A Novel Sparse Active Online Learning Framework for Fast and Accurate Streaming Anomaly Detection Over Data Streams

Zhong Chen[1], **Yi He**[2*], **Di Wu**[3], **Chen Zhao**[4] and **Meikang Qiu**[5]
[1]School of Computing, Southern Illinois University Carbondale
[2]Department of Data Science, William & Mary
[3]College of Computer and Information Science, Southwest University
[4]School of Engineering and Computer Science, Baylor University
[5]School of Computer and Cyber Sciences, Augusta University
zhong.chen@cs.siu.edu, yihe@wm.edu, wudi.cigit@gmail.com, chen_zhao@baylor.edu

Abstract

Online Anomaly Detection (OAD) is critical for identifying rare yet important data points in large, dynamic, and complex data streams. A key challenge lies in achieving accurate and consistent detection of anomalies while maintaining computational and memory efficiency. Conventional OAD approaches, which depend on distributional deviations and static thresholds, struggle with model update delays and catastrophic forgetting, leading to missed detections and high false positive rates. To address these limitations, we propose a novel Streaming Anomaly Detection (SAD) method, grounded in a sparse active online learning framework. Our approach uniquely integrates $\ell_{1,2}$-norm sparse online learning with CUR decomposition-based active learning, enabling simultaneous fast feature selection and dynamic instance selection. The efficient CUR decomposition further supports real-time residual analysis for anomaly scoring, eliminating the need for manual threshold settings about temporal data distributions. Extensive experiments on diverse streaming datasets demonstrate SAD's superiority, achieving a 14.06% reduction in detection error rates compared to five state-of-the-art competitors.

1 Introduction

Online Anomaly Detection (OAD) is a critical technique for identifying deviations from established behavioral patterns in real-time data streams, enabling timely responses to irregularities [O'Reilly *et al.*, 2014; Salehi *et al.*, 2016; Guha *et al.*, 2016; Manzoor *et al.*, 2018; Lai *et al.*, 2020; Chen *et al.*, 2023; Chen *et al.*, 2024b]. It has broad applications in various domains, including intrusion and fraud detection [Muhammad *et al.*, 2020], financial and web services [Pazarbasioglu *et al.*, 2020], network security management [Du, 2022], and public health or military surveillance [Burkle, 2020]. Early anomaly detection is crucial for minimizing operational disruptions, enhancing troubleshooting efficiency, and facilitating rapid corrective actions, thereby safeguarding system integrity and ensuring continuity in dynamic environments.

Subspace learning-based anomaly detection methods have recently gained significant attention by identifying orthogonal subspaces that capture latent patterns in data from unknown distributions [Li *et al.*, 2011; He *et al.*, 2017; Chen *et al.*, 2021; Zhang and Zhao, 2022]. These methods reduce the dimensionality of input data into low-dimensional subspaces, enhancing structural representation while mitigating noise. Prominent approaches such as online oversampling Principal Component Analysis (osPCA) [Lee *et al.*, 2012] and sketch-OAD [Huang and Kasiviswanathan, 2015] leverage matrix sketching, which maintains compact sets of orthogonal vectors to approximate data streams efficiently. Complementing these, non-parametric methods such as Very Fast Decision Tree (VFDT) [Tan *et al.*, 2011] and Hoeffding Anytime Tree (HATT) [Bifet *et al.*, 2017] address concept drift challenges, offering interpretable solutions for evolving data streams. Recent advancements integrate online deep learning frameworks [Sahoo *et al.*, 2018; Lian *et al.*, 2022], yielding robust models like the adaptive deep log anomaly detector (Ada) [Yuan *et al.*, 2020], deep autoencoding Gaussian mixture model (DAGMM) [Zong *et al.*, 2018], and an adaptive model pooling approach for online deep anomaly detection (ARCUS) [Yoon *et al.*, 2022], which adapt dynamically to complex, evolving data streams. These innovations collectively enhance detection accuracy, scalability, and adaptability in real-time anomaly identification.

Residual analysis has emerged as a promising approach for effective anomaly detection by examining discrepancies between observed data and their reconstructed estimates, where anomalies are identified through abnormally large residual errors resulting from deviations from dominant patterns [Tong and Lin, 2011; Peng *et al.*, 2018; Ding *et al.*, 2019]. A commonly employed strategy involves reconstructing data using representative instances, but challenges arise due to noisy or irrelevant attributes, necessitating the simultaneous selection of structurally relevant attributes and informative instances for accurate reconstruction. To address this, recent advancements pivot from traditional factorization methods

*Corresponding author: Dr. Yi He (yihe@wm.edu)

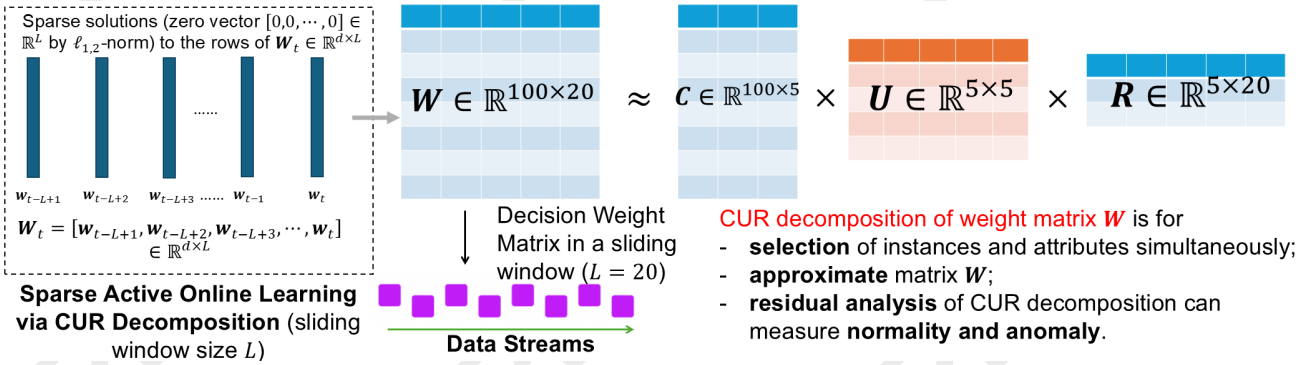


Figure 1: CUR decomposition identifies a subset of representative instances (matrix $\boldsymbol{C}$) and discriminative attributes (matrix $\boldsymbol{R}$) from the decision weight matrix $\boldsymbol{W}$ in a sliding window, constructing a low-rank approximation that minimizes reconstruction error while preserving interpretability. The components are formed by small subsets of actual columns and rows of $\boldsymbol{W}$. By dynamically selecting c instances and r attributes, CUR captures structural patterns in streaming data, where anomalies often deviate significantly from these representative subsets.

to online CUR decomposition, which offers enhanced interpretability by dynamically maintaining low-rank approximations through curated subsets of instances and attributes over time. This method not only improves noise resilience but also provides a transparent framework for temporal instance and attribute selection, enabling real-time anomaly detection in evolving data streams. By integrating residual analysis with interpretable online decomposition, the approach systematically isolates anomalies while preserving the integrity of underlying data structures, thereby enhancing detection accuracy and adaptability in dynamic environments.

