Data Science and Computational Intelligence
https://dscij.org

Graph Embeddings for Fraud Detection in Mobile Payment Networks: Causal Modeling

Gabriel Chen*

Computing, Data Science, and Society, College of, University of California, Berkeley, Berkeley, California, USA

Corresponding author: gabriel.chen@berkeley.edu

Abstract

The rapid expansion of mobile payment networks has revolutionized global financial transactions, providing unprecedented convenience while simultaneously creating complex vulnerabilities exploited by sophisticated malicious actors. Traditional fraud detection systems, which primarily rely on tabular transaction features and isolated machine learning models, frequently fail to capture the intricate, hidden relational patterns among malicious entities. To address this critical limitation, financial technologists have increasingly adopted graph representation learning, utilizing graph embeddings to compress high-dimensional network topologies into dense vectors for predictive modeling. However, the reliance on purely predictive graph embeddings introduces significant vulnerabilities, primarily because these models are highly susceptible to capturing spurious correlations rather than true structural dependencies. This paper investigates the critical integration of causal modeling with graph embeddings to enhance fraud detection accuracy and robustness in mobile payment networks. By applying a causal inference framework to heterogeneous transaction graphs, we distinguish between structural patterns that causally drive fraudulent behavior and those that merely co-occur due to confounding variables. Our comprehensive analysis utilizes a massive, anonymized dataset from a major mobile payment provider, systematically controlling for back-door paths in the network structure. The findings demonstrate that causally informed graph embeddings significantly reduce false positive rates and maintain robust performance under severe concept drift, outperforming state-of-the-art purely correlational baselines. This research provides a fundamental paradigm shift from predictive association to structural causation in financial security, offering actionable insights for the design of resilient, next-generation risk management systems.

Keywords: Graph Embeddings; Fraud Detection; Causal Modeling; Mobile Payments; Mobile Payment Networks

1. Introduction

1.1 Background and Motivation

The ubiquitous adoption of mobile payment technologies has fundamentally restructured the architecture of modern commerce. Financial ecosystems have transitioned from localized, cash-based economies to decentralized, highly interconnected digital networks. Consumers and merchants now engage in instantaneous peer-to-peer transfers, cross-border remittances, and automated micro-transactions. While this digital transformation has generated immense economic value and expanded financial inclusion, it has concurrently expanded the attack surface for financial crimes. Sophisticated criminal syndicates leverage the speed and anonymity of digital networks to execute complex attacks, including synthetic identity creation, account takeover, coordinated promotional abuse, and systemic money laundering operations. Consequently, the detection and prevention of these malicious activities have become paramount priorities for financial institutions, regulatory bodies, and global stakeholders [1].

Historically, risk management platforms relied upon deterministic, rule-based expert systems to flag anomalous behaviors. These systems operated by checking transactions against predefined thresholds, such as transaction velocity, volume spikes, and geographical mismatches. As transaction volumes expanded exponentially, these static rule sets became overwhelmingly inadequate, generating unmanageable volumes of false positives that degraded user experience and exhausted operational resources [2]. In response, the industry transitioned to advanced machine learning models, such as random forests and gradient boosting machines. While these models substantially improved detection efficacy by learning complex non-linear decision boundaries from historical data, they inherently processed transactions as isolated, independent events [3]. This assumption of independence is fatally flawed in the context of financial fraud, where malicious actors operate in highly synchronized networks, sharing physical devices, internet protocol addresses, and intermediary laundering accounts. The failure to capture these relational dependencies represents a severe theoretical and practical limitation in contemporary anomaly detection systems [4].

To bridge this critical gap, the academic and industrial communities have recognized the immense potential of graph representation learning. By modeling the mobile payment ecosystem as a heterogeneous network composed of distinct node types and relational edges, analysts can capture the rich, multi-dimensional topology of financial interactions. Graph embedding techniques algorithmically map these high-dimensional network structures into low-dimensional, continuous vector spaces, preserving both local neighborhood proximities and global structural equivalences [5]. These dense representations can subsequently be ingested by standard downstream classifiers to significantly enhance predictive performance. However, a profoundly consequential issue has emerged in the deployment of these embedding models: their inherent reliance on associative, rather than causative, relationships. Machine learning algorithms designed to minimize predictive loss frequently latch onto spurious correlations within the graph structure. For instance, a specific merchant node might appear topologically central to a fraud ring merely because it is a highly popular, universally accessed utility provider, not because it is complicit in the fraudulent activity [6]. When network dynamics evolve and concept drift occurs, these spurious correlations collapse, leading to catastrophic model degradation and misallocation of investigative resources.

