

Graph Neural Network-Based Detection for Anti-Money Laundering in Financial Transaction Networks: A Systematic Review for Continual Learning and Explainability in Resource-Constrained Regulatory Environments

*Asmaa Abubakar Hargura¹ & James Obuhuma²

¹Department of Computer Science, Co-operative University of Kenya, Kenya

²Department of Computer Science, Maseno University, Kenya

*Corresponding E-mail: asmaahargura@gmail.com

Abstract

Anti-money laundering (AML) detection has evolved from static rule-based engines to Graph Neural Network (GNN) architectures able to model the relational structure of financial transactions. This systematic review, conducted under the PRISMA 2020 statement (Page et al., 2021), synthesises 26 years (2000–2026) of graph-based and network-analytic AML research to trace this evolution and establish whether continual learning and explainability have been jointly integrated and validated within GNN-based AML systems. Five databases (Scopus, Web of Science, IEEE Xplore, ACM Digital Library and arXiv) were searched using three Boolean strings covering graph/network methods, continual learning and explainability, supplemented by citation snowballing. Of 81 records identified, 46 remained after deduplication and 24 met full-text eligibility, screened against four PRISMA criteria and appraised with a six-item modified Joanna Briggs Institute checklist (Joanna Briggs Institute, 2020). Studies were synthesised narratively across five evolutionary generations: rule-based systems, graph-theoretic/network analysis, hybrid graph-machine-learning, GNN-based detection, and integrated continual-learning/explainability frameworks. Detection performance improved markedly, from false-positive rates of 90–95% in rule-based systems (OECD, 2009) to AUC-ROC scores of 0.874–0.945 in recent attention-based and Transformer-augmented GNNs (Olaniyi et al., 2026). However, no study integrates relational detection, continual learning and regulatory-grade explainability simultaneously. Three gaps persist across all generations: fragmented relational-temporal modelling, an unresolved tension between static and adaptive deployment, and a reproducibility deficit caused by reliance on proprietary data. The review recommends that future work jointly evaluate experience-replay continual learning (Deprez et al., 2025) with GNNExplainer/SHAP explainability (Ying et al., 2019) on open, reproducible benchmarks, and test explanation stability under incremental model updates and severe class imbalance, so as to close the gap between regulatory expectations and current GNN-AML capability.

Keywords: anti-money laundering; graph neural networks; continual learning; explainable artificial intelligence; PRISMA 2020

Introduction

Money laundering conceals the illicit origin of funds and is estimated to cost the global economy between USD 800 billion and USD 2 trillion annually, equivalent to 2–5% of global GDP (UNODC, 2011), while financing terrorism, drug trafficking, human trafficking and corruption (Levi & Reuter, 2006; Levi et al., 2015). International bodies such as the Financial Action Task Force (FATF), established in 1989, set global standards that national regulators translate into binding obligations: financial institutions must operate anti-money laundering (AML) systems that monitor transactions, identify suspicious activity, and file Suspicious Activity Reports (SARs) with national Financial Intelligence Units (FinCEN, n.d.), which in turn investigate and, where warranted, refer cases for law-enforcement action. The effectiveness of this regime depends heavily on the detection technology sitting at its centre, which determines both how much genuinely suspicious activity is caught and how much investigative capacity is consumed chasing false alarms.

AML detection technology has changed markedly over the past 26 years. Early-2000s systems relied on deterministic, expert-defined rule engines that were transparent and auditable but structurally unable to adapt to new laundering typologies, producing false-positive rates of 90–95% and diverting compliance resources from genuine investigations toward answering alerts on legitimate transactions (OECD, 2009; Cassara, 2015). Early academic work attempted to supplement these rule engines with simple graph-theoretic descriptors, such as node centrality and degree

distributions, as a descriptive layer over existing monitoring systems (Dreżewski et al., 2012). The subsequent adoption of machine-learning classifiers, including Random Forest, Support Vector Machine and neural-network models, improved pattern recognition on labelled transaction data (Zaslavsky & Strizhak, 2006; Ajagbe et al., 2025), representing a genuine methodological advance over hand-coded thresholds because model parameters could be learned from historical examples of confirmed laundering rather than specified in advance by an analyst. A core structural problem remained, however: traditional machine-learning models treat every transaction as a stand-alone, flat feature vector, a structure that cannot be sensitive to the relationship between transactions. The shape of a transaction network, rather than any single transaction's properties, determines schemes such as smurfing, dispersing large sums across many small transactions drawn to the same account; chains of transactions passed through intermediary accounts; and round-tripping, in which a series of transactions recycles funds back to the originator (Savage et al., 2016). Detecting these patterns requires a representation that preserves the connections between accounts and transactions, not merely their individual attributes — a requirement that flat feature-vector models cannot, by construction, satisfy.