Despite advancements, OAD remains challenging due to several persistent issues. Firstly, the context-dependent nature of anomalies, where definitions vary across applications, precludes one-size-fits-all solutions, and while density-based anomaly scoring [O'Reilly *et al.*, 2014] is common, it struggles in high-dimensional, open feature spaces where distance metrics lose discriminative power [He *et al.*, 2019; Wu *et al.*, 2023; Schreckenberger *et al.*, 2023; He *et al.*, 2023; Chen *et al.*, 2024a]. Secondly, catastrophic forgetting in streaming models necessitates rapid anomaly identification and model updates before subsequent data arrives. Thirdly, severe class imbalance arises as normal data overwhelms anomalies, skewing detection accuracy. Lastly, balancing detection efficiency with minimizing false positives is critical to avoid alarm fatigue while ensuring timely anomaly capture. To tackle these issues, we propose integrating $\ell_{1,2}$-norm-based sparse online learning with CUR decomposition-based online active learning, enabling concurrent streaming feature selection and instance curation. This hybrid approach mitigates dimensionality challenges and catastrophic forgetting while enhancing interpretability. By dynamically prioritizing sparse, informative features and representative instances, the framework supports robust residual analysis for precise OAD, ensuring adaptability and scalability in evolving data streams.

1.1 Motivation

Effective OAD hinges on the dynamic selection of representative features and informative instances to mitigate catastrophic forgetting, ensuring incremental model updates retain critical patterns. While existing subspace selection methods [Li *et al.*, 2011; He *et al.*, 2017] isolate anomalies through preprocessing steps, their decoupled optimization from detection pipelines risks suboptimal performance. To overcome this, we propose a unified framework integrating $\ell_{1,2}$-norm-based sparse online learning for feature sparsity and CUR decomposition-based active learning for instance selection, synergistically optimizing both processes in streaming settings. This approach leverages residual analysis from online CUR decomposition to quantify instance normality, where anomalies exhibit disproportionately large reconstruction errors compared to normal data (Figure 1). By jointly prioritizing sparse, discriminative features and representative instances, the method enhances adaptability to evolving data streams while maintaining interpretability, addressing dimensionality challenges and alleviating catastrophic forgetting.

1.2 Contribution

In this paper, we propose an effective streaming anomaly detection (SAD) method from a sparse active online learning perspective. Technically, our main idea is to seamlessly integrate the $\ell_{1,2}$-norm based sparse online learning and the CUR decomposition based online active learning, leading to an effective residual analysis for SAD. On one hand, SAD is able to provide real-time feature interpretations by leveraging the benefits of the $\ell_{1,2}$-norm penalty. On the other hand, SAD can automatically detect anomalies in streaming data based on the CUR decomposition theory, which does not require hand-set thresholds and makes no assumption on the data distribution. The main contributions of our work are as follows:

- We introduce a new perspective that optimizes feature and instance selection and anomaly detection as a whole instead of treating them as multiple separate steps.
- We leverage the power of sparse online learning and CUR decomposition with residual analysis to learn and detect anomalies in an online fashion.
- We evaluate the performance of the proposed framework on real-world streaming datasets and compare SAD with multiple online anomaly detection competitors.

2 Related Work

Anomaly detection in data streams has been addressed through diverse learning strategies, including k-nearest neighbors (kNN), kernel density estimation (KDE), isolation forest (iForest) [Xiang *et al.*, 2023], extended isolation forest (EIF) [Hariri *et al.*, 2019], and locality-sensitive hashing (LSH) [Meira *et al.*, 2022]. kNN-based methods like NETS [Yoon *et al.*, 2019] (with set-based updates) and MD-UAL [Yoon *et al.*, 2021] (leveraging data-query duality) identify outliers through proximity queries, while local outlier factor variants such as MiLOF [Salehi *et al.*, 2016] (using summarization) and DILOF [Na *et al.*, 2018] (via sampling) focus on density deviations. KDE-based STARE [Yoon *et al.*, 2020] accelerates anomaly scoring by skipping stationary regions during density updates, and LSH-driven MStream [Yoon *et al.*, 2020] combines hashing with dimensionality reduction for efficient streaming. Isolation forest adaptations such as RRCF [Guha *et al.*, 2016] employ tree ensembles and sketching to handle concept drift. While these methods adopt window-based processing to manage evolving streams, their primary emphasis lies on minimizing computational costs for incremental model updates or preconfigured ensembles, often relying on manual feature engineering, such as random subsampling, linear transformations, or dimensionality reduction that restricts scalability and generalizability [Pang *et al.*, 2018]. This dependence on handcrafted features limits their adaptability to complex and high-dimensional data dynamics.

Recent advances in online anomaly detection span diverse methodologies to address evolving data streams. Xie et al. [2018] introduce OnlineBPCA, enhancing Bayesian PCA to jointly model row and column principal direction variations for precise real-time detection. Pei et al. [2023] leverage succinct tensor sketches to dynamically maintain subspaces representing non-anomalous historical data, enabling rapid outlier scoring for incoming streams. Siffer et al. [2017] apply extreme value theory to identify outliers in high-throughput time series, while Wang et al. [2020] propose incremental frameworks (FKDA-X, FKDA-CX, FKDA-C) for novelty detection in unlabeled chunk data streams. Boniol et al. [2021] develop SAND, a domain-agnostic method adaptive to distribution drifts, and Bhatia et al. [2023] extend higher-order sketches to preserve dense subgraph structures in streaming graphs. These approaches collectively advance real-time detection through adaptive subspace maintenance, statistical modeling, and incremental learning, though challenges persist in balancing accuracy, scalability, and interpretability. For a comprehensive overview, please refer to a recent survey [Bouman *et al.*, 2024].

Our proposed framework introduces a novel paradigm for OAD by synergizing $\ell_{1,2}$-mixed norm-based sparse online learning with CUR decomposition-driven active learning, enabling dynamic, assumption-free identification of anomalies in large-scale streaming data. Unlike existing methods that often rely on predefined data distributions or disjoint optimization of feature selection and detection, our approach unifies sparse feature regularization and representative instance curation in a single streaming workflow. The $\ell_{1,2}$-norm promotes row sparsity to discard redundant features in high-dimensional streams, while online CUR decomposition actively selects informative instances, maintaining a low-rank approximation that captures evolving patterns. This dual optimization facilitates robust residual analysis, where anomalies are flagged via large reconstruction errors derived directly from the sparse and interpretable CUR basis, eliminating reliance on distributional assumptions.

3 Proposed Method

3.1 Problem Formulation of OAD

In online anomaly detection for streaming data, we consider a continuous sequence $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \cdots\}$, where each instance $\boldsymbol{x}_t \in \mathbb{R}^d$ arrives at timestamp t in a d-dimensional feature space. The objective is to identify rare instances that deviate substantially from the majority of reference data points over time. At each timestamp t, the task involves computing an anomaly score, quantifying the "abnormality" of $\boldsymbol{x}_t$, and classifying the instance via predefined rules (e.g., thresholding), where $y_t = 1$ denotes an anomaly and $y_t = -1$ a normal instance. Our approach leverages CUR decomposition-based residual errors to derive these scores, measuring the discrepancy between observed and reconstructed data. Crucially, as both normal and anomalous data distributions evolve dynamically, the detection framework must continuously adapt to shifting patterns. This necessitates an online mechanism that updates its reference model incrementally, ensuring robust anomaly identification in non-stationary, high-dimensional streams while avoiding reliance on static assumptions about data behavior.

