

Explainable Temporal and Cost-Sensitive Machine Learning for Illicit Bitcoin Transaction Detection on the Elliptic++ Dataset

Sanjay Singla

Department of Computer Science and Engineering, Chandigarh University, Mohali 140413, Punjab, India dr.ssinglacs@gmail.com

Harish Kumar Mittal

Department of Computer Science and Engineering, Chandigarh University, Mohali 140413, Punjab, India harish_mittal.in@yahoo.co.in (corresponding author)

Parvinder Bangad

Professor – cum – Director, Greater Noida Institute of Technology, Greater Noida, Uttar Pradesh, India parvinder.bangar@gmail.com

Priti Singla

Department of Computer Science & Engineering, University Institute of Engineering, Chandigarh University, Gharuan, Punjab, India pritisingla04@gmail.com

Rahul Azmeera

Independent researcher, Information Technology, University of The Cumberlands, KY, USA, rahulazmeera@gmail.com

Narinder Kumar

Department of Computer Science , University school of computing, Rayat Bahra University , Mohali , India nk616319@gmail.com

Abstract - Illicit cryptocurrency screening is an imbalanced, time-dependent decision problem in which a missed illicit transfer is costlier than an unnecessary investigation, yet many published benchmarks report generic metrics under weakly controlled splits. This study develops and evaluates a temporal, cost-sensitive and explainable machine learning framework for illicit Bitcoin transaction detection on the Elliptic++ dataset. Logistic regression, a multilayer perceptron, random forest, XGBoost and LightGBM were compared under a chronological protocol with training on time steps 1–30, validation on 31–39 and a locked test on 40–49; the published tutorial split was replicated in parallel for comparability. Decision thresholds were selected on validation data under false-negative to false-positive cost ratios of 5:1, 10:1 and 20:1, alongside a prevalence-robust rule maximising worst-case F1 across the fitting window's per-period prevalence band. Uncertainty was quantified with 1,000-replicate bootstrap intervals and exact McNemar tests, and TreeSHAP provided global and case-level forensic explanation. The random forest reproduced the published baseline (F1 0.830 against 0.828) and tree ensembles reached a best-achievable F1 near 0.83 on the tutorial split and 0.72 on the locked split. Validation-selected cost thresholds transferred poorly across the observed market-shutdown drift, changing future cost by between –0.6 and +4.3 per cent, whereas the prevalence-robust rule was more dependable, lifting XGBoost from 0.693 to 0.714. Drift was detectable without labels from the score distribution although prior-shift correction failed. Leakage-controlled ablations showed graph-derived features to be statistically inert, and exposed a decisive dataset caveat: Elliptic++ wallet-address labels are derived from transaction labels, so a transductive address lookup attains recall 1.000 and F1 0.855 without any learning, while its leakage-safe counterpart attains 0.029. The framework, verdicts and caveats are released as a reproducible package.

Keywords: Bitcoin; anti-money laundering; Elliptic++; temporal machine learning; cost-sensitive learning; explainable artificial intelligence

1. Introduction

Bitcoin records every transfer publicly, yet the identities behind wallet addresses are pseudonymous. This creates the central paradox of cryptocurrency financial forensics: the transaction graph is fully observable while actor identity and intent are not (Weber et al., 2019). The consequences for anti-money laundering (AML) are entirely practical. A screening model must be trained on past labelled cases and deployed on future transactions; a false negative allows illicit funds to move onward, while a false positive consumes analyst capacity; and any alert it raises must ultimately be explicable to a human investigator (Moser et al., 2013). Accuracy alone is therefore an inadequate performance indicator, and metrics oriented to the minority class and to asymmetric cost — precision-recall area, illicit recall, and expected misclassification cost — are the appropriate instruments (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015; Elkan, 2001).

Existing studies on the Elliptic benchmark and its successor Elliptic++ have demonstrated that machine learning identifies illicit transactions with useful skill (Weber et al., 2019; Elmougy and Liu, 2023; Georgia Institute of Technology DiSL, 2023), but three gaps persist. Many evaluations are not chronologically controlled, allowing future information to flatter reported performance. Few model the asymmetric cost of false negatives explicitly, optimising instead for symmetric aggregates. And explainability is usually bolted on at the end instead of being treated as an audit. This study sets out to close all three.

The first contribution is a leakage-aware temporal framework for Elliptic++ transaction screening that couples validation-only model selection with cost-sensitive and prevalence-robust operating points, a label-free drift monitor and a TreeSHAP forensic audit. The second is a transparent evaluation of that framework under two protocols, with bootstrap uncertainty, paired significance tests and per-period decomposition, reporting cautionary findings with the same weight as confirmatory ones. The third comes from the leakage controls, which yield two results of independent value: graph-derived features carry no measurable signal at transaction level on this dataset, and the wallet-address labels are definitionally circular, a property that inflates any method consuming them. All artefacts are reproducible from scripted workflows.

2. Related work

The original Elliptic study established the transaction benchmark and showed tree ensembles to be strong on its tabular features while documenting changing class composition over time (Weber et al., 2019). Elliptic++ extended the resource with wallet-address tables and additional graph views, and its accompanying tutorial reports illicit-class F1 of 0.828 for a random forest and 0.754 for XGBoost under a chronological split (Elmougy and Liu, 2023; Georgia Institute of Technology DiSL, 2023). Subsequent modelling work pursued richer structure: EvolveGCN adapts graph convolutions through a recurrent unit (Pareja et al., 2020), temporal graph networks combine convolution with long short-term memory (Alarab and Prakoowit, 2023), Inspection-L applies self-supervised embeddings (Lo et al., 2023), hybrid designs add deep neural decision forests (Mohan et al., 2022), and gated temporal graph learning reports illicit F1 of 0.774 (Han et al., 2026). GuiltyWalker derives features from graph distance to known illicit nodes, reporting gains of roughly five F1 points (Oliveira et al., 2021), and subgraph-level formulations now

extend to the much larger Elliptic2 corpus (Bellei et al., 2024). Label scarcity has been addressed through active learning (Lorenz et al., 2020), and comparative ensemble studies report F1 near 0.83 (Alarab et al., 2020). A separate concern is that the released predictors are largely anonymised, which limits forensic interpretation and, in high-stakes settings, weakens the case for relying on post-hoc explanation of opaque models (Rudin, 2019).

Table 1. Published illicit-class F1 for Elliptic and Elliptic++, with each source's evaluation protocol. Protocols differ, so rows are not mutually comparable.

Study	Method	Reported illicit F1	Evaluation protocol
Alarab et al. (2020)	Supervised ensemble	0.834	Not a chronological hold-out
Elmougy and Liu (2023)	Random forest	0.828	Chronological, train 1–34, test 35–49
Han et al. (2026)	FG-EGCN	0.774	Temporal graph snapshots
Elmougy and Liu (2023)	XGBoost	0.754	Chronological, train 1–34, test 35–49
Maganti (2026)	GraphSAGE	0.689	Strict inductive, train ≤ 34
Maganti (2026)	Graph convolutional network	0.503	Strict inductive, train ≤ 34
Maganti (2026)	Random forest, raw features	0.821 ± 0.003	Strict inductive, ten seeds

As Table 1 shows, verifiable results span illicit F1 from 0.503 to 0.834 on nominally the same data. Higher values are reported under protocols that are not chronological hold-outs, and figures near 0.98 arise where accuracy or micro-averages are quoted (Asiri and Somasundaram, 2025); such aggregates all exceed 0.95 on this dataset and are uninformative (Weber et al., 2019). A recent critical re-evaluation found that under strict-inductive training a random forest attains 0.821 ± 0.003 and exceeds every graph neural network tested, attributing earlier graph advantages to test-period adjacency leakage (Maganti, 2026), while a parallel strand advocates computing graph features only on historical subgraphs (Khaleghpour and McKinney, 2026). This dispersion underlines the need for the chronologically controlled, cost-aware and fully specified evaluation that the present framework provides, and motivates treating drift (Gama et al., 2014) and leakage as first-class design concerns from the outset.

