RESEARCH PAPER

Causal-Temporal Graph Contrastive Learning for Robust Credit Card Fraud Detection under Concept Drift and Label Scarcity

D. Baskar[1], A. Nandha Kishore[2], C. Paul Daniel[3], M. Dakshinamoorthy[4]

[1]Assistant Professor, Department of Computer and Communication Engineering, V.S.B Engineering College, Karur, India.
Email: baskard123@gmail.com (Corresponding Author)

[2]Department of Computer and Communication Engineering, V.S.B Engineering College, Karur, India.
Email: nandhakishoremns@gmail.com

[3]Department of Computer and Communication Engineering, V.S.B Engineering College, Karur, India.
Email: vcmanohar5@gmail.com

[4]Department of Computer and Communication Engineering, V.S.B Engineering College, Karur, India.
Email: mjagannathankavitha@gmail.com

***Corresponding author: D. Baskar, Assistant Professor, Department of Computer and Communication Engineering, V.S.B Engineering College, Karur, India**
Email: baskard123@gmail.com

Received: 25th May, 2026; **Revised:** 7th June, 2026; **Accepted:** 12th June, 2026; **Available Online:** 14th June, 2026

ABSTRACT

Background

Credit card fraud detection systems operate under severe class imbalance, delayed feedback, and non-stationary data streams. Practical deployment therefore requires models that preserve high recall while controlling false positives and maintaining performance under temporal distribution shifts.

Objective

This paper proposes Causal-Temporal Graph Contrastive Learning (CT-GCL), a unified framework for fraud detection on dynamic, heterogeneous transaction graphs.

Materials and Methods

CT-GCL integrates temporal subgraph sampling and time-aware message aggregation to represent evolving user-merchant-device interactions, a causal/environmental disentanglement design with a counterfactual invariance objective that encourages predictions to be stable under environmental perturbations, and self-supervised contrastive representation learning to improve data efficiency under limited labels. We evaluate CT-GCL using a reproducible protocol based on strictly time-ordered train/validation/test splits, a fixed tuning budget per baseline, and repeated runs with multiple random seeds. Performance is assessed primarily with AUPRC, complemented by Recall@K and F1-score, which better reflect operational performance under extreme imbalance.

Results

Across three benchmark datasets (European Credit Card, IEEE-CIS, and BankSim), CT-GCL yields consistent gains over competitive tabular, sequential, and graph-based baselines and exhibits improved stability under temporal slicing of the test period.

Conclusion

The proposed CT-GCL framework demonstrates robust performance under concept drift and label scarcity, making it a promising solution for real-world credit card fraud detection systems.

Index Terms: Credit Card Fraud Detection, Financial Fraud Analytics, Heterogeneous Graphs, Temporal Graph Learning, Graph Neural Networks, Causal Invariance, Self-Supervised Learning, Contrastive Learning, Class Imbalance, Concept Drift, Precision–Recall (AUPRC).

How to cite this article: Baskar D, Nandha Kishore A, Paul Daniel C, Dakshinamoorthy M. Causal-Temporal Graph Contrastive Learning for Robust Credit Card Fraud Detection under Concept Drift and Label Scarcity. Int J Drug Deliv Technol. 2026;16(60s):468-483. DOI: 10.25258/ijddt.16.60s.57

Source of support: Nil.

Conflict of interest: None

I. INTRODUCTION

A. The Pervasiveness of Financial Fraud

The integration of information technology into the global financial ecosystem has streamlined commerce but simultaneously expanded the attack surface for malicious actors. As payment infrastructures transition toward frictionless, real-timesettlement systems, the window for effective fraud intervention has narrowed from days to milliseconds. In this environment, fraud detection systems must operate under stringent latency constraints while maintaining high predictive accuracy.

Recent studies highlight that fraud patterns evolve rapidly due to adversarial adaptation, seasonal behavioral shifts, and infrastructure changes, leading to non-stationary data distributions over time [1], [2]. Consequently, detection systems must balance high recall with operational constraints such as low

false positive rates and real-time decision requirements [3].

Beyond direct financial loss, false positives—legitimate transactions incorrectly flagged as fraudulent—create customer friction, reputational risk, and downstream operational cost. Therefore, the objective of a modern fraud detection system is inherently dual: maximize the capture rate of illicit activities (recall) while minimizing disruption to genuine users (preci- sion). Achieving this balance under extreme class imbalance and evolving behavioral patterns remains a central challenge in applied machine learning for financial security.

B. The Three-Body Problem of Fraud Detection

Despite significant advances in machine learning and graph- based modeling, credit card fraud detection remains con- strained by three tightly coupled structural challenges: extreme class imbalance, concept drift, and label scarcity. We refer to this triad as the "Three-Body Problem" of fraud detection because each factor dynamically interacts with the others, rendering isolated mitigation strategies insufficient.

Extreme Class Imbalance: Fraudulent transactions are intrinsically rare, often constituting less than 0.2%–4% of total transaction volume in widely used benchmarks. Under such skewed distributions, standard empirical risk minimization with cross-entropy loss tends to favor majority-class prediction, yielding high nominal accuracy but poor minority-class recall. While resampling techniques, cost-sensitive learning, and focal-style objectives partially address this issue, they may distort the natural data distribution or amplify outdated minority patterns under evolving conditions [1], [3].

Concept Drift and Non-Stationarity: Fraud is inherently adversarial. Attackers continuously adapt strategies in response to deployed defenses, leading to non-stationary data distribu- tions over time [2]. A model trained on historical fraud signa- tures may degrade when consumer behavior shifts seasonally or when fraud rings alter transaction velocity, device usage, or merchant targeting patterns. Correlation-driven models are par- ticularly vulnerable under such distribution shifts, especially when they rely on environment-specific artifacts rather than invariant behavioral mechanisms [4], [5].

Label Scarcity and Feedback Latency: In real-world deployment, fraud labels are rarely available instantaneously. Chargeback cycles and manual investigations introduce verifi- cation delays, meaning supervised models are trained on par- tially outdated or incomplete ground truth [1]. Meanwhile, the majority of transaction data remains unlabeled, representing a substantial but underutilized source of structural and temporal information that could inform representation learning.

These three factors interact in non-trivial ways. Imbalance exacerbates overfitting under drift; drift renders resampled minority examples obsolete; and label scarcity limits the ability of supervised systems to adapt rapidly. A robust fraud detection framework must therefore simultaneously address imbalance, non-stationarity, and limited supervision rather than treating them as independent challenges.

C. Limitations of the Current Paradigm

For over a decade, financial fraud detection has been dominated by supervised machine learning models operating on tabular representations of transactions. Classical approaches such as Logistic Regression, Random Forests, and Gradient Boosting Decision Trees (e.g., XGBoost, LightGBM) remain competitive baselines in empirical evaluations due to their efficiency and strong performance on structured data [1], [3]. These models typically treat each transaction as an indepen- dent and identically distributed (i.i.d.) instance, relying primar- ily on handcrafted features derived from historical statistics.

However, the i.i.d. assumption constitutes a fundamental limitation in realistic fraud settings. Fraudulent behavior is inherently relational: transactions may share devices, payment instruments, merchant identifiers, or participate in coordinated activity patterns. Purely tabular models cannot explicitly ex- ploit such higher-order dependencies. Graph-based learning frameworks, including Graph Convolutional Networks (GCN) [6], Graph Attention Networks (GAT) [7], and fraud-specific graph models such as SemiGNN [8], have demonstrated the importance of relational modeling for capturing cross-entity risk propagation.

Furthermore, supervised detectors remain vulnerable under operational non-stationarity. Realistic deployment must ac- count for evolving customer behavior, adaptive adversaries, verification latency, and delayed feedback, all of which induce distribution shift over time [1], [2]. Correlation-driven mod- els that rely on environment-specific artifacts can deteriorate rapidly when contextual signals change. Although imbalance- aware training strategies—such as resampling, cost-sensitive objectives, and calibrated thresholding—have been extensively studied [3], these techniques alone do not address the deeper issue of invariant generalization under dynamic relational environments.