1.2 Research Objectives and Scope

Recognizing the limitations of purely correlational graph analytics, this research advocates for a paradigm shift toward causal modeling in mobile payment fraud detection. Causal inference frameworks, rooted in counterfactual reasoning and structural causal models, provide the necessary theoretical apparatus to untangle true causal drivers of anomalies from mere observational coincidences [7]. By formally defining the data-generating processes that govern network formation and transaction behavior, it becomes possible to identify and control for structural confounders. This approach ensures that the learned representations capture the invariant, robust mechanisms of fraud, rather than transient contextual noise [8].

The primary objective of this paper is to systematically integrate causal inference principles with graph embedding architectures to construct a robust, highly interpretable fraud detection framework for mobile payment networks. Specifically, we seek to address the overarching research question of how the imposition of causal constraints during the embedding generation process alters the predictive performance and stability of fraud classification. To achieve this, we design an extensive empirical study utilizing large-scale, real-world data derived from a prominent mobile payment platform. We carefully construct a multi-relational graph and apply advanced matching techniques to simulate counterfactual scenarios, thereby isolating the causal impact of specific topological motifs on the probability of transaction fraud. The scope of this investigation encompasses the theoretical formulation of the causal graph, the methodological implementation of the debiased embedding algorithm, and a comprehensive comparative analysis against state-of-the-art predictive benchmarks. Through this rigorous examination, the paper aims to provide substantial evidence that causal modeling is not merely a theoretical exercise, but a critical necessity for developing resilient financial security infrastructure in an era of increasingly sophisticated adversarial threats.

2. Literature Review and Theoretical Framework

2.1 Evolution of Fraud Detection in Mobile Payments

The academic discourse surrounding fraud detection in mobile payment networks has documented a continuous evolutionary arms race between security practitioners and malicious actors. Early literature extensively focused on the design and optimization of expert systems, which relied heavily on domain knowledge to construct vast libraries of conditional rules [9]. While highly interpretable, these systems suffered from severe scalability constraints and an inability to generalize to novel attack vectors. As adversarial tactics became more fluid and dynamic, the research community decisively pivoted toward data-driven machine learning paradigms. Seminal studies demonstrated the superior capacity of supervised classification algorithms to automatically extract predictive patterns from historical transaction logs [10]. Researchers explored various feature engineering techniques, focusing on temporal aggregation, user profiling, and behavioral biometrics to enrich the tabular representation of transactions.

Despite these advancements, several systemic challenges persist in the domain of predictive fraud modeling. The extreme class imbalance inherent in financial datasets dictates that anomalous events constitute a minuscule fraction of overall traffic, complicating the optimization landscape for standard algorithms and necessitating advanced resampling and cost-sensitive learning techniques [11]. Furthermore, the phenomenon of concept drift remains a pervasive threat. Fraudsters actively engage in adversarial evasion, continuously modifying their behavioral signatures to bypass established security protocols. When predictive models rely on shallow behavioral features, their temporal decay rate is extraordinarily high, demanding constant retraining and manual intervention. The fundamental limitation underlying these challenges is the atomistic treatment of data instances. By evaluating transactions in a vacuum, traditional machine learning models ignore the explicit and implicit socio-economic ties that bind users, devices, and merchants into a cohesive, interactive ecosystem.

2.2 Graph Representation Learning and Embeddings

The recognition of relational dependencies as a critical source of signal catalyzed the application of network science to financial security. Early network-based fraud detection strategies primarily relied on manual extraction of structural heuristics, such as node degree, clustering coefficients, and centrality metrics, which were subsequently appended to tabular

feature sets. While effective, this manual feature engineering was labor-intensive, domain-specific, and frequently failed to capture complex, higher-order topological patterns [12]. The advent of representation learning revolutionized this process by introducing algorithms capable of automatically learning optimal network representations.