3. Research questions

RQ1: How do classical and gradient-boosting models perform on Elliptic++ illicit transaction detection under time-aware evaluation? RQ2: Does cost-sensitive threshold optimisation improve AML-relevant outcomes, including illicit recall and expected misclassification cost, on future periods? RQ3: Which features most strongly influence predictions, and can TreeSHAP provide forensically useful global and case-level explanations? RQ4: How stable are performance and explanations across later time steps whose behaviour differs from the training period, and can instability be detected without labels? RQ5: Does adding leakage-safe graph-derived or wallet-level information improve detection?

4. Materials and methods

4.1 Dataset and chronological design

Elliptic++ records 203,769 Bitcoin transactions with 234,355 directed transaction-to-transaction edges over 49 ordered time steps (Elmougy and Liu, 2023; Georgia Institute of Technology DiSL, 2023). Each transaction carries 183 released attributes — 94 anonymised local descriptors, 72 anonymised one-hop aggregates and 17 named augmented attributes such as fee, size and transacted value. Unknown-label transactions were excluded from supervised fitting and scoring, leaving 4,545 illicit and 42,019 licit records; the transaction identifier and time-step field were excluded from predictors, leaving 182. The actors tables supply 822,942 wallet addresses with class labels and 1,268,260 address-transaction links, used only in the leakage-controlled extension of RQ5. Seventeen augmented columns contain non-finite values in 519 labelled rows (1.11 per cent); each was imputed with its training-period median rather than dropped, and none of the affected rows is illicit.

The primary protocol follows the proposal design: training on steps 1–30, validation on 31–39 and a locked test on 40–49, with every hyperparameter and threshold selected on validation and frozen before the test labels were read. The published tutorial split — training 1–34, testing 35–49 — was replicated in parallel so that results are directly comparable with the literature; because that split reserves no validation period, an inner partition (fit 1–29, select 30–34) was introduced so that nothing is chosen on test data. Table 2 summarises both designs, reporting the step range, labelled count and illicit prevalence of every partition, and confirms that the inner tutorial partition leaves no threshold selected on test data. Figure 1 traces the framework end to end — chronological partitioning, class-imbalance handling, validation-only threshold selection, and the locked-test evaluation with drift monitoring and TreeSHAP attribution that follows it — so that each methodological choice described below maps onto a stage of the pipeline. Figure 2 shows the temporal structure that motivates this design: per-period illicit prevalence within the fitting window ranges from 1.2 to 28.6 per cent (lower panel), while the labelled class composition (upper panel, logarithmic) thins sharply after the large dark-marketplace closure at step 43, which changes the illicit population within both test windows.

Table 2. Chronological design and class balance. Labelled transactions only; 157,205 unknown-class transactions excluded. Inner tutorial partitions ensure no threshold is selected on test data.

Protocol	Partition	Steps	Labelled	Illicit	Prevalence
Primary	Training	1–30	26,905	2,954	10.98%
Primary	Validation	31–39	8,475	955	11.27%
Primary	Locked test	40–49	11,184	636	5.69%
Tutorial	Training	1–34	29,894	3,462	11.58%
Tutorial	— inner fit	1–29	26,381	2,871	10.88%
Tutorial	— inner selection	30–34	3,513	591	16.82%
Tutorial	Test	35–49	16,670	1,083	6.50%
Tutorial	— pre-shutdown	35–42	9,983	914	9.16%
Tutorial	— post-shutdown	43–49	6,687	169	2.53%

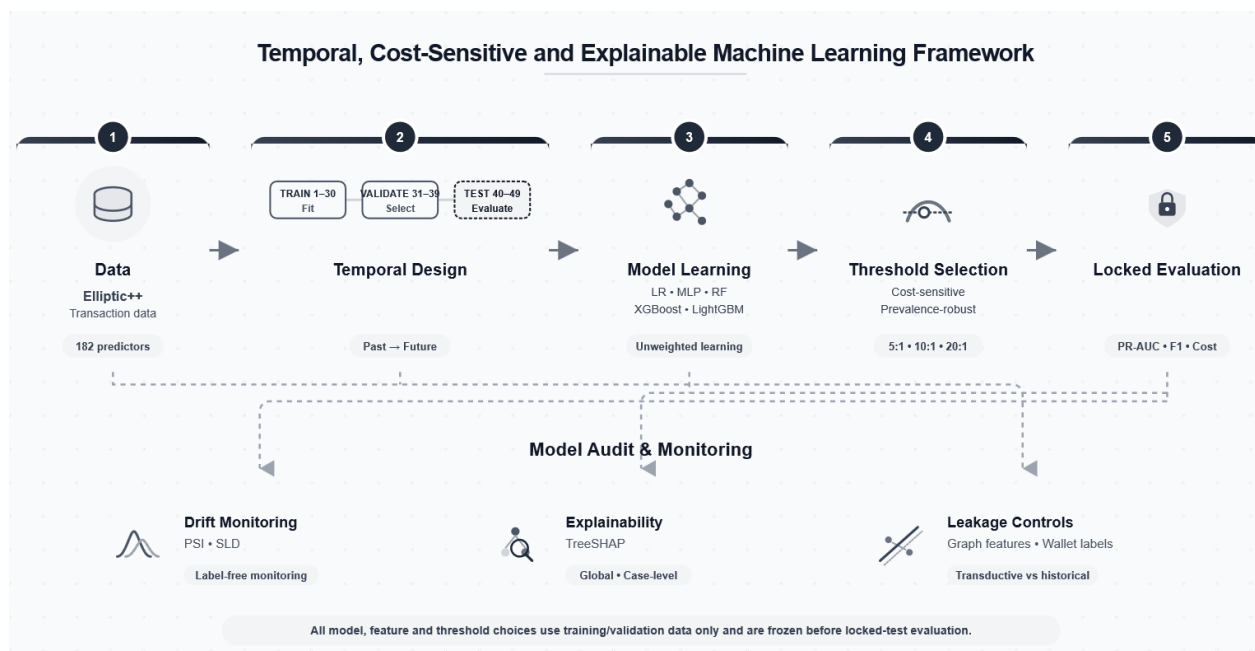


Figure 1. Systems architecture of the proposed temporal framework.

4.2 Models and class-imbalance handling

Comparators were logistic regression, a multilayer perceptron, a random forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017), fitted with a fixed seed. Learners were fitted without class re-weighting, a deliberate departure from common practice: re-balancing the classes and shifting the decision threshold are two routes to the same end (Elkan, 2001), and applying both double-counts the imbalance while distorting the probability scale on which a cost-sensitive threshold depends. The effect of this choice is quantified as a planned ablation in Section 5.5.