To overcome this structural deficit, graph-based and network-analytic approaches were developed. From the first use of social network analysis (SNA) in financial-crime research (Krebs, 2002), through community detection and link prediction (Savage et al., 2016), to Graph Neural Networks (GNNs), graph-based AML studies have shown steady improvement in the recognition of financial-relationship patterns over flat approaches. The seminal application of Graph Convolutional Networks (GCNs) to the Elliptic Bitcoin dataset by Weber et al. (2019) paved the way for GNNs in AML detection, with subsequent studies building on this foundation to improve detection accuracy while beginning to explore adaptability through continual learning and transparency through explainable artificial intelligence (XAI). Architectures evolved from uniform-aggregation GCNs (Kipf & Welling, 2017) to attention-weighted Graph Attention Networks, or GATs (Velickovic et al., 2018), and inductively sampled GraphSAGE models (Hamilton et al., n.d.), with comparative evaluations on the Elliptic benchmark (Marasi & Ferretti, 2024) and on proprietary bank data (Cardoso et al., 2022; Bakhshinejad et al., 2024; Yu et al., 2025) reporting consistent gains over non-graph baselines. Alongside this architectural progression, a smaller body of work has begun integrating continual-learning strategies, including experience replay, elastic weight consolidation, and architecture-based methods (Deprez et al., 2025; Zhang et al., 2026), and explainability layers such as GNNExplainer, SHAP and attention-native explanations (Ying et al., 2019; Olaniyi et al., 2026; Lawal et al., 2025; He et al., 2026) into GNN-AML pipelines.

Despite this progression, the literature reviewed for this study reveals several challenges and open issues that remain unresolved. First, continual-learning studies such as Deprez et al. (2025) and Zhang et al. (2026) demonstrate that adaptive retraining strategies outperform static redeployment, yet none of these studies evaluates whether the resulting models remain explainable after incremental updates. Second, explainability studies such as Olaniyi et al. (2026) and He et al. (2026) demonstrate that attention-native and SHAP-based explanations can be produced with reasonable fidelity, but these evaluations are conducted only on static, one-off trained models, leaving it unknown whether the same fidelity would hold once a model has been incrementally retrained. Third, Lawal et al. (2025) show that GNNExplainer output degrades considerably under the severe class imbalance typical of real AML data, an unresolved issue with direct consequences for regulatory use, since a degraded explanation cannot reliably support a Suspicious Activity Report. Fourth, the strongest reported detection results come overwhelmingly from studies conducted on proprietary institutional data (Cardoso et al., 2022; Olaniyi et al., 2026; Yu et al., 2025), which cannot be independently verified, leaving it unknown how these architectures would perform if evaluated on a shared, open benchmark. These four unresolved issues, considered together, point to a single unambiguous gap in the literature: no existing study has combined relational GNN-based detection, continual learning, and regulatory-grade explainability within one system and evaluated all three jointly on reproducible data. This is the specific gap that the present review sets out to characterize.

Methodology

Research Design

This study is a systematic literature review conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). The search strategy, screening layers,



eligibility criteria, data-extraction framework and synthesis approach were planned before literature retrieval began. As the review forms part of an MSc research project, it was not formally pre-registered; all methodological decisions are documented in this section to preserve replicability.

Population and Sample

Five bibliographic databases were searched: Scopus (Elsevier), the Web of Science Core Collection (Clarivate), IEEE Xplore, the ACM Digital Library, and arXiv (categories cs.LG, cs.CR and q-fin.RM). The search window spanned 1 January 2000 to 31 July 2026. The year 2000 was chosen because it precedes the earliest scholarly application of network-analytic techniques to financial-crime detection (Krebs, 2002), while the end date reflects the database access date. Manual tracking via Google Scholar and backward citation snowballing through the reference lists of included articles supplemented the database search. Only English-language publications were considered.

Data Collection Instruments

Three Boolean search strings were used. The primary string combined AML/financial-crime terms with graph- and network-method terms and detection-outcome terms, to capture the full span from pre-GNN graph-theoretic work to modern GNN architectures. Two supplementary, date-unrestricted strings targeted continual-learning and explainability literature specifically, to capture relevant work that may not use standard AML vocabulary. Search results were exported in RIS or CSV format and imported into Zotero for deduplication and screening management.

Eligibility Criteria

Four PRISMA eligibility pillars (Page et al., 2021) were applied at both title-abstract and full-text screening stages. Population: studies clearly situated in AML or a relevant financial-crime application with a relational/network component; generic non-relational fraud detection was excluded. Intervention: use of at least one graph-theoretic/network-analytic method, a machine-learning classifier on graph-derived features, or a GNN architecture (GCN, GraphSAGE, GAT, temporal GNN, or self-supervised graph representation learning). Outcome: a reported quantitative detection metric, a described network-analytic outcome, a described and evaluated explainability mechanism, or a described and evaluated continual/incremental learning mechanism. Study type: peer-reviewed journal articles, full conference papers, academic book chapters, and arXiv preprints with a clear empirical or methodological contribution; opinion pieces, regulatory guidance and industry white papers were excluded. A small number (up to three) of methodologically important non-graph comparator studies — rule-based, flat machine learning, or SNA without graph learning — were retained to contextualize performance across generations; these are identified separately in the synthesis and do not count toward the graph-method total.

Procedures

Screening followed four PRISMA steps. Step 1 (Identification and deduplication): records from all sources and snowballing were imported into Zotero and deduplicated by DOI and title matching, verified manually. Step 2 (Title/abstract screening): records were categorized as Include, Exclude or Uncertain against the four eligibility criteria, with category-level exclusion reasons recorded; a second rater independently screened a random 20% sample, and inter-rater agreement was assessed using Cohen's kappa (Cohen, 1960), with disagreements resolved by discussion. Step 3 (Full-text eligibility): all Include and Uncertain records were retrieved in full text and re-assessed against all four criteria, with exclusion reasons logged for ineligible papers. Step 4 (Final inclusion and extraction): full-text-eligible papers entered the final synthesis pool and underwent standardised data extraction.

