

A Self-Supervised Graph-Temporal Model for Fraud Detection in Digital Bank Account Opening: A Systematic Review

***Esther Mueni¹, David Muriuki¹ & Casper Shikali²**

¹Department of Computer Science, School of Computing and Mathematics,
Cooperative University of Kenya

² Department of Computer Science and Technology, School of Science and Computing
South Eastern Kenya University

***Corresponding E-mail:** esthermueni395@gmail.com

Abstract

Digital bank account opening has become the latest trend in bank application and account opening. With its growth it also becomes one of the focuses of financial fraud. 2024 alone saw the potential for an estimated US\$3.3 billion in losses at United States financial institutions from synthetic identity fraud. Most detection methods are based on supervised machine learning methods which are limited by the distance from the label source to the target server due to the properties of labels, extreme label-class imbalance and adversarial concept drift, and which can degrade in their detection capabilities by 15–30% within months of deployment. In this systematic review conducted as per PRISMA 2020 guidelines, the evidence related to fraud detection through supervised, graph-based approach, temporal approach, and self-supervised approach has been synthesised to determine if an integrated framework can cater to all the above three constraints. Selection of relevant literature was based on a search in six bibliographic databases, plus the proceedings of seven major machine-learning conferences, from January 2014 to April 2025. After deduplication and two-stage screening (with κ s of 0.81 and 0.87 respectively), 34 studies were included from 1,847 which were later appraised using a Mixed Methods Appraisal Tool adapted for computational research. We identify the literature as going through three stages (2014-2017 $n = 8$; 2018-2022 $n = 15$; 2023-2025 $n = 9$) with two foundational pre-2014 works and six themes: supervised learning limitations, graph relational modelling, temporal sequence modelling, self-supervised contrastive learning, and integrated hybrid architectures. Seven research gaps were found across the studies, most importantly none of the studies included assessed an integrated graph-temporal self-supervised model for simulated label delay use or for account-opening data. Based on the above evidence, a conceptual yet still unvalidated architecture Self-Supervised Graph-Temporal (SSL-GT) is proposed that combines heterophily-aware graph learning, transformer temporal encoding and graph contrastive pre-training. The major findings indicate persistent limitations in supervised learning under delayed labels and drift, while graph, temporal and contrastive methods offer complementary but incompletely validated capabilities for digital account opening. The review concludes that an integrated SSL-GT framework remains a research proposal rather than a demonstrated solution. It recommends evaluation under realistic label delays, fraud-preserving graph augmentation, drift testing, operational-latency benchmarking and fairness audits.

Keywords: Self-supervised learning, Graph neural networks, Temporal modelling, Account opening fraud, Label latency

Introduction

The shift from branch-based to fully digital customer onboarding has removed the friction of physical verification that once filtered out fraudulent applications. Digital channels now allow fraudsters to submit synthetic identities, which combine genuine and fabricated personally identifiable information, at a volume and velocity that traditional rule-based systems cannot match (Jesus et al., 2022). Synthetic identity fraud is estimated to have exposed United States lenders to US\$3.3 billion in losses on accounts opened through 2024, and approximately 8.3% of digital account creations were flagged as suspicious in the first half of 2025 (TransUnion, 2025). Every dollar lost to fraud costs financial institutions a further US\$4.41 in investigation, remediation and compliance (LexisNexis Risk Solutions, 2024), in addition to reputational damage and regulatory exposure under the Bank Secrecy Act and anti-money laundering (AML) frameworks.

Account-opening fraud differs materially from post-onboarding transaction fraud because it occurs before any account history exists. Any relational or sequential signal must therefore be constructed entirely from application-time metadata such as device identifiers, IP addresses, e-mail domains and submission timing. Organised rings reuse this infrastructure across many applications submitted within short time windows (Cheng et al., 2022; Liu et al., 2018), and synthetic identities are frequently nurtured for months before a large "bust-out" withdrawal (Jesus et al., 2022). The resulting coordination produces subtle relational and temporal signatures that cannot be detected by analysing applications in isolation. Evidence from transaction-fraud graphs therefore shows general viability, not direct transferability to account opening.

Detection is further complicated by three structural constraints. First, fraud labels are confirmed only after institutional investigation, so the most recent applications are unlabelled at decision time (Dal Pozzolo et al., 2015; Jesus et al., 2022). Second, fraud typically constitutes 0.1-1% of applications, and this extreme imbalance biases standard classifiers and inflates accuracy-based metrics (Saito & Rehmsmeier, 2015). Third, fraudsters adapt strategically to deployed controls, so models lose 15-30% of their detection effectiveness within months of deployment (Dal Pozzolo et al., 2018; Gama et al., 2014). These constraints reinforce one another: the few labels available are both delayed and rapidly outdated by drift.

Supervised classifiers have proven effective in fraud detection but their dependence on confirmed results is still an issue when the application doesn't have any previous transaction history, and labels from the previous investigations are not available (Dal Pozzolo et al., 2015; Jesus et al., 2022). Although graph-based approaches have been developed to detect common devices and synchronized applications, traditional graph models do not consider the timing of events and can be vulnerable to counterfeited camouflage and heterophilic relationships (Dou et al., 2020; Shi et al., 2022). Whilst the above methods can be applied to application sequences or self-supervised methods can be trained on unlabelled records, it remains unclear whether these methods will be equally effective at detecting anomalies during account opening (Cheng et al., 2022; Liu et al., 2022; Zhang et al., 2024).