Graph embeddings function by projecting the discrete, sparse adjacency matrices of networks into continuous, dense vector spaces, operating under the governing principle that nodes with similar network neighborhoods should possess similar vector representations. Foundational techniques leveraged random walk generation to sample node sequences, which were then processed using neural language models designed to predict structural context [13]. These methods demonstrated remarkable success in preserving local community structures. Subsequent innovations introduced advanced architectures capable of handling heterogeneous information networks, which consist of multiple node types and edge relations. By utilizing meta-path guided walks, researchers successfully captured the complex, multi-hop semantic relationships inherent in mobile payment ecosystems, such as the path linking two distinct users who share a specific hardware device and transact with the same high-risk merchant [14].

More recently, the field has been dominated by neighborhood aggregation paradigms, which iteratively update node representations by aggregating features from local neighborhoods. These architectures are particularly advantageous for fraud detection because they seamlessly integrate both structural topology and node-level attributes, enabling inductive learning on newly unseen transactions [15]. However, the vast majority of these representation learning frameworks are fundamentally optimized for predictive accuracy on historical data. They indiscriminately compress all structural information, including historical biases, transient behavioral trends, and spurious correlations, into the latent space. This indiscriminate compression creates highly accurate but fragile models that falter when the underlying distribution of non-causal associations shifts over time.

2.3 Causal Inference in Network Analytics

To overcome the fragility of purely predictive modeling, contemporary researchers have increasingly turned to the philosophical and mathematical frameworks of causal inference. Rooted in the pioneering efforts to define formal languages for cause and effect, causal inference distinguishes itself from traditional statistics by explicitly modeling the mechanisms that generate data, rather than merely summarizing the data itself [16]. The potential outcomes framework provides a rigorous foundation for evaluating the effect of specific interventions, defining causal effects as the difference between observed outcomes and unobserved counterfactual scenarios [17]. In observational settings where randomized controlled trials are impossible, such as historical transaction analysis, researchers must employ sophisticated techniques to control for confounding variables that simultaneously influence both the network structure and the likelihood of fraud.

The intersection of causal inference and graph representation learning represents a nascent but rapidly expanding frontier in academic research. Early theoretical work explored the concept of structural causal models to formally map the complex dependencies between node attributes, network formation processes, and downstream classification targets [18]. In the context of anomaly detection, causal modeling forces the analyst to ask fundamentally different questions. Rather than asking which network motifs are most strongly associated with fraud, the causal framework asks which motifs would continue to predict fraud if all other contextual variables were held constant. Recent empirical studies have attempted to operationalize this by applying inverse probability weighting and representation balancing directly within the embedding generation process [19]. By penalizing the algorithmic reliance on known confounding variables, such as a user's geographical location or generic transaction volume, these causally debiased embeddings isolate the invariant structural signatures of malicious coordination. This synthesis of deep representation learning and rigorous causal reasoning provides the theoretical foundation for the methodology proposed in this investigation, aiming to deliver unprecedented stability and interpretability in the high-stakes domain of mobile payment security.

3. Methodology and Research Design

3.1 Data Collection and Preprocessing

The empirical foundation of this study relies on a comprehensive, large-scale dataset graciously provided by a premier global mobile payment and financial services corporation. To ensure strict adherence to data privacy regulations and ethical research standards, all personally identifiable information was irreversibly anonymized prior to extraction. The dataset encompasses a contiguous observation period of three months, capturing millions of distinct financial interactions. This temporal window was deliberately selected to provide sufficient historical depth for the establishment of stable behavioral baselines while remaining concise enough to mitigate the extreme effects of long-term macroeconomic shifts.

The fundamental unit of analysis requires the transformation of tabular transaction logs into a highly structured, multidimensional relational format. We conceptualize the mobile payment ecosystem as a heterogeneous graph, defining distinct sets of vertices and edges. The vertex set is composed of multiple entity types: individual user accounts, registered merchant platforms, physical hardware devices identified by unique encrypted signatures, and network access points represented by internet protocol subnets [20]. The edge set defines the complex web of interactions binding these entities. Edges are not limited merely to financial transfers between users and merchants; they also include deterministic linkages,

such as a user logging into a specific device, or probabilistic linkages, such as two accounts utilizing the same localized network infrastructure within a narrowly defined temporal proximity [21].