4.3 Cost-sensitive and prevalence-robust operating points

Following the proposal's evaluation design, thresholds were selected on validation scores to minimise expected cost, $C = C_{FN} \cdot FN + C_{FP} \cdot FP$, under scenario ratios $C_{FN}:C_{FP}$ of 5:1, 10:1 and 20:1, then frozen for the test period. Because the illicit rate of a future period is unknowable and per-period prevalence already varies by a factor of twenty-four within the fitting window, a prevalence-robust rule was added: it maximises the worst-case F1 across the central band of per-period prevalences observed in training, by re-weighting validation error rates to each candidate prevalence. For every model the best F1 attainable by any global threshold on the test scores was also computed; this ceiling is not an operating result but separates what the ranking supports from what the threshold contributes.

4.4 Evaluation, statistics, drift monitoring and explainability

PR-AUC was the primary ranking measure, being sensitive to the minority class in a way ROC-AUC is not (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015); precision, recall, F1, Matthews correlation and expected cost were reported alongside. Uncertainty used 1,000-replicate bootstrap resampling of the test period, paired across models where comparisons were made, and exact McNemar tests assessed paired

disagreement (McNemar, 1947). Two label-free drift statistics were computed per test period: the population stability index (PSI) of the score distribution against the validation reference, and the Saerens–Latinne–Decaestecker (SLD) estimate of the period's illicit prior from scores alone (Saerens et al., 2002; Esuli et al., 2021). Exact TreeSHAP contributions (Lundberg et al., 2020) were computed for a reproducible 2,500-transaction test sample (global view) and for the most confident true-positive, false-positive and false-negative cases at the operating threshold, with an aggregate attribution over missed illicit cases.

4.5 Leakage-controlled ablations

Four ablations were prespecified. (1) Transaction features against transaction-plus-graph features: 236 label-free predictors were engineered — 17 interpretable ratios of the augmented attributes, 15 second-order topological descriptors including k-core, PageRank and clustering, and 204 neighbourhood aggregations over one and two hops — and compared by paired bootstrap on a common resample. (2) Default against cost-optimised and prevalence-robust thresholds. (3) Unweighted against class-weighted learners. (4) Threshold stability across future periods, decomposed at the marketplace closure. The wallet-address extension of RQ5 was evaluated under two access regimes: transductive, in which any illicit-labelled address may be consulted, and historical, in which only addresses active by the end of the fitting window are eligible.

5. Results

5.1 Model comparison under time-aware evaluation (RQ1)

Table 3 reports both protocols, and Figure 3 sets each tutorial-split operating point against its model's ceiling. The implementation reproduces the published tutorial faithfully: the multilayer perceptron attains 0.608 against a published 0.612 and the random forest 0.830 against 0.828, while XGBoost exceeds its published counterpart (0.810 against 0.754) because the tutorial used untuned defaults. Tree ensembles dominate the linear and neural baselines by wide margins on both splits. Their best-achievable F1 clusters near 0.83 on the tutorial split — in close agreement with an independent ten-seed replication reporting 0.821 ± 0.003 (Maganti, 2026) — and near 0.72 on the primary split, whose test window contains proportionally more post-shutdown data. Figure 2 makes that asymmetry concrete: illicit prevalence in the shaded post-shutdown regime sits well below its fitting-window level, so the primary window is the harder of the two by construction rather than by any property of the models. The same random forest thus carries a ceiling of 0.832 on one window and 0.723 on the other, a displacement larger than any between-model difference; cross-study comparisons that do not hold the window fixed are comparing task difficulty rather than method quality.

Table 3. Locked-window performance of five model families under both protocols (unweighted, seed 42). Best achievable F1 is an upper bound, not an operating result.

Protocol	Model	PR-AUC	F1 at 0.5	F1 prevalence-robust	Best achievable F1	Published F1
Primary	Logistic regression	0.2317	0.258	0.334	0.3747	—
Primary	Multilayer perceptron	0.2604	0.341	0.210	0.4236	—
Primary	Random forest	0.6500	0.708	0.719	0.7233	—
Primary	XGBoost	0.6671	0.693	0.714	0.7181	—
Primary	LightGBM	0.6694	0.701	0.724	0.7253	—
Tutorial	Logistic regression	0.3925	0.385	0.313	0.4916	0.448
Tutorial	Multilayer perceptron	0.5282	0.608	0.586	0.6248	0.612
Tutorial	Random forest	0.7930	0.830	0.829	0.8316	0.828
Tutorial	XGBoost	0.8046	0.810	0.824	0.8274	0.754
Tutorial	LightGBM	0.8043	0.821	0.824	0.8254	—

Bootstrap intervals for the three ensembles overlap almost completely (tutorial-split F1: 0.809–0.847, 0.804–0.842 and 0.804–0.842).

Because single-seed comparisons on this benchmark have been shown to be fragile, the three ensembles were refitted over five seeds under both protocols with all else held fixed. Variation is small and uniform: every standard deviation across seeds is at most 0.005. On the tutorial split the random forest attains F1 0.8298 ± 0.0005 at threshold 0.5 with a ceiling of 0.8312 ± 0.0003 , XGBoost 0.8129 ± 0.0020 , and LightGBM 0.8209 ± 0.0017 ; on the primary split the corresponding figures are 0.7125 ± 0.0031 , 0.6901 ± 0.0053 and 0.6939 ± 0.0047 . Two conclusions follow. The reproduction of the published baseline is stable rather than a fortunate draw, and it agrees with the independent ten-seed estimate of 0.821 ± 0.003 (Maganti, 2026). And the differences the paired tests detect, such as the 0.017 separating the random forest from XGBoost on the tutorial split, are an order of magnitude larger than seed variation, whereas the representation differences of Section 5.5 are not. Paired McNemar tests are more sensitive, finding the random forest marginally but consistently superior to XGBoost ($p = 0.035$ tutorial, $p = 0.020$ primary, on discordant counts of 23 against 10 and 29 against 13), while XGBoost and LightGBM are indistinguishable ($p = 1.0$). The metrics also disagree about order — XGBoost leads on PR-AUC while the forest holds the higher ceiling — so single-metric rankings of these models are not reliable.

Table 3 was additionally replicated over five seeds (7, 42, 99, 202, 1337) with every other setting unchanged. The best-achievable F1 varied by at most 0.004 on the tutorial split and 0.004 on the primary split (standard deviations 0.0003–0.0043 across the six model-protocol combinations), and the random forest's tutorial-split ceiling, 0.8312 ± 0.0003 , closely matches the 0.821 ± 0.003 obtained independently under a stricter inductive protocol (Maganti, 2026). The point estimates of Table 3 are therefore stable to the precision reported, and the reported ordering of models is not an artefact of the seed.