The question that remains is whether it is possible to combine relational modelling, temporal modelling, and self-supervised representation learning in a manner that would overcome the challenges of a delayed label, extreme class imbalance, and ever-changing fraud patterns within digital account-opening data. The surveyed reviews are mainly about the area of fraud prior to the present graph neural networks, transformers and contrastive learning (Phua et al., 2010; Akoglu et al., 2015), and the empirical studies reviewed here assess the approaches with respect to various fraud domains, data sets and performance indicators. In this regard, there is no direct evidence in the literature that combines the two approaches of graph and temporal and evaluates them in realistic label-delay and temporal-drift scenarios for account-opening fraud applications. The specific research gap that is addressed is that of the empirical validation of the proposed SSL-GT framework. The proposed SSL-GT framework is still conceptual and needs to be validated in practice, this is the specific research gap addressed.

Four theories are covered in the review. Graph convolutional networks (GCNs) are usually used to propagate features within the graph, assuming homophily, which fraudsters break by actively contacting with legitimate users, and thus, heterophily-aware graph learning (Shi et al., 2022; Tang et al., 2022) has been developed. Contrastive representation learning does a task to get to a point of being invariant to some semantically-preserving transformation, and discriminative of instances (Chen et al., 2020); and such representations should be more robust in the presence of intentionally introduced distributional shift. There exists a difference between virtual drift (input distribution) and real drift (feature-label mapping) (Gama et al., 2014) and adversarial fraud drift is similar to real drift but has a sudden and a gradual part (Hinder et al., 2024; Žliobaitė et al., 2016). Finally, transformer self-attention links all the events in a sequence, and the pre-train-then-fine-tune paradigm (Devlin et al., 2019) also underlies the graph pre-training approach studied here.

The review focuses on four theories. Typically, graph convolutional networks (GCNs) propagate features in the graph, under the assumption of homophily, which fraudsters violate by actively connecting to legitimate users, leading to the development of heterophily-aware graph learning (Shi et al., 2022; Tang et al., 2022). Contrastive representation learning performs a task to lead to representations that are invariant to certain semantics-preserving transformations, but discriminative between instances (Chen et al., 2020), and such representations should be more robust in the presence of intentionally introduced distributional shift. Virtual drift theory differentiates virtual drift (in the input distribution) from real drift (in the feature-label mapping) (Gama et al., 2014), while adversarial fraud drift is similar to real drift but with a sudden and a gradual part (Hinder et al., 2024; Žliobaitė et al., 2016). Lastly, transformer self-

attention connects all the events in a sequence, and the pre-train-then-fine-tune paradigm (Devlin et al., 2019) also forms the backbone of the graph pre-training strategy examined here.

The empirical literature considered in this review brings together graph-based, temporal and self-supervised approaches to fraud detection. Comparing their findings helps explain why these approaches may be complementary, while also revealing limits to their application in digital account opening.

Across the empirical literature, differences in datasets, available labels, temporal evaluation and performance measures complicate direct comparisons of graph, temporal and self-supervised approaches (Dal Pozzolo et al., 2015; Jesus et al., 2022; Liu et al., 2022). Evidence from transaction fraud and other anomaly-detection tasks therefore cannot, on its own, establish effectiveness for new digital bank account applications.

Relevant research is spread across several communities with inconsistent evaluation practices. The two closest prior reviews pre-date the methods now at the frontier: Akoglu et al. (2015) surveyed graph-based anomaly detection before the deep-learning and self-supervised era, and Phua et al. (2010) surveyed supervised data-mining approaches before graph neural networks (GNNs) and transformers existed. The reviewed literature does not provide a systematic synthesis of which combinations of methods are most suitable for the operational constraints of label latency and concept drift in account-opening fraud.

This review synthesized fraud-detection studies published from 2014 to 2025 to assess successive methodological approaches in relation to label latency, class imbalance and concept drift, and to delineate the gaps motivating each part of the proposed Self-Supervised Graph-Temporal (SSL-GT) model. Four research questions (RQs) guided the review. RQ1: What machine-learning methods have been adopted and how do they perform when dealing with label scarcity and class imbalance in financial fraud detection from 2014 to 2025? RQ2: What are the benefits of GNNs and temporal models for fraud detection, and what are the potential challenges of combining them in account-opening scenarios? RQ3: What evidence exists for self-supervised and contrastive learning in addressing label latency and concept drift in fraud detection? RQ4: What methodological and operational gaps remain in integrated graph-temporal fraud detection, and how might SSL-GT address them? This review does not test SSL-GT; rather, it provides the evidence base for its future development and evaluation.

Methodology

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guideline (Page et al., 2021). Before full-text screening, a written protocol that included the research questions, search strategy, eligibility criteria, data-extraction fields and synthesis approach was finalised. It was not registered in PROSPERO (which was not available at the time of the review), and the protocol was not deposited in the Open Science Framework (OSF) that is recognised as a reporting limitation. The protocol can be requested from the corresponding author and is planned to be registered in the future. The SSL-GT components were pre-specified to prevent the evidence from determining model design.

