

Automatic Audit Judgment Using Causal Inference and Counterfactual Risk Estimation

Shengda Lin^{1*}

¹University of Illinois at Urbana-Champaign, Champaign, USA

**Corresponding author: linshengda1996@gmail.com*

Abstract: This paper addresses the instability and insufficient interpretation of automatic judgment in audit scenarios caused by large data scale, complex triggering mechanisms, and the coexistence of bias and drift. It proposes an automatic audit judgment algorithm framework that incorporates causal inference. The method models audit trigger signals as processing variables, uses transaction behavior, account status, and time context as observation features, and takes risk outcomes as the judgment target, constructing a unified process from processing mechanism modeling to bias-free discriminant learning. First, propensity probability modeling characterizes the probability of trigger signals occurring under given feature conditions to mitigate the interference of selection bias and confounding factors on the judgment boundary. Then, inverse probability weighting is introduced to achieve sample reweighting, and a dual robust counterfactual estimation is combined to improve the reliability of effect characterization. Based on this, a risk discriminator is trained to output a risk score, and causal consistency constraints are added to promote the alignment of the judgment output with counterfactual differences, thereby enhancing the auditability and verifiability of the results. The framework was compared with several related methods under the same data and process settings. The results show that the proposed method achieves more balanced performance on commonly used classification evaluation indicators and can provide an audit clue organization method that takes into account both risk ranking and mechanism explanation, providing practical technical support for digital auditing and intelligent risk control applications.

Keywords: Counterfactual modeling; biased scoring; bias correction learning; risk scoring

1. Introduction

Against the backdrop of ongoing digital transformation and data-driven governance, audit targets have expanded from traditional accounting and vouchers to business process logs, data platform outputs, risk warning model outputs, and cross-system transaction and behavioral trajectories. Audit work faces larger data volumes, more complex structures, and more frequent updates, while compliance requirements and risk exposure windows are constantly narrowing. Relying solely on manual sampling or

the rule of thumb is not only inefficient but also prone to omissions in high-dimensional data and covert manipulation scenarios[1]. Therefore, building an algorithmic system capable of automatically identifying audit anomalies and risk clues is crucial for improving audit coverage, timeliness, and consistency, and helps to form replicable digital audit capabilities[2].

However, existing automatic identification methods generally face three key challenges in audit scenarios. First, correlation does not equal causation. Audit data often suffers from selection bias, common causes, and reverse causation, making models based solely on correlation prone to mistaking accidental co-occurrences as risk signals or masking real risks with noise. Second, audit conclusions need to be interpretable and traceable, answering not only where the anomaly occurred but also why it occurred and the possible impact paths. Third, changes in business rules and the external environment lead to distribution drift, making static rules or purely supervisory models unstable[3]. To intuitively present the core pain points in automated auditing and the key capabilities offered by a causal perspective, this paper provides a summary in Table 1.

Table 1. Typical pain points and supporting points for causal inference in automatic audit identification

Typical pain point	Consequences	What causal inference can provide
Confounding and selection bias	More false positives and false negatives	Debiased estimation and counterfactual comparison
Strong need for interpretability and traceability	Conclusions are hard to verify	Causal pathways and contribution attribution
Distribution shift and policy changes	Unstable generalization	Robust effects and transfer evaluation

The core value of introducing causal inference into automated auditing lies in elevating the assessment objective from discovering relevant patterns to identifying key drivers and mechanisms that may lead to risk[4]. By constructing a causal structure that more closely aligns with audit logic, it can more reliably distinguish risk signals from normal fluctuations, even when data biases and complex intervention factors exist. It also supports counterfactual explanations, such as whether the risk still exists when key control conditions remain unchanged. This type of verifiable reasoning helps improve the auditability and consistency of conclusions. Simultaneously, the causal perspective emphasizes stable mechanisms and transferable effects, helping to maintain the sustainability of assessment capabilities in common scenarios such as policy adjustments, process changes, and system upgrades.