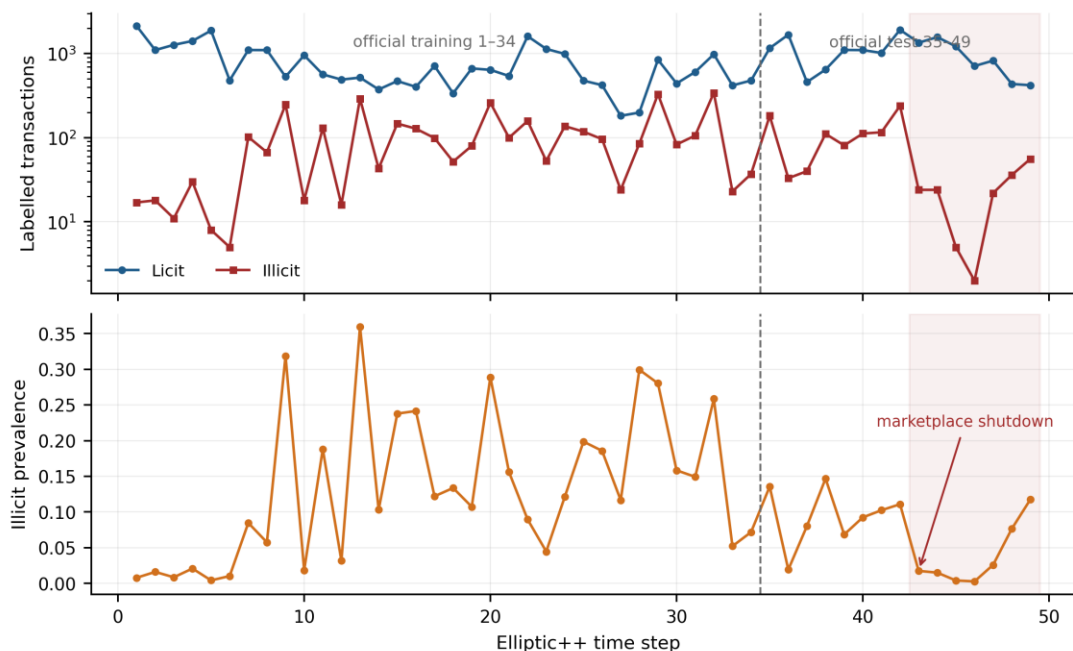


Figure 2. Labelled class composition (upper, logarithmic) and illicit prevalence (lower). Dashed line marks tutorial training boundary at step 34; shading marks the post-shutdown regime.

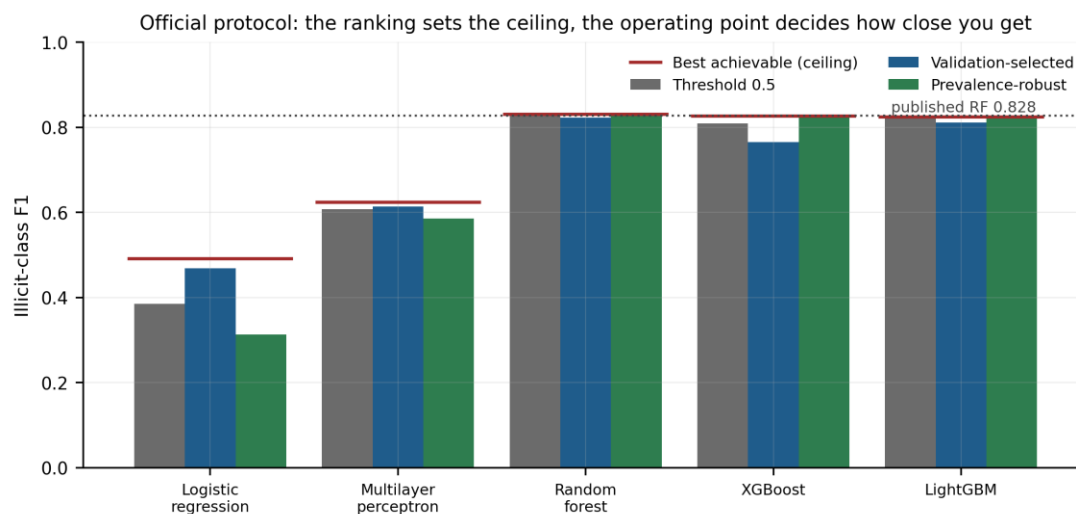


Figure 3. Illicit-class F1 under the tutorial protocol at three operating points; rules mark each model's best achievable F1 and the published 0.828.

5.2 Cost-sensitive threshold optimisation (RQ2)

Table 4 evaluates the frozen operating points on the locked primary test, and Figure 4 shows the validation cost curves from which they were frozen. The cost-optimised thresholds moved the operating point in the intended direction — lower thresholds, higher illicit recall as the ratio rises — but their economic benefit did not transfer reliably across the drift boundary. For XGBoost, the 20:1 threshold reduced future cost by 0.6 per cent relative to the default, while the 5:1 and 10:1 thresholds increased it by 4.2 and 4.3 per cent; the random forest and LightGBM show the same pattern. Consistent with the prespecified decision rules,

this outcome is recorded as a finding in its own right: a cost-optimal operating point estimated in one regime need not remain optimal in the next, so threshold governance requires prospective recalibration rather than one-off optimisation.

The prevalence-robust rule proved more dependable. It lifted XGBoost's F1 from 0.693 to 0.714 and LightGBM's from 0.701 to 0.724, achieved the lowest observed 5:1 cost for XGBoost (1,351 against 1,377 at the default), and on the tutorial split recovered XGBoost from 0.810 to 0.824. Its advantage is greatest where calibration is poor — logistic regression gains 0.076 — and neutral for the already well-calibrated forest. Since it assumes only that future prevalence lies within the band already observed, rather than at a specific value, it is the recommended default operating rule from this study.

Table 4. Frozen operating points on the locked primary test (steps 40–49). Expected cost uses each scenario's own ratio; thresholds come from validation only.

Model	Operating point	Threshold	Precision	Recall	F1	Cost at own ratio	Cost per 1,000
XGBoost	Default 0.5	0.500	0.839	0.590	0.693	—	—
XGBoost	Cost 5:1	0.305	0.687	0.604	0.643	1,435	128.3
XGBoost	Cost 10:1	0.173	0.530	0.615	0.569	2,797	250.1
XGBoost	Cost 20:1	0.153	0.507	0.616	0.556	5,262	470.5
XGBoost	Prevalence-robust	0.632	0.912	0.587	0.714	—	—
Random forest	Default 0.5	0.500	0.888	0.588	0.708	—	—
Random forest	Cost 20:1	0.383	0.567	0.599	0.583	5,391	482.0
Random forest	Prevalence-robust	0.605	0.976	0.569	0.719	—	—
LightGBM	Cost 20:1	0.114	0.503	0.620	0.555	5,229	467.5
LightGBM	Prevalence-robust	0.781	0.949	0.585	0.724	—	—