Search strategy: Seven sources were searched to cover computer science, financial technology and applied machine learning: IEEE Xplore, ScienceDirect, ACM Digital Library, Scopus, Google Scholar, arXiv (cs.LG, cs.AI, stat.ML) and the archived proceedings of NeurIPS, ICML, ICLR, AAAI, KDD, WWW and CIKM. The search covered January 2014 to April 2025 and was applied to titles, abstracts and keywords, adapted to each database syntax. The conceptual query was:

("fraud detection" OR "anomaly detection" OR "account fraud") AND ("graph neural network" OR "GNN" OR "graph convolutional" OR "GraphSAGE" OR "graph attention" OR "self-supervised learning" OR "contrastive learning" OR "representation learning" OR "temporal model" OR "transformer" OR "LSTM") AND ("financial fraud" OR "bank account" OR "digital onboarding" OR "synthetic identity") AND ("concept drift" OR "label latency" OR "class imbalance" OR "delayed supervision")

In IEEE Xplore the same structure was implemented with "Document Title" and "Abstract" field tags and filtered to journals and conferences (2014-2025). Supplementary keywords (e.g., "XGBoost", "heterogeneous graph", "NT-Xent loss", "KYC", "BAF dataset") captured terminological variation, and reference lists of included studies were hand-searched.

A priori, inclusion and exclusion criteria (Table 1) were set and applied on all screening stages. Much of the latest hybrid-architecture research (2023-2025) is published as preprints, before or in lieu of, journal publication; the exclusion of arXiv preprints would have caused a lag bias against the newest and most relevant research.

Table 1. Inclusion and exclusion criteria

Category	Criterion	Specification
Inclusion	Topic	Addresses at least one of: fraud or anomaly detection methodology; graph-based relational learning for fraud; temporal sequence modelling for fraud; self-supervised or contrastive learning for financial anomaly detection; label-scarce financial classification
Inclusion	Study type	Empirical studies, methodological proposals with experimental evaluation, benchmark dataset papers and reviews with quantitative synthesis
Inclusion	Publication type	Peer-reviewed journals and conference proceedings; arXiv preprints from recognised venues (NeurIPS, ICML, ICLR, AAAI, KDD, WWW, CIKM, IEEE Transactions)
Inclusion	Language and date	English; January 2014 - April 2025; full text accessible
Exclusion	Scope	Post-onboarding transaction fraud only, with no applicability to account opening
Exclusion	Methodology	Position papers, opinion pieces and editorials without empirical evaluation
Exclusion	Quality	No reportable performance metrics or undescribed experimental method
Exclusion	Duplication	Earlier versions superseded by a later published version
Exclusion	Domain	Domains unrelated to financial fraud with no transferable methodological contribution

Study selection: Screening was conducted by the primary researcher, and a random 20% sample was independently reviewed by a second researcher, with disagreements resolved by consensus. The 1,847 records retrieved were deduplicated to 1,294 unique records. Title and abstract screening excluded 1,089 records (irrelevant topic, $n = 634$; wrong domain, $n = 211$; no empirical evaluation, $n = 147$; non-English, $n = 97$), with substantial agreement ($\kappa = 0.81$). Of the 205 full texts assessed, 171 were excluded (insufficient methodological detail, $n = 54$; no fraud-detection focus, $n = 48$; superseded version, $n = 31$; full text unavailable, $n = 25$), with strong agreement ($\kappa = 0.87$). Thirty-four studies were included in the synthesis (Figure 1).

Each study was appraised with a Mixed Methods Appraisal Tool (MMAT) checklist adapted for computational and applied machine-learning research. Five dimensions were scored as satisfied (2), partially satisfied (1) or not satisfied (0), giving a maximum of 10: (1) clarity of research question; (2) dataset appropriateness and description; (3) experimental rigour, including train-test separation, temporal ordering and leakage prevention; (4) metric appropriateness for imbalanced settings, penalising studies reporting only accuracy or AU-ROC without AU-PRC; and (5) reproducibility, including code availability, hyperparameter reporting and significance testing. Both researchers independently scored a 20% calibration sample (seven studies) and resolved discrepancies by consensus; the second researcher re-scored a further 20% as a check. No study was excluded on quality alone; scores weighted findings in the synthesis.

Data extraction and synthesis: A standardised form, piloted on five studies, captured bibliographic details, objective, datasets, architecture, metrics, limitations and relevance to each RQ; 20% of entries were verified against the source records. Heterogeneity in datasets and metrics precluded meta-analysis, so bibliometric analysis (publication trends, keyword co-occurrence) and thematic content analysis were used. Certainty of evidence was rated qualitatively (high, moderate or low) using GRADE-style reasoning adapted to non-randomised computational studies, based on consistency, risk of bias and directness to account opening.

