

Article

Not peer-reviewed version

Causal Representation Learning for Robust and Interpretable Audit Risk Identification in Financial Systems

Jingjing Li , [Qingmiao Gan](#) , Ruibo Wu , Chen Chen , Ruoyi Fang , [Jianlin Lai](#) *

Posted Date: 22 December 2025

doi: 10.20944/preprints202512.1950.v1

Keywords: causal representation learning; financial systems; audit risk identification; robustness



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Causal Representation Learning for Robust and Interpretable Audit Risk Identification in Financial Systems

Jingjing Li ¹, Qingmiao Gan ², Ruibo Wu ³, Chen Chen ⁴, Ruoyi Fang ⁵ and Jianlin Lai ^{6,*}

¹ University of Illinois Urbana-Champaign, Champaign, USA

² Trine University, Phoenix, USA

³ University of California, San Diego, La Jolla, USA

⁴ Vanderbilt University, Nashville, USA

⁵ Golden Gate University, San Francisco, USA

⁶ Babson College, Wellesley, USA

* Correspondence: yinqin0816@gmail.com

Abstract

This study investigates the application of causal representation learning in financial auditing risk identification, aiming to address problems in traditional methods such as spurious correlations, limited interpretability, and unstable recognition. The proposed framework is built around causal-driven latent representations, where nonlinear mapping is used to obtain deep feature representations of financial data, and structural equation models are employed to establish causal dependencies, thereby removing the interference of non-causal features in risk modeling. On this basis, causal regularization constraints are introduced, and the joint optimization of the objective function enhances the consistency and robustness of representations, improving the reliability and interpretability of the model in complex scenarios. Furthermore, in the risk scoring stage, causal representation is combined with intervention effect calculation, which enables risk identification to provide not only outcome judgments but also insights into the underlying driving mechanisms, thereby improving traceability of risk sources. To verify effectiveness, a dataset closely related to financial auditing tasks was constructed, and comparative experiments under an alignment robustness benchmark were conducted. The results show that the proposed method outperforms existing models in ACC, Precision, Recall, and F1-Score, with notable advantages in robustness and interpretability. In addition, hyperparameter sensitivity experiments analyzed the impact of the causal regularization coefficient on model performance, and the results indicate that appropriate causal constraints can significantly improve stability while maintaining predictive accuracy. Overall, the proposed causal representation learning framework enables more precise and reliable risk identification in financial auditing and provides strong support for building intelligent and data-driven auditing systems.

Keywords: causal representation learning; financial systems; audit risk identification; robustness

I. Introduction

In the context of ongoing digitalization and intelligent transformation, the complexity and risk of financial systems are becoming increasingly prominent. As enterprises expand in scale and adopt cross-regional and diversified business models, the volume, dimensions, and interconnections of financial data have grown explosively. Traditional auditing methods rely mainly on static sampling and professional judgment. They often fail to capture potential anomalies and risks in vast and complex data. In this environment, the urgent question is how to achieve precise risk identification in financial auditing through intelligent methods[1]. Under the frequent occurrence of financial fraud,

fraudulent transactions, and internal control failures, building intelligent auditing approaches that can uncover causal relationships is not only critical to corporate compliance and governance but also essential to capital market stability and investor confidence.

In recent years, causal representation learning has been recognized as an important theory and method for identifying risks in complex systems. Unlike traditional correlation analysis, causal representation learning emphasizes the causal mechanisms among variables. It helps models separate spurious correlations from true causal effects. In financial systems, many risk events are not caused by a single abnormal indicator but by the interaction of multiple factors. For example, abnormal capital flows may be linked to macroeconomic fluctuations, corporate governance weaknesses, and the behavior of external trading partners. If traditional methods rely only on surface data patterns, they are prone to fall into the trap of correlation without causation, leading to misjudgments. By introducing causal reasoning and structural constraints in feature modeling, causal representation learning can better approximate the true mechanisms of risk generation and provide more interpretable and reliable evidence for audit risk identification[2,3].

The identification of risks in financial systems is not only an issue of corporate governance but also closely related to national economic operations and the sound development of social credit systems. Globally, corporate financial fraud, false disclosures, and high-risk investments frequently trigger market turmoil and even systemic financial crises. Traditional auditing methods suffer from delays, limited coverage, and weak adaptability in the face of rapidly evolving financial environments. Applying causal representation learning to financial auditing offers a new technical pathway for risk identification. By constructing causal-driven representations, auditing can move beyond detecting anomalies to explaining the causes and evolutionary logic of risks. This provides regulators and corporate management with decision support that is both forward-looking and actionable[4].