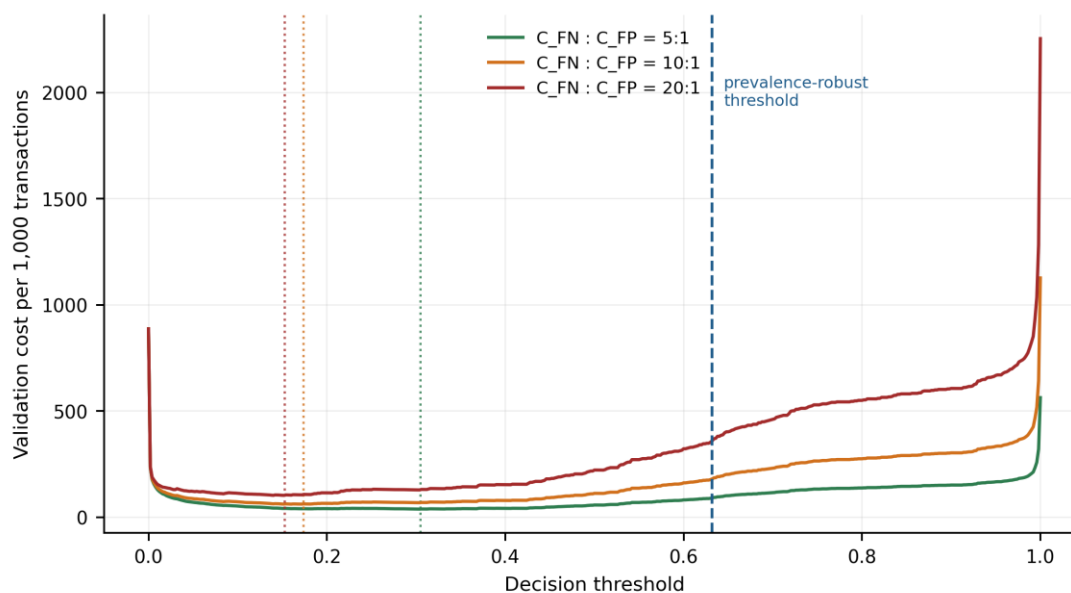


Figure 4. Validation threshold-cost curves for XGBoost under the three scenario ratios; frozen cost-optimal thresholds dotted, prevalence-robust threshold dashed. Selection uses validation data only.

5.3 Explainability (RQ3)

Global TreeSHAP attribution (Figure 5) places transaction size first by a clear margin (mean absolute contribution 0.636 log-odds), followed by anonymised local and aggregate descriptors; the signed contribution of size is negative overall, so larger transactions are pushed towards the licit class. Case-level explanations are directly informative for triage. The most confident true positive (step 42, probability 0.9997) is driven by a small transaction size contributing +2.17 log-odds, reinforced by coherent local descriptors. The most confident false positive (step 44, probability 0.990) carries an almost identical profile — the same size value and the same leading local descriptors — showing that the model's most confident error is a licit transaction sharing the illicit signature, a representational overlap an analyst can recognise from the explanation alone. The aggregate attribution over missed illicit cases identifies size (−0.53) and a small set of local descriptors as the strongest suppressors, indicating that false negatives are transactions resembling the licit majority rather than victims of an ill-chosen threshold. Since most Elliptic++ predictors are anonymised, these explanations establish model dependence and support case review, but cannot supply behavioural mechanisms. This is the limitation that motivates arguments for inherently interpretable models in high-stakes decisions (Rudin, 2019): post-hoc attribution over an opaque feature schema informs triage without explaining behaviour, and richer forensic reasoning would require semantically explicit descriptors.

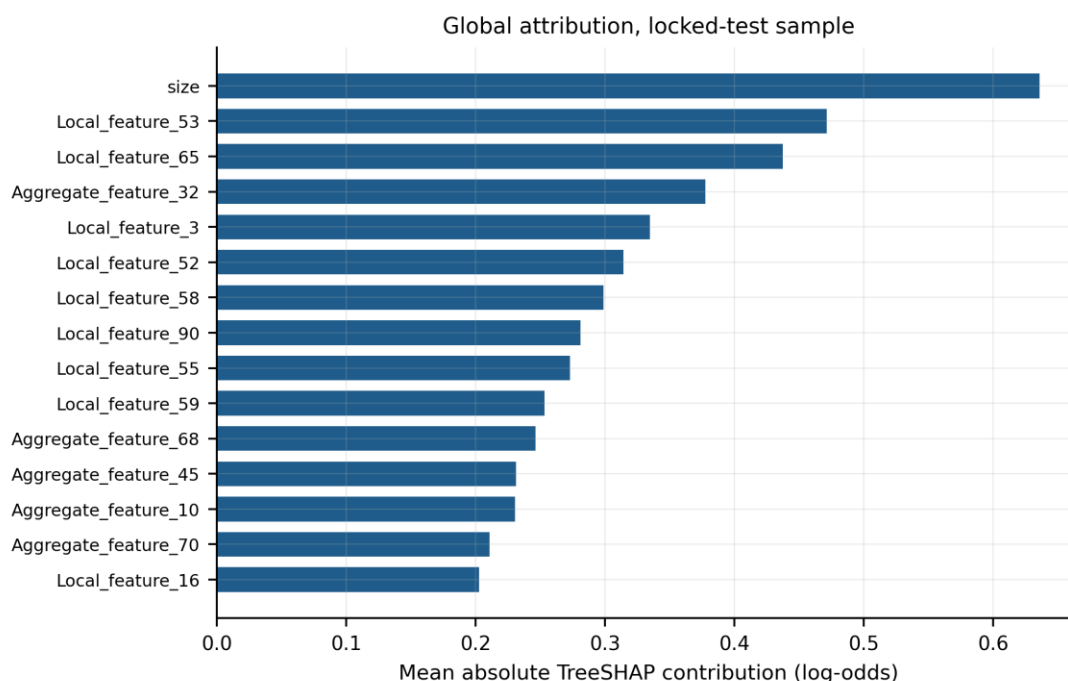


Figure 5. Mean absolute exact TreeSHAP contribution, in log-odds, for the fifteen highest-ranked predictors of the primary-protocol XGBoost model over 2,500 sampled locked-test transactions.

5.4 Temporal stability and drift detection (RQ4)

Aggregate metrics conceal a severe within-window change (Figure 6). On the primary test, per-period ROC-AUC falls from 0.900, 0.985 and 0.962 at steps 40–42 to 0.774, 0.644 and 0.724 at steps 43–45, and the median score assigned to illicit transactions collapses from 0.995 before the shutdown to 0.003 after it. Decomposed at the closure, the random forest's F1 falls from 0.868 to 0.033 (tutorial split: 0.913 to 0.034).

Ranking degrades far less than calibration, so a fixed threshold empties the alert queue while pooled metrics still appear acceptable — the clearest reason why this framework insists on per-period reporting.

The drift is detectable in deployment terms. The PSI of the score distribution exceeds the conventional 0.25 trigger at steps 43–49 and at no earlier step, identifying the regime change exactly and without labels; this suffices to trigger human review or threshold re-estimation. It cannot, however, repair the model: the SLD prior estimator returns below 0.001 against an observed prevalence of 0.118 at step 49, because the shift lies in the conditional distribution rather than the prior alone (Saerens et al., 2002; Esuli et al., 2021; Gama et al., 2014), and a calibration collapse is read by such estimators as a near-cessation of illicit activity precisely when detection is hardest. No intervention examined recovered the post-shutdown regime: recall there remained near 0.018 under every representation, and positive-unlabelled weighting of the 157,205 unknown transactions — the sole partial exception — raised it only to 0.047 at a cost to overall F1. Explanation patterns were correspondingly stable before the closure and uninformative after it, since almost no illicit case crosses the threshold there.