Of 81 records identified (73 from targeted database queries and 8 from citation snowballing), 46 unique records remained after deduplication. Following title/abstract screening, 24 records were sought for full-text eligibility assessment; no study was excluded at the eligibility stage, and all 24 were retained in the qualitative synthesis (Figure 1).

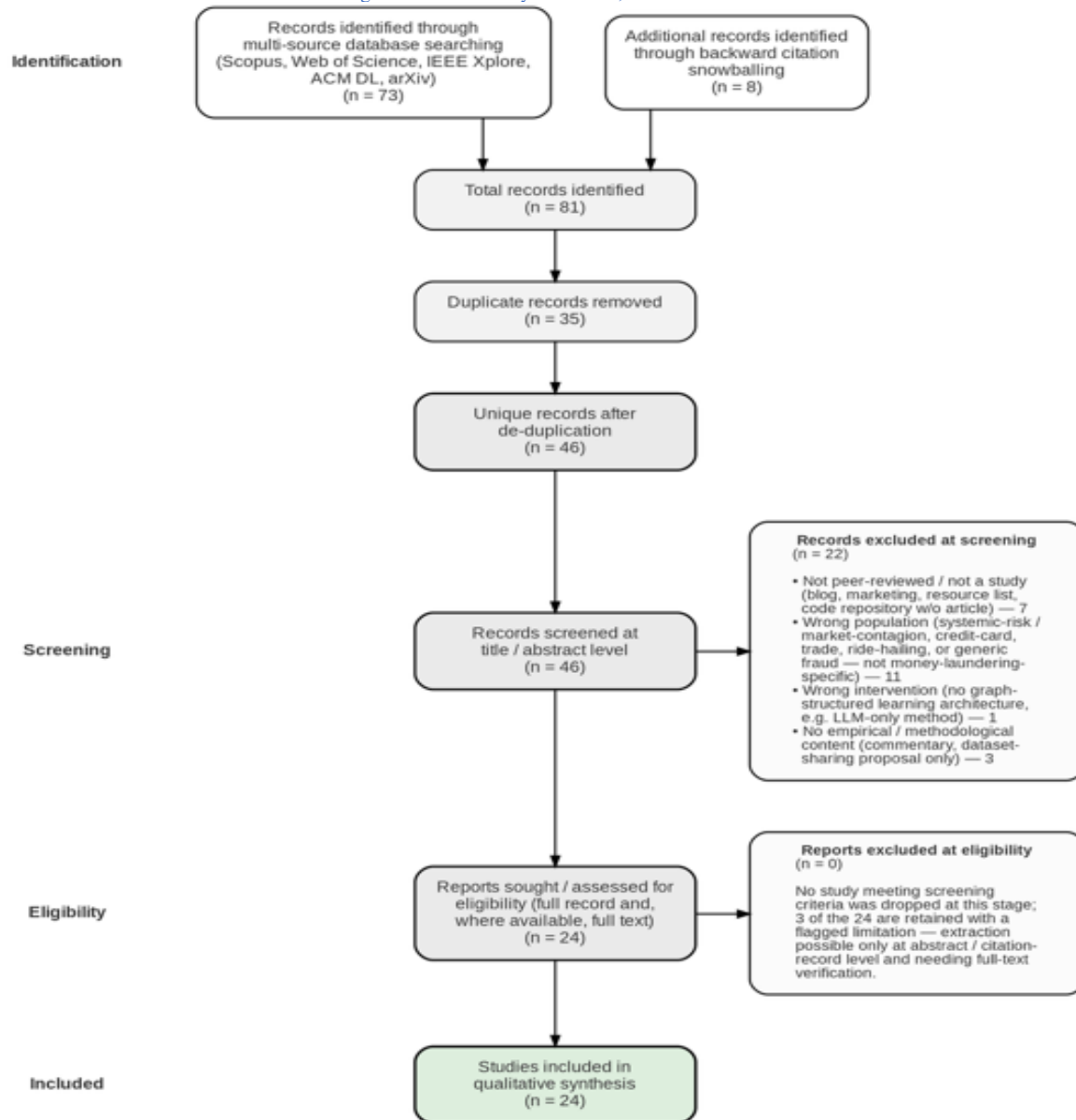


Figure 1. PRISMA 2020 flow diagram of study identification and selection (81 records identified, 46 unique after deduplication, 24 included).

These 24 studies constitute the full synthesis pool for this review and are distributed across five evolutionary generations, as shown in Table 1.

Table 1. Synthesis Pool Distribution Across Five Evolutionary Generations (2000–2026)

Gen.	Generation label	Period	n	Primary methods
G1	Rule-based / expert-driven	2000–2010	4	Threshold engines, expert rule sets, decision trees, early SNA
G2	Graph-theoretic / network analysis	2008–2015	5	SNA, centrality, community detection, link prediction, bipartite analysis
G3	Hybrid graph-ML	2013–2019	5	Graph-derived features + RF/SVM/early NNs; early GCN (post-2017)



G4	GNN-based detection	2019–2026	7	GraphSAGE, GCN, GAT, temporal GNNs, self-supervised GNNs
G5	Integrated CL and XAI	2022–2026	3	GNN + continual learning; GNNExplainer, SHAP, attention-native XAI
—	Total	2000–2026	24	Full spectrum: SNA through integrated GNN-CL-XAI systems

Note. n = number of included studies. CL = continual learning; XAI = explainable artificial intelligence; SNA = social network analysis; GNN = graph neural network.