D. The Graph Neural Network Revolution and Its Gaps

To address the relational limitations of tabular models, the research community has increasingly adopted graph-based learning frameworks. Foundational architectures such as Graph Convolutional Networks (GCN) [6] and Graph Attention Net- works (GAT) [7] enable message passing over structured data, allowing node representations to incorporate neighborhood signals. In financial fraud detection, semi-supervised graph models such as SemiGNN demonstrate that relational informa- tion—shared devices, merchants, or transaction patterns—can significantly enhance risk propagation modeling [8].

Subsequent advances in dynamic graph modeling further

extend this paradigm to temporal settings. Architectures such as Temporal Graph Networks (TGN) [9] and Temporal Graph Attention Networks (TGAT) [10] enable representation learn- ing over evolving graph streams, which is particularly relevant for fraud scenarios characterized by sequential behavioral shifts.

Despite these advancements, critical gaps remain. Many graph models operate primarily as correlation-driven engines and may overfit to environment-specific contexts (e.g., fre- quently used merchants or shared infrastructure) that shift over time. Recent research in graph out-of-distribution gen- eralization highlights the vulnerability of message-passing ar- chitectures under structural distribution shifts [5], [11]. These findings underscore the need for invariant and robustness- aware graph learning under deployment constraints [12].

Moreover, graph-based fraud detection remains label-limited. Self-supervised and contrastive learning objectives have emerged as promising strategies to improve representa- tion quality under scarce supervision. In particular, contrastive learning on heterogeneous or temporal graphs has demon- strated improved robustness and data efficiency [13]–[15]. However, integrating temporal dynamics, causal invariance, and self-supervision within a unified fraud detection frame- work remains underexplored.

E. Research Contributions

This work bridges the identified gaps by introducing **Causal-Temporal Graph Contrastive Learning (CT-GCL)**, a unified framework tailored to the operational realities of modern fraud detection. The primary contributions of this study are summarized as follows:

1) **Unified Causal-Temporal Architecture:** We propose an architecture that integrates causal intervention mecha-

   nisms with temporal attention, enabling the model to separate invariant fraud-related signals from transient environmental context while adapting to non-stationary transaction dynamics.
2) **Dual-View Contrastive Representation Learning:** We introduce a contrastive learning objective tailored for heterogeneous and temporal graphs. By leveraging struc- tural and temporal augmentations, the framework learns robust node representations from partially labeled data, mitigating label scarcity and improving representation stability.
3) **Drift-Aware Empirical Evaluation:** We conduct a comprehensive experimental study comparing CT-GCL against representative baselines spanning tabular mod- els, sequential architectures, and graph-based approaches (e.g., optimized boosting, temporal attention models, and semi-supervised graph frameworks) [8], [15]–[17]. Re- sults demonstrate consistent improvements in AUPRC and reduced degradation under temporally segmented evaluation protocols.
4) **Structured Methodological Positioning:** We provide a structured synthesis of prior work, identifying limitations in correlation-driven, imbalance-focused, and purely self-supervised approaches, and positioning CT-GCL as an integrated solution that combines temporal modeling, causal invariance, and contrastive regularization within a single coherent framework.

II. STRUCTURED LITERATURE REVIEW

The domain of credit card fraud detection is characterized by a rapid turnover of methodologies. This section provides a structured analysis of recent peer-reviewed literature, categorizing approaches into three distinct generations: Tabu- lar Learning, Sequence Modeling, and Graph Representation Learning.

A. Generation I: Tabular Machine Learning and Ensemble Methods

The foundational layer of fraud detection research relies on feature engineering and supervised classification applied to tabular transaction records. Classical statistical models and ensemble-based learners remain competitive due to their efficiency, interpretability, and strong performance under structured data constraints.

- **Ensemble and Deep Baselines:** Tree-based ensembles such as Gradient Boosting Decision Trees (e.g., XGBoost, LightGBM) and Random Forests consistently demon- strate strong performance in fraud benchmarks when combined with careful preprocessing and imbalance- aware evaluation protocols [1], [18]. Deep feedforward baselines have also been explored, though their advantage often depends on feature richness and dataset scale [3].

- **Class Imbalance as the Central Constraint:** Fraud detection is fundamentally characterized by extreme label imbalance. Accuracy is therefore a misleading metric, and evaluation must emphasize metrics such as AUPRC,recall at fixed precision, or cost-sensitive risk mini- mization [2], [3]. Imbalance-aware strategies—including resampling, cost-sensitive learning, and calibrated thresh- olding—have been extensively studied in both academic and applied contexts [12], [18].
- **Limitations:** Purely tabular models treat each transaction as independent and identically distributed, ignoring rela- tional dependencies between entities. As a result, they struggle to capture collusion, shared devices, synthetic identity rings, or coordinated merchant usage that mani- fest as structured patterns across accounts and infrastruc- ture [8], [12]. This independence assumption limits their ability to model cross-entity risk propagation.

B. Generation II: Deep Sequence Modeling

Recognizing that user behavior is inherently sequential, sub- sequent research models transactions as time-ordered events to capture temporal signatures such as bursts, velocity shifts, spending periodicity, and anomalous short-term deviations [3]. By explicitly encoding temporal order, sequence models move beyond static feature aggregation and enable behavioral pattern tracking across transaction histories.

- **Temporal Representation:** Recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) architec- tures, and attention-based sequence encoders have been applied to fraud detection to capture evolving behav- ioral trajectories. Attention mechanisms in particular al-low models to prioritize recent or contextually relevant transactions when forming risk estimates [17]. These approaches are effective in detecting abrupt short-horizon behavioral shifts that are difficult to represent using static tabular aggregates.
- **Limitations:** While sequence models successfully cap- ture *intra-user* temporal dynamics, they typically operate at the individual account level and do not naturally model *inter-user* or *inter-entity* dependencies (e.g., shared devices, coordinated merchant usage, fraud rings). As a result, they remain limited in detecting organized or cross-entity fraud patterns. Moreover, purely sequential architectures may still suffer under distribution shift if temporal correlations are environment-specific rather than causally stable [2], [4].

C. Generation III: Graph Representation Learning (GNNs)

Graph learning reframes fraud detection as classification over a multi-entity interaction network, where nodes represent accounts, devices, merchants, or transactions and edges encode behavioral or infrastructural relationships. Foundational archi- tectures such as Graph Convolutional Networks (GCN) [6] and Graph Attention Networks (GAT) [7] enable message passing across relational structures, allowing node representations to incorporate neighborhood information.

Fraud-specific graph models demonstrate that relational propagation significantly enhances detection performance compared to purely tabular or sequential baselines. Semi- supervised graph attentive learning, for example, fuses node attributes with neighborhood signals to capture shared-risk structures [8]. More recent work emphasizes cost-sensitive op- timization and group-level structure, including reinforcement- driven weighting strategies to identify influential fraud actors in sparse regimes [19].

Despite these advances, most GNN-based fraud detectors remain correlation-driven. Standard message-passing mech-anisms aggregate neighboring information without explic- itly distinguishing invariant behavioral mechanisms from environment-specific artifacts. As a result, these models may overfit to contextual signals—such as temporarily compro- mised merchants or regional infrastructure—that shift under temporal distribution drift.

D. Generation IV: Self-Supervised Graph Contrastive Learn- ing

To mitigate label scarcity and improve representation robust-ness, recent research has pivoted toward Graph Contrastive Learning (GCL) frameworks. These methods leverage self-supervised objectives to learn structural representations from unlabeled data by maximizing agreement between augmented views of the same graph instance.

Contrastive learning in representation modeling was for-malized through the InfoNCE objective in Contrastive Pre-dictive Coding [20]. In graph domains, mutual-information-based approaches such as Deep Graph Infomax [13] and augmentation-based Graph Contrastive Learning (GraphCL) [14] demonstrate that contrastive objectives can significantly improve embedding quality without dense supervision.

In fraud detection, Temporal Heterogeneous Graph Con-trastive Learning (TH-GCL) constructs structural and temporal perturbations of transaction graphs and aligns embeddings across views [15]. By leveraging unlabeled interactions, these approaches enhance robustness under sparse labeling and partial observability.