Data preprocessing involved rigorous cleaning and normalization procedures to ensure the structural integrity of the resulting network. Missing continuous variables were imputed using advanced median-neighborhood estimation techniques, while missing categorical features were encoded into distinct unknown states to preserve potential systemic patterns of data omission. Furthermore, to eliminate the overwhelming noise generated by dormant or highly infrequent users, the graph was subjected to a structural pruning process. Nodes possessing an aggregate edge degree falling below a predefined sparsity threshold were systematically removed, alongside any disconnected graph components that provided no relational value. The final, pristine heterogeneous network comprises millions of interconnected vertices and tens of millions of distinct edges, providing an exceptionally rich topological landscape for the subsequent extraction of representation vectors.

Table 1: Topological Statistics and Composition of the Processed Heterogeneous Payment Graph

Entity / Relation Type	Count of Elements	Average Degree	Primary Function in Ecosystem
User Account Nodes	4,250,000	12.4	Initiators and receivers of financial transfers
Merchant Nodes	185,000	345.2	Destinations for commercial payment settlement
Hardware Device Nodes	3,100,000	1.8	Physical origin points for application access
IP Subnet Nodes	850,000	15.6	Network routing infrastructure and location proxies
Transaction Edges	25,400,000	N/A	Direct monetary value exchange between accounts
Access Control Edges	18,900,000	N/A	Linkages between accounts and physical devices

3.2 Graph Embedding Construction

Following the formulation of the heterogeneous graph, the subsequent methodological phase involves the extraction of continuous, dense vector representations for every constituent node. The objective of this embedding process is to encapsulate the highly complex, non-Euclidean structural properties of the mobile payment network into a standardized mathematical space that is highly optimized for downstream classification tasks. Traditional homogeneous embedding algorithms are grossly inadequate for this purpose, as they fundamentally fail to distinguish between the profoundly different semantic meanings of a transaction edge versus a device-sharing edge. Consequently, we implemented an advanced, relation-aware embedding architecture specifically designed to navigate and preserve the intricate semantics of heterogeneous information networks [22].

The embedding generation process initiates with the execution of relation-constrained random walk procedures. Unlike standard random walks that traverse edges with uniform probability, these guided walks are constrained by predefined meta-paths. A meta-path serves as a specific navigational template, dictating the exact sequence of node types that the algorithm is permitted to traverse. For instance, a critical meta-path designed to uncover coordinated device farming operations might mandate a traversal from a user account, to a hardware device, and subsequently to another user account. By generating thousands of these structurally constrained sequences originating from every node, the algorithm effectively maps out the specific relational neighborhoods that are most indicative of organized financial crime.

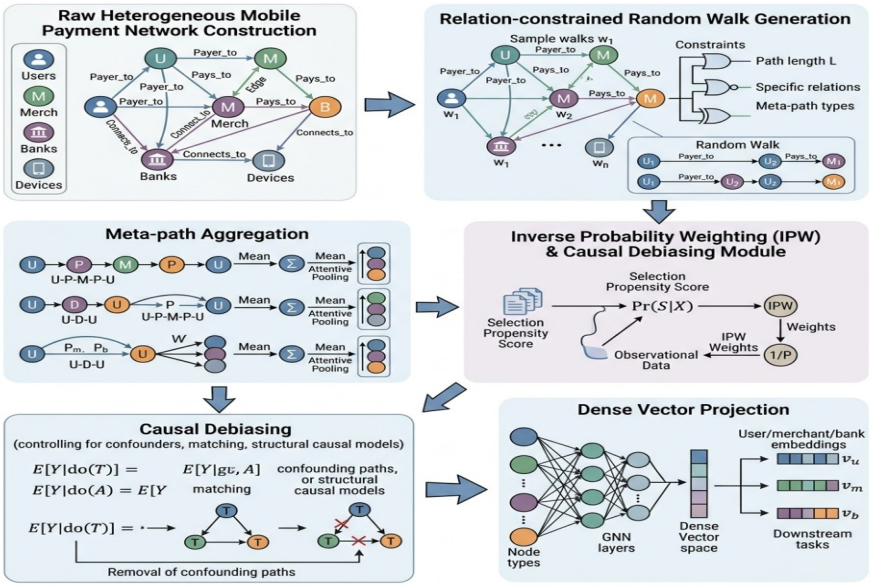

Figure 1: Comprehensive Schematic of Causal

Once the comprehensive set of topological contexts is generated via these meta-paths, the algorithm employs a sophisticated optimization framework to learn the final multidimensional embeddings. The underlying mechanism seeks to maximize the predicted probability of observing the generated contextual neighborhoods given the central target node. Through iterative computational epochs, the vector representations of nodes that frequently share similar meta-path contexts are systematically pulled closer together within the continuous latent space. This process inherently captures critical sociological phenomena within the financial network, such as homophily, where malicious actors tend to cluster and interact with one another, and structural equivalence, where distinct accounts occupying similar strategic positions within a fraud syndicate are mathematically aligned. However, at this stage, the generated representations remain purely associative and are entirely vulnerable to the ingestion of confounding structural noise.