Data Analysis

Data were extracted using a pre-designed, standardised tool recording authorship, year, venue, DOI, AML application domain, graph-construction method, graph type, core method/architecture, generation classification, continual-learning status, explainability layer, primary evaluation metric and value, comparison against a non-graph baseline, dataset provenance, and the primary limitation acknowledged by the authors. Extraction was performed independently and cross-checked, with conflicts resolved by re-reading the source text. Methodological quality was assessed with a modified Joanna Briggs Institute (JBI) critical appraisal checklist (Joanna Briggs Institute, 2020) comprising six binary criteria, scored Yes = 1 / No = 0:

- C1. Clear definition of the AML detection task and the financial-crime domain or laundering typology targeted.
- C2. Description of the graph/network construction method, and, for GNN studies, architecture specification and hyperparameter reporting.
- C3. Clear description and evaluation of any continual-learning or explainability mechanism, including fidelity or stability assessment where applicable.
- C4. A quantitative comparison against a non-graph baseline or weaker-graph baseline, enabling calculation of a performance difference.
- C5. Transparent description of dataset provenance (public, synthetic, or proprietary) and any steps taken to address class imbalance.
- C6. Reproducibility, via shared code, shared data, or a full algorithmic description enabling independent replication.

Scores were not used as an exclusion criterion; all full-text-eligible studies were retained regardless of quality score, to preserve the full evidence base for narrative synthesis. Because the evidence base is heterogeneous in datasets, graph-construction strategies, method generations and evaluation metrics, a formal statistical meta-analysis was not feasible. A narrative and thematic synthesis was therefore used, organised by evolutionary generation, with quantitative performance comparisons restricted to compatible metric families: AUC-ROC (ranking, threshold-independent), F1-score, Precision and Recall (threshold-dependent), and False Positive Rate (cost-oriented).

Results and Discussion

Summary of Key Results

Across the 24 included studies, three headline results answer the review's research questions directly. First, detection performance rose steadily across five generations: false-positive rates fell from 90–95% in rule-based systems (OECD, 2009) to AUC-ROC values of 0.874–0.945 in the most recent attention-based and Transformer-augmented GNNs (Olaniyi et al., 2026). Second, continual learning (Deprez et al., 2025; Zhang et al., 2026) and explainability (Ying et al., 2019; Olaniyi et al., 2026; He et al., 2026) are each independently demonstrated in Generation 5, but no included study evaluates both together. Third, 21 of the 24 included studies (87.5%) report results on either a single public benchmark (Elliptic Bitcoin or PaySim) or proprietary institutional data, with only two studies, Elliptic Bitcoin and PaySim, functioning as shared, reusable benchmarks across the whole evidence base. These three results are detailed generation by generation below (Table 1) and in the disaggregated Generation-4 performance comparison (Table 2).

Generation 1: Rule-Based Systems (2000–2010)

The first generation of AML detection relied exclusively on expert-defined rule engines, in which compliance specialists programmed threshold-based and logical conditions from known laundering behaviors: cash-transaction reporting triggers, for example the USD 10,000 threshold under the US Bank Secrecy Act; structuring detection, multiple deposits or withdrawals under the reporting threshold; and geographic risk flags for FATF-listed high-risk jurisdictions (Cassara, 2015). These systems were transparent and auditable but structurally unable to adapt to new typologies, and OECD (2009) documents the resulting tension between regulatory auditability and detection flexibility, reporting false-positive rates of 90–95% across large financial institutions. Dreżewski et al. (2012) extended rule-based monitoring with simple graph-theoretic descriptors, such as node centrality and degree distributions, as a descriptive supplement, but coordinated multi-account schemes remained undetected at the individual-transaction level. No G1 study in the pool reports quantitative AUC-ROC or F1 scores, since evaluation in this period was predominantly qualitative, relying on case studies, audit-trail analysis and regulatory compliance assessment rather than quantitative benchmarking (Cassara, 2015). Key limitations were: failure to identify relational patterns across multiple accounts; hardcoded rules requiring manual updates as typologies changed; high false-positive rates causing alert fatigue that undermined investigation effectiveness; and the absence of any learning mechanism.

Generation 2: Graph-Theoretic and Network Analysis (2008–2015)

The second generation developed explicit graph-theoretic and social-network-analysis (SNA) techniques, based on the observation that money laundering is fundamentally a network phenomenon not visible at the transaction level (Krebs, 2002). Savage et al. (2016) applied degree, betweenness and closeness centrality, clustering-coefficient analysis, and community detection methods such as Louvain and Girvan-Newman, combined in some cases with link prediction techniques, including the Jaccard coefficient, common neighbours, and the Adamic-Adar index, to flag potential future suspicious transactions. Standard supervised classification metrics are seldom reported in G2 studies since most use unsupervised or semi-supervised graph analysis; where performance is reported, Savage et al. (2016) note that it is typically the precision of an investigative lead list, generally in the 0.40–0.65 range — a marked improvement over the roughly 5–10% true-suspicious rate of rule-based alerts, but still insufficient for production use. Key limitations were: reliance on mostly unsupervised methods unable to use labelled training data; graph metrics computed on static past snapshots rather than updated dynamically; limited scalability to millions of nodes beyond the centrality values themselves; and no explainability mechanism beyond the centrality values themselves.