However, while contrastive objectives improve data effi-ciency and representation stability, they primarily enforce consistency across augmented views. They do not explicitly disentangle invariant causal mechanisms from spurious en-vironmental correlations that may shift across deployment contexts.

E. Generation V: Causal and Invariant Graph Learning

The emerging frontier integrates causal inference principles with graph neural networks to improve out-of-distribution (OOD) generalization. Rather than directly modeling $P(Y|G)$, invariant learning frameworks seek to separate stable causal mechanisms from environment-dependent correlations.

Invariant Risk Minimization (IRM) formalizes the objec- tive of learning representations whose predictive relationships remain stable across environments [4]. More recent work ex- tends this idea to graph domains, proposing invariant learning strategies under structural distribution shift [5], [11]. Causal representation learning further emphasizes disentangling stable mechanisms from contextual variation [21].In fraud detection, this paradigm is particularly relevant because environmental factors—such as merchant category, geographic region, or seasonal shopping patterns—may cor- relate with fraud in one time period but shift in another. A causally grounded model seeks to capture invariant mecha- nisms such as abnormal velocity, coordinated device usage, or dense relational cliques rather than memorizing compromised entities.

While promising, causal graph learning remains underex- plored in financial fraud detection. Existing methods often address either temporal modeling or self-supervised learning independently. A unified framework that integrates temporal dynamics, contrastive self-supervision, and causal invariance remains limited in current literature.

F. Identified Research Gaps

Despite the methodological evolution across five genera- tions, several structural limitations persist.

First, most Generation I and II approaches operate under implicit i.i.d. or purely sequential assumptions that fail to capture multi-entity relational dependencies. While sequence models improve intra-user behavioral modeling, they lack mechanisms to propagate risk signals across shared devices, merchants, or coordinated fraud rings.

Second, Generation III graph-based methods signifi- cantly enhance relational modeling but remain predominantly correlation-driven. Standard message-passing architectures ag- gregate neighboring information without explicitly distin- guishing invariant behavioral mechanisms from environment- specific artifacts. Consequently, these models may overfit to contextual signals—such as temporarily compromised mer- chants or regional infrastructure—that shift under temporal distribution drift [5], [11].

Third, although Generation IV contrastive learning improves data efficiency under label scarcity by aligning augmented views [13], [14], it primarily enforces representational con- sistency rather than causal stability. Self-supervised objectives enhance robustness to noise but do not explicitly block spuri- ous causal pathways that vary across environments.

Finally, Generation V causal and invariant graph learning introduces intervention-based training strategies to promote stability across environments [4], [21]. However, existing frameworks often address either causal disentanglement or temporal dynamics independently. An integrated design that simultaneously combines (i) temporal modeling for drift adap- tation, (ii) self-supervised contrastive learning for label effi- ciency, and (iii) explicit causal invariance objectives for robust out-of-distribution generalization remains underexplored in financial fraud detection.

These gaps motivate the development of a unified Causal-Temporal Graph Contrastive Learning framework that jointly addresses relational modeling, temporal non-stationarity, label scarcity, and environmental invariance within a single coherent architecture.

TABLE I
TEMPORAL GRAPH CONSTRUCTION AND LEAKAGE PREVENTION

Item	Guideline
Time split	Train/validation/test are strictly time-ordered; no shuffling across periods.
Feature fitting	Any scaling/encoding is fit on the training pe- riod only and applied forward.
Edge gating	Edges for a transaction at time t may only use events with timestamp $\le t$.
Neighbor sampling	Sampling probability is time-decayed (recent neighbors prioritized).
Similarity graphs	When entity IDs are missing, similarity edges are computed within the same time window to avoid future information.

TABLE II
SUMMARY OF CORE NOTATION USED IN THE CT-GCL FRAMEWORK

Symbol	Meaning
$G_t = (V_t, E_t)$	Transaction graph at time t
$\tau(v)$	Node type of node v
x_v	Feature vector of node v
$y \in \{0, 1\}$	Label (legitimate/fraud)
G_v	Sampled k-hop subgraph for target v
H_C, H_E	Causal / environmental representations
L_{cl}	Contrastive (InfoNCE) loss
L_{inv}	Invariance loss (counterfactual consistency)

III. PROBLEM FORMULATION

We formalize the credit card fraud detection task as a node

Transaction $\xrightarrow{\quad r \quad}$ *User*. Each relation induces a type-specific adjacency matrix $A^{(t)}_r$.

- **Features (X_t):** Each node $v \in V_t$ is associated with a feature vector $x_v \in \mathbb{R}^d$, derived from historical aggregates and contextual attributes computed up to time t.

b) Local Subgraph Extraction.: For a target node v at time t, prediction is performed over a sampled k-hop ego-subgraph $G_v \subseteq G_t$. Let

$$N_k(v) = \{u \mid \text{dist}(u, v) \le k\} \tag{3}$$

denote the k-hop neighborhood of v. The induced subgraph

$$G_v = G_t[N_k(v)] \tag{4}$$

classification problem within a dynamic, heterogeneous graph.

A. Dynamic Heterogeneous Graph

Let $G_t = (V_t, E_t)$ denote the dynamic transaction graph at time t. Fraud detection operates over a continuous event stream rather than a static graph snapshot. Formally, let

serves as the local relational context for classification. This localized formulation ensures computational tractability while preserving higher-order structural dependencies critical for fraud detection.

B. The Classification Task

Given a target transaction node v occurring at time t, the objective is to learn a predictive mapping

$$S = \{(u_i, r_i, v_i, t_i)\}_{i=1}^{N} \quad (1)$$

denote a sequence of time-stamped interactions, where $u_i, v_i \in V$ are entities (e.g., users, merchants, devices, transactions), $r_i \in R$ is the relation type, and $t_i \in R^+$ is the transaction timestamp. The graph at time t is constructed as

$$G_t = \{(u_i, r_i, v_i) \mid t_i \le t\}. \quad (2)$$

This formulation enforces temporal causality: representations at time t are computed exclusively from information available up to t. Such causally consistent graph construction prevents look-ahead bias, which can otherwise inflate performance under offline evaluation and distort robustness assessment under distribution shift.

a) Heterogeneous Structure.: The graph is heterogeneous, containing multiple node and relation types:

- **Nodes (V_t):** The node set consists of entity types A, with type mapping $\tau : V_t \to A$. Typical types include *Account*, *Device*, *Merchant*, and *Transaction*.
- **Edges (E_t):** The edge set includes typed relations R, such as $User \xrightarrow{buys} Item$, $User \xrightarrow{uses} Device$, and

$$f_\vartheta : G_v \to [0, 1], \quad (5)$$

where G_v denotes the k-hop ego-subgraph centered at v, and $f_\vartheta(G_v)$ estimates the probability of fraud. The binary label is defined as $y_v \in \{0, 1\}$, where 1 indicates fraudulent activity and 0 indicates legitimate behavior.

a) Delayed and Partial Supervision.: In operational environments, labels are not immediately observable. Let Δ denote the feedback delay between transaction occurrence and fraud confirmation (e.g., chargeback or manual review). At training time, the learner observes a partially labeled dataset

$$D_L = \{(G_{v_i}, y_{v_i})\}_{i \in L}, \quad (6)$$

where $L \subset \{1, \ldots, N\}$ indexes transactions with confirmed labels. The remaining transactions constitute an unlabeled set D_U.

