

PolicyShift-AML: Auditing Selective-Label Policy Dependence in Anti-Money-Laundering Transaction Monitoring

Muhammetalp Erdem

alperdemdev@gmail.com

Trabzon University

Research Article

Keywords: anti-money laundering, positive–unlabeled learning, selective labels, audit policy, graph-history features, operational evaluation

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10968989/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

PolicyShift-AML: Auditing Selective-Label Policy Dependence in Anti-Money-Laundering Transaction Monitoring

Muhammetalp Erdem^{1*}

^{1*}Department of Computer Engineering, Trabzon University, Söğütli,
Trabzon, 61000, Türkiye.

Corresponding author(s). E-mail(s): alperdemdev@gmail.com;

Abstract

Anti-money-laundering models learn from cases produced by selective investigation, so offline performance can reflect an institution’s scrutiny policy rather than latent event status. We introduce PolicyShift-AML, a controlled protocol that holds a synthetic transaction stream and latent labels fixed while varying the policy that reveals positive labels. A re-audit of the initial implementation corrected an invalid class-prior estimator, case-control bias in propensity and risk estimation, right-censoring under confirmation delay, unequal policy capacity, and latent-label access in model feedback. On 3,029,170 TransXion transactions, ten-seed experiments show a 26.1%–29.3% best-to-worst training-policy spread in latent Recall@500; this is training-policy sensitivity, not deployment transfer. Under model-feedback labels, apparent Recall@500 exceeds latent recall by 0.1085 for a neural visible-label baseline. A deployable non-negative positive–unlabeled estimator, using audit propensities fitted on the complete historical audit log and a Horvitz–Thompson class prior, reaches 0.4021 Recall@500 at a 30-day delay under rule-based selection, versus 0.2352 for LightGBM and 0.4021 for an oracle diagnostic. Graph-history ablations and a stratified 216,315-row IBM AMLWorld stress test further delimit when shared structure supports inductive generalization and when positivity failure prevents identification. The results support policy-aware evaluation, while not establishing efficacy on real institutional investigations.

Keywords: anti-money laundering, positive–unlabeled learning, selective labels, audit policy, graph-history features, operational evaluation

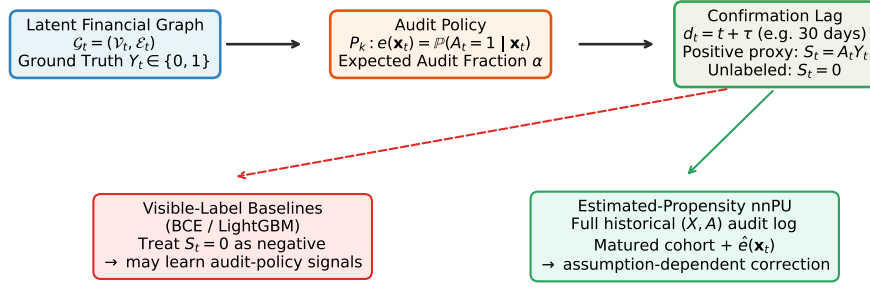


Fig. 1 Revised PolicyShift-AML protocol. The audit decision A is made at triage time from pre-decision features; the simulated confirmed-positive proxy $S = AY$ becomes usable only after delay τ . Audit policies share the same expected capacity. Propensities are fitted from the complete (X, A) audit log, whereas prevalence and PU risk use fully matured cohorts.

1 Introduction

Transaction-monitoring systems do not observe a random sample of illicit and legitimate activity. They observe outcomes of alerts selected by rules, thresholds, analysts, or incumbent models. Treating every unobserved transaction as benign can therefore teach a model the historical scrutiny mechanism rather than the latent target. This is a selective-label problem, distinct from a change in transaction behavior.

A bank can change an alert threshold while the event stream remains fixed; conversely, criminals can change behavior while the alert policy remains fixed. Conventional temporal drift evaluations mix these processes. Recent work emphasizes ranking under review capacity [1], temporal graph consistency [2], and feedback limitations in fraud operations [3, 4], but controlled evidence on the label-generating policy remains limited.