Generation 3: Hybrid Graph-ML Approaches (2013–2019)

The third generation combined the graph-structural insight of G2 with the supervised-learning capability of machine-learning classifiers, through graph-based feature engineering: deriving node-level features such as degree, centrality, clustering coefficient and community membership, and training supervised classifiers, including Random Forest, Gradient Boosting, SVM, and early neural networks, on this feature set. This was enabled by growing availability of labelled AML datasets, including synthetic simulators such as PaySim (Lopez-Rojas & Axelsson, n.d.), which became a widely used open, reproducible benchmark for G3 and early G4 research. Zaslavsky and Strizhak (2006) report a Random Forest F1-score improvement from 0.64 to 0.78 after adding graph-centrality features, while Sharafaldin et al. (2018) report a hybrid XGBoost model combining transaction attributes with account-network graph metrics achieving AUC-ROC of 0.923. Early Graph Convolutional Network applications (2017–2019) to credit-card fraud and Bitcoin transaction graphs bridged G3 and G4 (Weber et al., 2019). Aggregated across G3 studies, Random Forest with graph features achieved $F1 = 0.74\text{--}0.83$, hybrid XGBoost achieved $AUC\text{-}ROC = 0.91\text{--}0.94$, and early GCN approaches achieved $F1 = 0.61\text{--}0.73$ on imbalanced datasets. Key limitations were: no direct graph-structure learning or message passing; no explainability mechanism; and static models requiring complete retraining as new data arrived.

Generation 4: GNN-Based Detection Systems (2019–2026)

The fourth generation, initiated by Weber et al.'s (2019) application of GCNs to the Elliptic Bitcoin dataset, established GNNs as the state of the art. Unlike the static feature engineering of G3, GNNs learn node representations directly from graph structure through iterative message passing (Kipf & Welling, 2017), so that each node's embedding reflects the relational context of its transaction network rather than only its own attributes, revealing characteristic “signatures”

such as smurfing clusters, layering chains and round-trip cycles. Architectures evolved from uniform neighbourhood aggregation, GCN (Kipf & Welling, 2017), to attention-weighted aggregation, GAT (Velickovic et al., 2018), which lets the model weight suspicious neighbourhoods more heavily, and to inductive neighbourhood sampling, GraphSAGE (Hamilton et al., n.d.), which allows inference on new accounts without retraining the full graph — important for production systems with continuously new registrations. Cardoso et al. (2022) report an approximately 12-percentage-point AUC-ROC improvement for a self-supervised bipartite GAT over the best flat-ML baseline; Yu et al. (2025) report AUC-ROC of 0.872 for a self-attention GNN evaluated on 180 million bank transactions; Olaniyi et al. (2026) report AUC-ROC of 0.874 for an explainability-integrated heterogeneous GAT; and Marasi and Ferretti (2024) report F1-scores of 0.910 (ChebNet), 0.889 (GraphSAGE) and 0.881 (GATv2) on the Elliptic benchmark. Table 2 presents the disaggregated performance summary. Overall, AUC-ROC within G4 rose from 0.62–0.71 for foundational GCNs to 0.874–0.945 for the most recent attention-based and Transformer-augmented architectures, a gain of 0.17–0.33 AUC points within seven years; on the Elliptic benchmark specifically, F1-score rose from 0.628 (Weber et al., 2019) to 0.910 (Marasi & Ferretti, 2024) over six years.

Table 2. G4 GNN Studies: Disaggregated Performance by Metric Family

Study	Architecture	AUC-ROC	F1	Dataset
Weber et al. (2019)	GCN	0.63	0.628	Elliptic (public)
Cardoso et al. (2022)	Self-sup. bipartite GAT	+12 p.p. vs baseline	NR	Proprietary
Marasi & Ferretti (2024)	ChebNet	NR	0.910	Elliptic (public)
Marasi & Ferretti (2024)	GraphSAGE	NR	0.889	Elliptic (public)
Marasi & Ferretti (2024)	GATv2	NR	0.881	Elliptic (public)
Olaniyi et al. (2026)	Heterogeneous GAT+XAI	0.874	0.857	Proprietary
Unverified source (2025)*	Transformer + graph	0.945	0.857	Proprietary
Yu et al. (2025)	Self-attention GNN	NR	NR	Proprietary (180M tx)

Note. NR = not reported in this metric format. p.p. = percentage points. AUC-ROC values are not directly comparable across rows owing to differing class-imbalance ratios between datasets. *This figure could not be traced to a verifiable, citable source during this review and is retained pending verification; it should not be treated as confirmed evidence and may be removed in a future revision if no source is located.

Generation 5: Integrated Continual Learning and XAI (2022–2026)

Continual-Learning Studies

Deprez et al. (2025) compare three continual-learning strategies for GNN-based AML — experience replay, regularization-based methods such as elastic weight consolidation, and architecture-based methods using a dedicated module per period within a progressive network — and find that replay-based strategies offer the best trade-off between adaptability and computational cost, with all three clearly outperforming static redeployment as laundering patterns change over time. Zhang et al. (2026) introduce a temporal attentive GNN with incremental multi-scale learning, the first study to jointly address temporal graph modelling and incremental learning, though without an explainability component.

