applied Soft Computing*, 102:107079, 2021.

[Cheng *et al.*, 2024] Shitong Cheng, Xinyu Su, Baiyang Chen, Hongmei Chen, Dezhong Peng, and Zhong Yuan. Gbmod: A granular-ball mean-shift outlier detector. *Pattern Recognition*, page 111115, 2024.

[Ernst, 2024] R. Ernst. Counterfactual generating variational autoencoder for anomaly detection. In *xAI 2024 Late-breaking Work, Demos and Doctoral Consortium*, volume 3793 of *CEUR Workshop Proceedings*, pages 353–360. CEUR-WS, 2024.

[Gunning and Aha, 2019] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

[Han *et al.*, 2023] Xiao Han, Lu Zhang, Yongkai Wu, and Shuhan Yuan. Achieving counterfactual fairness for anomaly detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 55–66. Springer, 2023.

[Kanamori *et al.*, 2020] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2855–2862, 2020.

[Kraiem *et al.*, 2021] Inès Ben Kraiem, Faiza Ghozzi, André Péninou, Geoffrey Roman-Jimenez, and Olivier Teste. Human-interpretable rules for anomaly detection in time-series. In *International Conference on Extending Database Technology*, pages 457–462. OpenProceedings. org, 2021.

[Li *et al.*, 2023a] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H. Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2023.

[Li *et al.*, 2023b] Zhong Li, Yuxuan Zhu, and Matthijs Van Leeuwen. A survey on explainable anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 18(1):1–54, 2023.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems*, volume 30, pages 1–10, 2017.

[Panagiotou *et al.*, 2024] Emmanouil Panagiotou, Manuel Heurich, Tim Landgraf, and Eirini Ntoutsi. Tabcf: Counterfactual explanations for tabular data using a transformer-based vae. In *5th ACM International Conference on AI in Finance*, pages 274–282, 2024.

[Pang *et al.*, 2021] 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.

[Tang *et al.*, 2023] Wenbing Tang, Jing Liu, Yuan Zhou, and Zuohua Ding. Causality-guided counterfactual debiasing for anomaly detection of cyber-physical systems. *IEEE Transactions on Industrial Informatics*, 20(3):4582–4593, 2023.

[Tsiourvas *et al.*, 2024] Asterios Tsiourvas, Wei Sun, and Georgia Perakis. Manifold-aligned counterfactual explanations for neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 3763–3771. PMLR, 2024.

[Verma *et al.*, 2024] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12):1–42, 2024.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[Xiao *et al.*, 2023] Chunjing Xiao, Xovee Xu, Yue Lei, Kunpeng Zhang, Siyuan Liu, and Fan Zhou. Counterfactual graph learning for anomaly detection on attributed networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10540–10553, 2023.

[Zhou *et al.*, 2025] Peng Zhou, Qihui Tong, Shiji Chen, Yunyun Zhang, and Xindong Wu. Eace: Explain anomaly via counterfactual explanations. *Pattern Recognition*, 164:111532, 2025.