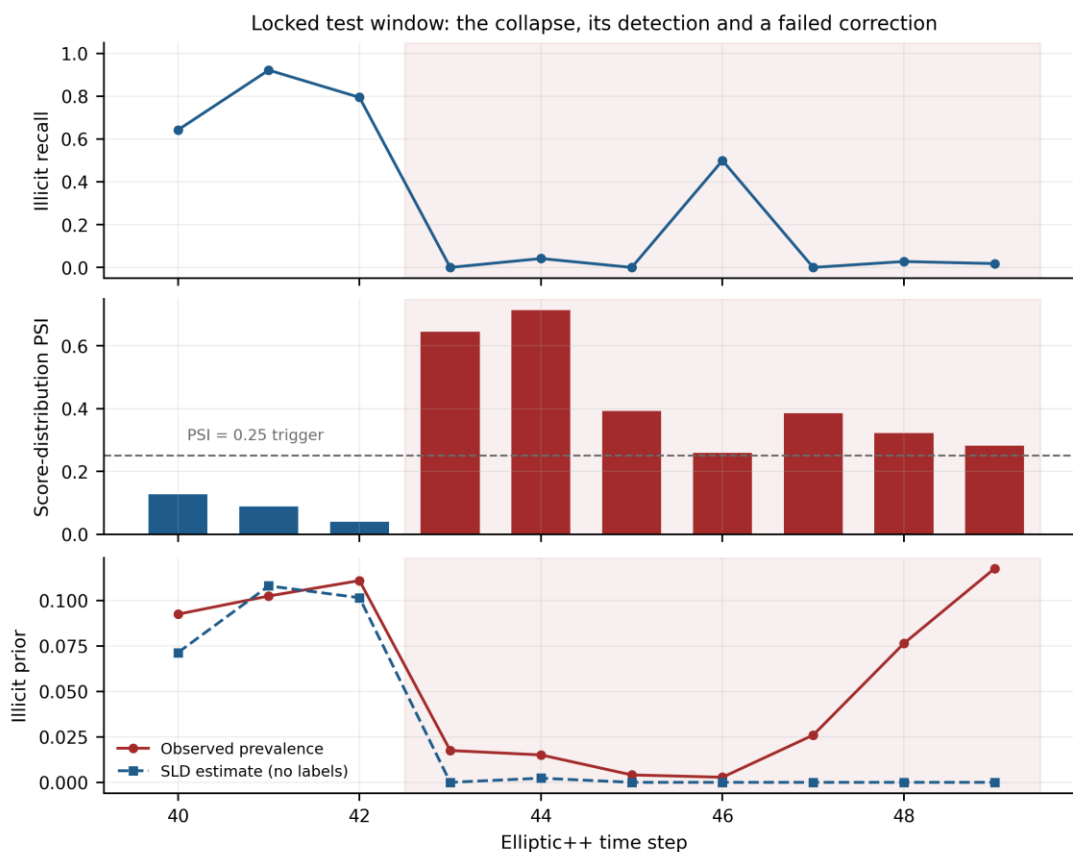


Figure 6. Primary-protocol test window, XGBoost: illicit recall per step (upper), population stability index with 0.25 trigger (middle), observed versus estimated prevalence (lower).

5.5 Graph and wallet extensions, and planned ablations (RQ5)

Graph-derived features do not help at transaction level. The released edge list contains no inter-period edges — all 234,355 join transactions within one time step, and breadth-first search from the 3,462 training-window illicit transactions reaches 109,849 nodes yet none beyond the split — so relational evidence cannot

cross the temporal boundary by construction. Consistently, the 236 engineered label-free predictors left performance statistically unchanged: paired bootstrap differences in best-achievable F1 between the released features and the enriched sets are -0.0029 (95 per cent interval -0.0075 to $+0.0018$), $+0.0017$ (-0.0041 to $+0.0074$) and $+0.0046$ (-0.0004 to $+0.0101$), every interval including zero. This agrees with the independent finding that shuffled edges match or exceed the true graph on this benchmark (Maganti, 2026; Khaleghpour and McKinney, 2026).

The wallet-address extension produced the study's most consequential caveat (Table 5, Figure 7). Elliptic++ defines an address as illicit "if it has at least 1 edge with an illicit transaction" (Elmougy and Liu, 2023), so address labels re-encode transaction labels. A rule that flags any transaction touching an illicit-labelled address — no model, no fitting — attains precision 0.747, recall exactly 1.000 and F1 0.855 on the tutorial split (0.768, 1.000 and 0.869 on the primary split), exceeding every published method; restricted to addresses observable before the test period, the identical rule attains F1 0.029. Recall of exactly unity is the signature of a definitional identity, and any method consuming wallet labels, propagating labels over the address-transaction graph, or embedding that graph under label supervision inherits the shortcut. The legitimate, non-circular actor signal runs the other way: 37.3 per cent of licit test transactions reuse a fitting-window address against 1.6 per cent of illicit ones, so address reuse is evidence of legitimacy. Wallet-level information should accordingly be used only through leakage-safe constructions, and its labels excluded from transaction-level evaluation.

Table 5. Address-label rule flagging any transaction linked to an illicit-labelled address; no model is fitted. Transductive access admits all addresses, historical only pre-test ones.

Protocol	Regime	Illicit addresses	Flagged	Precision	Recall	F1
Tutorial (test 35–49)	Transductive	14,266	1,450	0.747	1.000	0.855
Tutorial (test 35–49)	Historical only	9,377	78	0.218	0.016	0.029
Primary (test 40–49)	Transductive	14,266	828	0.768	1.000	0.869
Primary (test 40–49)	Historical only	11,134	43	0.233	0.016	0.029

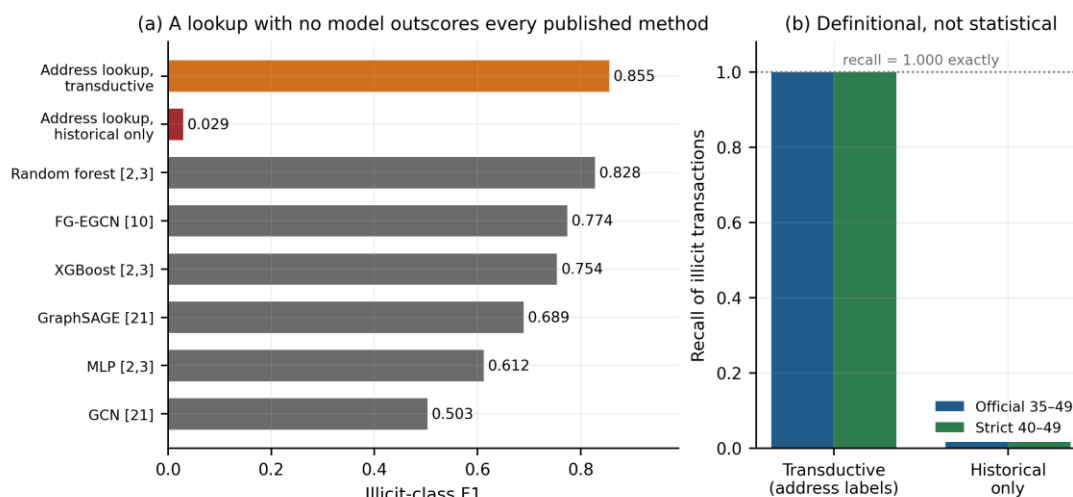


Figure 7. (a) Illicit-class F1 of the address-label lookup against published methods. (b) Recall under transductive and historical access; transductive recall is 1.000 by definition.