This study contributes an executable stress-test protocol rather than a claim of real-bank causal recovery. It (i) separates a latent synthetic stream from five audit policies with a common expected capacity; (ii) defines training-policy utility and visible-label optimism on a fixed latent holdout; (iii) evaluates a selected-at-random positive-unlabeled (PU) learner whose propensity model, class prior, and marginal risk are estimated on representative historical cohorts; and (iv) tests graph-history ablations, confirmation delays, and complete typology support loss.

The observable positive proxy is not equated with a legal determination of money laundering. Financial institutions file Suspicious Activity Reports when activity is known or suspected to be suspicious; filing is not proof that a crime occurred [5].

2 Related work

AML benchmarks and representations.

AMLWorld provides realistic synthetic transactions and typology annotations [6]. TransXion was designed to reduce static-pattern shortcuts in dynamic benchmarks [7]. Network-aware AML research includes structural-entropy representations [8], sub-graph benchmarks [9], ranking models [1], and pseudo-label graph attention [10]. Our models use point-in-time graph-history aggregates, not message-passing GNNs.

Positive-unlabeled learning.

PU learning distinguishes observed positives from an unlabeled mixture [11, 12]. Non-negative PU estimation prevents an empirical negative-risk component from becoming arbitrarily negative [13]. Under selected-at-random labeling, inverse propensity correction extends beyond constant labeling probability [14]. Related work studies online PU learning under drift [15] and PU learning for financial graphs [16]. We focus on implementation conditions often hidden by oracle evaluation: which cohort fits $e(X)$, which cohort estimates π , and which sample represents the marginal feature distribution.

Policy feedback and temporal drift.

Continual graph learning addresses changes in data and tasks [17–20]. PolicyShift-AML instead holds the evaluation population fixed and alters only the mechanism revealing training positives. This isolates training-label dependence but does not emulate simultaneous criminal adaptation.

3 Problem formulation

For event i , let X_i be triage-time features, $Y_i \in \{0, 1\}$ the latent benchmark label, $A_i \in \{0, 1\}$ the audit decision, and $S_i \in \{0, 1\}$ the simulated positive proxy. The benchmark assumes

$$S_i = A_i Y_i. \quad (1)$$

This is a controlled-data assumption of perfect specificity and complete confirmation among audited latent positives, not a claim about operational SARs.

Let $e(x) = P(A = 1 \mid X = x)$. Inverse-propensity identification requires consistency, conditional exchangeability $Y \perp A \mid X$, positivity $e(X) > 0$, and an adequate propensity model. Where an assumption fails, the estimator is descriptive rather than causally identifying.

An audit decision is available immediately, but S_i is available only at $t_i + \tau$. At cutoff T , prevalence and risk estimation use

$$\mathcal{M}_{T,\tau} = \{i : t_i + \tau \leq T\}. \quad (2)$$

The class prior is estimated on this representative matured cohort by

$$\hat{\pi}_{\text{HT}} = \frac{1}{|\mathcal{M}_{T,\tau}|} \sum_{i \in \mathcal{M}_{T,\tau}} \frac{S_i}{\widehat{e}(X_i)}, \quad (3)$$

because $E[S/e(X)] = E[Y]$ under the stated assumptions. This replaces the invalid ratio $E[S]/E[e(X) \mid S = 1]$ in the initial pipeline.

3.1 Estimands

Let f_P be trained from labels generated by policy P and let $\mathcal{H}$ be a fixed latent holdout. Define

$$U_K(P) = \text{Recall@}K(f_P; \mathcal{H}, Y), \quad \Gamma_K(P_a, P_b) = U_K(P_a) - U_K(P_b). \quad (4)$$

The bounded Training-Policy Sensitivity Index is

$$\text{TPSI}_K(P_a, P_b) = \frac{|U_K(P_a) - U_K(P_b)|}{\max\{U_K(P_a), U_K(P_b), \epsilon\}}. \quad (5)$$

TPSI is not the population-stability index and does not measure deployment-population shift: the holdout is unchanged. Visible-label optimism is $\text{Recall@}K$ against policy-visible positives minus $\text{Recall@}K$ against latent labels on the same holdout.