This setting naturally induces a semi-supervised learning problem over dynamic graphs, where structural and temporal information from D_U can be exploited to improve representation learning despite incomplete supervision.

b) Class Imbalance.: The empirical class prior satisfies

$$P(Y = 1) \ll P(Y = 0), \quad (7)$$

often by several orders of magnitude. Let $\pi = P(Y = 1)$ denote the minority prior. Under extreme imbalance ($\pi \to 0$), minimizing unweighted empirical risk may lead to degenerate predictors biased toward majority-class prediction. Therefore, the learning objective must preserve minority-class sensitivity while maintaining probabilistic calibration.

c) Risk Minimization Under Temporal Constraints.: Let $P_t(G, Y)$ denote the joint distribution at time t. The supervised objective seeks to minimize empirical risk:

$$L_{sup}(\vartheta) = E_{(G_v, y_v)\sim P_t}[\ell(f_\vartheta(G_v), y_v)], \quad (8)$$

where $\ell(\cdot)$ is a binary classification loss (e.g., cross-entropy). However, inference at time t is performed under partial observability: only interactions with timestamps $\leq t$ are available, and labels for some transactions may only be revealed at $t + \Delta$. Consequently, the predictor must operate under temporally causal feature construction while remaining robust to delayed supervision and evolving data distributions.

C. Temporal Distribution Shift

Unlike static node classification tasks, fraud detection operates under non-stationary data distributions driven by adversarial adaptation and evolving user behavior.

Let $P_t(G, Y)$ denote the joint distribution of graph structure and labels at time t. In adversarial environments,

$$P_t(G, Y) \neq P_{t+\delta}(G, Y), \quad (9)$$

where $\delta > 0$ denotes a temporal offset. Such temporal distribution shift arises from changes in transaction patterns, merchant infrastructure, device usage, and fraud strategies.

a) Environment-Based Formulation.: We model each time slice or data segment as an environment $e \in E$ with corresponding distribution $P_e(G, Y)$. Standard empirical risk minimization (ERM) optimizes performance on the training environment e_{train}:

$$\min_\vartheta E_{(G_v, y_v)\sim P_{e_{train}}}[\ell(f_\vartheta(G_v), y_v)]. \quad (10)$$

However, under distribution shift, the deployment environment e_{test} may differ:

$$P_{e_{train}}(G, Y) \neq P_{e_{test}}(G, Y). \quad (11)$$

This motivates minimizing worst-case risk across environments:

$$\min_\vartheta \max_{e\in E} E_{(G_v, y_v)\sim P_e}[\ell(f_\vartheta(G_v), y_v)], \quad (12)$$

c) Implication for Fraud Detection.: The central challenge is therefore to learn a predictor that generalizes across time-dependent environmental shifts while remaining sensitive to invariant fraud mechanisms. This requires representations that are robust to contextual variation yet responsive to structural behavioral anomalies indicative of fraudulent intent.

D. Causal Structural Model (CSM)

To ground our methodology in causal theory, we propose the following Structural Causal Model for the data generation process.

Let:

- C denote **causal factors** (e.g., stolen credentials, abnormal transaction velocity) that directly determine fraudulent behavior.
- E denote **environmental factors** (e.g., merchant category, geographic region, seasonal activity) that may correlate with fraud but do not causally generate it.

We assume the structural relationships:

$$C \to G, \quad E \to G, \quad C \to Y.$$

Causal factors influence graph structure (e.g., fraudsters forming dense relational cliques), while environmental factors also shape connectivity patterns (e.g., users clustering around popular merchants). The fraud label Y is determined by causal mechanisms:

$$Y = h(C),$$

where $h(\cdot)$ is invariant across environments.

Standard graph neural networks implicitly model:

$$P(Y|G) = P(Y|C, E), \quad (14)$$

thereby entangling the causal path

$$C \to G \to Y$$

with the spurious path

$$E \to G \to Y.$$

Under temporal distribution shift,

$$P_t(E) \neq P_{t+\delta}(E), \quad (15)$$

a formulation aligned with domain generalization principles [22], [23].

b) Invariant Representation Learning.: Invariant Risk Minimization (IRM) proposes learning representations $h_v = \phi_\vartheta(G_v)$ such that the optimal classifier remains stable across environments [4]. Formally, we seek representations satisfying:

$$P_e(Y \mid h_v) = P_{e'}(Y \mid h_v), \quad \forall e, e' \in E. \tag{13}$$

In graph-structured data, recent studies demonstrate that message-passing architectures are particularly sensitive to structural distribution shifts when relational context entangles causal and environmental signals [5], [11]. Aggregating all neighbors indiscriminately may propagate spurious correlations tied to transient contextual factors.while the structural mechanism $P(Y \mid C)$ remains stable. Consequently, predictors that rely on environmental correla- tions risk performance degradation when contextual variables change.

a) Invariant Representation Objective.: The objective of CT-GCL is to approximate the invariant conditional distribu- tion $P(Y \mid C)$ by learning a representation decomposition:

$$h_v = \phi_\vartheta(G_v) = H_C \oplus H_E, \tag{16}$$

for perturbed H'_E drawn from alternative environments. By encouraging invariance to such perturbations, CT-GCL approximates an invariance constraint without requiring explicit identification of structural equations.

IV. METHODOLOGY: CAUSAL-TEMPORAL GRAPH CONTRASTIVE LEARNING (CT-GCL)

The proposed CT-GCL framework is composed of four integrated modules: (1) Temporal Heterogeneous Subgraph Construction, (2) Causal Disentanglement via Counterfactual Intervention, (3) Dual-View Contrastive Learning, and (4) Multi-Objective Optimization.

A. Module 1: Temporal Heterogeneous Subgraph Construction

Processing the global graph is computationally intractable and may propagate obsolete structural dependencies. For each target transaction v, we construct a localized k-hop subgraph G_v centered at v.

Temporal Sampling. We employ a time-decay sampling strategy. For neighbor u with time difference $\Delta t = t_v - t_u$, the sampling probability follows:

$$P(u|v) \propto e^{-\beta \Delta t} \tag{19}$$

where $\beta > 0$ controls decay strength. This prioritizes recent interactions while fading out obsolete structural patterns, thereby implicitly handling concept drift.

Heterogeneous Projection. Since nodes may have different feature dimensions, we project them into a common latent space:

where H_C captures stable causal mechanisms and H_E encodes environment-specific variation. We seek representations satisfying:

$$P(Y \mid H_C, H_E) = P(Y \mid H_C), \tag{17}$$

ensuring prediction depends only on invariant causal features.

b) Counterfactual Consistency.: This formulation aligns with causal representation learning, which seeks to disentan- gle stable mechanisms from environment-dependent variation [21]. The principle of invariant prediction further asserts that causal relationships remain stable across environments, whereas spurious correlations do not [24]. Operationally, this motivates enforcing prediction consistency under counterfac- tual perturbations of environmental components:

$$f_\vartheta(H_C, H_E) \approx f_\vartheta(H_C, H'), \tag{18}$$

$$H = \text{Combine}(H_C, H_E) = W_{comb}[H_C || H_E] \tag{23}$$

where W_{comb} is a learnable projection matrix.

Causal Intervention. To ensure H_C captures invariant mechanisms, we approximate counterfactual intervention by generating:

$$H' = \text{Combine}(H_C, H_E^{rand}) \tag{24}$$

where H_E^{rand} is sampled from another instance in the batch.

Consistency Constraint. The model is trained to produce consistent predictions:

$$f_\vartheta(H) \approx f_\vartheta(H') \tag{25}$$

This intervention mechanism approximates counterfactual stability. If predictions vary significantly after replacing H_E, the model is relying on spurious environmental correlations. By enforcing consistency, the optimization objective penalizes dependence on unstable contextual signals and concentrates predictive information within H_C.

This design operationalizes the principle that fraud intent should remain invariant to superficial environmental changes such as merchant identity or geographic region.

C. Module 3: Dual-View Contrastive Learning

To address label scarcity, we construct a self-supervised objective using two augmented views of G_v:

- **Structural Perturbation:** Stochastic edge dropout with rate $p_s = 0.2$, prioritizing low-centrality edges.
- **Temporal Masking:** Masking features of the most recent 10% of transactions to encourage historical reasoning.

Let z_1 and z_2 denote embeddings of transaction v from the two views. We minimize the InfoNCE loss:

$$h_v^{(0)} = W_{\tau(v)} x_v \tag{20}$$

where $W_{\tau(v)}$ is a learnable projection matrix specific to node type $\tau(v)$. This ensures type-aware representation alignment before message passing.