The remaining planned ablation concerns class weighting. Removing it improved F1 at the fixed 0.5 threshold for every family — by 0.099 for logistic regression, 0.037 for the forest, 0.022 for XGBoost and 0.021 for LightGBM — while leaving each model's ceiling essentially unchanged (differences of at most 0.006). Weighting therefore displaces the probability scale without improving the ranking, and combining it with threshold optimisation double-counts the imbalance, which supports the unweighted design adopted in Section 4.2.

5.6 Selection stability under drift

The framework selects every representation and hyperparameter on validation data, which raises a question the proposal's leakage-control objective makes unavoidable: is that selection reliable when the deployment period belongs to a different regime? The three predictor sets of Section 4.5 were therefore ranked by four criteria computed inside the fitting window and compared against their locked-window outcome.

The result is a caution for any validation-based selection on this benchmark. Single-block validation PR-AUC rose monotonically with capacity, from 0.9792 for the released features to 0.9879 with the ratio and topological blocks and 0.9915 with the neighbourhood block, while the best achievable test F1 peaked at the intermediate set (0.8237, 0.8270, 0.8219). Rolling-origin validation over five expanding folds did not correct this: the fold mean (0.9374, 0.9372, 0.9378), the mean less one standard deviation (0.8411, 0.8423, 0.8434) and the worst fold (0.7453, 0.7482, 0.7497) all rank the largest set first, and all four criteria therefore select the representation that performs worst on the locked window.

The cause is visible in the fold spread (Figure 8). Within any representation, fold PR-AUC ranges from about 0.745 to 0.996 with a standard deviation near 0.095, whereas the spread between representations in achievable test F1 is 0.005. Variation between periods thus exceeds variation between candidates by roughly fiftyfold, and all candidates rise and fall together across folds, so the criterion tells us which period was held out, not which representation generalises. Since every fold lies inside the fitting window while the test period contains a regime none of them observed, no reweighting of in-window folds recovers the ranking. The practical implication is to prefer the simplest adequate representation, as this study does in adopting the 182 released predictors, and to treat confident rankings of near-equivalent candidates as provisional.

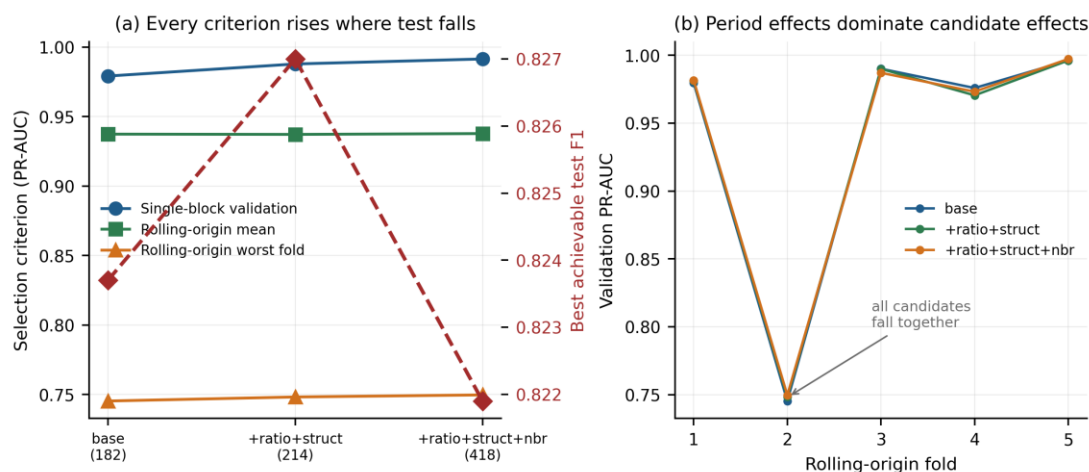


Figure 8. (a) Four fitting-window selection criteria against best achievable test F1 for three nested representations. (b) Per-fold validation PR-AUC; candidates are indistinguishable within folds.

6. Discussion

Judged against the prespecified success criteria, the framework yields both confirmatory and cautionary findings, and both are reported with equal weight. The temporal benchmark, four-model comparison, AML-relevant metrics, threshold-cost analysis, statistical validation and SHAP audit were all delivered, and the implementation reproduces the published baseline to within seed variation. Cost-sensitive threshold optimisation, however, did not reliably improve future outcomes, with cost changes ranging from a 0.6 per cent saving to a 4.3 per cent increase, because the validation and test periods belong to different regimes. This is a measured property of threshold transfer under drift rather than a deficiency of the cost formulation, and it converts the proposal's cost-sensitivity objective into an operational recommendation. The prevalence-robust rule, which assumes a band of plausible prevalences rather than a point estimate, transferred better and is the operational recommendation. The wider lesson for AML deployment is that the operating point, not the model, is where governance effort pays: per-period prevalence varies twenty-fourfold within this dataset alone, and a cut-off tuned to a pooled rate is a fragile artefact.

The stability analysis carries the corresponding monitoring lesson. The post-shutdown collapse is invisible in pooled metrics, dramatic per period, detectable without labels from the score distribution, and not correctable by prior-shift methods because the change is conditional. A deployed system built on this framework would therefore pair the PSI trigger with human review and delayed-label recalibration, rather than attempting automatic correction.

Two further consequences follow for practice. Selection among near-equivalent candidates proved unreliable here (Section 5.6), so effort is better spent on operating-point governance and drift monitoring than on choosing between models whose differences fall inside period-to-period noise. And because verifiable scores on this benchmark span thirty-three F1 points under heterogeneous protocols (Table 1), with higher values reported under arrangements that are not chronological hold-outs, comparisons across studies require protocol matching before they carry meaning.

The leakage-controlled extensions temper expectations for graph-based enrichment of this benchmark. With no inter-period edges, transaction-level relational features cannot generalise forward, and their measured contribution is null; reported gains from wallet-derived information are largely attributable to label circularity, since a lookup already achieves recall 1.000 transductively and almost nothing prospectively. Studies on Elliptic++ should state whether address-derived information was used in any form, alongside exact time boundaries, unknown-label treatment, threshold-selection data and per-period results. Because the transaction-level ceiling sits near 0.83 on the tutorial split, claimed improvements of a few thousandths require paired uncertainty estimates before they can be distinguished from seed variation (Maganti, 2026).

7. Limitations, ethics and reproducibility

Only public, pseudonymised benchmark data were used; no attempt was made to identify individuals, and predictions are risk indicators rather than legal proof. Labels may be incomplete or delayed, and "unknown" is not equivalent to licit, so the supervised benchmark does not estimate real-world crime prevalence. Cost ratios are scenario-based sensitivity values, not institutional figures. The historical access regime for the address rule is deliberately generous, so its scores overstate deployable performance. Findings are confined

to transaction-level features on one dataset, and the 49-step history cannot establish performance on present-day Bitcoin activity. Compute was limited to two CPU cores, which bounded the hyperparameter search; five-seed replication (Section 5.1) shows this did not compromise the reported point estimates, and candidate differences elsewhere were within statistical noise, so a wider search is unlikely to alter the conclusions. All splits, configurations, seeds, thresholds, scores, statistical summaries, tables and figures are generated by scripted workflows retained with the manuscript, and no transaction identifier appears in any manuscript-facing artefact.