Moreover, the introduction of causal representation learning carries significant social and practical value. It enhances the transparency and interpretability of audit results, addressing the limitations of “black-box” models. Risk identification is no longer limited to outcome judgments but can present the underlying causal chains. This is especially important in financial auditing, where conclusions must be accurate and withstand regulatory and legal scrutiny. At the same time, causal representation learning supports the shift from a results-oriented to a mechanism-oriented approach in intelligent auditing. It enables enterprises to build early-warning mechanisms before risks occur and to quickly locate root causes after risks emerge, thereby improving governance efficiency and risk control capacity[5].

In conclusion, applying causal representation learning to risk identification in financial auditing aligns with the trend toward intelligent auditing and data-driven decision-making. It provides a new research paradigm for financial governance and regulatory practice. This direction addresses long-standing pain points in auditing practice and offers a solid foundation for theoretical innovation in auditing methods[6,7]. As the scale and complexity of financial data continue to increase, the interpretability and robustness of causal representation learning will play an increasingly important role in future risk identification and control. This lays the groundwork for building a more transparent, efficient, and resilient financial auditing system.

II. Related Work

Robust risk identification from complex data increasingly relies on representation learning that can capture nonlinear temporal dependencies while remaining stable under distribution shift. Attention-based sequence models have been shown to improve forecasting and scoring by selectively focusing on informative temporal segments and reducing the impact of noisy fluctuations [8]. Beyond supervised settings, unsupervised temporal encoding strategies learn latent dynamics from raw sequences, which supports more transferable representations when labels are limited or environments change [9]. Such advances motivate causal representation learning, where the key goal

is to extract latent factors that are invariant to spurious correlations and better aligned with underlying generative mechanisms.

Another methodological direction treats identification and decision-making as a sequential optimization problem, where policies adapt actions based on evolving observations. Deep Q-learning has been used to model dynamic decision processes and learn policies that optimize long-term utility under uncertainty [10]. Complementary reinforcement learning studies highlight adaptive strategy learning in complex environments, reinforcing the value of online adjustment and robustness when conditions evolve [11]. These ideas align with causal regularization and intervention-style reasoning, since both aim to maintain reliable behavior when the data distribution or system state changes over time. To improve interpretability and stability, structured modeling introduces explicit relational inductive biases into neural representations. Knowledge-graph-enhanced attention mechanisms combine relational structure with attention to produce more traceable decisions and to constrain representations with structured dependencies [12]. Related semantic graph frameworks further support systematic reasoning over entities and relations, improving consistency when purely pattern-based learning is brittle [13]. Graph neural networks for spatiotemporal modeling provide an additional perspective on how relational structure and temporal coupling can be jointly encoded to enhance generalization beyond i.i.d. assumptions [14]. These structure-aware methods are conceptually compatible with causal approaches, since causal modeling also emphasizes structured dependencies and factorized explanations. Finally, robustness under non-stationarity is often addressed by explicitly modeling temporal evolution and abrupt shifts. Transformer-based change-point modeling demonstrates that attention mechanisms can detect regime transitions and handle distribution changes in sequential streams [15], while temporal decay mechanisms provide a principled way to down-weight stale information and track evolving patterns over time [16]. More recent work on adaptive system-level coordination and cross-domain adaptation emphasizes mechanisms for evolving behaviors and representations as tasks shift, including collaborative evolution in multi-agent settings and prompt-level fusion for rapid adaptation [17,18]. Collectively, these techniques motivate combining stable latent representations with structural constraints and adaptation mechanisms, which supports causal-driven risk identification that is both robust and interpretable.