4 Methods

4.1 Audit policies

We simulate uniform, amount-focused, rule-based, model-feedback, and typology-blind Bernoulli audit mechanisms. Each is calibrated to an expected audit fraction of 0.15 over the stream. The feedback incumbent is trained only from an initial uniformly audited seed set and observed confirmed positives; it never sees latent Y . Equal capacity does not equalize positive yield, which remains part of the mechanism under study.

4.2 Deployable PU risk

For score g and positive/negative losses ℓ_+ and ℓ_- ,

$$R(g) = E[\ell_-(g(X))] + E\left[\frac{S}{e(X)}\{\ell_+(g(X)) - \ell_-(g(X))\}\right]. \quad (6)$$

The implementation uses a stable nnPU decomposition. Observed positives receive self-normalized inverse-propensity weights

$$w_i = \frac{1/\hat{e}(X_i)}{\sum_{j:S_j=1} 1/\hat{e}(X_j)}. \quad (7)$$

Positive risk is $\hat{\pi} \sum_i w_i \ell_+(g(X_i))$. Negative risk is the loss on an independent uniform sample from the matured cohort minus the corresponding weighted positive loss, clipped at zero. The independent marginal sample matters because the classifier set

includes all visible positives plus at most 120,000 sampled unlabeled events and is not representative.

We fit $\hat{e}(X)$ with a 60-tree LightGBM classifier on the full pre-cutoff (X, A) audit log and clip predictions to $[0.005, 0.99]$; the stochastic nnPU weight calculation applies a further 0.01 floor for variance control. Oracle-propensity nnPU is retained only as a simulation diagnostic. Baselines are LightGBM on visible labels, a two-layer MLP with visible-label binary cross-entropy (BCE), and inverse-propensity-weighted BCE.

4.3 Features and leakage controls

Transaction-local features include logged amounts, amount ratio, cross-bank and self-loop indicators, payment format, hour, and day of week. Graph-history features add sender/receiver prior counts, totals, means, activity span, and new-node flags computed from earlier events only. A property test recomputed 25,000 rows after perturbing future events and required exact invariance of past feature rows.

5 Experimental design

TransXion.

The cache contains 3,029,170 transactions and 4,641 positives. Training precedes 1 April 2025; the fixed April holdout contains 213,662 events and 423 positives. Stage 1 compares label policies at $\tau = 30$ days. Stage 2 uses $\tau \in \{0, 7, 14, 30, 60\}$ under amount, rule, and feedback selection. Stage 3 compares transaction-only with graph-history features under uniform, amount, and rule policies.

IBM AMLWorld HI-Small.

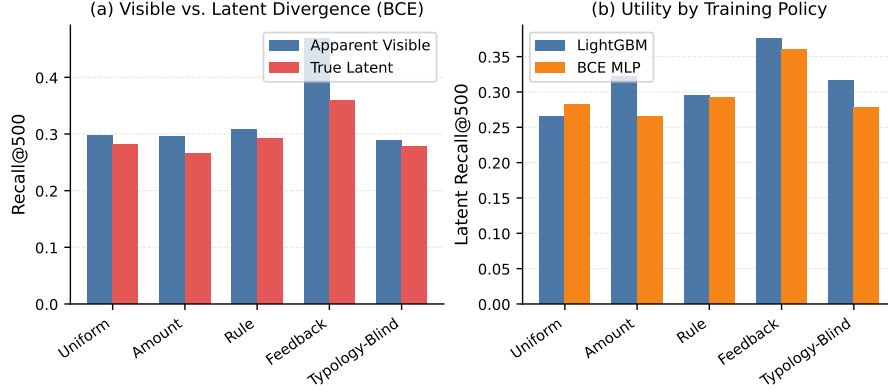
The source benchmark contains 5,081,134 transactions and 5,177 laundering events [6]. Stage 4 does not use every background row: it uses a 216,315-row stratified cache retaining all positives and sampled background. After restricting to days 1–10, 215,636 rows and 4,522 positives remain. Days 1–6 train and days 7–10 test. CYCLE and BIPARTITE events receive audit probability $p_{\text{explore}} \in \{0, 0.01, 0.05, 0.10\}$ while other propensities are rescaled to preserve total expected capacity.