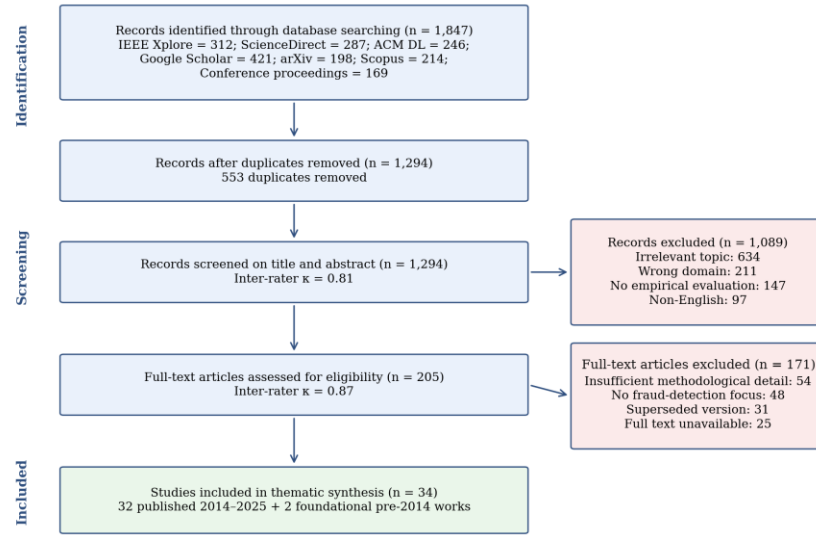


Figure 1. PRISMA 2020 flow diagram for study identification, screening and inclusion

Results and discussions

Study characteristics and quality: The per-source yield is shown in Figure 1. The 34 included studies comprise NeurIPS, ICML, ICLR and AAAI papers ($n = 9$), other peer-reviewed journal articles ($n = 11$), IEEE and ACM conference papers ($n = 6$), IEEE Transactions articles ($n = 4$), arXiv preprints ($n = 2$) and industry reports ($n = 2$). Thus, 30 studies (88%) are formally peer-reviewed and 4 (12%) are grey literature. The two preprints were retained as the most direct evidence for graph contrastive pre-training on fraud graphs and are weighted accordingly. Quality scores ranged from 7 to 10, indicating generally high methodological quality (Table 2). Reproducibility was the weakest dimension: among the 12 representative studies, six fully and six only partially satisfied it, typically because headline metrics were reported without released code or complete hyperparameters.

Table 2. Quality assessment scores of representative studies (adapted MMAT; 2 = satisfied, 1 = partially satisfied, 0 = not satisfied)

Study	RQ clarity	Dataset quality	Exp. rigour	Metric approp.	Reproducibility	Total /10
Gama et al. (2014)	2	1	1	1	2	7
Saito & Rehmsmeier (2015)	2	2	2	2	1	9
Hamilton et al. (2017)	2	2	2	1	2	9
Kipf & Welling (2017)	2	2	2	1	2	9
Liu et al. (2018)	2	2	2	2	1	9
Velickovic et al. (2018)	2	2	2	1	2	9
Wang et al. (2019)	2	2	2	2	1	9
Dou et al. (2020)	2	2	2	2	1	9
Liu et al. (2022)	2	2	2	2	1	9
Vinay et al. (2023)	2	2	2	2	2	10
Zhang et al. (2024)	2	2	2	2	2	10
Olaniyan et al. (2025)	2	2	1	1	1	7

Publication trends and thematic clusters: The 32 studies within the review window show three phases (Figure 2). Phase 1 (2014-2017, $n = 8$) established supervised baselines, concept-drift theory and evaluation practice for imbalanced classification. Phase 2 (2018-2022, $n = 15$), the most productive period, adapted GNNs to fraud graphs and culminated in the release of the NeurIPS 2022 Bank Account Fraud (BAF) benchmark. Phase 3 (2023-2025, $n =$

9) shows convergence of self-supervised frameworks with graph-temporal hybrid architectures and early attention to deployment constraints. Two pre-2014 works (Bolton & Hand, 2002; Phua et al., 2010) provide historical context. Industrial groups were influential, notably Ant Financial (Liu et al., 2018; Wang et al., 2019) and Feedzai, which released BAF (Jesus et al., 2022).

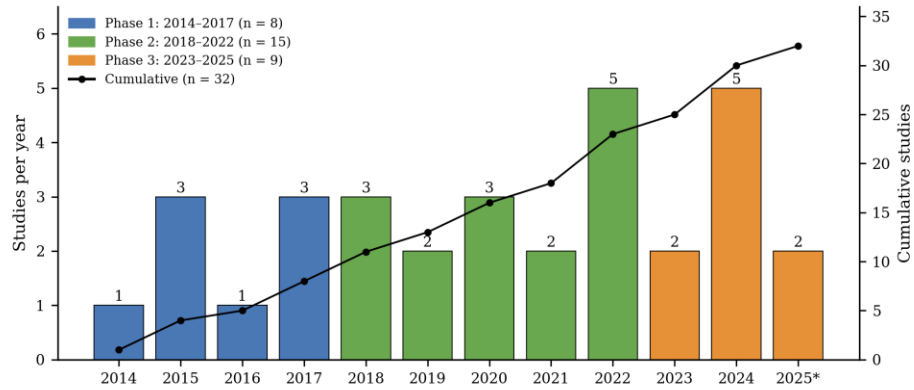


Figure 2. Publication trend of included studies by year and developmental phase (*2025 to April)

Keyword co-occurrence revealed five clusters: (1) supervised learning; (2) GNNs; (3) temporal modelling; (4) self-supervised and contrastive learning; and (5) the account-opening domain (synthetic identity, KYC, AML, BAF). Co-occurrence among Clusters 2, 3 and 4 rose sharply in Phase 3, whereas links between these clusters and Cluster 5 remained weak, showing that account-opening fraud has been largely overlooked by the self-supervised and hybrid-modelling literature.