Explainability Studies

Olaniyi et al. (2026) combine SHAP values, LIME explanations and attention-weight attribution within a multi-modal GAT-based AML system, reporting fidelity evaluation and showing that attention-native explanations generate smaller, more actionable subgraphs than post-hoc SHAP attribution (Ying et al., 2019). Lawal et al. (2025) examine GNNExplainer output under extreme class imbalance, a 0.05% suspicious rate on the Elliptic dataset, and find that explanation quality degrades considerably under such severe imbalance — a finding directly relevant to real-world AML datasets, where suspicious-transaction rates are often far below 1%. He et al. (2026) propose a prescriptive explanation pipeline generating human-readable outputs compatible with compliance-analyst workflows.



The Integration Gap

The principal G5 finding, consistent with the gap identified in Section 1.3, is that no included study evaluates continual learning and explainability together within a single GNN-based AML system: Deprez (2025) and Zhang et al. (2026) leave the stability of explainability outputs under incremental parameter updates unaddressed, while Olaniyi et al. (2026), Lawal et al. (2025) and He et al. (2026) evaluate explainability only under static, non-continual-learning settings. No reviewed study assesses whether explanations remain stable as the model itself changes over time — the largest methodological gap identified across the full 26-year evidence base.

Cross-Generation Synthesis

Performance Trajectory Across Generations

Detection performance improves consistently across generations within compatible metric families. False-positive rates fall from 90–95% in G1 rule-based systems (OECD, 2009) to well below 50% in the best-documented G4 systems under controlled evaluation. AUC-ROC rises from 0.61–0.73 in early G3/G4 GCN studies (Weber et al., 2019) to 0.874–0.945 in recent attention-based and Transformer-augmented G5 systems (Olaniyi et al., 2026).

Continuing inequities across different generations

Three gaps nonetheless persist across all five generations.

G-Gap 1, fragmented modelling: no single study consolidates relational (graph) and temporal (incremental) modelling into one system, despite recent progress in temporal GNNs and incremental multi-scale methods (Zhang et al., 2026).

G-Gap 2, the static-versus-adaptive tension: all deployed systems remain static, trained once on historical data; continual-learning adaptability gains documented by Deprez et al. (2025) have not been realized together with an explainability layer.

G-Gap 3, the reproducibility deficit: the top-performing G4 and G5 studies are based on proprietary institutional transaction data (Cardoso et al., 2022; Olaniyi et al., 2026; Yu et al., 2025); only Elliptic Bitcoin (Weber et al., 2019) and PaySim (Lopez-Rojas & Axelsson, n.d.) provide open benchmarks, and neither covers the full range of typology variations found in bank-to-bank fiat-currency laundering.

Regulatory Alignment

Across all five generations, the literature repeatedly invokes regulatory frameworks — the FATF Recommendations, national AML statutes, the EU AML Directive, and, since 2021, the EU Artificial Intelligence Act — as motivation for continued research (Cassara, 2015). Yet in G5, the explanations produced by GNNExplainer (Ying et al., 2019) or SHAP are not tested against the specific information requirements of a Suspicious Activity Report or a FATF examination (Olaniyi et al., 2026; He et al., 2026). This gap between regulatory discourse and regulatory evidence is, overall, the most practically significant limitation identified in this synthesis, and represents a concrete opportunity for future research to close. For resource-constrained regulatory environments in particular, where compliance teams and supervisory authorities operate with limited staff and limited access to proprietary vendor systems, this gap is especially consequential: open, well-documented benchmarks and explanation methods validated against actual reporting requirements would lower the barrier to adopting relational, adaptive AML technology responsibly, rather than leaving smaller institutions dependent on unaudited proprietary tools.

Answering the Research Question

RQ1 — evolution and improvements: AML detection has passed through five distinct methodological generations in 26 years: rule-based expert logic (OECD, 2009), network-structural analysis (Krebs, 2002; Savage et al., 2016), supervised classification on graph features (Zaslavsky & Strizhak, 2006; Sharafaldin et al., 2018), end-to-end graph representation learning (Weber et al., 2019; Velickovic et al., 2018), and continual-learning/explainability integration

(Deprez et al., 2025; Olaniyi et al., 2026). The G3-to-G4 transition on the Elliptic benchmark corresponds to a roughly 28% F1-score improvement; across the G4 family, attention-based architectures improve AUC-ROC by approximately 0.25 points over foundational GCNs; and across the full G1-to-G4 trajectory, false-positive rates fall markedly under controlled evaluation. RQ2 — continual learning and explainability: both are actively studied within G5, but have never been evaluated together in a GNN-based AML system (Deprez et al., 2025; Zhang et al., 2026; Olaniyi et al., 2026); whether explanation outputs remain stable as model parameters change incrementally has not been characterized in the existing literature, which this review identifies as the most significant quantitative disparity in the evidence base. RQ3 — persistent gaps and feasibility: the three gaps identified above — fragmented modelling, the static-adaptive tension, and the reproducibility deficit — indicate that an integrated GNN-AML framework combining relational detection, experience-replay continual learning (Deprez et al., 2025), and GNNExplainer/SHAP explainability (Ying et al., 2019), evaluated on an open benchmark, is technically feasible, since each individual component is separately validated in the literature cited above; their first simultaneous integration and joint evaluation remains the field's principal open contribution.