3.2 Online Convex Optimization for OAD

Online convex optimization (OCO) [Shalev-Shwartz and others, 2012] provides a formal framework for online learning and OAD, modeled as an iterative interaction between a learner and a dynamic environment. In this setup, the learner iteratively selects a model parameter $\boldsymbol{w}_t$ from a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ at each timestep t. The process unfolds as follows: upon receiving an instance $\boldsymbol{x}_t$, the learner predicts, incurs a convex loss $\ell_t(\boldsymbol{w}_t, (\boldsymbol{x}_t, y_t))$ upon observing the true label y_t, and updates $\boldsymbol{w}_t$ to $\boldsymbol{w}_{t+1}$ via mechanisms such as online gradient descent: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \lambda_t \nabla \ell_t(\boldsymbol{w}_t, (\boldsymbol{x}_t, y_t))$, where $\lambda_t > 0$ governs the step size. This framework enables adaptive decision-making in adversarial or non-stationary environments by balancing immediate regret minimization with long-term model stability, making it foundational for streaming applications requiring continuous adaptation.

In general, the goal of OCO is to iteratively select model parameters $\{\boldsymbol{w}_t\}$ that minimize cumulative loss over time, despite adversarially chosen convex loss functions ℓ_t revealed only after each prediction. The performance metric is regret, defined as the difference between the learner's total loss and the minimal loss achievable by a fixed optimal model $\boldsymbol{w}^\star \in \mathcal{S}$ in hindsight: $Regret_T = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t, (\boldsymbol{x}_t, y_t)) - \sum_{t=1}^{T} \ell_t(\boldsymbol{w}^\star, (\boldsymbol{x}_t, y_t))$, where the first term represents the learner's actual loss and the second term the best possible loss with perfect hindsight. Common loss functions include the least squares loss for regression, and the logistic loss or hinge loss for classification ($y_t \in \{-1, +1\}$). A core objective is to

design algorithms with sub-linear regret ($Regret_T = o(T)$), ensuring the average regret diminishes to zero as $T \to \infty$. This no-regret property guarantees the learner's performance asymptotically matches the best fixed strategy in hindsight, even in non-stationary environments [Siddiqui *et al.*, 2018]. Such frameworks underpin robust OAD systems, as outlined in Algorithm 1, by enabling adaptive model updates that balance immediate loss minimization with long-term stability.

Algorithm 1 Online Learning (OL) Algorithm for OAD

Require: the learning rate λ.
Ensure: the parameters of the classifier $\boldsymbol{w}_{T+1}$.
Initialization: $\boldsymbol{w}_0 = \mathbf{0}$
for $t = 1, 2, \cdots, T$ **do**
  Receive $\boldsymbol{x}_t \in \mathbb{R}^d$;
  Optimize $\boldsymbol{w}_t = \text{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{w}_{t-1} - \boldsymbol{w}\|_2^2 + R(\boldsymbol{w})$, $R(\cdot)$ is the regularization term;
  Predict $\hat{y}_t = \text{sign}(\boldsymbol{w}_t^T \boldsymbol{x}_t)$;
  Receive true label $y_t \in \{-1, 1\}$ of $\boldsymbol{x}_t$;
  Suffer the loss $\ell_t(\boldsymbol{w}_t) = [1 - y_t \boldsymbol{w}_t^T \boldsymbol{x}_t]_+$;
  if $\ell_t(\boldsymbol{w}_t) > 0$ **then**
    Update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \lambda_t \nabla \ell_t(\boldsymbol{w}_t)$;
  else if $\ell_t(\boldsymbol{w}_t) \leq 0$ **then**
    Update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$;
  end if
end for
Return: $\boldsymbol{w}_{T+1}$

3.3 $\ell_{1,2}$ Regularization based SOL

In this section, we formalize the online anomaly detection problem and propose a sparse online learning method using $\ell_{1,2}$-mixed regularization to achieve structured sparsity. We first observe a key property of ℓ_2 regularization: when $|\boldsymbol{w}|_2 \leq \lambda$, the solution reduces to the zero vector (Theorem 1). While a zero weight vector lacks generalization power—potentially undermining its utility as a regularizer—this behavior proves advantageous in incremental online settings with grouped weights. In sparse online learning, for instance, each sliding window l ($l = 1, \ldots, L$) employs a distinct weight vector $\boldsymbol{w}^l \in \mathbb{R}^d$. The prediction for an instance $\boldsymbol{x}$ is the vector $\big(\langle \boldsymbol{w}^1, \boldsymbol{x}\rangle, \ldots, \langle \boldsymbol{w}^L, \boldsymbol{x}\rangle\big)$, with the final class determined by $\text{argmax}_l \langle \boldsymbol{w}^l, \boldsymbol{x}\rangle$. Since all $\boldsymbol{w}^l$ operate on the same feature space, sparsity should be enforced collectively across corresponding features. Specifically, we seek to zero entire rows of the weight matrix $w_i^1, w_i^2, \ldots, w_i^L$ for each feature i ($i = 1, \ldots, d$). Here, the ℓ_2 regularizer's tendency to suppress entire weight vectors becomes valuable. By applying $\ell_{1,2}$ regularization—combining the ℓ_1 sparsity and the ℓ_2 shrinkage—we simultaneously achieve feature-wise sparsity while preserving non-zero weights' discriminative power.

Formally, let $\boldsymbol{W} \in \mathbb{R}^{d\times L}$ represent a $d \times L$ matrix where the l-th ($l = 1, 2, \cdots, L$) column of the matrix is the weight vector $\boldsymbol{w}^l$, where d is the total number of all evolvable features. Thus, the i-th ($i = 1, 2, \cdots, d$) row corresponds to the weight of the i-th feature with respect to all instances. The mixed $\ell_{1,2}$-norm of $\boldsymbol{W}$, denoted $\|\boldsymbol{W}\|_{\ell_{1,2}}$, is obtained by computing the ℓ_2-norm of each row of $\boldsymbol{W}$ and then applying the ℓ_1-norm to the resulting d dimensional vector, i.e., $\|\boldsymbol{W}\|_{\ell_{1,2}} = \sum_{i=1}^d \|\boldsymbol{w}_i\|_2$. Thus, in a mixed-norm regularized optimization problem, we seek the minimizer of the objective function,

$$f(\boldsymbol{W}) + \lambda \|\boldsymbol{W}\|_{\ell_{1,2}} \tag{1}$$

where $f(\boldsymbol{W})$ is a loss function, we define specifically $f(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{W}_t\|_F^2$ in our study.