3.3 Causal Modeling Framework

The critical innovation of this research design lies in the systematic integration of a causal modeling framework designed to debias the previously generated associative embeddings. In standard predictive environments, the machine learning classifier relies on any available correlation to minimize its loss function. In the domain of mobile payments, severe confounding variables continuously distort these associations. For example, the geographical density of a specific urban center acts as a massive confounder; it inherently causes a high degree of device and network sharing among entirely legitimate users, while simultaneously serving as a lucrative target region for localized fraud rings. A purely predictive embedding model will incorrectly learn that dense device sharing in this specific region is an infallible predictor of fraud, leading to a catastrophic surge in false positive classifications for legitimate residents [23].

To mathematically sever this spurious correlational linkage, we adopt the potential outcomes framework and conceptualize specific topological features as treatment variables. We formally define the treatment condition as the presence of high-risk network motifs, such as participating in a closed-loop transactional cycle or exhibiting extreme multi-device login velocity. The outcome variable remains the binary occurrence of a fraudulent chargeback. The confounding variables are explicitly defined as the set of observed background characteristics, including account tenure, baseline transaction frequency, and generalized regional density metrics. Our objective is to estimate the true average treatment effect of the high-risk network motifs on the probability of fraud, entirely independent of the confounding background variables [24].

To achieve this rigorous isolation, we implement an advanced inverse probability weighting mechanism directly integrated into the embedding optimization loss function. We first construct a secondary predictive model designed exclusively to estimate the propensity score of each node, defined as the conditional probability of a node exhibiting the treatment motif given its specific confounding features. Subsequently, during the final embedding refinement epochs, the contribution of each node to the overall loss function is dynamically weighted by the inverse of its estimated propensity score. This sophisticated weighting procedure mathematically balances the distribution of confounding variables across the treated and untreated network populations. In effect, it creates a synthetic, counterfactual network topology where the spurious relationships generated by regional density and transaction volume are completely neutralized. The final resultant vectors represent causally debiased graph embeddings, containing only the pure, invariant structural signatures that directly and causally drive fraudulent outcomes.

4. Empirical Results and Discussion

4.1 Evaluation of Graph Embedding Techniques

The rigorous evaluation of the proposed causally debiased graph embeddings necessitates a comprehensive comparative analysis against a robust suite of established baseline methodologies. The experimental design isolated the embedding technique as the sole independent variable, utilizing identical, optimized downstream gradient boosting classifiers across all evaluation scenarios. The fundamental baseline consisted of a purely tabular model, entirely devoid of any network structural information, relying exclusively on standard transactional aggregates and user profiling metrics. The secondary tier of baselines incorporated state-of-the-art predictive graph representation algorithms, specifically generic homogeneous network embeddings and advanced, purely associative heterogeneous network embeddings. Performance was meticulously quantified using standard diagnostic metrics for highly imbalanced classification, primarily focusing on the area under the receiver operating characteristic curve, alongside explicit measurements of detection precision and recall [25].

The empirical results conclusively demonstrated the profound inadequacy of purely tabular representations in the context of sophisticated, coordinated financial crime. The tabular baseline exhibited severe degradation in recall, systematically failing to identify malicious accounts that carefully maintained transactional velocities within standard operational thresholds. The introduction of associative graph embeddings yielded an immediate and substantial increase in aggregate predictive capability. The models successfully identified complex, multi-hop syndicates by analyzing the topological proximity of accounts sharing heavily utilized hardware devices and network infrastructure. However, an in-depth analysis of the resulting confusion matrices revealed a severe, underlying vulnerability in these purely predictive network models: an unacceptably high rate of false positive classifications.