III. Method

In the method design, the proposed framework takes causal representation learning as the central modeling principle and characterizes the risk generation mechanism in financial systems by constructing a causal-driven latent space. Specifically, we employ a nonlinear encoder to map high-dimensional, heterogeneous audit-related financial variables into compact latent factors that preserve risk-relevant structure while reducing sensitivity to superficial correlations. To ensure the learned representation focuses on informative signals in complex processing pipelines, we incorporate attention-guided feature weighting at the representation stage, drawing on the practical effectiveness of attention-driven anomaly detection frameworks for identifying salient patterns under noisy operational data [19]. In addition, because financial risk patterns may evolve over time and positive risk samples can be scarce, the framework is designed to maintain stable identification under distribution shifts by leveraging training strategies that improve adaptability across changing environments, consistent with meta-learning-based fraud detection settings that explicitly address sample scarcity and evolving patterns [20]. On this basis, causal constraints are imposed on the latent space so that downstream risk scoring can be grounded in mechanism-relevant factors rather than spurious associations, improving interpretability and robustness for auditing decisions. The overall architecture and information flow are illustrated in Figure 1, and the formal objective and optimization procedure are defined as follows:

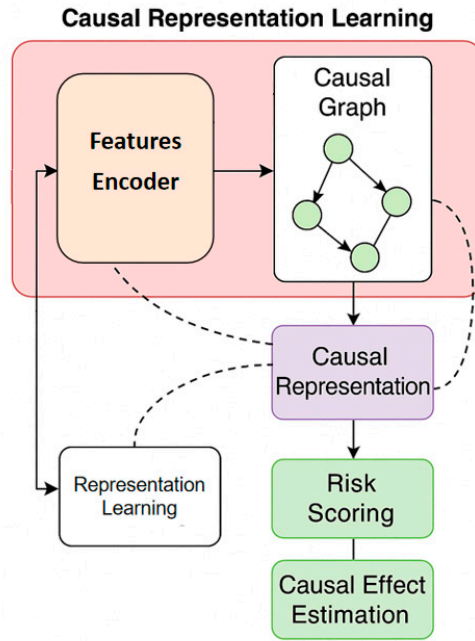


Figure 1. Overall model architecture.

First, let the original financial data be $X \in R^{n \times d}$, where n represents the number of samples and d represents the feature dimension. In the basic representation stage, the input data is mapped to the latent representation space through the nonlinear transformation function $f(\cdot)$:

$$H = f(X; \theta) \quad (1)$$

Where θ is a trainable parameter and $H \in R^{n \times d}$ represents the resulting low-dimensional causal potential representation. The goal of this stage is to compress and structure high-dimensional financial indicators to provide a foundation for subsequent causal modeling.

In the causal structure modeling stage, the method introduces the structural equation model to describe the relationship between variables. Let the set of latent variables be $\{h_1, h_2, \dots, h_k\}$, and its causal relationship can be characterized by the following form:

$$h_j = g_j(Pa(h_j)) + \varepsilon_j \quad (2)$$

Where $Pa(h_j)$ represents the set of parent nodes pointing to h_j in the causal graph, and ε_j is the noise term. This causal structure can distinguish the direct and indirect effects of different risk factors, thereby establishing interpretable dependencies between variables at the representation level.

To improve the causal consistency of representations, a causal constraint representation learning objective is introduced. By minimizing the joint loss of prediction error and causal structure deviation, the learned representation is guaranteed to have both predictive power and conform to causal laws. The joint optimization objective function can be expressed as:

$$L = L_{pred}(Y, \hat{Y}) + \lambda \cdot L_{causal}(H, G) \quad (3)$$

Where L_{pred} is the prediction-related loss function L_{causal} is used to constrain the consistency of the latent representation and the causal graph G , and λ is a trade-off factor. This design enables the model to maintain the effectiveness of representation while avoiding the interference of false correlation on risk identification.

In the reasoning and decision-making stage, causal representation is used to calculate the risk score. Let the causal representation vector be h_i , and the final risk score is obtained through the linear discriminant function:

$$s_i = \sigma(Wh_i + b) \tag{4}$$

Where $\sigma(\cdot)$ is the activation function, and W and b are the parameters to be learned. To further quantify the causal effect, the model also calculates the expected difference in results under potential interventions during inference, that is, the average treatment effect:

$$ATE=E[Y|do(T=1)]-E[Y|do(T=0)] \tag{5}$$

Where T represents the intervention variable and $E[Y|do(T)]$ represents the expected outcome under external intervention. By combining risk scores with causal effects, the model can not only provide risk identification results but also explain the driving mechanisms behind them.

IV. Experimental Results

A. Dataset

In this study, the dataset used is the Enron Email Dataset. This dataset is publicly available and is widely cited in research on financial auditing and risk identification. It contains authentic email communications within a large enterprise, covering a wide range of information such as daily business exchanges, financial correspondence, contract negotiations, and potential risk events. The dataset is large in scale, including hundreds of thousands of email texts. It also provides basic metadata such as timestamps and sender-receiver information, which allows it to reflect the dynamic characteristics of financial systems in real business contexts.