B. Module 2: Causal Disentanglement

This module separates the node representation into a causal component H_C and an environmental component H_E using two parallel GNN encoders:

$$H_C = \mathrm{GNN}_C(G_v) \tag{21}$$

$$H_E = \mathrm{GNN}_E(G_v) \tag{22}$$

The combined representation is defined as:

$$L_{cl} = -\log \sum_{z' \in B} \frac{\exp(\mathrm{sim}(z_1, z_2)/\tau)}{\exp(\mathrm{sim}(z_1, z')/\tau)} \tag{26}$$

where sim denotes cosine similarity, τ is a temperature parameter, and B represents batch negatives.

From an information-theoretic perspective, this maximizes a lower bound on mutual information [20]. Graph adaptations include Deep Graph Infomax [13] and GraphCL [14]. Although contrastive objectives may suffer from false negatives [25], the concurrent invariance constraint reduces excessive dependence on instance-specific environmental context.

By aligning embeddings across structural and temporal perturbations, the encoder learns invariant semantic features that persist under edge sparsification, temporal masking, and variation.

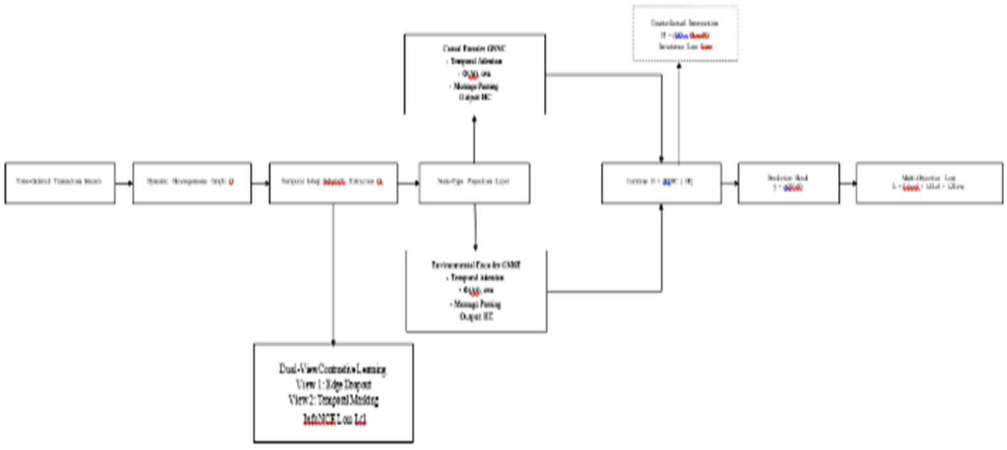

Fig. 1. Architecture of the proposed CT-GCL framework.

D. Module 4: Temporal Attention Mechanism

To explicitly model temporal evolution, message passing incorporates relative time encoding:

$$\alpha_{vu} = \text{Softmax}\ \text{LeakyReLU}\ a^T [h_v \| h_u \| \Phi(t_v - t_u)] \tag{27}$$

Unlike static graph attention, this mechanism captures velocity-based fraud signatures. Sequences of geographically distant transactions occurring within seconds may indicate automated fraud, whereas identical transactions spaced over weeks may be benign.

Temporal attention also mitigates concept drift by down-weighting outdated historical interactions and emphasizing recent behavioral evidence.

a) Prediction Layer: The final fraud probability is computed as:

$$\hat{y} = f_\vartheta(H) = \sigma(W_o H) \tag{28}$$

where W_o is a learnable weight matrix and $\sigma(\cdot)$ denotes the sigmoid function.

E. Optimization Objective

The final multi-task loss is:

$$\mathbf{L} = \mathbf{L}_{focal} + \lambda_1 \mathbf{L}_{cl} + \lambda_2 \mathbf{L}_{inv} \tag{29}$$

- L_{focal}: Focal loss for imbalanced classification [26].
- L_{cl}: Contrastive InfoNCE loss.
- L_{inv}: Invariance loss defined as

$$\mathbf{L}_{inv} = \text{JS}\ p_\vartheta(y|H) \,\|\, p_\vartheta(y|H') \tag{30}$$

The focal loss emphasizes minority-class discrimination. The contrastive loss regularizes representation learning using unlabeled data. The invariance loss enforces counterfactual stability under environmental perturbations.

These objectives operate at complementary levels: prediction layer, representation layer, and structural invariance layer. Their joint optimization produces representations that are discriminative, data-efficient, and robust under temporal distribution shift.

V. IMPLEMENTATION DETAILS AND REPRODUCIBILITY

This section provides a detailed description of the experimental protocol, hyperparameter configuration, fairness controls, and computational characteristics of CT-GCL to ensure transparency and reproducibility.

Training Protocol

We adopt a strictly chronological temporal split protocol aligned with realistic fraud deployment settings [1], [27]. The dataset is partitioned into three consecutive time windows: an early window for training, an intermediate window for validation, and a later window for testing. This design exposes the model to temporal distribution shift and delayed label confirmation effects that naturally arise in operational fraud detection systems.

Unlike random splits frequently used in static graph benchmarks, chronological partitioning prevents leakage of future structural signals into the training process. In dynamic financial networks, graph connectivity patterns and transactional statistics evolve over time due to changing user behavior and adversarial adaptation [2]. Random partitioning may therefore introduce relational information from future periods into training, leading to optimistic offline estimates that fail to transfer to deployment [1]. The temporal protocol ensures that all predictive signals used during model fitting would have been available at deployment time.

a) Drift-Aware Evaluation.: Fraud detection systems operate under non-stationary distributions, where both legitimate behavior and fraud strategies evolve continuously — a phenomenon commonly referred to as concept drift [2]. By enforcing strict chronological separation between training, validation, and testing periods, the evaluation framework captures natural drift effects and provides a realistic estimate of temporal generalization capacity, consistent with contemporary fraud evaluation practices [12].

b) Feedback Latency Simulation.: Fraud labels are often confirmed only after non-trivial delay (e.g., chargeback cycles or manual review), creating verification latency challenges [1]. By separating validation and test periods chronologically,

evaluation mimics delayed supervision and prevents implicit access to future label confirmations. This setup provides a more faithful estimate of performance under feedback latency, which is a central issue in real-world fraud detection systems [27].

TABLE III
DEFAULT HYPERPARAMETERS (UNLESS TUNED)

Hyperparameter	Value
Embedding dimension d	128
GNN layers	2
Neighborhood fanout	(15, 10)
Batch size	1024
Optimizer	Adam
Learning rate	10^{-3}
Temperature τ (InfoNCE)	0.2
λ_1 (contrastive)	0.1
λ_2 (invariance)	1.0
Early stopping metric	Validation AUPRC

c) Determinism and Random Seeds.: To ensure experimental determinism, all stochastic processes are controlled via fixed random seeds. These include:

- Parameter initialization,
- Mini-batch construction and shuffling,
- Neighborhood sampling during temporal subgraph extraction,
- Edge perturbation for contrastive views,
- Counterfactual environmental substitution in causal intervention.

Each experiment is repeated over five independent runs. Reported metrics correspond to the mean and standard deviation across runs, reducing variance-induced bias and improving robustness of reported performance.

d) Temporal Leakage Prevention.: Temporal splits strictly follow chronological ordering without any shuffling across periods. All preprocessing operations—including normalization, categorical encoding, and feature scaling—are fitted exclusively on the training split and applied forward to validation and test splits. No statistics from validation or test windows are used during model fitting.

Furthermore, subgraph construction respects temporal causality: for a transaction occurring at time t, only events with timestamps $\leq t$ are included in the corresponding subgraph. This guarantees that no future edges or relational signals are incorporated during representation learning.

This protocol eliminates look-ahead bias, which can otherwise inflate offline evaluation in fraud detection benchmarks [1], [2].

B. Hyperparameters and Optimization

Unless otherwise specified, we use the default configuration summarized in Table III. The embedding dimension is set to $d = 128$, and two GNN layers are employed. This configuration balances representational capacity with the risk of over-smoothing that may arise in deeper graph architectures [6], [7].