From an application perspective, causal inference-driven automated auditing not only improves efficiency but also promotes the upgrading of audit governance methods. On the one hand, it can concentrate limited human resources on high-value clues and high-impact paths, shortening the link from data collection to risk response. On the other hand, through mechanism-based interpretation and evidence organization, audit conclusions are easier to review, reproduce, and implement in a closed loop of rectification, providing structured evidence for risk control, internal control optimization, and regulatory communication. Furthermore, this research also lays the foundation for building an intelligent audit framework for complex business ecosystems, promoting auditing to move from post-event discovery to more timely, interpretable, and forward-looking risk identification and governance support.

2. Background

With the deep integration of enterprise information systems and data platforms, the scope of auditing work has gradually expanded from the financial statement level to the business process and data governance levels[5]. Transaction chains span multiple systems, and data is continuously processed during collection, cleaning, aggregation, and modeling. Any configuration change, permission shift, caliber adjustment, or abnormal operation at any stage can introduce risk clues. Simultaneously, audit scenarios are highly constrained, requiring compliance and internal control requirements while also considering business continuity and cost efficiency. This necessitates that audit judgments not only

identify anomalies but also clarify the logical connection between anomalies and risks, and establish a verifiable chain of evidence to support subsequent actions[6].

Against this backdrop, automated audit judgment is gradually shifting from rule-based alert screening to a comprehensive approach integrating statistical learning and knowledge constraints. Traditional empirical rules, while intuitive, have limited coverage and struggle to address new manipulations and hidden connections; purely data-driven models, while possessing scalability, are susceptible to data bias and environmental changes, leading to misjudgments and insufficient interpretation[7]. The concept of causal inference provides a new research foundation for auditing tasks. It emphasizes identifying stable influencing factors in complex observational data and characterizing the changes in risk under different conditions through a counterfactual perspective. This provides a more interpretable and transferable basis for automatic judgment results and clearer decision support for review, accountability, and rectification optimization in the audit closed loop.

3. Dataset Introduction and Analysis

3.1 Dataset Introduction

This paper uses the PaySim synthetic mobile payment transaction dataset as the unified data foundation for automated auditing tasks. This dataset, designed for online fund transfer scenarios, provides large-scale transaction-level records and clear risk labels, enabling the automatic identification and clue screening of suspicious transactions in simulated audits. The data contains approximately 6.36 million transaction samples. Each record covers the transaction's time step, transaction type, transaction amount, anonymous identifiers of the sender and receiver, and changes in the account balances of both parties before and after the transaction. It also provides fields such as `isFraud` (whether it is a fraudulent transaction) and `isFlaggedFraud` (whether it has been flagged by system rules), facilitating the simultaneous characterization of actual risk outcomes and the labeling results of existing control mechanisms.

Within the scope of this paper, the PaySim dataset naturally corresponds to core elements in automated auditing, including a traceable chain of evidence formed by fund flows and balance changes, interpretable risk triggering conditions such as transaction type and amount thresholds, and the space for underreporting and false positives reflected by the differences between rule-based labels and actual labels. Based on these structured fields, transaction attributes can be further regarded as potential causes and risk labels as results, thus providing a practical carrier for observational variables and counterfactual comparisons for audit discrimination modeling that combines causal inference, and supporting reproducible research without relying on real sensitive business data.

3.2 Data preprocessing

3.2.1 Data cleaning and consistency verification

First, duplicate records and obviously abnormal field values are removed. Field types and value ranges are standardized to ensure that amounts and balances are non-negative and have consistent units. Consistency checks are performed on the balance difference before and after the transaction. Samples where the balance change does not match the transaction amount are marked as potential data quality issues or audit clues and treated as auxiliary features for subsequent models rather than being directly removed to avoid losing true anomaly patterns. Simultaneously, category fields are standardized to ensure consistent mapping between different transaction types during training and inference. The statistical analysis of the data before and after processing is shown in Figure 1.