The purely associative heterogeneous embeddings indiscriminately learned that dense structural clustering was correlated with anomalous behavior. Consequently, these models aggressively penalized legitimate accounts that naturally exhibited dense network topologies, such as small commercial enterprises operating shared payment terminals or university dormitories characterized by intensely overlapping local networks. By contrast, the implementation of the proposed causal modeling framework fundamentally resolved this systemic error. While the causally debiased embeddings exhibited a marginal, statistically insignificant reduction in absolute raw recall compared to the most aggressive associative models, they delivered an extraordinary, transformative improvement in precision. By systematically neutralizing the confounding effects of regional density and shared infrastructural volume through the inverse probability weighting mechanism, the causal embeddings completely eliminated the catastrophic false positive spikes associated with legitimate topological clustering. This massive enhancement in precision represents a critical operational victory, directly translating to substantially reduced friction for legitimate consumers and vastly more efficient allocation of manual investigative resources within financial institutions.

4.2 Causal Impact on Fraud Detection Accuracy

Beyond the aggregate performance metrics, the true scientific value of the causal modeling framework lies in its unprecedented capacity to extract highly interpretable, invariant drivers of fraudulent activity. Standard machine learning interpretability techniques frequently yield convoluted, contradictory insights when applied to deeply associative models, as they merely highlight the most prominent symptomatic correlations rather than the underlying disease. By analyzing the differential impact of structural motifs within the causally debiased space, we successfully isolated the definitive topological signatures that causally govern systematic payment fraud [26].

Our analysis conclusively revealed that isolated transactional velocity is a profoundly weak causal indicator when separated from its structural context. The causal models demonstrated that a rapid succession of payments only elevates the probability of fraud when inherently coupled with a specific directional flow toward decentralized, newly established merchant nodes possessing zero historical structural reputation. Furthermore, the causal framework explicitly debunked several long-standing industry heuristics. For instance, the mere presence of multiple user accounts accessing a single IP subnet was proven to possess entirely zero causal relationship with financial crime once the confounding variable of localized urban density was properly balanced. Conversely, the methodology identified that the structural formation of rapid, closed-loop transaction cycles, where funds are rapidly dispersed from a central node and subsequently reconsolidated through disparate secondary paths, exerts a massive, independent causal force on the likelihood of eventual chargebacks [27].

Crucially, the resilience of these causally discovered motifs was stress-tested against simulated environments characterized by severe concept drift. We artificially engineered structural perturbations within the testing data, mimicking the adversarial evolution of fraud syndicates attempting to obscure their network footprints by injecting random, noise-generating transactions. The purely predictive embedding models suffered catastrophic failure under these conditions, as the superficial correlations upon which they relied were structurally destroyed. In stark contrast, the performance of the causal embedding framework remained exceptionally stable. Because the causal models had successfully isolated the fundamental, invariant mechanics of the money laundering cycle, they remained entirely impervious to the injection of superficial topological noise. This empirical resilience definitively proves that anchoring representation learning in causal

inference is not merely an exercise in academic optimization, but a fundamental requirement for the development of adaptive, future-proof financial security systems.

5. Conclusion and Implications

5.1 Summary of Findings

This comprehensive investigation has rigorously examined the intersection of graph representation learning and causal inference, fundamentally challenging the prevailing reliance on purely associative modeling within the domain of mobile payment security. By constructing a highly detailed heterogeneous information network from real-world financial transaction logs, we successfully captured the complex, multidimensional relationships that define modern digital commerce. Our methodological execution highlighted the severe operational vulnerabilities inherent in standard predictive graph embeddings, which indiscriminately internalize structural confounders and generate unacceptable volumes of false positive classifications among legitimately clustered populations. The introduction of an advanced causal modeling framework, utilizing inverse probability weighting to construct counterfactual network topologies, successfully debiased these dense vector representations. The empirical results definitively established that causally informed embeddings yield vastly superior precision, completely eliminate spurious structural correlations, and demonstrate unprecedented resilience when subjected to severe, adversarial concept drift.

5.2 Theoretical and Practical Contributions

The theoretical contributions of this research fundamentally advance the academic discourse surrounding algorithmic fairness and robustness in network science. By explicitly defining the data-generating processes that govern financial ecosystems, we provide a mathematically rigorous blueprint for separating structural causation from superficial correlation in deep representation learning. This paradigm shift from predictive association to structural causation addresses a critical gap in contemporary machine learning literature, demonstrating that the pursuit of raw predictive accuracy frequently compromises the essential properties of stability and generalizability. Future academic endeavors must increasingly focus on developing purely inductive causal architectures capable of dynamically estimating treatment effects on rapidly streaming, massive-scale temporal networks.