Statistics and reproducibility.

Definitive experiments use ten paired seeds. We report mean paired differences, 95% percentile bootstrap intervals over seed pairs (5,000 resamples), and exact two-sided sign-flip tests. With ten non-zero pairs, the minimum attainable two-sided value is $2/2^{10} = 0.001953$. These summaries have different interpretations and need not agree mechanically. Manifests store seeds, software versions, cohorts, audit rates, prior error, and propensity error.

Table 1 Stage 1 mean Recall@500 over ten seeds.

Training policy	LightGBM		BCE MLP	
	Apparent	Latent	Apparent	Latent
Uniform	0.2616	0.2655	0.2974	0.2818
Amount	0.3613	0.3220	0.2959	0.2660
Rule	0.2849	0.2946	0.3071	0.2929
Feedback	0.4539	0.3757	0.4685	0.3600
Typology blind	0.3323	0.3165	0.2888	0.2780

**Fig. 2** Stage 1. Left: apparent versus latent Recall@500. Right: latent Recall@500 by training policy and learner on the same April holdout. The right panel is deliberately not a transfer matrix.

6 Results

6.1 Selective labels distort model comparison

Table 1 replaces the original repeated transfer matrix. On the fixed latent holdout, LightGBM ranges from 0.2655 Recall@500 after uniform-policy training to 0.3757 after feedback training; BCE ranges from 0.2660 after amount-policy training to 0.3600 after feedback training. Relative to each method’s best policy, the spreads are 29.3% and 26.1%.

Uniform-minus-feedback latent recall is -0.1102 for LightGBM (95% interval $[-0.1875, -0.0437]$; exact $p = 0.0137$) and -0.0783 for BCE ($[-0.1284, -0.0326]$; $p = 0.0137$). The direction is opposite the old collapse headline: feedback labels improve latent ranking here. Yet visible evaluation is optimistic under feedback by 0.0783 for LightGBM and 0.1085 for BCE. Policy dependence remains; cross-policy deployment collapse is not established.

Table 2 Stage 2 latent Recall@500 at $\tau = 30$ days (ten-seed means).

Method	Rule	Amount	Feedback
LightGBM visible	0.2352	0.3411	0.3955
BCE visible	0.1787	0.1995	0.2144
IPW-BCE	0.1804	0.2144	0.2213
nnPU estimated	0.4021	0.4014	0.4123
nnPU oracle	0.4021	0.4019	0.4125

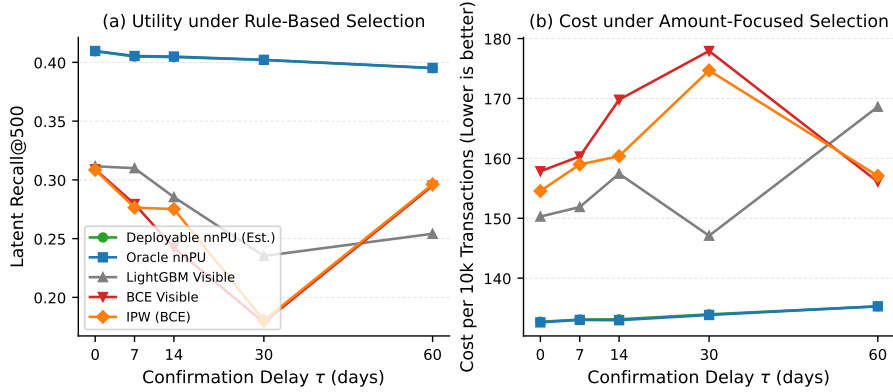


Fig. 3 Stage 2 delay sensitivity. Curves show ten-seed means; non-monotonic visible-baseline behavior reflects label loss and changing temporal composition of the matured cohort.