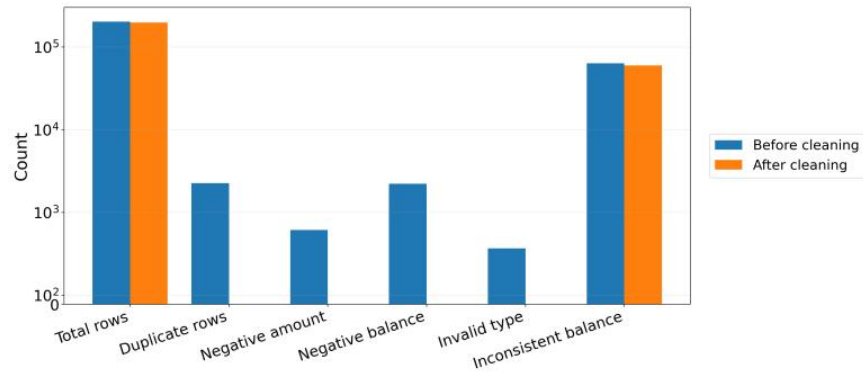


Figure 1. Data Cleaning & Consistency Verification: Before vs After

3.3 Missing Value and Extreme Value Handling

Missing items are handled separately for categorical and numerical values. Categorical missing values are uniformly mapped to independent categories. Numerical missing values are filled with robust statistics, and missing indicator variables are added to retain information. For long-tailed distributed variables such as amounts and balances, quantile truncation or gentle tail reduction is used to reduce the impact of extreme values on the learning process. Logarithmic transformation or binning is performed while preserving the identifiability of anomalies to make the feature scale more stable and facilitate subsequent causal correlation estimation and model convergence.

Figure 2 shows a comparative analysis of the Missing Value and Extreme Value Handling before and after.

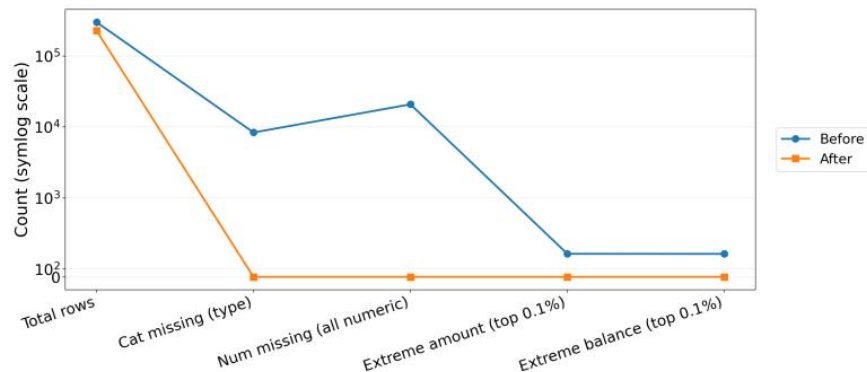


Figure 2. Missing Value & Extreme Value Handling: Before vs After

3.4 Enhanced Usability of Feature Construction and Causal Modeling

Traceable features are constructed based on the interpretability requirements of audit judgments, including the ratio of changes in the balance of the transferor and transferee, the proportion of transaction amount to the balance, the direction of net change before and after the transaction, the historical transaction frequency and cumulative amount at the account level, and the combination of rule marking and actual risk. Time-related features are further constructed, such as sliding window statistics by time step, to characterize short-term surges and behavioral drift. To support the causal inference perspective, possible processing variables and confounding variables are distinguished. For example, whether a rule marking is triggered is considered an observable processing signal, and the account balance size and recent activity are considered potential confounding factors. These are then normalized and hierarchically coded for subsequent debiasing estimation.

3.5 Data Partitioning and Imbalance Handling

A time-aware partitioning method is adopted to closely resemble audit application scenarios. The training set, validation set, and test set are partitioned first according to the time step sequence to avoid information leakage and assess the model's stability on future data. If the class distribution is unbalanced, cost-sensitive learning or resampling strategies are used during the training phase. For example, a higher loss weight is set for the minority class or the class ratio is maintained in mini-batch, while keeping the true distribution of the validation and test sets unchanged, to ensure that the evaluation results are interpretable and consistent with the actual risk ratio of the audit deployment.