Thematic synthesis

Theme 1: Limitations of supervised learning (RQ1). Ensembles such as random forest and XGBoost consistently outperform logistic regression and single trees on known fraud patterns (Carcillo et al., 2018; Dal Pozzolo et al., 2015), and deep networks removed much of the manual feature engineering of rule-based predecessors (Bolton & Hand, 2002). However, three limitations are common in all 14 supervised studies: label latency 30-90 days (9 studies), prevalence 0.1-1% (11 studies) and performance loss 15-30% under drift (7 studies). Each supervised study has at least one restriction. Under temporally ordered evaluation, the strongest ensembles deteriorate the fastest, and studies with AU-PRC report greater deterioration than those with AU-ROC, as AU-ROC hides the degradation of the minority class (Saito & Rehmsmeier, 2015). Label latency is a characteristic of the investigation process and not a feature-engineering issue and therefore this cannot be solved in the supervised paradigm alone.

Theme 2: Graph neural networks (RQ2). Across 18 graph studies, GNNs outperform tabular baselines by 10–20% on relational fraud benchmarks, but the gain is not uniform. The upper end (15-20%) is derived from dense, heterogeneous graphs from industry, like account-device graph from Alipay (Liu et al., 2018), while the lower end (10-12%) is from sparser academic benchmarks like Yelp and Amazon review graphs. No study has yet reported the gain of using AU-PRC or a temporally ordered split, so the benefit of account-opening data is yet to be quantified. GraphSAGE is able to support inductive inference on new applications (Hamilton et al., 2017), while CARE-GNN is capable of resisting camouflage (Dou et al., 2020). There are eight studies that find heterophily to be crucial, and the computation costs, static graphs temporal blindness, and dependencies of the inherited labels are not solved.

Theme 3: Temporal modelling (RQ2). In 11 temporal studies, transformer-based models were better able to capture long-range dependencies than LSTM-based models. This result was confirmed in a study with multi-day ring window in Cheng et al. (2022), a study with sub-hour application rhythm in digital banking in Wang et al. (2023) and a study with session-level velocity in transaction networks in Kim et al. (2024). Fusing GNN structure with temporal attention improved AUC by 12-18% over GNN-only baselines (Cheng et al., 2022). Graph and temporal models detect partly non-overlapping subsets of fraud (Cheng et al., 2022; Olaniyan et al., 2025), yet none of the 11 studies reported inference latency or computational cost under production-representative load.

Theme 4: Self-supervised and contrastive learning (RQ3). Most self-supervised fraud methods adapt SimCLR (Chen et al., 2020) and GraphCL (You et al., 2020), using the NT-Xent loss to contrast augmented views. Across 10 studies, graph contrastive learning outperformed supervised baselines when 10% or fewer labels were available, and the benefit was greatest when only labels older than 30 days were used, mirroring operational latency. Vinay et al. (2023) showed that contrastive representations generalise to unseen fraud tactics, and Zhang et al. (2024) reported precision of 0.87 and recall of 0.83 with robustness to temporal shift. These studies rely on rich entity histories that do not exist at account opening, so they are proof of concept rather than direct evidence of transfer. Six studies showed strong sensitivity to augmentation design, and none validated a fraud-specific augmentation that preserves shared-device clusters or coordinated IP usage.

Theme 5: Integrated hybrid architectures (RQ4). Six hybrid studies achieve 12-25% improvement with respect to the single-component baseline of joint graph-temporal modelling. The most comprehensive one is the Graph-Temporal Contrastive Transformer (GTCT) proposed by Olaniyan et al. (2025), which proved to enhance novel fraud-type detection by 15-25%. The six studies, however, on five different data sets, all use four different main metrics, of which none is BAF, so the range is a qualitative indication of complementarity rather than a transferable estimate. None of the hybrid studies performed label delay, used the BAF data set, or reported the inference latency. Table 3 below summarizes the key studies.

Comparison with the empirical literature further clarifies the limits of transferring reported results to digital account opening. Liu et al. (2018) examined heterogeneous account-device relationships, and Dou et al. (2020) addressed camouflage in graph-based fraud detection, supporting the relevance of relational structure while leaving application-time label delay unresolved. Cheng et al. (2022) demonstrated the value of spatial-temporal attention, but its reported performance does not by itself establish generalisation to new account applications without transaction histories. Similarly, Liu et al. (2022) supports contrastive learning for attributed-network anomalies, whereas Zhang et al. (2024) examines contrastive learning in the insurance domain. These empirical contributions support individual components of SSL-GT, but differ in data availability, fraud setting and evaluation protocol; they should not be treated as direct performance evidence for the integrated architecture.