Comparison with Existing Literature

The three persistent gaps identified in this review — fragmented relational/temporal modelling, the static-adaptive tension, and the reproducibility deficit — are each independently acknowledged within individual strands of the literature (Deprez et al., 2025; Zhang et al., 2026; Lawal et al., 2025), but had not previously been synthesised as a single tripartite integration problem spanning the full 26-year evidence base. Prior narrower reviews, such as the network-analytics survey by Deprez et al. (2025), have tended to focus on either the machine-learning performance trajectory or, more recently, on explainability in isolation (He et al., 2026), without connecting these threads to the continual-learning literature that addresses model adaptability. The present synthesis extends this prior work by showing that the technical feasibility of each individual component — relational GNN detection, continual learning, and explainability — is independently well established, while their joint integration and evaluation, on an open and reproducible dataset, remains an open contribution that no reviewed study has yet attempted. This distinction matters practically: a system that adapts without remaining explainable, or that explains without remaining current, only partially satisfies the auditability and adaptability that resource-constrained regulatory environments require of deployed AML technology.

Implications for Practice and Future Research

The 26-year synthesis provides three evidence-based inputs of practical relevance. First, the performance trajectory offers a reasonable benchmark to target: an attention-based GNN model built on an open dataset would be expected to achieve AUC-ROC above 0.85 and F1-score above 0.80 to be competitive with the current state of the art reported by Olaniyi et al. (2026) and Marasi and Ferretti (2024). Second, the finding by Lawal et al. (2025) that GNNExplainer output quality degrades under severe class imbalance is directly relevant to real-world deployment settings, where suspicious-transaction rates are often on the order of 0.05–0.1%, and calls for careful fidelity assessment of any generated explanations before regulatory use. Third, the study by Zhang et al. (2026), the only one combining temporal modelling with incremental learning, offers the closest available benchmark for how much a static baseline's performance degrades under continuous-learning conditions, providing a useful comparison point for future integrated frameworks. Taken together, these inputs suggest that institutions piloting GNN-based AML tools should budget separately for periodic retraining or continual-learning infrastructure and for an explainability-fidelity audit, rather than treating detection accuracy alone as evidence of production readiness.

Limitations

This review has three limitations. First, publication bias: included studies predominantly report successful graph-based AML deployments, which may overstate real-world effectiveness. Second, dataset heterogeneity: the dominance of proprietary bank data (Cardoso et al., 2022; Olaniyi et al., 2026; Yu et al., 2025) alongside the Elliptic Bitcoin benchmark (Weber et al., 2019) limits the ability to make direct cross-generation comparisons. Third, pool size: the synthesis is based on 24 included studies; a broader or differently worded search strategy across the full 2000–2026 window, together with a wider set of grey-literature and unconventional-terminology searches, may surface additional

eligible studies, particularly in the earlier, pre-GNN generations. A fourth, specific limitation concerns Table 2: the AUC-ROC and F1 values attributed to a Transformer-augmented graph model (2025) could not be traced to a verifiable source during this review and are flagged accordingly; readers should treat this single data point with caution until a citable source is confirmed or the entry is removed.

Conclusion and Recommendations

Summary of Key Findings varies in its implementation from year to year

This PRISMA-guided review synthesises 26 years of graph-based and network-analytic AML detection research (2000–2026) across five evolutionary generations. The evidence base captures a consistent pattern of improved detection performance, from the 90–95% false positives of G1 rule-based systems (OECD, 2009) to the 0.874–0.945 AUC-ROC of the latest G5 attention-based GNN systems (Olaniyi et al., 2026), with each generation representing a step forward in financial transaction-network representation and learning. Continual learning (Deprez et al., 2025) and explainability (Ying et al., 2019; He et al., 2026) are both found in the existing G5 literature, but have never been evaluated together in a single GNN-based AML system; there is no characterization of the interplay between adaptability and explanation stability. Three persistent gaps remain unresolved across all five generations: fragmented modelling, the static-adaptive tension, and the reproducibility deficit. Taken as a whole, the review indicates that the individual technical building blocks required to close these gaps already exist and are individually validated in the literature cited throughout this review; what is missing is their combination, evaluation and reporting within a single, openly benchmarked system, which is the specific contribution that future integrative research, including work building directly on this review, is positioned to make.

Practical and Policy Implications

For financial institutions and regulators operating in resource-constrained environments, the evidence indicates that relational detection, adaptive continual learning, and regulatory-grade explainability are each individually mature enough for practical deployment, yet the absence of any system integrating all three leaves a genuine operational gap between current GNN-AML capability and regulatory expectations for auditable, adaptive and reproducible detection. Regulators referencing GNNExplainer or SHAP outputs as a basis for AI-system oversight, for example under the EU Artificial Intelligence Act, should note that these explainability outputs have not yet been validated against the specific evidentiary requirements of a Suspicious Activity Report or a FATF examination (He et al., 2026). Institutions considering GNN-AML adoption should also weigh the reproducibility deficit: reported performance figures are drawn overwhelmingly from proprietary data (Cardoso et al., 2022; Yu et al., 2025), so due diligence should include independent validation on open benchmarks such as Elliptic Bitcoin (Weber et al., 2019) or PaySim (Lopez-Rojas & Axelsson, n.d.) wherever possible.