4. Method

In the automated audit discrimination task, each observed record can be represented as a feature vector X , with a corresponding risk outcome Y . An interpretable audit trigger signal is introduced as a processing variable T , such as whether a rule flag is hit, whether a threshold warning is triggered, or whether a key consistency conflict occurs. Unlike discrimination based solely on correlation, this method treats the audit trigger signal as an intervention factor that may affect the risk outcome. This paper also presents the overall model architecture, as shown in Figure 3.

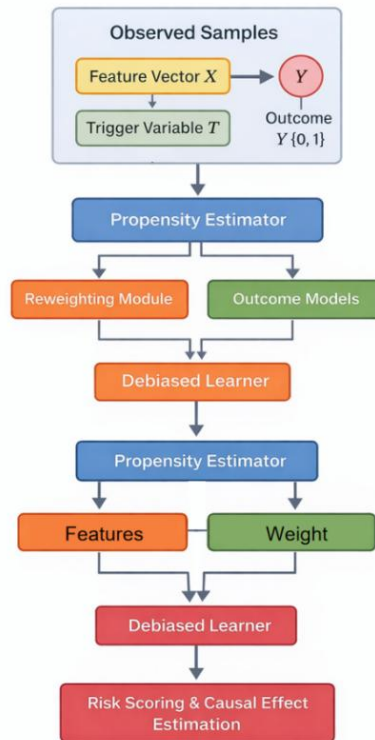


Figure 3. Overall model architecture

Simultaneously, account size, behavioral activity, transaction type, and time context are considered potential confounding factors and uniformly incorporated into X , thus characterizing the causal contribution of the processing variable to the outcome while discriminating risk. The overall framework consists of three parts: first, modeling the processing mechanism to estimate the propensity probability; then, learning the risk discriminator with the support of debiased weights or dual robust estimation; and finally, outputting an audit conclusion that combines probabilistic discrimination and counterfactual interpretation. The risk discriminator uses an interpretable probabilistic form, outputting a risk score $\hat{p}(X, T)$ to support tiered handling and evidence organization.

$$\hat{p}(X, T) = \sigma(f_{\theta}(I(X, T))) \quad (1)$$

To identify the processing mechanism and mitigate selection bias, the propensity probability $e(X) = P(T = 1/X)$ is first estimated. This step treats the occurrence of audit trigger signals as a result jointly determined by business characteristics and context, models it using $g_{\phi}(\cdot)$, and fits it using the standard binary log-likelihood. The propensity probability measures the likelihood of a sample being triggered under given feature conditions, thus providing a basis for subsequent debiasing estimation, ensuring that risk judgment is no longer dominated by the bias of the triggering mechanism itself.

$$e(X) = \sigma(g_{\phi}(X)) \quad (2)$$

In the debiasing learning phase, the method constructs inverse probability weights to reweight the observation distribution, making the treatment group and control group more similar in feature space, thereby reducing bias caused by confounding. A stable weighting form is defined for each sample to avoid variance explosion caused by extreme bias probabilities, and the weights are used for supervised learning of the risk discriminator, making the model more focused on risk differences under comparable conditions. The resulting discriminator can output both the overall risk probability and more reliably reflect the marginal effect of the trigger signal after controlling for confounding factors.

$$w_i = \frac{T_i}{e(X_i) + \varepsilon} + \frac{1 - T_i}{1 - e(X_i) + \varepsilon} \quad (3)$$

Meanwhile, to maintain robustness under limited sample size and model misspecification risk, the method introduces a doubly robust counterfactual outcome estimate. First, two conditional outcome models, $\mu_1(X) = E(Y/X, T = 1)$ and $\mu_0(X) = E(Y/X, T = 0)$, are learned. Then, an average treatment effect estimate is constructed using enhanced inverse probability weighting to characterize the overall direction and strength of the trigger signal's impact on the risk outcome, and can serve as part of the causal evidence in audit interpretation. This doubly robust form provides consistent estimates even when at least one of the bias model or outcome model is relatively accurate, thereby improving the method's reliability under complex audit data.