6.2 Estimated nnPU under confirmation delay

At 30 days (Table 2), estimated nnPU reaches 0.4021 Recall@500 under rule selection, versus 0.2352 for LightGBM, 0.1787 for BCE, and 0.1804 for IPW-BCE. The nnPU–LightGBM difference is 0.1669 ($[0.1092, 0.2348]$; exact $p = 0.001953$). Under amount selection it is 0.0603 ($[0.0400, 0.0804]$; $p = 0.001953$). Under feedback it is 0.0168 ($[-0.0045, 0.0418]$; $p = 0.25$), so superiority to LightGBM is not supported there.

Estimated and oracle nnPU differ by at most 0.0005 in these means. This is a benchmark diagnostic, not validation on bank data. Mean absolute prior error is 1.00×10^{-4} under rule selection, 1.05×10^{-4} under amount selection, and 3.01×10^{-4} under feedback selection; feedback also has the largest propensity error (0.0172). Across delays, estimated-nnPU recall remains between 0.3950 and 0.4168, whereas visible baselines vary non-monotonically as the matured cohort changes.

6.3 Graph-history sensitivity and zero-support stress test

Graph-history nnPU maintains mean Recall@500 of 0.4024, 0.4012, and 0.4007 under uniform, amount, and rule training policies, giving mean-based TPSI values 0.0029 and 0.0041 for the two selective policies. Transaction-only nnPU scores 0.0014 under

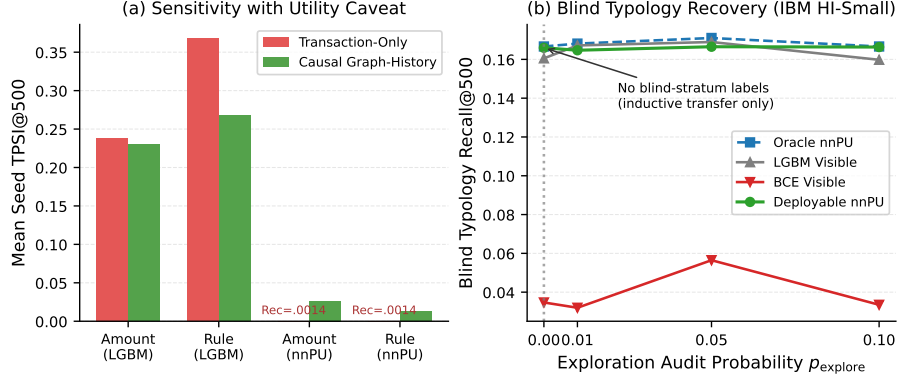


Fig. 4 Left: Stage 3 training-policy sensitivity. Right: Stage 4 blind-typology Recall@500. At zero support, non-zero recall is inductive transfer from shared features and does not identify a blind-stratum response.

every policy. Its TPSI is therefore zero for the wrong reason: uniformly failed models are superficially invariant. Sensitivity must be reported with absolute utility. For LightGBM, graph-history TPSI is 0.2323 (amount) and 0.0898 (rule), only slightly below transaction-only values 0.2486 and 0.0907. Thus the old claim of 3.8–6.3-fold greater graph invariance is not reproduced. The supported result is that graph-history features are necessary for useful nnPU ranking in this configuration, not that they generally guarantee policy invariance.

In Stage 4, the mean number of visible blind-typology positives rises from 0.0 at $p_{\text{explore}} = 0$ to 2.9, 15.8, and 29.0 as exploration increases. Nevertheless, blind Recall@500 is non-monotonic: estimated nnPU scores 0.1660, 0.1647, 0.1665, and 0.1664; LightGBM scores 0.1605, 0.1672, 0.1690, and 0.1598. BCE is lower and highly variable (0.0347 at zero support; seed SD 0.0540). Thus the titration establishes additional labeled support but does not demonstrate that nominal exploration improves ranking within this range. At zero support, nnPU’s 0.1660 recall is best understood as inductive transfer from features learned on other typologies. It cannot validate blind-stratum identification. Estimated nnPU also does not dominate LightGBM overall: at zero exploration their overall recalls are 0.1662 and 0.1748.