Table 3. Comparative analysis of key studies (2017-2025)

Study (model)	Key findings	Limitations	Relevance to SSL-GT
Hamilton et al. (2017): GraphSAGE	Inductive learning that generalises to unseen nodes at scale	May dilute minority fraud signals; sampling-sensitive	Backbone for real-time scoring of new applications
Velickovic et al. (2018): GAT	Per-edge attention; gains on heterogeneous graphs	Costly on dense graphs; no temporal dimension	Per-neighbour attention weighting
Liu et al. (2018): GEM	AU-ROC 0.94; 15–20% over feature baselines on Alipay account-device graphs	Label-hungry; no temporal modelling	Heterogeneous graph construction
Wang et al. (2019): SemiGNN	Label-efficient semi-supervised hierarchical attention	Not evaluated under label latency	Semi-supervised precursor
Dou et al. (2020): CARE-GNN	Resists camouflaged fraudsters mimicking legitimate users	Fully supervised; static graph	Heterophily-aware construction
Liu et al. (2022): CoLA	Contrastive SSL beat supervised baselines under label scarcity	No latency or drift evaluation	Validates graph SSL
Zhang et al. (2024): POCL	Precision 0.87, recall 0.83; robust to temporal shift	No graph structure; insurance domain	Contrastive pre-training for drift
Reynisson et al. (2024): GraphGuard	Transferable embeddings on dynamic credit-card fraud graphs	No transformer; transaction-level; preprint	Closest graph contrastive precedent

Study (model)	Key findings	Limitations	Relevance to SSL-GT
Olaniyan et al. (2025): GTCT	GNN–transformer fusion; 15–25% over single-component models	No label-latency evaluation; not account opening	Primary hybrid reference

Reconciling apparent contradictions

Two apparent contradictions can be reconciled. First, self-supervised learning outperforms supervised baselines under low labels (Liu et al., 2022; Vinay et al., 2023), but the advantage diminishes as labels increase (Zhang et al., 2024); under banking label delay, recent labels are always scarce, so the low-label regime applies. Second, heterophily-aware models outperform standard GNNs in some studies (Shi et al., 2022; Tang et al., 2022), yet a standard attention GNN performs well in Cheng et al. (2022). Differences in camouflage intensity plausibly explain this, although no study has varied camouflage directly.

Certainty of evidence and publication bias

Funnel plots and Egger tests were not meaningful because the studies share no common effect-size metric, but publication bias is plausible because machine-learning venues favour positive results; the reported 10-25% gains are best read as optimistic upper bounds. Certainty was rated as follows: GNNs outperform tabular baselines on relational fraud benchmarks (moderate: consistent direction, but effect size varies with graph density and metric, and no study uses BAF); self-supervised methods outperform supervised baselines under label scarcity (moderate: consistent across domains but indirect to 90-day account-opening delay); hybrid models outperform single-component models (low to moderate: six studies, none with label-delay simulation or latency measurement); and SSL-GT simultaneously resolves label latency, relational blindness and concept drift (no direct evidence: an untested hypothesis).

Table 4. Research gaps identified across thematic clusters

Gap	Title	Description	Supporting evidence
1	Label dependency in supervised systems	Investigation delays of 30-90 days make recent fraud behaviour invisible to training pipelines	Jesus et al. (2022); Dal Pozzolo et al. (2015); Carcillo et al. (2018)
2	Absent SSL–graph–temporal integration for account opening	GNNs and SSL have advanced separately; their integration for account-opening fraud under label scarcity is absent	Olaniyan et al. (2025); Reynisson et al. (2024); Liu et al. (2022)
3	Insufficient drift-robustness evidence	Theory supports SSL robustness to shift, but evidence under realistic 30-90-day delays is limited	Gama et al. (2014); Hinder et al. (2024); Dal Pozzolo et al. (2018)
4	Neglect of deployment feasibility	Inference latency, resource use and real-time constraints are rarely measured	Olaniyan et al. (2025); Hamilton et al. (2017)
5	Unrealistic evaluation benchmarks	Benchmarks rarely simulate label latency or adversarial drift; BAF has not been used for any graph-temporal SSL model	Jesus et al. (2022); Cheng et al. (2022)
6	Augmentation design for fraud graphs	Random augmentation may destroy shared-device and coordinated-IP signals; no fraud-specific guidance exists	Liu et al. (2022); Reynisson et al. (2024); You et al. (2020)
7	Algorithmic fairness	No study analyses disparate impact across protected demographic groups	All themes

The SSL-GT conceptual framework

In response to these gaps, SSL-GT is proposed as a conceptual architecture rather than a validated solution. Table 5 links each of its six components to the supporting evidence. SSL-GT extends GTCT (Olaniyan et al., 2025) with graph contrastive pre-training and a delayed-supervision protocol, adds a transformer temporal encoder to GraphGuard (Reynisson et al., 2024), and adds relational structure to the tabular POCL design (Zhang et al., 2024). To the best of the evidence reviewed, no prior work combines all six components; the contribution is a novel combination of separately validated ideas.