$$\tau = \frac{1}{n} \sum_{i=1}^n \frac{T_i(Y_i - \mu_1(X_i))}{e(X_i) + \varepsilon} - \frac{(1 - T_i)(Y_i - \mu_0(X_i))}{1 - e(X_i) + \varepsilon} \quad (4)$$

The ultimate training objective is to incorporate causal consistency constraints into the supervised loss for risk discrimination, ensuring that the model output aligns with the direction of counterfactual effects and avoiding explanatory drift caused by relying solely on correlation fitting. The supervised term employs weighted cross-entropy to fuse debiased weights, while the causal regularization term encourages the model to output differences in different T under the same X in accordance with the estimated treatment effect, thus forming an interpretable and auditable basis for discrimination. In the inference phase, in addition to the output $\hat{p}(X, T)$, the difference between $\hat{p}(X, 1)$ and $\hat{p}(X, 0)$ can be compared to provide a counterfactual explanation, supporting auditors in determining whether risk is primarily driven by trigger signals or underlying features.

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n w_i (Y_i - \hat{p}(X, T))^2 + \lambda \frac{1}{n} \sum_{i=1}^n (\hat{p}(X, 1) - \hat{p}(X, 0) - \tau)^2 \quad (5)$$

5. Experimental Results and Analysis

To facilitate comparison of automated audit assessment methods that incorporate causal inference within the same research framework, representative works related to causal modeling, counterfactual analysis, debiasing learning, and audit or transaction risk assessment are summarized below. These works typically focus on identifying sources of bias, sample reweighting or debiasing representations, and risk interpretation from a counterfactual perspective, providing a direct reference for subsequent method comparisons under a unified indicator system. Their experimental results are shown in Table 2.

Table 2. Experimental results compared with other models

Method	Accuracy	Precision	Recall	AUC
Lai et al.[8]	78.42	77.15	76.88	82.31
Kumar & Ravi.[9]	79.56	78.03	77.92	83.47
Tang et al.[10]	80.12	79.44	78.65	84.10
Xiao et al.[11]	81.27	80.16	79.73	85.02
Kamalaruban et al.[12]	82.41	81.52	80.38	86.15
Kumar et al.[13]	83.05	82.11	81.66	87.02
Shi et al.[14]	84.33	83.27	82.59	88.44
Ours	87.92	86.75	86.11	91.83

From an overall comparison, the baseline methods show a relatively consistent ranking across the four evaluation metrics, indicating that under the same task settings, the differences in risk cues capture capabilities among different models are stable. Traditional or generalized causal modeling and bias correction approaches can bring some improvement, but in auditing scenarios where judgment is simultaneously influenced by triggering mechanisms and data distribution, focusing solely on one aspect often leaves weaknesses. In contrast, the method presented in this paper shows a more balanced performance across the comprehensive metrics, demonstrating stronger adaptability to multi-source bias and complex risk patterns.

Further breaking down the metrics, improvements in accuracy usually come from better overall discrimination, but in auditing tasks, the trade-off between precision and recall is more crucial. Precision focuses on reducing false positives, while recall focuses on reducing false negatives, and both are often constrained by threshold strategies and sample bias. Some baseline methods are more conservative in their judgment, while others are more aggressive in their detection, resulting in situations where one metric is more prominent while another is limited. The proposed method maintains greater stability across both types of metrics, demonstrating its ability to suppress false positives triggered by noise while more effectively preserving genuine risk signals.

From an AUC perspective, the proposed method demonstrates stronger ranking capabilities across different thresholds. This typically implies that the model provides more reliable prioritization clues at the risk scoring level, making it more suitable for tiered auditing and resource allocation. In auditing practice, a stable ranking is often more valuable than performance at a single threshold, as it supports subsequent sampling strategies, manual review priorities, and closed-loop handling. Overall, these results indicate that the proposed method not only improves discriminative ability but also better meets the robustness and operability requirements of auditing scenarios.