7 Discussion

The corrected results support three conclusions. Selective observation can change model ranking and apparent performance even when the latent holdout is fixed. A carefully implemented PU estimator can be competitive under benchmark assumptions, but naive IPW-BCE is not an adequate substitute. Finally, evaluation design is as consequential as model choice: unequal audit capacity and a mislabeled estimand produced a stronger but false initial headline.

Validation teams should retain audit-decision logs, distinguish decision from confirmation time, and report stress tests under plausible label policies. Random exploration

may help establish support, but this study does not prescribe a universal percentage. Real auditing carries legal, privacy, workload, and fairness constraints requiring domain-specific governance.

8 Limitations

Both benchmarks are synthetic. The assumptions $S = AY$, no unmeasured audit confounding, and positivity may fail in practice. The propensity model is not cross-fitted or formally calibrated. Ten seeds yield coarse exact tests. Graph-history features are engineered aggregates rather than a GNN. The IBM cache does not preserve the source background distribution without sampling weights. At zero blind-typology support, statistical correction is unidentified; recall reflects inductive extrapolation and may fail under a different feature–typology relation.

9 Conclusion

After correcting the estimator, cohorts, policy generator, and estimand, the defensible result is a 26.1%–29.3% training-policy utility spread and policy-dependent visible-label optimism—not the previously reported 39.7% deployment collapse. Estimated nnPU is robust in this synthetic delayed-feedback benchmark and closely tracks an oracle diagnostic, but real-world use depends on auditable selection logs and strong identification assumptions. The broader contribution is a reproducible way to expose those assumptions and failures before deployment claims are made.

Acknowledgments. The author thanks the creators of the TransXion and IBM AMLWorld benchmarks.

Funding. No external funding was received.

Competing Interests. The author declares no competing financial or non-financial interests.

Ethics Approval. Not applicable. The experiments use synthetic benchmark data.

Consent to Participate. Not applicable.

Consent for Publication. Not applicable.

Availability of Data and Materials. TransXion is available at <https://github.com/chaos-max/TransXion>; IBM AMLWorld data are available at <https://github.com/IBM/AML-Data>. Cache provenance and sampling are recorded in supplementary scripts and manifests.

Code Availability. Corrected scripts, compact results, and manifests accompany the manuscript as supplementary material. No public repository is claimed at submission.

Authors’ Contributions. M.E. conceived the study, designed the benchmark, implemented the experiments, analyzed results, and wrote the manuscript.

Use of Generative AI. OpenAI Codex assisted code auditing, experiment orchestration, and language revision. The author is responsible for verifying implementation, results, citations, and final text.