Neighborhood sampling fanout is set to $(15, 10)$ for two-hop subgraph construction, following inductive mini-batch training strategies commonly used in scalable graph learning [28]. This provides sufficient structural context while maintaining computational tractability.

The temperature parameter $\tau = 0.2$ for the InfoNCE objective follows empirical stability observations in graph contrastive learning [14], [20]. Lower temperatures increase gradient sharpness but may introduce instability, whereas excessively large values reduce discriminative power between positive and negative pairs.

The contrastive coefficient $\lambda_1 = 0.1$ regulates the contribution of the self-supervised alignment objective relative to supervised classification. This value ensures that representation regularization complements, rather than dominates, fraud detection.

The invariance coefficient $\lambda_2 = 1.0$ enforces counterfactual consistency between original and environment-perturbed representations, aligning with invariant learning principles [4]. Smaller values weaken the invariance constraint, while excessively large values may constrain representational flexibility.

Optimization is performed using Adam with learning rate 10^{-3}. We employ early stopping based on validation AUPRC with a patience of 20 epochs. AUPRC is selected as the stopping criterion due to extreme class imbalance in fraud detection, where AUROC may provide overly optimistic estimates of minority-class performance [29].

a) Hyperparameter Sensitivity.: We observe that CT-GCL remains stable across moderate variations in embedding dimension and contrastive weight. In particular, embedding sizes within the range $[64, 256]$ produce consistent performance trends without severe degradation. Similarly, contrastive coefficients within $[0.05, 0.2]$ maintain effective auxiliary regularization. Extremely small invariance weights reduce robustness under temporal shift, whereas excessively large values may slow convergence. These observations indicate that the model does not rely on fragile hyperparameter tuning for competitive performance.

C. Fair Tuning and Baseline Alignment

To ensure fair comparison, identical tuning budgets are enforced across all baselines. Hyperparameter search is conducted exclusively on the validation period, and no test-period information is used for model selection. This protocol aligns with realistic fraud evaluation practices that emphasize strict temporal separation to avoid optimistic bias [27].

Search spaces follow recommendations from respective original publications where available, including learning rates, embedding dimensions, and dropout rates. All models share:

- Identical temporal splits,
- Identical preprocessing pipelines,
- Identical early stopping criteria,
- Identical evaluation metrics (AUPRC, AUROC).

The use of consistent evaluation metrics is particularly important under severe class imbalance, where inappropriate

metric selection may distort comparative conclusions [29]. Furthermore, enforcing identical chronological splits ensures robustness under concept drift and prevents leakage of future information [2].

This design ensures that performance differences reflect architectural characteristics rather than disparities in tuning effort, thereby supporting reproducible and deployment-aligned evaluation [12].

D. Computational Complexity and Scalability

Let m denote the number of sampled edges in the k-hop subgraph G_v, and let d denote the embedding dimension.

a) Message Passing Cost.: For each GNN layer, message passing incurs computational complexity

$$O(md), \tag{31}$$

which is consistent with standard graph convolution and attention-based architectures [6], [7].

CT-GCL employs two encoders during training for causal–environment disentanglement. Therefore, the forward- pass cost scales as a constant multiple of $O(md)$ while remaining linear in the number of sampled edges.

b) Contrastive Overhead.: Dual-view augmentation requires additional forward passes for structurally and temporally perturbed subgraphs. Consequently, the per-batch training cost increases by a small constant factor, similar to prior graph contrastive learning frameworks [14]. Importantly, overall complexity remains linear in m.

c) Neighborhood Sampling and Scalability.: Subgraph construction follows mini-batch neighborhood sampling strategies commonly used in scalable inductive graph learning [28]. This ensures that both computation and memory scale with the sampled subgraph size rather than the full graph.

d) Inference Cost.: At deployment time, only the causal encoder GNN_C and classification head are retained. Therefore, inference complexity remains $O(md)$ per layer, with no additional contrastive or counterfactual computations.

e) Memory Complexity.: Memory usage scales linearly with the number of sampled nodes and edges per batch. By restricting propagation to localized temporal subgraphs, CT-GCL avoids full-graph computation and remains practical for large-scale financial graphs.

f) Practical Deployment Considerations.: Fraud detection systems operate under strict latency constraints. Since contrastive learning and counterfactual intervention are used only during training, real-time deployment requires a single forward pass through a temporal GNN and a lightweight classifier. This maintains efficiency while preserving robustness under temporal distribution shift.

VI. ROBUSTNESS AND DRIFT EVALUATION

A. Temporal Degradation Curves

In addition to aggregate test metrics, we evaluate temporal stability by partitioning the test period into equal-length chronological windows (e.g., weekly segments) and reporting AUPRC within each window. This evaluation protocol ismotivated by prior work demonstrating that distribution drift may be concealed by a single aggregate performance score [27].

Fraud detection operates in non-stationary environments where both legitimate behavior and adversarial strategies evolve over time. Concept drift has been extensively studied in streaming and adaptive learning literature [2]. In financial fraud settings, even small shifts in behavioral baselines or transaction distributions can significantly alter model calibration [1].

Temporal slicing therefore provides a more realistic assessment of deployment stability than a single aggregate AUPRC computed over the entire test horizon. A global metric may obscure gradual degradation caused by shifting contextual variables or emerging fraud patterns. By evaluating performance across sequential windows, we approximate a deployment scenario in which a model is trained once and applied over an extended future period without retraining.

Formally, let $P_t(G, Y)$ denote the joint distribution in window t. Temporal robustness can be assessed by examining the variation of performance metric M_t (e.g., AUPRC) across windows:

$$M_t = \mathrm{AUPRC}(f_\vartheta; P_t(G, Y)). \tag{32}$$

Significant downward trends or high variance in M_t indicate sensitivity to distributional shift. In contrast, limited performance fluctuation suggests that the learned representation captures relatively invariant fraud mechanisms.

Since CT-GCL explicitly enforces counterfactual stability through the invariance loss, improved temporal stability is expected when compared to purely correlation-driven models. This evaluation directly tests whether causal disentanglement and contrastive regularization contribute to robustness under temporal drift.

B. Backtesting Under Feedback Latency

To emulate verification latency, we optionally delay the availability of fraud labels by Δ days in an online replay setting. This protocol follows realistic modeling strategies proposed for credit card fraud detection, where confirmation of fraudulent activity may occur significantly after transaction time [27]. Similar feedback-aware evaluation mechanisms have been explored in streaming fraud systems [30].

Formally, let a transaction occur at time t. Its corresponding label y_t becomes available only at time $t + \Delta$. During replay simulation, the model is updated using only labels satisfying:

$$t_i + \Delta \le t, \tag{33}$$

ensuring that no future confirmation information is incorporated into the learning process.

This setting reflects operational investigation workflows in which only a subset of alerts is confirmed promptly, while others remain unresolved for extended periods. As a result, the training signal may be partially outdated or incomplete, increasing reliance on structural regularities rather than recent

confirmation cues.

Performance stability under delayed-feedback replay indicates that the learned representation captures durable fraud mechanisms rather than short-lived correlations tied to immediate label availability. In particular, models that overfit to transient environmental signals may degrade significantly when supervision is temporally misaligned.

This evaluation complements temporal degradation analysis by introducing explicit label uncertainty into the deployment scenario. While temporal slicing measures robustness under distribution shift, delayed replay evaluates resilience under supervision latency. Together, these analyses provide a comprehensive assessment of robustness in realistic financial fraud environments.

VII. EXPERIMENTAL SETUP

A. Datasets

We evaluate CT-GCL on three benchmark datasets representing complementary fraud detection regimes.

TABLE IV
DATASET STATISTICS

Dataset	Transactions	Fraud %	Characteristics
European CC	284,807	0.17%	Static / PCA features
IEEE-CIS	590,540	3.5%	Heterogeneous / Rich metadata
BankSim	594,643	1.2%	Synthetic / Agent-based simulation

a) Preprocessing.: For the European dataset, which lacks explicit user identifiers, we construct a similarity-based transaction graph by connecting transactions that are close in time, transaction amount, or PCA feature profiles. This induces relational structure despite anonymization.