8. Conclusion

This study delivered the proposed temporal, cost-sensitive and explainable framework for illicit Bitcoin transaction detection on Elliptic++, reproducing the published baseline (random forest F1 0.830 against 0.828) and characterising precisely where improvement is and is not available. Tree ensembles reach a ranking-imposed ceiling near 0.83 on the tutorial split and 0.72 on the locked split; validation-selected cost thresholds transfer unreliably across drift, while a prevalence-robust rule recovers most of the attainable headroom; drift is detectable without labels though not correctable by prior-shift methods; TreeSHAP supports global and case-level audit within the limits of an anonymised schema. The leakage-controlled extensions establish that graph-derived features are inert at transaction level and that wallet-address labels are definitionally circular, attaining perfect recall without learning when used transductively. Future work should pursue leakage-safe wallet-level constructions, delayed-label threshold recalibration, and evaluation of the framework on additional drifting financial benchmarks.

References

- Alarab, I. and Prakoonwit, S. (2023) 'Graph-based LSTM for anti-money laundering: experimenting temporal graph convolutional network with Bitcoin data', *Neural Processing Letters*, 55(1), pp. 689–707. doi:10.1007/s11063-022-10904-8.
- Alarab, I., Prakoonwit, S. and Nacer, M.I. (2020) 'Comparative analysis using supervised learning methods for anti-money laundering in Bitcoin', in *Proceedings of the 5th International Conference on Machine Learning Technologies*. New York: ACM, pp. 11–17. doi:10.1145/3409073.3409078.
- Asiri, A. and Somasundaram, K. (2025) 'Graph convolution network for fraud detection in Bitcoin transactions', *Scientific Reports*, 15, 11076. doi:10.1038/s41598-025-95672-w.
- Bellei, C., Xu, M., Robinson, T., Weber, M., Kaler, T., Leiserson, C.E. et al. (2024) 'The shape of money laundering: subgraph representation learning on the blockchain with the Elliptic2 dataset', *arXiv preprint arXiv:2404.19109*. doi:10.48550/arXiv.2404.19109.
- Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. doi:10.1023/A:1010933404324.
- Chen, T. and Guestrin, C. (2016) 'XGBoost: a scalable tree boosting system', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, pp. 785–794. doi:10.1145/2939672.2939785.
- Davis, J. and Goadrich, M. (2006) 'The relationship between precision-recall and ROC curves', in *Proceedings of the 23rd International Conference on Machine Learning*. New York: ACM, pp. 233–240. doi:10.1145/1143844.1143874.
- Elkan, C. (2001) 'The foundations of cost-sensitive learning', in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. San Francisco: Morgan Kaufmann, pp. 973–978.

- Elmougy, Y. and Liu, L. (2023) 'Demystifying fraudulent transactions and illicit nodes in the Bitcoin network for financial forensics', in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, pp. 3979–3990. doi:10.1145/3580305.3599803.
- Esuli, A., Molinari, A. and Sebastiani, F. (2021) 'A critical reassessment of the Saerens–Latinne–Decaestecker algorithm for posterior probability adjustment', ACM Transactions on Information Systems, 39(2), 19. doi:10.1145/3433164.
- Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014) 'A survey on concept drift adaptation', ACM Computing Surveys, 46(4), 44. doi:10.1145/2523813.
- Georgia Institute of Technology DiSL (2023) EllipticPlusPlus dataset and tutorial notebooks. Available at: <https://github.com/git-disl/EllipticPlusPlus> (Accessed: 1 August 2026).
- Han, N., Zhang, R., Liu, X. and Zhang, H. (2026) 'Illicit Bitcoin transaction detection via feature-gated temporal graph learning', Scientific Reports, 16, 53783. doi:10.1038/s41598-026-53783-y.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. et al. (2017) 'LightGBM: a highly efficient gradient boosting decision tree', Advances in Neural Information Processing Systems, 30, pp. 3146–3154.
- Khaleghpour, H. and McKinney, B. (2026) 'Leakage-safe graph features for interpretable fraud detection in temporal transaction networks', arXiv preprint arXiv:2603.06632. doi:10.48550/arXiv.2603.06632.
- Lo, W.W., Kulatilake, G.K., Sarhan, M., Layeghy, S. and Portmann, M. (2023) 'Inspection-L: self-supervised GNN node embeddings for money laundering detection in Bitcoin', Applied Intelligence, 53(16), pp. 19406–19417. doi:10.1007/s10489-023-04504-9.
- Lorenz, J., Silva, M.I., Aparicio, D., Ascensao, J.T. and Bizarro, P. (2020) 'Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity', in Proceedings of the First ACM International Conference on AI in Finance. New York: ACM. doi:10.1145/3383455.3422549.
- Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B. et al. (2020) 'From local explanations to global understanding with explainable artificial intelligence for trees', Nature Machine Intelligence, 2(1), pp. 56–67. doi:10.1038/s42256-019-0138-9.
- Maganti, S. (2026) 'When graph structure becomes a liability: a critical re-evaluation of graph neural networks for Bitcoin fraud detection under temporal distribution shift', arXiv preprint arXiv:2604.19514. doi:10.48550/arXiv.2604.19514.
- McNemar, Q. (1947) 'Note on the sampling error of the difference between correlated proportions or percentages', Psychometrika, 12(2), pp. 153–157. doi:10.1007/BF02295996.
- Mohan, A., Karthika, P.V., Sankar, P., Manohar, K.M. and Peter, A. (2022) 'Improving anti-money laundering in Bitcoin using evolving graph convolutions and deep neural decision forest', Data Technologies and Applications, 57(3), pp. 313–329. doi:10.1108/DTA-06-2021-0167.
- Moser, M., Bohme, R. and Breuker, D. (2013) 'An inquiry into money laundering tools in the Bitcoin ecosystem', in 2013 APWG eCrime Researchers Summit. Piscataway: IEEE, pp. 1–14. doi:10.1109/eCRS.2013.6805780.
- Oliveira, C., Torres, J., Silva, M.I., Aparicio, D., Ascensao, J.T. and Bizarro, P. (2021) 'GuiltyWalker: distance to illicit nodes in the Bitcoin network', arXiv preprint arXiv:2102.05373. doi:10.48550/arXiv.2102.05373.
- Pareja, A., Domeniconi, G., Chen, J., Ma, T., Suzumura, T., Kanezashi, H. et al. (2020) 'EvolveGCN: evolving graph convolutional networks for dynamic graphs', Proceedings of the AAAI Conference on Artificial Intelligence, 34(4), pp. 5363–5370. doi:10.1609/aaai.v34i04.5984.
- Rudin, C. (2019) 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', Nature Machine Intelligence, 1(5), pp. 206–215. doi:10.1038/s42256-019-0048-x.
- Saerens, M., Latinne, P. and Decaestecker, C. (2002) 'Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure', Neural Computation, 14(1), pp. 21–41. doi:10.1162/089976602753284446.

- Saito, T. and Rehmsmeier, M. (2015) 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', PLoS ONE, 10(3), e0118432. doi:10.1371/journal.pone.0118432.
- Weber, M., Domeniconi, G., Chen, J., Weidele, D.K.I., Bellei, C., Robinson, T. et al. (2019) 'Anti-money laundering in Bitcoin: experimenting with graph convolutional networks for financial forensics', arXiv preprint arXiv:1908.02591. doi:10.48550/arXiv.1908.02591.