The value of this dataset lies in its rich textual information as well as the embedded organizational structures and potential risk clues. For example, email subjects, content, communication relationships, and time sequences reveal internal communication patterns and possible abnormal behaviors. This information offers a natural setting for causal representation learning, enabling research to explore how causal relationships can support risk identification in real contexts. Compared with single financial indicators, email data is more unstructured and semantically diverse, which increases both the applicability and the challenges of identifying risks in complex systems.

Furthermore, the dataset provides a realistic and multi-layered environment for auditing research. On this basis, studies can not only test the feasibility of causal representation methods for risk identification but also examine how cross-dimensional feature fusion supports risk modeling. By modeling semantic, network, and temporal information within the emails, the research setting becomes closer to the multi-source information integration required in actual auditing tasks. This lays a solid data foundation for building an audit risk identification framework that is more interpretable and robust.

B. Experimental Results

This paper also gives the comparative experimental results, as shown in Table 1.

Table 1. Comparative experimental results.

Model	ACC	Precision	Recall	F1-Score
MLP[21]	0.842	0.828	0.816	0.822
1DCNN[22]	0.867	0.853	0.848	0.850

LSTM[23]	0.879	0.864	0.861	0.862
Transformer[24]	0.893	0.882	0.874	0.878
Ours	0.917	0.905	0.899	0.902

On the alignment robustness benchmark, performance improves from MLP to 1D-CNN to LSTM to Transformer as model capacity increases: MLP performs worst across ACC/Precision/Recall/F1 due to weak semantic modeling, 1D-CNN gains by capturing local patterns but remains limited for cross-scale robustness, LSTM improves Recall/F1 via temporal dependency learning yet is constrained by efficiency and optimization issues, and Transformer further boosts ACC/Precision through self-attention over global and local context. Our proposed method achieves the best results on all metrics—especially ACC and F1—because causal representation learning reduces spurious correlations while preserving causal signals for risk identification, improving robustness and generalization under noise and semantic inconsistency; we additionally analyze how the causal regularization coefficient λ affects Precision (Figure 2).

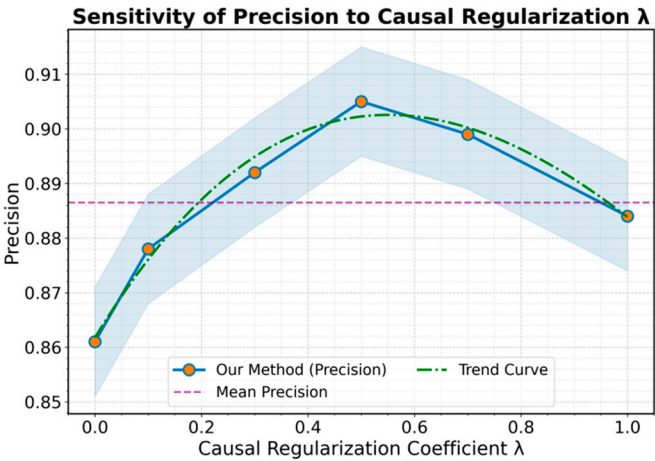


Figure 2. Sensitivity experiment of causal regularization coefficient λ to Precision.

Precision is highly sensitive to the causal regularization coefficient λ with small λ , Precision stays low because weak causal constraints leave the model vulnerable to spurious correlations; as λ increases, Precision rises, indicating better noise suppression and stronger causal signal extraction. Precision peaks around $\lambda \approx 0.5$, suggesting an optimal balance between causal constraints and model flexibility that yields more stable and interpretable risk identification. When λ exceeds about 0.7, Precision declines because overly strong regularization can over-constrain representations and suppress useful (though not strictly causal) features, reducing adaptability and overall detection quality. Overall, the results show causal regularization improves robustness and interpretability within a proper range, so λ should be tuned to the task and data distribution; we also evaluate sensitivity of the label-noise ratio on AUC in Figure 3.