Given the specific variants of various norms, the model update for the $\ell_{1,2}$ mixed-norm is readily available. Let $\boldsymbol{w}^l \in \mathbb{R}^d$ denote the l-th ($l = 1, 2, \cdots, L$) column of the matrix $\boldsymbol{W} \in \mathbb{R}^{d\times L}$, i.e., $\boldsymbol{W} = [\boldsymbol{w}^1, \boldsymbol{w}^2, \cdots, \boldsymbol{w}^L]$, and $\bar{\boldsymbol{w}}^i \in \mathbb{R}^L$ denote the i-th ($i = 1, 2, \cdots, d$) row of the matrix $\boldsymbol{W} \in \mathbb{R}^{d\times L}$, i.e., $\boldsymbol{W} = [\bar{\boldsymbol{w}}^1; \bar{\boldsymbol{w}}^2; \cdots; \bar{\boldsymbol{w}}^d]$. Analogously to the standard norm-based regularization, we let $\boldsymbol{W}_t = [\boldsymbol{w}_{t-L+1}, \boldsymbol{w}_{t-L+2}, \cdots, \boldsymbol{w}_t] \in \mathbb{R}^{d\times L}$ be the incremental matrix with all good feature alignment. For the $\ell_{1,2}$ mixed-norm, we need to solve the problem,

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times L}} \{\frac{1}{2}\|\boldsymbol{W} - \boldsymbol{W}_t\|_F^2 + \lambda \|\boldsymbol{W}\|_{\ell_{1,2}}\} \tag{2}$$

where $\|\cdot\|_F^2$ is the Frobenius norm of a matrix and $\lambda > 0$ is the regularization parameter.

This problem is equivalent to

$$\min_{\boldsymbol{W}=[\bar{\boldsymbol{w}}^1;\bar{\boldsymbol{w}}^2;\cdots;\bar{\boldsymbol{w}}^d]\in\mathbb{R}^{d\times L}} \sum_{i=1}^d \{\frac{1}{2}\|\bar{\boldsymbol{w}}^i - \bar{\boldsymbol{w}}_t^i\|_2^2 + \lambda \|\bar{\boldsymbol{w}}^i\|_2\} \tag{3}$$

where $\bar{\boldsymbol{w}}_t^i$ is the i-th row of $\boldsymbol{W}_t$. It is immediate to see that the problem given in Eq. (5) is decomposable into d separate problems of dimension L in Eq. (6), each of which can be solved by the procedures described in the following Theorem 1. The end result of solving these types of mixed-norm problems is a sparse matrix with numerous zero rows. In this way, this method can not only alleviate the curse of dimensionality by the incremental learning strategy, but also promote the sparsity of decremental and incremental features by considering feature correlations over time.

Theorem 1 (Closed-form Solution)**.** *The closed-form solution of the following ℓ_2-norm minimization: $\bar{\boldsymbol{w}}_\star^i = \text{argmin}_{\bar{\boldsymbol{w}}^i\in\mathbb{R}^L}\{\frac{1}{2}\|\bar{\boldsymbol{w}}^i - \bar{\boldsymbol{w}}_t^i\|_2^2 + \lambda\|\bar{\boldsymbol{w}}^i\|_2\}$, where $i = 1, 2, \cdots, d$, is:*

$$\bar{\boldsymbol{w}}_\star^i = \begin{cases} \mathbf{0} & \text{if } \|\bar{\boldsymbol{w}}_t^i\|_2 \leq \lambda \\ (1 - \frac{\lambda}{\|\bar{\boldsymbol{w}}_t^i\|_2})\bar{\boldsymbol{w}}_t^i & \text{if } \|\bar{\boldsymbol{w}}_t^i\|_2 > \lambda \end{cases} \tag{4}$$

Remark 1: Theorem 1 can be proved with the proximal operator of the ℓ_2-norm (Euclidean norm). It is worth noting that the ℓ_2 regularization results in a zero weight vector under the condition that $\|\bar{\boldsymbol{w}}_t^i\|_2 \leq \lambda$. This condition is rather more stringent for sparsity than the condition for ℓ_1 (where a weight is sparse based only on its value, while here, sparsity happens only if the entire weight vector has ℓ_2-norm less than λ), so it is unlikely to hold in high dimensions. However, it does constitute a very important building block when using a mixed ℓ_1/ℓ_2-norm as the regularization function.

Dataset/Method	osPCA	sketch-OAD	VFDT	HATT	ARCUS	Ours
a9a	21.726±0.103	18.792±0.079	32.710±1.436	31.633±0.306	18.792±0.079	**16.270±0.046**
codrna	16.812±0.383	21.782±0.151	16.714±0.096	16.601±0.176	16.696±1.073	**15.510±0.466**
german	36.294±0.973	**36.185±1.079**	36.361±1.472	41.130±1.284	37.130±1.246	39.125±1.305
ijcnn1	3.653±0.036	3.603±0.033	3.114±0.15	**2.973±0.030**	3.658±0.027	3.063±0.043
MITFace	5.336±0.230	11.333±1.285	5.331±0.225	5.097±0.264	5.275±0.272	**4.725±0.183**
PCMAC	32.391±1.183	39.954±2.064	32.344±1.220	32.643±1.147	32.020±1.261	**27.012±1.272**
spambase	26.352±0.561	36.072±13.322	28.201±0.691	28.959±0.398	27.376±0.373	**21.383±0.459**
splice	34.450±1.908	39.735±1.498	35.895±1.775	35.870±1.684	34.470±1.134	**33.925±1.244**
svmguide3	20.981±0.233	22.949±0.655	26.448±0.797	24.401±0.541	20.986±0.230	**20.909±0.232**

Table 1: Accumulated anomaly detection mistake rate (%) of different algorithms on all datasets. The best results are highlighted in bold.

3.4 OAL through CUR Decomposition

The CUR decomposition [Mahoney and Drineas, 2009] provides a low-rank approximation to a data matrix $\boldsymbol{W} \in \mathbb{R}^{d\times L}$. In particular, CUR decomposes the data matrix $\boldsymbol{W}$ into the form of a product of three matrices as $\boldsymbol{W} \approx \boldsymbol{CUR}$, where $\boldsymbol{C} \in \mathbb{R}^{d\times c}$, $\boldsymbol{U} \in \mathbb{R}^{c\times r}$, and $\boldsymbol{R} \in \mathbb{R}^{r\times L}$ ($c < L$ and $r < d$). Unlike other low-rank approximations such as Singular Value Decomposition (SVD), CUR extracts $\boldsymbol{C}$ and $\boldsymbol{R}$ as small numbers of the column and row vectors of $\boldsymbol{W}$, respectively. In other words, $\boldsymbol{C}$ and $\boldsymbol{R}$ are subsets of c columns and r rows of the original data matrix $\boldsymbol{W}$, respectively. This property helps practitioners to interpret the result more easily than that in the case of SVD.