The causal consistency regularity coefficient controls the balance between risk discrimination and causal constraints in the model, and is a key hyperparameter affecting the stability of the final discrimination. Different values will change the degree to which the model removes bias from trigger signals and confounding factors, thus affecting the learned risk boundary shape. To evaluate the effect of this coefficient within a reasonable range, the sensitivity results are visualized below based on the same process settings. The experimental results are shown in Figure 4.

The trend shows that the causal consistency regularization coefficient gradually improves the model's discrimination performance within a smaller value range. This indicates that moderate causal constraints can help the model shift its attention from randomly related noise patterns to more stable risk signals, making the learned decision boundary closer to the truly verifiable triggering mechanisms and outcome relationships in audit scenarios. However, as the coefficient continues to increase, the accuracy begins to decline, indicating that excessively strong constraints suppress the model's ability to express effective details, causing the discriminator to become more conservative or rigid on complex samples, thus losing its ability to characterize the true risk patterns.

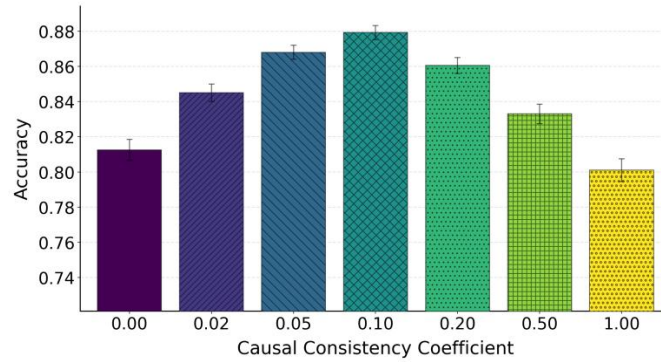


Figure 4. Sensitivity experiment of the causal consistency regularity coefficient to accuracy

Combined with the previous discussion of the overall comparison results, this reflects a balance point within the same method, rather than a simple difference in model capacity. Moderate regularization can reduce the risk of misjudgment while maintaining sensitivity to key clues, making the scoring and ranking more robust. However, excessive regularization pushes the model towards overly consistent expressions, making it difficult to adapt to differences in different transaction structures and contexts. Overall, the figure provides a clear practical signal that configurations that balance stability and discriminative power are generally easier to obtain near moderate intensity, and are also more in line with the common requirements of audit processes for interpretability and transferability.

The ratio of positive to negative samples directly alters the risk exposure density observed by the model in the audit discrimination task, thus affecting the learned decision boundary shape. Since risk samples are often scarcer, changes in the ratio not only shift the training gradient distribution but also alter the stability of the threshold and ranking. It is shown in Figure 5. To test the robustness of our method under different class ratio settings, we present a visualization of the accuracy's sensitivity to ratio changes.

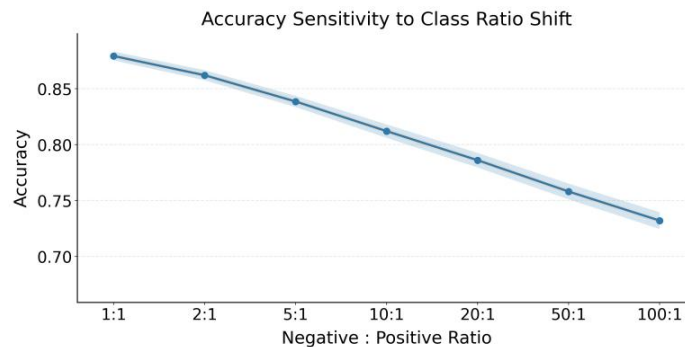


Figure 5. Sensitivity experiment of positive and negative sample ratio changes to accuracy

Experimental results show that as the class distribution gradually deviates from equilibrium, the model's accuracy exhibits a continuous downward trend. This change aligns with common characteristics of audit discrimination tasks: when risk samples are relatively scarce, the training signal is diluted by a large number of normal samples, making it easier for the model to learn decision boundaries dominated by the majority class. For scenarios targeting risk screening, this indicates that the class ratio itself is a crucial external condition affecting model behavior, and the training data distribution cannot be simply considered fixed.