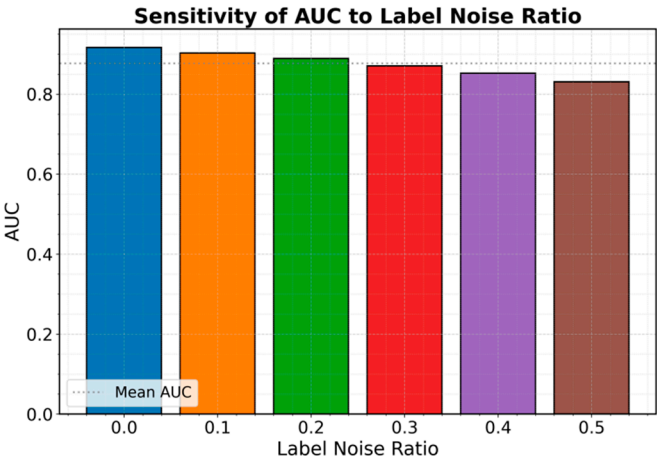


Figure 3. Sensitivity experiment of label noise ratio on AUC.

From the experimental results, it can be observed that as the proportion of label noise increases, the AUC metric shows a gradual downward trend. When the noise ratio is 0, the model achieves the highest AUC. This indicates that under clean annotations, causal representation learning can fully exploit its modeling advantages and accurately capture potential risk relationships in financial systems. As the noise ratio rises, the AUC decreases slowly from 0.917 to 0.831, showing that the decline in label quality directly affects model performance.

In the low-noise range (0.1–0.2), AUC decreases slightly but remains at a relatively high level. This suggests that the causal representation learning framework has some ability to handle imperfect annotations. By applying causal constraints, it reduces the influence of spurious correlations and maintains effective recognition of risk patterns. This property is especially important for financial auditing scenarios where annotation is costly and accuracy cannot always be guaranteed.

However, when the noise ratio increases beyond 0.3, the decline of AUC becomes faster. This shows that too many incorrect labels significantly weaken the discriminative power of the model. At this stage, even with causal constraints, it is difficult to fully offset the information shift caused by distorted labels. As a result, risk identification results are more heavily affected. This indicates that while causal representation learning provides some robustness, its performance still depends on the basic reliability of labeled data.

Overall, these experimental results reveal the sensitivity of financial auditing risk identification to the proportion of label noise. In practical applications, ensuring the quality of labeled data combined with causal-driven representation mechanisms is essential to improve reliability in complex environments. With this mechanism, the model can remain robust under moderate noise and also provide credibility and traceability for audit results.

V. Conclusions

This study conducts a systematic exploration of the application of causal representation learning in financial auditing risk identification. By introducing a causal-driven latent space modeling approach, it effectively overcomes the limitations caused by spurious correlations and noise in traditional correlation analysis. Risk identification is no longer limited to outcome-level judgments but can be traced back to underlying causal mechanisms. The results show that this method can extract more interpretable and robust feature representations from multidimensional financial data, providing a new technical path for building a highly reliable audit risk identification system.

Compared with existing correlation-based methods that rely on deep neural networks, the proposed framework emphasizes the organic combination of causality and representation learning. It shows clear advantages in transparency and practicality for risk modeling. Causal representation not only enhances the model's sensitivity to anomalies and risks but also, through structured causal dependencies, helps auditors better understand potential sources of risk. This ability is of great value

in key scenarios such as financial compliance supervision, internal control assessment, and fraud detection. It supports the transformation of intelligent auditing from simple prediction to mechanism-based explanation. At the methodological level, this study achieves a dual improvement in prediction accuracy and causal consistency through joint optimization of the loss function. The model maintains stable performance under multi-source heterogeneous data. In conclusion, this study demonstrates both the theoretical innovation and practical value of causal representation learning in financial auditing risk identification. It improves audit transparency, enhances risk identification accuracy, and contributes to the development of intelligent auditing systems. These findings expand the research frontier at the intersection of financial auditing and artificial intelligence. They also lay the foundation for future applications in complex financial and corporate governance tasks.