Since the $\boldsymbol{R}$ has been determined by the $\ell_{1,2}$ constraint (r rows of $\boldsymbol{W}$ will be zero vectors in Section 3.3), which imposes sparse rows of the incremental matrix $\boldsymbol{W} \in \mathbb{R}^{d\times L}$. For the selection of $\boldsymbol{C}$, the optimization problem is defined as:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{L\times L}} \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}\|_F^2 + \eta \sum_{i=1}^{L} \|\boldsymbol{X}_{(i)}\|_2, \tag{5}$$

where $\boldsymbol{X} \in \mathbb{R}^{L\times L}$ is the parameter matrix, and $\eta > 0$ is a regularization parameter. Given the matrix $\boldsymbol{W}$, $\boldsymbol{W}_{(i)} \in \mathbb{R}^{1\times L}$ and $\boldsymbol{W}^{(i)} \in \mathbb{R}^{d\times 1}$ denote the i-th row vector and i-th column vector of $\boldsymbol{W}$, respectively. Similarly, given a set of indices $\mathcal{I}$, $\boldsymbol{W}_{\mathcal{I}}$ and $\boldsymbol{W}^{\mathcal{I}}$ denote the submatrices of $\boldsymbol{W}$ containing only $\mathcal{I}$ rows and columns, respectively. The term $\|\boldsymbol{X}_{(i)}\|_2$ induces $\boldsymbol{X}_{(i)}$ to be a zero vector, where $\boldsymbol{X}_{(i)} \in \mathbb{R}^{1\times L}$ is the i-th row vector of $\boldsymbol{X}$. The regularization constant η controls the degree of sparsity of the parameter matrix $\boldsymbol{X}$. If $\boldsymbol{X}_{(i)} = \boldsymbol{0}$ is a zero vector, the corresponding column of the data matrix $\boldsymbol{W}^{(i)}$ can be considered as an unimportant column for problem (5). On the other hand, $\boldsymbol{W}^{(i)}$ is important when the corresponding $\boldsymbol{X}_{(i)}$ is a nonzero vector. Therefore, we can select columns $\boldsymbol{C}$ as $\boldsymbol{W}^{\mathcal{I}}$, where $\mathcal{I} \subseteq [L] = \{1, 2, \cdots, L\}$ represents the indices corresponding to the nonzero row vectors of $\boldsymbol{X}$. Hence, the proposed algorithm can select more informative instances in the sliding window incrementally.

Problem (5) can be solved using the coordinate descent [Bien *et al.*, 2010]. The algorithm iteratively updates each parameter vector $\boldsymbol{X}_{(i)}$ corresponding to each row of the parameter matrix $\boldsymbol{X}$ until $\boldsymbol{X}$ converges. Then, the following equation is used to update $\boldsymbol{X}_{(i)} \in \mathbb{R}^{1\times L}$:

$$\boldsymbol{X}_{(i)} = \begin{cases} \boldsymbol{0} & \text{if } \|\boldsymbol{u}_i\|_2 \le \eta \\ (1 - \frac{\eta}{\|\boldsymbol{u}_i\|_2})\boldsymbol{u}_i & \text{if } \|\boldsymbol{u}_i\|_2 > \eta \end{cases} \tag{6}$$

where $\boldsymbol{u}_i \in \mathbb{R}^{1\times L}$ is computed by:

$$\boldsymbol{u}_i = \frac{(\boldsymbol{W}^{(i)})^\top}{\|\boldsymbol{W}^{(i)}\|_2}(\boldsymbol{W} - \sum_{j=1, j\neq i}^{L} \boldsymbol{W}^{(j)}\boldsymbol{X}_{(j)}). \tag{7}$$

where $\boldsymbol{W}^{(i)} \in \mathbb{R}^{d\times 1}$ denote the i-th column vector of $\boldsymbol{W}$. Algorithm 2 shows the pseudocode of coordinate descent. The inner loop (lines 3–4) performs Equation (6) to update each row of $\boldsymbol{X}$, and the outer loop (lines 2–5) repeats the update process until $\boldsymbol{X}$ converges. The computation cost of Equation (7) is $\mathcal{O}(L^2 d)$ time. Therefore, Equation (6) also requires $\mathcal{O}(L^2 d)$ time. Equation (6) can be modified to have $\mathcal{O}(Ld)$ time by updating the CUR every L rounds.

Algorithm 2 The CUR Decomposition Algorithm

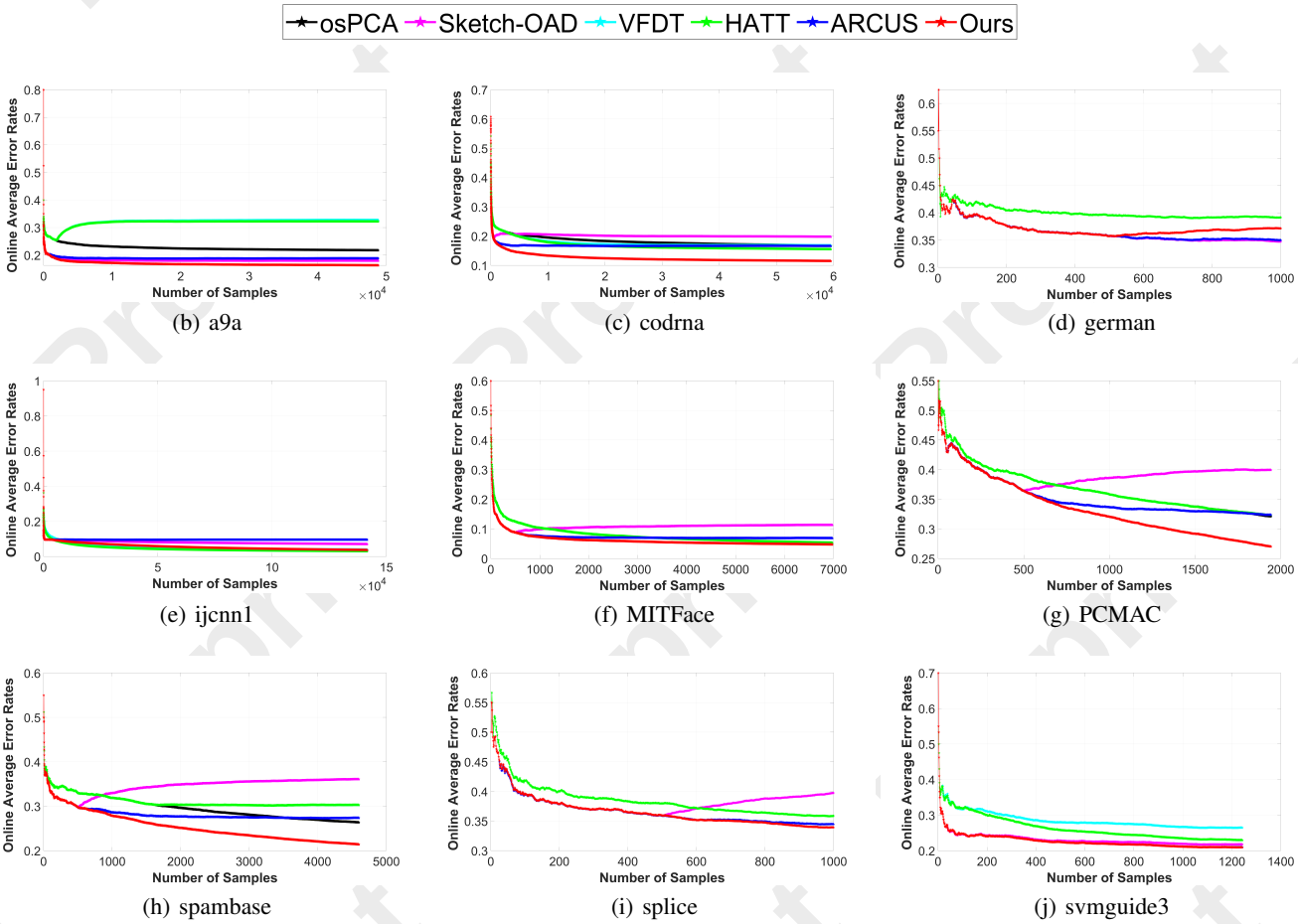
1: $[L] = \{1, 2, \cdots, L\}$, $\boldsymbol{X} \leftarrow \boldsymbol{0} \in \mathbb{R}^{L\times L}$;
2: **repeat**
3: **for** $i \in [L]$ **do**
4: Update $\boldsymbol{X}_{(i)}$ by Equation (6);
5: **end for**
6: **until** $\boldsymbol{X}$ converges.