For IEEE-CIS and BankSim, we use explicit entity relations (e.g., User–Transaction, Merchant–Transaction) to construct heterogeneous graphs.

b) Dataset Complementarity.: The selected datasets represent distinct evaluation regimes. IEEE-CIS provides a heterogeneous real-world environment with rich metadata and observable temporal variation, enabling evaluation of relational modeling and drift robustness. BankSim offers a controlled synthetic environment with simulated fraud strategies, allowing assessment of structural pattern recovery. The European dataset, characterized by extreme imbalance and anonymized PCA features, tests resilience when explicit relational identifiers are unavailable.

Evaluating across these diverse regimes reduces the risk of dataset-specific conclusions and supports broader generalization claims.

B. Baselines

We benchmark against representative baselines that are (i) widely used in practice and (ii) documented in peer-reviewed venues:

- **XGBoost**: A strong tabular baseline commonly reported in credit card fraud detection studies [18].
- **Optimized LightGBM**: A boosting-based fraud detection approach with imbalance-aware tuning [16].
- **Time-Aware Attention Gated Network**: A sequential behavioral model that extracts transaction patterns using temporal attention [17].
- **SemiGNN**: A graph-attentive semi-supervised baseline for financial fraud detection [8].
- **TH-GCL**: A temporal heterogeneous graph contrastive learning model designed for label-scarce fraud detection [15].

Baselines are selected to span architectural paradigms rather than incremental variants of a single model family. This includes tabular boosting methods, sequential neural encoders, attention-based GNNs, and contrastive graph learning frameworks. Such diversity ensures that improvements observed for CT-GCL reflect cross-paradigm advantages rather than narrow architectural comparisons.

Where possible, implementations follow official or widely adopted configurations. Hyperparameter search is conducted under equal tuning budgets, and identical temporal splits and preprocessing pipelines are enforced to ensure fair comparison.

C. Evaluation Metrics

Given the extreme class imbalance inherent in fraud detection, we do not report Accuracy. Instead, we evaluate using:

- **AUPRC**: The primary metric for imbalanced classification.
- **Recall@K**: The fraction of true fraud cases detected among the top K transactions ranked by predicted risk.
- **F1-Score**: The harmonic mean of precision and recall.

AUPRC is prioritized because it reflects performance under severe class imbalance, where true negatives dominate the dataset [29]. Unlike ROC-AUC, which may remain artificially high under skewed distributions, AUPRC directly captures precision–recall trade-offs in the minority-class regime.

Recall@K provides operational interpretability by simulating investigation capacity constraints. It answers the practical question: *Given a fixed review budget, how much fraud can be intercepted?* This aligns evaluation with real-world fraud screening workflows.

F1-Score complements AUPRC by summarizing the precision–recall balance at a chosen decision threshold, providing threshold-dependent interpretability.

VIII. RESULTS AND DISCUSSION

A. Comparative Performance

The quantitative results on the test sets are summarized below. All reported metrics are computed on the temporally held-out test period. Unless otherwise stated, each number corresponds to the mean over 5 random seeds, and we recommend reporting standard deviations in a supplementary table for full

reproducibility.

Discussion: On the IEEE-CIS dataset, graph-based methods (SemiGNN, TH-GCL, CT-GCL) significantly outperform strong tabular baselines, indicating that multi-entity relations provide complementary signals beyond i.i.d. features [8], [15].

TABLE V
PERFORMANCE COMPARISON (AUPRC / F1-SCORE)

Model	European	IEEE-CIS	BankSim
XGBoost	0.852 / 0.860	0.784 / 0.792	0.941 / 0.950
Opt. LightGBM	0.861 / 0.868	0.801 / 0.808	0.947 / 0.955
TA-Gated (Seq)	0.874 / 0.878	0.820 / 0.825	0.958 / 0.962
SemiGNN	0.885 / 0.887	0.835 / 0.840	0.965 / 0.968
TH-GCL	0.898 / 0.895	0.855 / 0.860	0.973 / 0.976
CT-GCL (Ours)	**0.910 / 0.905**	**0.872 / 0.880**	**0.982 / 0.985**

CT-GCL further improves AUPRC from 0.855 (TH-GCL) to 0.872 on IEEE-CIS. As confirmed by the ablation study, removing the causal module results in a 3.7-point performance drop, validating the contribution of causal invariance under temporal distribution shift [27].

Although the absolute gains appear numerically modest, such improvements are operationally meaningful under extreme imbalance, where small AUPRC increases correspond to substantial additional fraud capture at scale.

The consistent improvement across heterogeneous (IEEE-CIS), synthetic (BankSim), and similarity-constructed (European) graphs suggests that the advantage of CT-GCL is not tied to a specific graph construction strategy. The smaller margin on BankSim aligns with its controlled generative structure, where contextual confounding is reduced. In contrast, real-world datasets such as IEEE-CIS contain richer environmental variability, amplifying the benefit of causal disentanglement.

B. Ablation Study

To isolate the contribution of each proposed module, we conduct an ablation study on the IEEE-CIS dataset under the same temporal split protocol. The full CT-GCL model achieves an AUPRC of 0.872. We evaluate the impact of removing individual components while keeping all other settings unchanged.

TABLE VI
ABLATION RESULTS ON IEEE-CIS (AUPRC)

Model Variant	AUPRC
Full CT-GCL	0.872
w/o Causal Module	0.835
w/o Contrastive Loss	0.841
w/o Temporal Attention	0.820

a) Effect of Causal Disentanglement.: Removing the causal intervention module reduces AUPRC by 3.7 percentage points. This degradation indicates that separating invariant behavioral mechanisms from environment-dependent context contributes substantially to temporal generalization. The result supports the hypothesis that causal regularization mitigates overfitting to unstable contextual correlations.

b) Effect of Contrastive Self-Supervision.: Eliminating the contrastive objective decreases AUPRC by 3.1 points. This suggests that self-supervised alignment improves representation robustness under limited labeled data, consistent with prior graph contrastive learning findings [14].*Effect of Temporal Attention.:* Removing temporal attention leads to the largest degradation (5.2 points), high- lighting the importance of modeling event timing in non- stationary fraud environments. This result aligns with prior evidence that temporal encoding improves sequential fraud detection [17].

c) Complementarity of Components.: Notably, removal of any single component results in non-trivial performance decline. This indicates that improvements arise from the interaction of temporal modeling, contrastive regularization, and causal invariance rather than from a single architectural enhancement. Temporal attention provides dynamic adaptation, contrastive learning enhances representation quality, and causal disentanglement improves stability under distribution shift.

C. Robustness to Concept Drift

To evaluate robustness under temporal distribution shift, we partition the IEEE-CIS test period into four consecutive weekly windows and compute AUPRC within each window. This windowed evaluation follows established concept drift analysis practices in fraud detection [2], [27].

Static tabular models and correlation-driven graph models exhibit substantial degradation across time. By Week 4, XG-Boost and attention-based GNN baselines show performance drops exceeding 15% relative to Week 1 performance. In contrast, CT-GCL exhibits a significantly smaller degradation of 4.5%, indicating improved temporal stability.

The observed degradation gap suggests that invariance-aware representation learning enhances robustness under evolving fraud distributions. Models that implicitly encode environment-specific correlations are more vulnerable to shifts in contextual variables. When these correlations weaken or change over time, predictive reliability deteriorates.

By contrast, the causal disentanglement module reduces dependence on unstable environmental context, while temporal attention emphasizes recent behavioral evidence. Together, these mechanisms produce a smoother degradation trajectory across sequential windows, reflecting improved out-of-distribution generalization under adversarial drift.

IX. THREATS TO VALIDITY AND LIMITATIONS

A. Graph Construction Bias

For datasets without explicit entity relations (e.g., similarity-

graph construction), there is a risk that graph-building heuris- tics may amplify spurious similarity and inadvertently leak information across splits if not carefully time-gated.