Table 5. SSL-GT components and their evidence base

Component	Design specification	Gap addressed and evidence
Heterophily-aware graph construction	Multi-relational application graph with CARE-GNN-style neighbourhood filtering to resist camouflage	Gap 2; Dou et al. (2020); Shi et al. (2022)
Inductive GraphSAGE encoder	Neighbourhood-sampled encoding of unseen application nodes without full graph re-traversal, with attention-weighted aggregation	Gap 4; Hamilton et al. (2017); Velickovic et al. (2018)
Transformer temporal encoder	Multi-head self-attention over ordered application events to capture velocity, regularity and cross-session dependencies	Temporal blindness of graph models; Vaswani et al. (2017)
Graph contrastive pre-training	NT-Xent objective on augmented graph views that preserve shared-device and coordinated-IP subgraphs while applying feature and edge dropout	Gaps 1, 2 and 6; You et al. (2020); Reynisson et al. (2024)
Delayed-supervision fine-tuning	Fine-tuning with a simulated 90-day label delay using BAF temporal ordering; only labels confirmed before the cut-off are used	Gap 3; Zhang et al. (2024)
Joint spatial-temporal fusion	Late fusion of graph and temporal embeddings before the classification head	Gap 2; Cheng et al. (2022); Olaniyan et al. (2025)

The hypothesised benefits (label efficiency under delay, detection of novel strategies, resistance to camouflage and capture of complementary graph and temporal signals) are extrapolations from adjacent settings. The principal risks are augmentation sensitivity, inference cost and uncertain transfer of augmentations from other fraud domains.

The proposed evaluation protocol uses the NeurIPS 2022 BAF dataset (Jesus et al., 2022) as the primary benchmark. Unlike the Yelp and Amazon review graphs, which have static, fully observed structure and no temporal label ordering, BAF encodes application timestamps, enforces a temporally consistent split and reproduces realistic imbalance and drift. Labels for the final 90 days of training data will be withheld to simulate investigation delay (Dal Pozzolo et al., 2015; Jesus et al., 2022). AU-PRC and fraud-class F1 will be the primary metrics, because AU-ROC overstates performance under severe imbalance (Saito & Rehmsmeier, 2015), and inference latency under production-representative load will be a secondary metric addressing Gap 4.

Comparison with prior reviews

Unlike the narrative reviews of Akoglu et al. (2015) and Phua et al. (2010), this review applies PRISMA 2020 with reported inter-rater reliability and formal quality appraisal, covers GNNs, transformers and contrastive learning, focuses explicitly on account-opening fraud and derives its agenda from specific gaps. Each methodological generation resolved one weakness while adding another: supervised learning became label-dependent, GNNs added cost, self-supervised learning added augmentation uncertainty, and hybrids moved evaluation further from operational reality. The central challenge is now demonstrated success under realistic operating conditions rather than benchmark performance. Theoretically, heterophily in fraud graphs is best understood as an adversarial graph-learning problem, and domains with scarce labels but abundant structured unlabelled data are strong candidates for self-supervised adaptation.

Conclusion and recommendations

This systematic review of 34 studies provides 5 conclusions. In the first place, research has been in stages, with each stage addressing the primary limitation of the previous one, and introducing additional limitations and compromises. Second, supervised learning alone can't be used to detect digital account-opening fraud, due to label latency, class imbalance and adversarial concept drift. Third, GNNs can significantly enhance relational fraud detection and are label-dependent in current forms, yet need to be heterophily-aware. Fourth, self-supervised and contrastive methods are truly label efficient and distributional robust, but have not been tested in a 30–90-day label delay in account opening. Fifth, integrated graph-temporal hybrids are the state of the art but have not yet been evaluated in an operational setting. This evidence corroborates the proposed, yet to be validated, SSL-GT model - heterophily-aware graph construction, inductive GraphSAGE encoder, transformer temporal encoder, graph contrastive pre-training, delayed-supervision fine-tuning and joint spatial-temporal fusion. It remains to be seen in the next step of this research

whether SSL-GT can indeed achieve these goals in practice. Table 6 outlines the six recommended research priorities and calls on financial institutions and regulators to facilitate privacy-preserving data-sharing schemes that would allow for such an evaluation.

Table 6. Recommended research priorities

No.	Priority	Recommendation	Gaps addressed
P1	SSL-GT integration for account-opening fraud	Develop and evaluate an integrated self-supervised graph-temporal model on the BAF dataset under simulated 90-day label delay	1, 2, 3, 5
P2	Fraud-graph augmentation design	Develop theoretically grounded augmentations that preserve fraud-relevant structure while producing effective view diversity	2, 3, 6
P3	Operational feasibility benchmarking	Report inference latency, resource consumption and deployment cost alongside detection metrics	4
P4	Cross-domain generalisability	Build multi-dataset protocols and cross-institutional data-sharing frameworks	5
P5	Fairness and bias auditing	Make disparate-impact analysis across protected groups a standard evaluation component	7
P6	Adaptive retraining	Incorporate newly confirmed labels into pre-trained representations without full retraining	1, 3, 4

ACKNOWLEDGEMENTS

Thankful to the Cooperative University of Kenya for institutional support and the editors and reviewers for their guide and support during the publication process.