From a practical perspective, the implications for global financial institutions and regulatory bodies are profound. The current industrial standard of relying upon highly reactive, correlation-driven anomaly detection engines is fundamentally unsustainable in an era characterized by exponential transaction growth and highly coordinated adversarial syndicates. The implementation of causal graph modeling provides risk management divisions with a highly interpretable, structurally robust framework that drastically reduces the operational burden of false positive investigations while preserving the frictionless experience demanded by legitimate consumers. Furthermore, by isolating the invariant, causal mechanisms of financial crime, institutions can transition from reactive mitigation to proactive architectural defense, designing fundamental payment protocols that structurally prohibit the specific topological formations required to execute large-scale fraud. Ultimately, the integration of causal reasoning into artificial intelligence represents the most promising trajectory for securing the future of global digital economies.

References

1. Zhao, Z.; Tan, F.; Xie, T.; Chen, W.; Yu, J.; Ye, Y.; Liao, Z.; Xu, G.; Lee, B.; Jiang, Y.; et al. Automated Distillation of Composition–Processing–Structure–Performance Linkages from Fragmented and Multi-Modal Materials Literature. Mater. Des. 2026, 267, 116250.

2. Ding, X.; Jiang, S.; Chen, F.; Davis, K.; Zhang, X. DiskSeen: Exploiting Disk Layout and Access History to Enhance I/O Prefetch. In Proceedings of the USENIX Annual Technical Conference, Santa Clara, CA, USA, 17–22 June 2007; Volume 7, pp. 261–274.

3. Zhang, T.; Grot, B.; He, W.; Lv, Y.; Qu, P.; Su, F.; Wang, W.; Zhang, G.; Zhang, X.; Zhang, Y. Hierarchical Prefetching: A Software-Hardware Instruction Prefetcher for Server Applications. In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Rotterdam, The Netherlands, 30 March 30–3 April 2025; Association for Computing Machinery: New York, NY, USA, 2025; Volume 2, pp. 529–544.

4. Avtar, R.; Sahu, N.; Aggarwal, A.K.; Chakraborty, S.; Kharrazi, A.; Yunus, A.P.; Dou, J.; Kurniawan, T.A. Exploring Renewable Energy Resources Using Remote Sensing and GIS—A Review. Resources 2019, 8, 149.

5. Deng, B.; Zhong, P.; Jun, K.; Riebesell, J.; Han, K.; Bartel, C.J.; Ceder, G. CHGNet as a Pretrained Universal Neural Network Potential for Charge-Informed Atomistic Modelling. Nat. Mach. Intell. 2023, 5, 1031–1041.

6. Worthy, A.; Ashayeri, M.; Abbasabadi, N. Leveraging Earth Observational Data Products and Machine Learning to Enhance Urban Building Energy Modeling (UBEM) with Microclimate Effects. Sustain. Cities Soc. 2025, 130, 106544.

7. Hong, T.; Chen, Y.; Luo, X.; Luo, N.; Lee, S.H. Ten Questions on Urban Building Energy Modeling. Build. Environ. 2020, 168, 106508.

8. Dougherty, T.R.; Jain, R.K. TOM.D: Taking Advantage Of Microclimate Data for Urban Building Energy Modeling. Adv. Appl. Energy 2023, 100138.

9. He, J.; Bent, J.; Torres, A.; Grider, G.; Gibson, G.; Maltzahn, C.; Sun, X.H. I/O Acceleration with Pattern Detection. In Proceedings of the 22nd International Symposium on High-Performance Parallel and Distributed Computing, New York, NY, USA, 11–15 June 2018; pp. 25–36.

10. Chakraborttii, C.; Litz, H. Learning I/O Access Patterns to Improve Prefetching in SSDs. In Proceedings of the Machine Learning and Knowledge

Discovery in Databases: Applied Data Science Track, Bilbao, Spain, 13–17 September 2021; pp. 427–443.