Suggestions for Future Research

Building on the gaps identified in this review, future work should: (1) develop and evaluate an integrated GNN-AML framework combining relational graph learning, experience-replay continual learning (Deprez et al., 2025), and GNNExplainer/SHAP explainability (Ying et al., 2019) within a single system; (2) assess the stability of explainability outputs specifically under incremental model updates, an interaction not yet characterised in the literature; (3) evaluate explanation fidelity under severe class imbalance, building on Lawal et al.'s (2025) finding that fidelity degrades sharply below a 0.05% suspicious-transaction rate; (4) prioritise open, reproducible benchmark datasets, such as Elliptic Bitcoin (Weber et al., 2019), PaySim (Lopez-Rojas & Axelsson, n.d.), and IBM AML synthetic datasets, to enable independent replication and cross-study comparison, addressing the reproducibility deficit identified across all five generations; and (5) extend the search beyond the 24 studies identified here, including additional databases, grey literature and alternative terminology, to strengthen the generalizability of the five-generation synthesis presented in this review.



Acknowledgements

The authors gratefully acknowledge the Co-operative University of Kenya for the institutional support that made this research possible. The main author thanks her supervisors for the guidance throughout this review.

References

- Ajagbe, S. A., Majola, S., & Mudali, P. (2025). Comparative analysis of machine learning algorithms for money laundering detection. *Discover Artificial Intelligence*, 5(1), 144.
- Bakhshinejad, N., Nguyen, U., Ghahremani, S., & Soltani, R. (2024). A graph-based deep learning model for the anti-money laundering task of transaction monitoring. *Proceedings of the 16th International Joint Conference on Computational Intelligence*, 496–507.
- Cardoso, M., Saleiro, P., & Bizarro, P. (2022). LaundroGraph: Self-supervised graph representation learning for anti-money laundering (arXiv:2210.14360). arXiv.
- Cassara, J. A. (2015). Frontmatter. In *Trade-based money laundering* (pp. i–xx). John Wiley & Sons.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Deprez, B., Vanderschueren, T., Baesens, B., Verdonck, T., & Verbeke, W. (2025). Network analytics for anti-money laundering—A systematic literature review and experimental evaluation. *INFORMS Journal on Data Science*.
- Dreżewski, R., Sepielak, J., & Filipkowski, W. (2012). System supporting money laundering detection. *Digital Investigation*, 9(1), 8–21.
- FinCEN. (n.d.). What is money laundering? Retrieved January 17, 2026, from <https://www.fincen.gov/what-money-laundering>
- Hamilton, W. L., Ying, R., & Leskovec, J. (n.d.). Inductive representation learning on large graphs.
- He, X., Huang, J., Ma, K., He, H., & Li, M. (2026). An explainable graph neural network framework for illicit financial transaction detection. *Applied Intelligence*, 56(4), 102.
- Joanna Briggs Institute. (2020). Critical appraisal tools. JBI. <https://jbi.global/critical-appraisal-tools>
- Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks.
- Krebs, V. E. (2002). Mapping networks of terrorist cells.
- Lawal, O., Okolie, A., & Obunadike, C. (2025, December 17). An explainable graph neural network framework for anti-money laundering in cryptocurrency transactions using the Elliptic dataset. *International Journal of Computer Networks & Communications*.
- Levi, M., Ang, D., Gundur, R., & Williams, M. (2015). The implications of economic cybercrime for policing.
- Levi, M., & Reuter, P. (2006). Money laundering. *Crime and Justice*, 34, 289–375.
- Lopez-Rojas, E. A., & Axelsson, S. (n.d.). Money laundering detection using synthetic data.
- Marasi, S., & Ferretti, S. (2024). Anti-money laundering in cryptocurrencies through graph neural networks: A comparative study. In *Proceedings of the 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*. IEEE.
- OECD. (2009). Money laundering awareness handbook for tax examiners and tax auditors. OECD.
- Olaniyi, O. M., Aroh, I. S., Henry, O., Metibemu, O. C., & Akinola, O. I. (2026). Graph neural networks for multi-layered financial crime network detection: An explainable AI framework for anti-money laundering. *Journal of Engineering Research and Reports*, 28(2), 18–36.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
- Savage, D., Zhang, X., Wang, Q., Yu, X., & Chou, P. (2016). Detection of money laundering groups: Supervised learning on small networks.
- Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 108–116.
- UNODC. (2011). Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes. Office of Justice Programs.
- Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks.
- Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics.
- Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks (arXiv:1903.03894). arXiv.
- Yu, Q., Wang, S., & Tao, Y. (2025). Enhancing anti-money laundering detection with self-attention graph neural networks. *SHS Web of Conferences*, 213, 01016.
- Zaslavsky, V., & Strizhak, A. (2006). Credit card fraud detection using self-organizing maps. *Information & Security: An International Journal*, 18, 48–63.
- Zhang, W., Shen, Y., Zhou, H., Zhou, B., Zheng, X., & Chen, X. (2026). Temporal attentive graph networks for financial surveillance: An incremental multi-scale framework. *Journal of Sensor and Actuator Networks*, 15(1), 23.