References

- [1] Lin, C.-C., Li, G., Hung, C.-C., Kuo, Y.-C., Peng, W.-C.: Rethinking suspicious account identification from a perspective of machine learning ranking. *International Journal of Machine Learning and Cybernetics* **17**, 84 (2026) <https://doi.org/10.1007/s13042-025-02945-x>
- [2] Abolfadl, A.: KONTOTRAPH: Verified point-in-time feature consistency and amortised explanation for real-time anti-money laundering under a 200 ms decision budget. *arXiv preprint arXiv:2608.22389* (2026)
- [3] Deng, J.: When can fraud operations authorize automation? a decision-support framework for fresh audit evidence and review workload. *arXiv preprint arXiv:2608.08577* (2026)
- [4] Dhama, G.: The fundamental limits of fraud detection in card payment networks. *arXiv preprint arXiv:2605.27557* (2026)
- [5] Financial Crimes Enforcement Network: Frequently Asked Questions Regarding Suspicious Activity Reporting Requirements. <https://www.fincen.gov/resources/statutes-regulations/guidance/frequently-asked-questions-regarding-suspicious-activity>. Accessed 7 September 2026 (2025)
- [6] Altman, E., Blanuša, J., Niederhäusern, L., Egressy, B., Anghel, A., Atasu, K.: Realistic synthetic financial transactions for anti-money laundering models. In: *Advances in Neural Information Processing Systems*, vol. 36, pp. 52635–52652 (2023)
- [7] Chen, K., Jiang, M., Zhao, Y., Li, Z., Chen, Z., Luo, W., Li, Z., Liu, S., Jing, Y., Ye, G., Wu, X., Chai, H.: Transxion: A high-fidelity graph benchmark for realistic anti-money laundering. In: *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 8707–8718 (2026). <https://doi.org/10.1145/3770855.3817550> . Code available at <https://github.com/chaos-max/TransXion>
- [8] Wang, S., Wang, P., Wu, B., Zhu, Y., Luo, W., Pan, Y.: Structural entropy minimization combining graph representation for money laundering identification. *International Journal of Machine Learning and Cybernetics* **15**, 3951–3968 (2024) <https://doi.org/10.1007/s13042-024-02129-z>
- [9] Berti, A., Restelli, M., Penz, D., Morini, M.: Elliptic2: A subgraph-level anti-money laundering benchmark on financial transaction networks. In: *Proceedings*

of the ACM Web Conference, pp. 3412–3422 (2024)

- [10] Li, M., Jia, L., Su, X.: Global-local graph attention with cyclic pseudo-labels for bitcoin anti-money laundering. *Scientific Reports* **15**, 22668 (2025) <https://doi.org/10.1038/s41598-025-08365-9>
- [11] Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220 (2008)
- [12] Bekker, J., Davis, J.: Learning from positive and unlabeled data: A survey. *Machine Learning* **109**(4), 719–760 (2020)
- [13] Kiryo, R., Niu, G., Plessis, M.C., Sugiyama, M.: Positive-unlabeled learning with non-negative risk estimator. In: *Advances in Neural Information Processing Systems*, vol. 30, pp. 1675–1685 (2017)
- [14] Bekker, J., Robberechts, P., Davis, J.: Beyond the selected completely at random assumption for learning from positive and unlabeled data. In: *Machine Learning and Knowledge Discovery in Databases*, pp. 71–85. Springer, ??? (2019). https://doi.org/10.1007/978-3-030-46147-8_5
- [15] Liang, X., Zhu, K., Xiao, A., Wen, Y., Zhang, K., Wang, S., Jian, L.: ROPU: A robust online positive-unlabeled learning algorithm. *Knowledge-Based Systems* **309**, 112808 (2025) <https://doi.org/10.1016/j.knosys.2024.112808>
- [16] Ingemarsson, F., Karlsson, E.: Positive-unlabeled graph learning for financial crime detection. *DiVA Academic Archive* (2026). *diva2:2072977*
- [17] Deprez, B., Wei, W., Verbeke, W., Baesens, B., Mets, K., Verdonck, T.: Advances in continual graph learning for anti-money laundering systems: A comprehensive review. *arXiv preprint arXiv:2503.24259* (2025)
- [18] Zhang, Q., Liu, L., Lan, T., Chen, W., Chen, R., Zhang, X.: Role-aware continual learning for robust anti-money laundering detection. *Expert Systems with Applications* **299**, 130136 (2026) <https://doi.org/10.1016/j.eswa.2025.130136>
- [19] Fan, H., Jiao, Y., Wang, M., Chen, J., Yang, S.: Fraud learns too: Continual graph learning under strategic adversarial drift in dynamic networks. *Scientific Reports* **16** (2026) <https://doi.org/10.1038/s41598-026-60997-7> . Advance online publication
- [20] Liu, K.: Driftguard-hcl: Concept-drift-aware continual heterogeneous graph contrastive learning for evolving financial fraud detection. *Journal of Mathematical Finance and Risk Management* **1**(2), 1–7 (2026) <https://doi.org/10.54097/x fj4xz68>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PolicyShiftAMLSupplementaryMaterial.zip](#)