Additionally, similarity-based graph construction may intro- duce inductive bias by encoding proximity assumptions that do not necessarily reflect true relational dependencies. While temporal gating reduces leakage risk, inferred edges may still reflect heuristic similarity rather than verified interaction, potentially influencing representation learning dynamics.

B. Temporal Leakage and Feedback Latency

Fraud labels may arrive with delay, and naive preprocessing can inadvertently use information from the future. We therefore recommend strictly time-ordered splits and training-only fitting of any scalers/encoders.

Moreover, while strict chronological splitting mitigates explicit look-ahead bias, subtle forms of temporal dependency may persist if feature engineering encodes aggregate statistics over extended windows. Ensuring strict causality in feature computation remains critical for faithful deployment simulation.

C. Baseline Fairness

Reported improvements can be inflated if baselines are under-tuned. We mitigate this by using an equal tuning budget and selecting models by validation AUPRC.

Although equal tuning budgets are enforced, absolute fairness across heterogeneous model families cannot be guaranteed. Differences in architectural capacity, inductive bias, and optimization dynamics may influence comparative outcomes. Future studies may explore larger-scale hyperparameter sweeps or standardized evaluation frameworks to further reduce such confounding effects.

D. Causal Modeling Assumptions

The proposed causal disentanglement framework assumes that fraud intent (causal factors) and environmental context can be meaningfully separated at the representation level. In practice, complete disentanglement may be imperfect, and some environmental signals may partially overlap with genuine behavioral risk indicators.

Furthermore, the framework approximates causal intervention through representation-level perturbation rather than explicit structural equation modeling. While this approach improves invariance empirically, it does not constitute formal causal identification in the strict interventional sense. Future work may investigate more explicit causal discovery techniques or domain-informed structural priors.

X. CONCLUSION AND FUTURE WORK

This work addressed a central challenge in modern credit card fraud detection: how to design models that remain accurate under extreme class imbalance, temporal distribution shift, and limited supervision. Traditional tabular and correlation-driven graph approaches often rely implicitly on environment-specific artifacts that degrade under adversarial adaptation. In contrast, we proposed Causal-Temporal Graph Contrastive Learning (CT-GCL), a unified framework that integrates causal invariance, temporal attention, and self-supervised representation learning within a dynamic heterogeneous graph setting.

The proposed architecture combines three complementary principles. First, temporal heterogeneous subgraph construc- tion ensures causally consistent modeling of evolving trans- action streams. Second, causal disentanglement with counter- factual consistency encourages the model to rely on invariant behavioral mechanisms rather than transient environmentalcontext. Third, dual-view contrastive learning enhances repre- sentation robustness under label scarcity by aligning structural and temporal perspectives of each transaction.

Empirical evaluation across IEEE-CIS, BankSim, and European Credit Card datasets demonstrates consistent improvements in AUPRC and F1-score relative to strong tabular, sequential, and graph-based baselines. More importantly, CT-GCL exhibits significantly reduced performance degradation under temporal slicing, suggesting improved out-of-distribution generalization in non-stationary environments. Ablation studies confirm that each module—temporal attention, causal intervention, and contrastive regularization—contributes meaningfully to overall stability and predictive quality.

Beyond quantitative gains, this study emphasizes the importance of invariance-aware modeling in adversarial financial domains. Fraud detection systems operate in settings where environmental correlations shift rapidly due to seasonal behavior, infrastructure changes, or adaptive attacker strategies. By explicitly separating causal and environmental representations and enforcing counterfactual stability, CT-GCL moves toward models that capture why fraud occurs rather than merely what correlated patterns appear in historical data.

From a deployment perspective, the framework remains computationally feasible for real-time screening pipelines, as inference relies solely on the causal encoder and temporal aggregation mechanisms. The architecture is compatible with mini-batch training and can be extended to streaming or incremental learning scenarios with periodic retraining.

Several avenues for future research emerge from this work. First, more explicit causal identification techniques could be explored to complement representation-level intervention, potentially incorporating domain knowledge or structural priors. Second, continual learning mechanisms may be integrated to further mitigate long-term distribution drift without full retraining. Third, improved interpretability techniques are needed to map latent causal representations to analyst-understandable concepts, enhancing trust and regulatory transparency. Finally, federated or privacy-preserving extensions of CT-GCL may enable collaborative fraud modeling across institutions without centralized data sharing.

In summary, CT-GCL represents a step toward a new

generation of fraud detection systems that integrate temporal reasoning, causal robustness, and self-supervised learning. By unifying these principles within a coherent graph-based architecture, the proposed framework advances both method- ological rigor and practical resilience in dynamic financial environments.

XI. REFERENCES

[1] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Adaptive machine learning for credit card fraud detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[2] J. Gama, I. Zˇ liobaitеˇ, A. Bifet *et al.*, "A survey on concept drift adaptation," *ACM Computing Surveys*, 2014.

[3] P. Sarna *et al.*, "Ai-driven credit card fraud detection: A comprehensive survey," *IEEE Access*, 2025.

[4] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[5] Q. Wu, X. Huang, H. Zhang *et al.*, "Handling distribution shifts on graphs: An invariant learning approach," in *International Conference on Learning Representations (ICLR)*, 2022.

[6] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[7] P. Velicˇkovic´, G. Cucurull, A. Casanova, A. Romero, P. Lio`, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[8] D. Wang, J. Lin, P. Cui, Q. Jia, Z. Wang, Y. Fang, Q. Yu, J. Zhou, S. Yang, and Y. Qi, "A semi-supervised graph attentive network for financial fraud detection," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2019, pp. 598–607.

[9] E. Rossi, B. P. Chamberlain, F. Frasca *et al.*, "Temporal graph networks for deep learning on dynamic graphs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, and K. Achan, "Inductive representation learning on temporal graphs," in *International Conference on Learning Representations (ICLR)*, 2020.

[11] S. Gui, X. Wang *et al.*, "Good: A graph out-of-distribution benchmark," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[12] N. J. Sarna, F. A. Rithen, U. S. Jui, S. Belal, A. Amin, T. K. Oishee, and A. K. M. M. Islam, "Ai driven fraud detection models in financial networks: A comprehensive systematic review," *IEEE Access*, vol. 13, pp. 141 204–141 233, 2025.

[13] P. Velicˇkovic´, W. Fedus, W. L. Hamilton *et al.*, "Deep graph infomax," in *International Conference on Learning Representations (ICLR)*, 2019.

[14] Y. You, T. Chen, Y. Sui *et al.*, "Graph contrastive learning with augmentations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] J. Wang, J. Liu, W. Zheng, and Y. Ge, "Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions," *IEEE Access*, vol. 13, pp. 145 754–145 771, 2025.

[16] A. A. Taha and S. J. Malebary, "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine," *IEEE Access*, vol. 8, pp. 25 579–25 587, 2020.

[17] Y. Xie, G. Liu, C. Yan, C. Jiang, and M. Zhou, "Time-aware attention-based gated network for credit card fraud detection by extracting transactional behaviors," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 3, pp. 1004–1016, 2023.

[18] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms," *IEEE Access*, vol. 10, pp. 39 700–39 715, 2022.

[19] P. Gao, Z. Li, D. Zhou, and L. Zhang, "Reinforced cost-sensitive graph network for detecting fraud leaders in telecom fraud," *IEEE Access*, vol. 12, pp. 173 638–173 646, 2024.

[20] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[21] B. Schoˇlkopf, F. Locatello, S. Bauer *et al.*, "Toward causal representation learning," in *Proceedings of the IEEE*, 2021.

[22] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[23] N. Ye, K. Li, S. Bai *et al.*, "Out-of-distribution generalization via risk extrapolation (rex)," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[24] J. Peters, P. Buˇhlmann, and N. Meinshausen, "Causal inference by using invariant prediction," in *Journal of the Royal Statistical Society: Series B*, 2016.

[25] F.-Y. Sun, J. Hoffmann, V. Verma, and J. Tang, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," in *International Conference on Learning Representations (ICLR)*, 2020.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolla´r, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[27] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learningstrategy," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3784–3797, 2018.

[28] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[29] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93 010–93 022, 2019.

[30] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3637–3647, 2018.