REFERENCES

- Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery*, 29(3), 626–688. <https://doi.org/10.1007/s10618-014-0365-y>
- Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3), 235–255. <https://doi.org/10.1214/ss/1042727940>
- Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with Spark. *Information Fusion*, 41, 182–194. <https://doi.org/10.1016/j.inffus.2017.09.005>
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (PMLR 119, pp. 1597–1607).
- Cheng, D., Wang, X., Zhang, Y., & Zhang, L. (2022). Graph neural network for fraud detection via spatial-temporal attention. *IEEE Transactions on Knowledge and Data Engineering*, 34(8), 3800–3813. <https://doi.org/10.1109/TKDE.2020.3025588>
- Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection and concept-drift adaptation with delayed supervised information. In *2015 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. <https://doi.org/10.1109/IJCNN.2015.7280527>
- Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784–3797. <https://doi.org/10.1109/TNNLS.2017.2736643>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
- Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (pp. 315–324). <https://doi.org/10.1145/3340531.3411903>
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44. <https://doi.org/10.1145/2523813>
- Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems* 30 (pp. 1024–1034).
- Hinder, F., Vaquet, V., & Hammer, B. (2024). One or two things we know about concept drift: A survey on monitoring in evolving environments. Part B: Locating and explaining concept drift. *Frontiers in Artificial Intelligence*, 7, Article 1330258. <https://doi.org/10.3389/frai.2024.1330258>
- Jesus, S., Pombal, J., Alves, D., Cruz, A. F., Saleiro, P., Ribeiro, R. P., Gama, J., & Bizarro, P. (2022). Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation. In *Advances in Neural Information Processing Systems 35, Datasets and Benchmarks Track*.
- Kim, J., Lee, S., & Park, C. (2024). Temporal attention mechanisms for sequential fraud detection in financial transaction networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3), 1842–1856.
- Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*.
- LexisNexis Risk Solutions. (2024). *True Cost of Fraud study: Financial services and lending*. LexisNexis Risk Solutions.



- University of Eldoret, Kenya, Mount Kenya University, Kenya, Chukwuemeka Odumegwu Ojukwu University, Nigeria, Kyambogo University, Uganda and University of Makeni, Sierra Leone
- Liu, Y., Li, Z., Pan, S., Gong, C., Zhou, C., & Karypis, G. (2022). Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6), 2378–2392.
- Liu, Z., Chen, C., Yang, X., Zhou, J., Li, X., & Song, L. (2018). Heterogeneous graph neural networks for malicious account detection. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 2077–2085). <https://doi.org/10.1145/3269206.3272010>
- Olaniyan, J., Olaniyan, D., Obagbuwa, I. C., & Ngafeeson, M. (2025). Graph-temporal contrastive transformer for financial fraud detection using transaction behavior modeling. *Algorithms*, 18(12), Article 770. <https://doi.org/10.3390/a18120770>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). *A comprehensive survey of data mining-based fraud detection research* (arXiv:1009.6119). arXiv.
- Reynisson, K., Schreyer, M., & Borth, D. (2024). *GraphGuard: Contrastive self-supervised learning for credit-card fraud detection in multi-relational dynamic graphs* (arXiv:2407.12440). arXiv.
- Saito, T., & 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. <https://doi.org/10.1371/journal.pone.0118432>
- Shi, F., Cao, Y., Shang, Y., Zhou, Y., Zhou, C., & Wu, J. (2022). H2-FDetector: A GNN-based fraud detector with homophilic and heterophilic connections. In *Proceedings of the ACM Web Conference 2022* (pp. 1486–1494). <https://doi.org/10.1145/3485447.3512195>
- Tang, J., Li, J., Gao, Z., & Li, J. (2022). Rethinking graph neural networks for anomaly detection. In *Proceedings of the 39th International Conference on Machine Learning* (PMLR 162).
- TransUnion. (2025). *State of omnichannel fraud report: H1 2025*. TransUnion.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30* (pp. 5998–6008).
- Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*.
- Vinay, M. S., Yuan, S., & Wu, X. (2023). Robust fraud detection via supervised contrastive learning. In *Proceedings of the 2023 IEEE International Conference on Big Data* (pp. 1279–1288). <https://doi.org/10.1109/BigData59044.2023.10386925>
- Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., Yu, Q., Zhou, J., Yang, S., & Qi, Y. (2019). A semi-supervised graph attentive network for financial fraud detection. In *Proceedings of the 2019 IEEE International Conference on Data Mining* (pp. 598–607). <https://doi.org/10.1109/ICDM.2019.00070>
- Wang, Z., Liu, Y., & Chen, H. (2023). Velocity-based temporal features for coordinated fraud detection in digital banking systems. *Expert Systems with Applications*, 211, Article 118571.
- You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2020). Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems 33*.
- Zhang, R., Cheng, D., Yang, J., Ouyang, Y., Wu, X., Zheng, Y., & Jiang, C. (2024). Pre-trained online contrastive learning for insurance fraud detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20), 22511–22519. <https://doi.org/10.1609/aaai.v38i20.30259>
- Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. In N. Japkowicz & J. Stefanowski (Eds.), *Big data analysis: New algorithms for a new society* (pp. 91–114). Springer. https://doi.org/10.1007/978-3-319-26989-4_4