3.5 Streaming Anomaly Detection through CUR Decomposition based Residual Errors

Residual analysis offers a principled approach to anomaly detection by quantifying deviations between observed data and their reconstructed estimates, where anomalies manifest as instances with disproportionately large residual errors due to divergence from dominant patterns. Traditional reconstruction methods rely on representative instances but face challenges from noisy or irrelevant attributes, necessitating joint selection of discriminative features and informative instances to accurately rebuild the underlying structure of the data matrix $\boldsymbol{W}$. To address this, we extend beyond standard factorizations to CUR decomposition, which explicitly selects a subset of representative instances ($\boldsymbol{C}$) and attributes ($\boldsymbol{R}$) to form a low-rank approximation of $\boldsymbol{W}$, enhancing interpretability and noise resilience. Mathematically, our proposed framework computes the residual matrix by

$$\min \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{CUR} - \tilde{\boldsymbol{R}}\|_F^2, \tag{8}$$

where $\tilde{\boldsymbol{R}}$ is the residual matrix of the weight matrix $\boldsymbol{W}$, and anomalies are identified by analyzing column-wise residuals:



(b) a9a (c) codrna (d) german

(e) ijcnn1 (f) MITFace (g) PCMAC

(h) spambase (i) splice (j) svmguide3

Figure 2: Dynamic learning curves in terms of online average error rates of all competing online algorithms.

a larger mixed norm $\|\tilde{\boldsymbol{R}}^\top\|_{2,0}$ of $\tilde{\boldsymbol{R}}(:, i)$ indicates that the i-th instance deviates significantly from the CUR-approximated structure, signaling higher anomaly likelihood. By integrating residual analysis with CUR's explicit instance-attribute selection, the method simultaneously mitigates noise interference and provides interpretable criteria for real-time anomaly ranking in dynamic streams.

4 Experimental Evaluation

4.1 Datasets and Evaluation Metrics

To validate the effectiveness of the proposed method, we conduct the experiments on nine streaming datasets and 10% of anomalies/outliers are randomly added in such datasets. Table 2 summarizes the attributes including number of samples, features, and classes. The online average error rate of the anomalies is utilized as the evaluation metric.

Dataset	#Samples	#Features	#Classes
a9a	48,842	123	2
codrna	59,535	8	2
german	1,000	24	2
ijcnn1	141,691	22	2
MITFace	6,977	361	2
PCMAC	1,943	3,290	2
spambase	4,601	57	2
splice	1,000	60	2
svmguide3	1,243	21	2

Table 2: Summary of the real streaming datasets in the experiments.

4.2 Experimental Settings

To evaluate the proposed SAD algorithm, we compare it with five state-of-the-art online anomaly detection algorithms: osPCA [Lee *et al.*, 2012], sketch-OAD [Huang and Kasiviswanathan, 2015], VFDT [Tan *et al.*, 2011], HATT [Bifet *et al.*, 2017], and ARCUS [Yoon *et al.*, 2022]. We implement all the competing algorithms in Matlab. For a fair comparison, the same experimental setup is applied to all algorithms. After the preliminary studies, we set the parameters by $L = 50$, $\lambda = 10$, and $\eta = 1$. All the other parameter values are determined based on [Lee *et al.*, 2012; Huang and Kasiviswanathan, 2015; Tan *et al.*, 2011; Bifet *et al.*, 2017; Yoon *et al.*, 2022]. Twenty independent runs are performed and the average results of each method are reported.

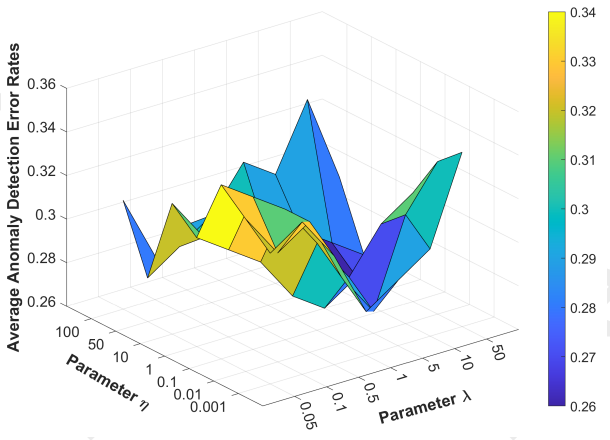


Figure 3: The sensitivity analysis of parameters λ and η of SAD on the "PCMAC" dataset.

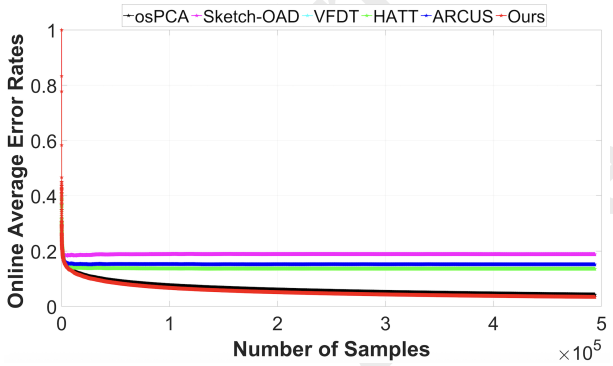


Figure 4: Dynamic learning curves (online average error rates) of all competing online algorithms on a real world dataset KDDCUP.

4.3 Overall Comparison

Table 1 presents the overall performance including average detection errors ($\pm$standard deviation) of all competing algorithms. Overall, SAD achieves the lowest classification errors in all the nine streaming datasets. The quantitative reductions (on average over the nine datasets) using the proposed SAD method and competing algorithms osPCA, sketch-OAD, VFDT, HATT, and ARCUS are 10.00%, 22.67%, 14.44%, 14.16%, and 9.04%, respectively.

4.4 Dynamic Error Rate Comparison

As shown in Figure 2, we investigate the dynamic error rate of all algorithms with the progression of a data stream. The online average error rate curves of SAD dominate the corresponding curves of all other algorithms (without much variation) on seven datasets except "german" and "ijcnn1". The superiority of SAD over others is evident on "codrna", "PCMAC", and "spambase" datasets. These results validate the efficiency of SAD compared to other competing algorithms.