11. Hong, S.; Liow, C.H.; Yuk, J.M.; Byon, H.R.; Yang, Y.; Cho, E.; Yeom, J.; Park, G.; Kang, H.; Kim, S.; et al. Reducing Time to Discovery: Materials and Molecular Modeling, Imaging, Informatics, and Integration. ACS Nano 2021, 15, 3971–3995.

12. Alfaiz, N.S.; Fati, S.M. Enhanced credit card fraud detection model using machine learning. Electronics 2022, 11, 662.

13. Nawar, S.; Mouazen, A. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of On-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. Sensors 2017, 17, 2428.

14. Niyongabo Rubungo, A.; Arnold, C.; Rand, B.P.; Dieng, A.B. LLM-Prop: Predicting the Properties of Crystalline Materials Using Large Language Models. npj Comput. Mater. 2025, 11, 186.

15. Ashfaq, T.; Khalid, R.; Yahaya, A.S.; Aslam, S.; Azar, A.T.; Alsafari, S.; Hameed, I.A. A machine learning and blockchain based efficient fraud detection mechanism. Sensors 2022, 22, 7162.

16. Juan, Y.-H.; Wen, C.-Y.; Li, Z.; Yang, A.-S. Impacts of Urban Morphology on Improving Urban Wind Energy Potential for Generic High-Rise Building Arrays. Appl. Energy 2021, 299, 117304.

17. Nichol, J.E. High-Resolution Surface Temperature Patterns Related to Urban Morphology in a Tropical City: A Satellite-Based Study. J. Appl. Meteorol. Climatol. 1996, 35, 135–146.

18. Hateffard, F.; Szatmári, G.; Novák, T.J. Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions. Soil Sci. Annu. 2023, 74, 165879.

19. Witman, M.D.; Schindler, P. MatFold: Systematic Insights into Materials Discovery Models' Performance through Standardized Cross-Validation Protocols. Digit. Discov. 2025, 4, 625–635.

20. Lu, S.; Zhou, Q.; Guo, Y.; Zhang, Y.; Wu, Y.; Wang, J. Coupling a Crystal Graph Multilayer Descriptor to Active Learning for Rapid Discovery of 2D Ferromagnetic Semiconductors/Half-Metals/Metals. Adv. Mater. 2020, 32, e2002658.

21. Nalajala, A.; Ragunathan, T.; Naha, R. Efficient Prefetching and Client-Side Caching Algorithms for Improving the Performance of Read Operations in Distributed File Systems. IEEE Access 2022, 10, 126232–126252.

22. Ko, T.W.; Deng, B.; Nassar, M.; Barroso-Luque, L.; Liu, R.; Qi, J.; Thakur, A.C.; Mishra, A.R.; Liu, E.; Ceder, G.; et al. Materials Graph Library (MatGL), an Open-Source Graph Deep Learning Library for Materials Science and Chemistry. npj Comput. Mater. 2025, 11, 253.

23. Li, X.; Zhou, Y.; Yu, S.; Jia, G.; Li, H.; Li, W. Urban Heat Island Impacts on Building Energy Consumption: A Review of Approaches and Findings. Energy 2019, 174, 407–419.

24. Ouyang, Z.; Sciusco, P.; Jiao, T.; Feron, S.; Lei, C.; Li, F.; John, R.; Fan, P.; Li, X.; Williams, C.A.; et al. Albedo Changes Caused by Future Urbanization Contribute to Global Warming. Nat. Commun. 2022, 13, 3800.

25. Kim, K.; Kang, S.; Yoo, J.; Kwon, Y.; Nam, Y.; Lee, D.; Kim, I.; Choi, Y.-S.; Jung, Y.; Kim, S.; et al. Deep-Learning-Based Inverse Design Model for Intelligent Discovery of Organic Molecules. npj Comput. Mater. 2018, 4, 67.

26. Yuan, H.; Zeng, J.; Zuo, J.; Wang, X.; Xu, D. A General LLM-Powered Text Mining Framework: Applied to Extract High Entropy Alloys. Comput. Mater. Sci. 2026, 264, 114476.

27. McDonald, R.I.; Biswas, T.; Sachar, C.; Housman, I.; Boucher, T.M.; Balk, D.; Nowak, D.; Spotswood, E.; Stanley, C.K.; Leyk, S. The Tree Cover and Temperature Disparity in US Urbanized Areas: Quantifying the Association with Income across 5,723 Communities. PLoS ONE 2021, 16, e0249715.