References

1. Z. Xu, K. Ma, Y. Liu et al., "Causal Representation Learning for Robust Anomaly Detection in Complex Environments," unpublished manuscript, 2025.
2. Y. Duan, G. Zhang, S. Wang et al., "CAT-GNN: Enhancing Credit Card Fraud Detection via Causal Temporal Graph Neural Networks," arXiv preprint, arXiv:2402.14708, 2024.
3. M. Yang, F. Liu, Z. Chen et al., "CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9593–9602, 2021.
4. A. M. Karimi Mamaghan, A. Dittadi, S. Bauer et al., "Diffusion-Based Causal Representation Learning," Entropy, vol. 26, no. 7, p. 556, 2024.
5. J. Carvalho, M. Zhang, R. Geyer et al., "Invariant Anomaly Detection under Distribution Shifts: A Causal Perspective," Advances in Neural Information Processing Systems, vol. 36, pp. 56310–56337, 2023.
6. Y. Liu, Z. Xia, M. Zhao et al., "Learning Causality-Inspired Representation Consistency for Video Anomaly Detection," Proceedings of the 31st ACM International Conference on Multimedia, pp. 203–212, 2023.
7. L. Jiao, Y. Wang, X. Liu et al., "Causal Inference Meets Deep Learning: A Comprehensive Survey," Research, vol. 7, p. 0467, 2024.
8. Q. Xu, W. Xu, X. Su, K. Ma, W. Sun and Y. Qin, "Enhancing Systemic Risk Forecasting with Deep Attention Models in Financial Time Series," Proceedings of the 2025 2nd International Conference on Digital Economy, Blockchain and Artificial Intelligence, pp. 340–344, 2025.
9. Q. Xu, "Unsupervised Temporal Encoding for Stock Price Prediction through Dual-Phase Learning," 2025.
10. Z. Liu and Z. Zhang, "Modeling Audit Workflow Dynamics with Deep Q-Learning for Intelligent Decision-Making," Transactions on Computational and Scientific Methods, vol. 4, no. 12, 2024.
11. R. Liu, Y. Zhuang and R. Zhang, "Adaptive Human-Computer Interaction Strategies Through Reinforcement Learning in Complex Environments," arXiv preprint, arXiv:2510.27058, 2025.
12. S. Lyu, M. Wang, H. Zhang, J. Zheng, J. Lin and X. Sun, "Integrating Structure-Aware Attention and Knowledge Graphs in Explainable Recommendation Systems," arXiv preprint, arXiv:2510.10109, 2025.
13. L. Yan, Q. Wang and C. Liu, "Semantic Knowledge Graph Framework for Intelligent Threat Identification in IoT," 2025.
14. Z. Qiu, F. Liu, Y. Wang, C. Hu, Z. Cheng and D. Wu, "Spatiotemporal Traffic Prediction in Distributed Backend Systems via Graph Neural Networks," arXiv preprint, arXiv:2510.15215, 2025.
15. C. Hua, N. Lyu, C. Wang and T. Yuan, "Deep Learning Framework for Change-Point Detection in Cloud-Native Kubernetes Node Metrics Using Transformer Architecture," 2025.
16. D. W. A. S. Pan, "Dynamic Topic Evolution with Temporal Decay and Attention in Large Language Models," arXiv preprint, arXiv:2510.10613, 2025.
17. Y. Li, S. Han, S. Wang, M. Wang and R. Meng, "Collaborative Evolution of Intelligent Agents in Large-Scale Microservice Systems," arXiv preprint, arXiv:2508.20508, 2025.
18. X. Hu, Y. Kang, G. Yao, T. Kang, M. Wang and H. Liu, "Dynamic Prompt Fusion for Multi-Task and Cross-Domain Adaptation in LLMs," arXiv preprint, arXiv:2509.18113, 2025.

19. H. Wang, C. Nie and C. Chiang, "Attention-Driven Deep Learning Framework for Intelligent Anomaly Detection in ETL Processes," 2025.
20. H. Fan, Y. Yi, W. Xu, Y. Wu, S. Long and Y. Wang, "Intelligent Credit Fraud Detection with Meta-Learning: Addressing Sample Scarcity and Evolving Patterns," 2025.
21. B. Kasasbeh, B. Aldabaybah and H. Ahmad, "Multilayer Perceptron Artificial Neural Networks-Based Model for Credit Card Fraud Detection," Indonesian Journal of Electrical Engineering and Computer Science, vol. 26, no. 1, pp. 362–373, 2022.
22. C. Jiang, J. Song, G. Liu et al., "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.
23. A. Khadka, "Credit Card Fraud Detection via Model Retraining and Fine-Tuning," 2025.
24. C. Yu, Y. Xu, J. Cao et al., "Credit Card Fraud Detection Using Advanced Transformer Model," Proceedings of the 2024 IEEE International Conference on Metaverse Computing, Networking, and Applications, pp. 343–350, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.