4.5 Sensitivity Studies

To execute SAD, the regularization parameters λ and η need to be specified. We examine their impact on performance via grid search, using the large-scale "PCMAC" dataset with $\lambda \in [10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 2\times10^{1}, 5\times10^{1}, 10^{2}]$ and $\eta \in [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$. As shown in Figure 3, which compares dynamic average detection error rates across parameter values, SAD exhibits stable performance for $\eta \in [10^{-2}, 10^{2}]$ but fluctuates significantly with λ. When η is fixed, error rates increase sharply as λ decreases from 10^2 to 10^{-2}: smaller λ values promote higher sparsity but simultaneously degrade detection accuracy. Overall, SAD demonstrates robustness to η yet sensitivity to λ.

4.6 Empirical Studies on the Real-world Dataset

We evaluate our proposed approach using the KDDCUP intrusion detection dataset[1], comprising 494,020 records across 34 dimensions with an anomaly rate of 1.77% (8,752 anomalies vs. 485,268 normal instances). All features are normalized via RobustScaler [Lusito *et al.*, 2023], chosen for its outlier robustness. Figure 4 demonstrates the dynamic performance of competing algorithms under incremental data streaming. Both osPCA and our SAD method significantly outperform four baseline methods in detection error. Notably, SAD achieves superior performance with 3.38% average detection error compared to osPCA's 4.25%, confirming its enhanced scalability and competitive efficacy for real-world large-scale streaming data applications.

5 Conclusion

To address the challenge of consistently and accurately detecting unexpected anomalies in large-scale streaming settings without compromising computational and memory efficiency, this paper introduces an innovative Streaming Anomaly Detection (SAD) method rooted in sparse active online learning. Our primary contribution is the seamless integration of $\ell_{1,2}$-norm based sparse online learning with CUR decomposition based online active learning, which facilitates effective residual analysis for anomaly detection. This dual approach enables SAD to not only provide real-time feature interpretations through the $\ell_{1,2}$-norm penalty but also to automatically identify anomalies in streaming data. Crucially, this method eliminates the need for manually setting thresholds and does not rely on assumptions about data distribution. Our extensive evaluation across multiple streaming datasets underscores the efficacy of SAD, demonstrating significant performance enhancements over existing OAD competitors.

As part of future work, we plan to expand the applicability of our SAD method to more complex data streams encountered in open environments, including those involving feature evolution and weak supervision.

Acknowledgments

This work done by Y. He has been supported in part by the National Science Foundation (NSF) under Grant Nos. IIS-2236578, IIS-2441449, IOS-2446522, and the Commonwealth Cyber Initiative (CCI). The work done by Z. Chen has been supported in part by an Illinois Innovation Network (IIN) sustaining Illinois seed funding grant. The work done by other authors were not supported by any of these funds.

[1]http://archive.ics.uci.edu/ml/datasets.php

References

[Bhatia *et al.*, 2023] Siddharth Bhatia, Mohit Wadhwa, Kenji Kawaguchi, Neil Shah, Philip S Yu, and Bryan Hooi. Sketch-based anomaly detection in streaming graphs. In *KDD*, pages 93–104, 2023.

[Bien *et al.*, 2010] Jacob Bien, Ya Xu, and Michael W Mahoney. Cur from a sparse optimization viewpoint. *NeurIPS*, 23, 2010.

[Bifet *et al.*, 2017] Albert Bifet, Jiajin Zhang, Wei Fan, Cheng He, Jianfeng Zhang, Jianfeng Qian, Geoff Holmes, and Bernhard Pfahringer. Extremely fast decision tree mining for evolving data streams. In *KDD*, pages 1733–1742, 2017.

[Boniol *et al.*, 2021] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J Franklin. Sand: streaming subsequence anomaly detection. *Proceedings of the VLDB Endowment*, 14(10):1717–1729, 2021.

[Bouman *et al.*, 2024] Roel Bouman, Zaharah Bukhsh, and Tom Heskes. Unsupervised anomaly detection algorithms on real-world data: how many do we need? *Journal of Machine Learning Research*, 25(105):1–34, 2024.

[Burkle, 2020] Frederick M Burkle. Declining public health protections within autocratic regimes: impact on global public health security, infectious disease outbreaks, epidemics, and pandemics. *Prehospital and Disaster Medicine*, 35(3):237–246, 2020.

[Chen *et al.*, 2021] Zhong Chen, Zhide Fang, Victor Sheng, Jiabin Zhao, Wei Fan, Andrea Edwards, and Kun Zhang. Adaptive robust local online density estimation for streaming data. *International Journal of Machine Learning and Cybernetics*, 12(6):1803–1824, 2021.

[Chen *et al.*, 2023] Zhong Chen, Victor Sheng, Andrea Edwards, and Kun Zhang. An effective cost-sensitive sparse online learning framework for imbalanced streaming data classification and its application to online anomaly detection. *Knowledge and Information Systems*, 65(1):59–87, 2023.

[Chen *et al.*, 2024a] Zhong Chen, Yi He, Di Wu, Huixin Zhan, Victor Sheng, and Kun Zhang. Robust sparse online learning for data streams with streaming features. In *SDM*, pages 181–189. SIAM, 2024.

[Chen *et al.*, 2024b] Zhong Chen, Victor Sheng, Andrea Edwards, and Kun Zhang. Cost-sensitive sparse group online learning for imbalanced data streams. *Machine Learning*, 113(7):4407–4444, 2024.

[Ding *et al.*, 2019] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *SDM*, pages 594–602. SIAM, 2019.

[Du, 2022] Meiyan Du. Application of information communication network security management and control based on big data technology. *International Journal of Communication Systems*, 35(5):e4643, 2022.

[Guha *et al.*, 2016] Sudipto Guha, Nina Mishra, Gourav Roy, and Okke Schrijvers. Robust random cut forest based anomaly detection on streams. In *ICML*, pages 2712–2721. PMLR, 2016.

[Hariri *et al.*, 2019] Sahand Hariri, Matias Carrasco Kind, and Robert J Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1479–1489, 2019.

[He *et al.*, 2017] Yongfu He, Yu Peng, Shaojun Wang, Datong Liu, and Philip HW Leong. A structured sparse subspace learning algorithm for anomaly detection in uav flight data. *IEEE Transactions on Instrumentation and Measurement*, 67(1):90–100, 2017.

[He *et al.*, 2019] Yi He, Baijun Wu, Di Wu, Ege Beyazit, Sheng Chen, and Xindong Wu. Online learning from capricious data streams: a generative approach. In *IJCAI*, 2019.

[He *et al.*, 2023] Yi He, Christian Schreckenberger, Heiner Stuckenschmidt, and Xindong Wu. Towards utilitarian online learning-a review of online algorithms in open feature space. In *IJCAI*, pages 6647–6655, 2023.

[Huang and Kasiviswanathan, 2015] Hao Huang and Shiva Prasad Kasiviswanathan. Streaming anomaly detection using randomized matrix sketching. *Proceedings of the VLDB Endowment*, 9(3):192–203, 2015.

[Lai *et al.*, 2020] Chieh-Hsin Lai, Dongmian Zou, and Gilad Lerman. Robust subspace recovery layer for unsupervised anomaly detection. In *ICLR*, 2020.

[Lee *et al.*, 2012] Yuh-Jye Lee, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Anomaly detection via online oversampling principal component analysis. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1460–1470, 2012.

[Li *et al.*, 2011] Jie Li, Guan Han, Jing Wen, and Xinbo Gao. Robust tensor subspace learning for anomaly detection. *International Journal of Machine Learning and Cybernetics*, 2:89–98, 2011.

[Lian *et al.*, 2022] Heng Lian, John Scovi Atwood, Bo-Jian Hou, Jian Wu, and Yi He. Online deep learning from doubly-streaming data. In *ACM-MM*, pages 3185–3194, 2022.

[Lusito *et al.*, 2023] Salvatore Lusito, Andrea Pugnana, and Riccardo Guidotti. Solving imbalanced learning with outlier detection and features reduction. *Machine Learning*, pages 1–58, 2023.

[Mahoney and Drineas, 2009] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[Manzoor *et al.*, 2018] Emaad Manzoor, Hemank Lamba, and Leman Akoglu. xstream: Outlier detection in feature-evolving data streams. In *KDD*, pages 1963–1972, 2018.

[Meira *et al.*, 2022] Jorge Meira, Carlos Eiras-Franco, Verónica Bolón-Canedo, Goreti Marreiros, and Amparo Alonso-Betanzos. Fast anomaly detection with

locality-sensitive hashing and hyperparameter autotuning. *Information Sciences*, 607:1245–1264, 2022.

[Muhammad *et al.*, 2020] Ghulam Muhammad, M Shamim Hossain, and Sahil Garg. Stacked autoencoder-based intrusion detection system to combat financial fraudulent. *IEEE Internet of Things Journal*, 10(3):2071–2078, 2020.

[Na *et al.*, 2018] Gyoung S Na, Donghyun Kim, and Hwanjo Yu. Dilof: Effective and memory efficient local outlier detection in data streams. In *KDD*, pages 1993–2002, 2018.

[O'Reilly *et al.*, 2014] Colin O'Reilly, Alexander Gluhak, and Muhammad Ali Imran. Adaptive anomaly detection with kernel eigenspace splitting and merging. *IEEE Transactions on Knowledge and Data Engineering*, 27(1):3–16, 2014.

[Pang *et al.*, 2018] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *KDD*, pages 2041–2050, 2018.

[Pazarbasioglu *et al.*, 2020] Ceyla Pazarbasioglu, Alfonso Garcia Mora, Mahesh Uttamchandani, Harish Natarajan, Erik Feyen, and Mathew Saal. Digital financial services. *World Bank*, 54:1–54, 2020.

[Pei *et al.*, 2023] Shuyu Pei, Jigang Wen, Kun Xie, Gaogang Xie, and Kenli Li. On-line network traffic anomaly detection based on tensor sketch. *IEEE Transactions on Parallel and Distributed Systems*, 2023.

[Peng *et al.*, 2018] Zhen Peng, Minnan Luo, Jundong Li, Huan Liu, and Qinghua Zheng. Anomalous: A joint modeling approach for anomaly detection on attributed networks. In *IJCAI*, volume 18, pages 3513–3519, 2018.

[Sahoo *et al.*, 2018] Doyen Sahoo, Quang Pham, Jing Lu, and Steven CH Hoi. Online deep learning: learning deep neural networks on the fly. In *IJCAI*, pages 2660–2666, 2018.

[Salehi *et al.*, 2016] Mahsa Salehi, Christopher Leckie, James C Bezdek, Tharshan Vaithianathan, and Xuyun Zhang. Fast memory efficient local outlier detection in data streams. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3246–3260, 2016.

[Schreckenberger *et al.*, 2023] Christian Schreckenberger, Yi He, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. Online random feature forests for learning in varying feature spaces. In *AAAI*, volume 37, pages 4587–4595, 2023.

[Shalev-Shwartz and others, 2012] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[Siddiqui *et al.*, 2018] Md Amran Siddiqui, Alan Fern, Thomas G Dietterich, Ryan Wright, Alec Theriault, and David W Archer. Feedback-guided anomaly discovery via online optimization. In *KDD*, pages 2200–2209, 2018.

[Siffer *et al.*, 2017] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. Anomaly detection in streams with extreme value theory. In *KDD*, pages 1067–1075, 2017.

[Tan *et al.*, 2011] Swee Chuan Tan, Kai Ming Ting, and Tony Fei Liu. Fast anomaly detection for streaming data. In *IJCAI*. Citeseer, 2011.

[Tong and Lin, 2011] Hanghang Tong and Ching-Yung Lin. Non-negative residual matrix factorization with application to graph anomaly detection. In *SDM*, pages 143–153. SIAM, 2011.

[Wang *et al.*, 2020] Yi Wang, Yi Ding, Xiangjian He, Xin Fan, Chi Lin, Fengqi Li, Tianzhu Wang, Zhongxuan Luo, and Jiebo Luo. Novelty detection and online learning for chunk data streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2400–2412, 2020.

[Wu *et al.*, 2023] Di Wu, Shengda Zhuo, Yu Wang, Zhong Chen, and Yi He. Online semi-supervised learning with mix-typed streaming features. In *AAAI*, volume 37, pages 4720–4728, 2023.

[Xiang *et al.*, 2023] Haolong Xiang, Xuyun Zhang, Hongsheng Hu, Lianyong Qi, Wanchun Dou, Mark Dras, Amin Beheshti, and Xiaolong Xu. Optiforest: optimal isolation forest for anomaly detection. In *IJCAI*, pages 2379–2387, 2023.

[Xie *et al.*, 2018] Kun Xie, Xiaocan Li, Xin Wang, Jiannong Cao, Gaogang Xie, Jigang Wen, Dafang Zhang, and Zheng Qin. On-line anomaly detection with high accuracy. *IEEE/ACM Transactions on Networking*, 26(3):1222–1235, 2018.

[Yoon *et al.*, 2019] Susik Yoon, Jae-Gil Lee, and Byung Suk Lee. Nets: extremely fast outlier detection from a data stream via set-based processing. *Proceedings of the VLDB Endowment*, 12(11):1303–1315, 2019.

[Yoon *et al.*, 2020] Susik Yoon, Jae-Gil Lee, and Byung Suk Lee. Ultrafast local outlier detection from a data stream with stationary region skipping. In *KDD*, pages 1181–1191, 2020.

[Yoon *et al.*, 2021] Susik Yoon, Yooju Shin, Jae-Gil Lee, and Byung Suk Lee. Multiple dynamic outlier-detection from a data stream by exploiting duality of data and queries. In *SIGMOD*, pages 2063–2075, 2021.

[Yoon *et al.*, 2022] Susik Yoon, Youngjun Lee, Jae-Gil Lee, and Byung Suk Lee. Adaptive model pooling for online deep anomaly detection from a complex evolving data stream. In *KDD*, pages 2347–2357, 2022.

[Yuan *et al.*, 2020] Yali Yuan, Sripriya Srikant Adhatarao, Mingkai Lin, Yachao Yuan, Zheli Liu, and Xiaoming Fu. Ada: Adaptive deep log anomaly detector. In *INFOCOM*, pages 2449–2458. IEEE, 2020.

[Zhang and Zhao, 2022] Zheng Zhang and Liang Zhao. Unsupervised deep subgraph anomaly detection. In *ICDM*, pages 753–762. IEEE, 2022.

[Zong *et al.*, 2018] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.