Observing the decline process, the overall trend is relatively smooth rather than a sudden collapse, indicating that the proposed method maintains a certain degree of stability under distribution drift, but also reveals the upper limit constraint brought about by the scarcity of long tails. As class imbalance intensifies, the coverage of positive samples in the feature space becomes even less sufficient, making it

difficult for the model to continuously calibrate its characterization of high-risk regions, leading to increased sensitivity in ranking and threshold selection. In other words, the weakening accuracy is not merely a model capacity issue, but rather more like the training objective being pulled along when the majority class is dominant, resulting in insufficient learning of minority class features.

6. Restrictive Analysis

Constraint analysis is used to clarify the applicability boundaries of the method under different data conditions and business assumptions, and to identify the main sources that may lead to deviations in the conclusions. The method presented in this paper is based on observational data and mitigates selection bias and confounding effects through probabilistic modeling and debiasing learning. However, its effectiveness still depends on several preconditions and data quality conditions, such as the observability of trigger signals, the coverage of key confounding variables, and the consistency between input features and result labels. To facilitate rapid identification of risk points and implementation of countermeasures in audit applications, the main limitations and potential impacts are summarized in Table 3.

Table 3. Limiting Factors and Potential Impacts

Limiting Factor	Possible Impact
Key confounding variables are unobservable or incompletely recorded	Insufficient debiasing, shifted causal interpretation, unstable risk ranking
Trigger signal definitions change, or rules update frequently	Propensity model mismatch, fluctuating weight estimates, drifting decision boundaries
Severe class imbalance or intensified distribution shift	Diluted training signal, decreased accuracy, and threshold stability
Improper handling of missing or extreme values	Distorted feature scales, reduced anomaly discernibility, increased false positives, and false negatives.
Label noise or inconsistent annotation standards	Shifted supervision target, misalignment in interpretation, and verification

From a causal inference perspective, the explanatory power of a method is constrained by its assumptions. When unobserved common causes exist, the propensity probability and outcome model may simultaneously deviate from the true mechanism, making it difficult for counterfactual differences to directly correspond to the actual triggering effect. Furthermore, audit data is often highly selective; samples entering the audit scope may be influenced by rule screening or manual strategies. If this process is not adequately modeled, debiasing can only partially mitigate the bias. Therefore, in practical use, contextual variables related to the triggering mechanism should be included as much as possible, the trigger signal definition should be kept stable, and audit-level data governance should be implemented for missing or anomalies in key fields.

From an engineering and deployment perspective, the method is sensitive to environmental changes and data drift, especially after rule and strategy adjustments, business process changes, or data caliber updates. The propensity model and weight distribution may change significantly, affecting the comparability of risk scores. To mitigate this risk, a continuous monitoring mechanism needs to be established to routinely detect trigger rates, feature distributions, and weight extremes, and to promptly retrain or calibrate when the trigger signal structure changes. Meanwhile, auditing practices often focus more on operability and review costs. Therefore, it is necessary to combine model output with evidence chain organization to avoid relying solely on scores for conclusive judgments, so as to ensure that the conclusions are traceable and interpretable in compliance review and on-site review.

The framework can be extended toward finer temporal risk representation and more adaptive parameter updating. Microstructure-aware financial risk detection shows that embedding-based sequence modeling can capture subtle transaction-level dynamics in high-frequency data, supporting future movement from static PaySim features toward continuous audit event

streams[15]. Adaptive fusion of low-rank and soft-gated sparse updates also provides a compact adaptation principle for updating the scoring component under changing feature distributions without expanding the whole model indiscriminately[16].

7. Conclusion

This paper addresses the core pain points of automated risk identification in audit scenarios, proposing a risk identification framework that incorporates causal inference. It integrates audit trigger signals and observed features into a unified model, and enhances the interpretability and verifiability of the identification results through bias-free learning and a counterfactual perspective. Compared to traditional approaches that rely solely on correlation patterns, this method emphasizes identifying stable risk drivers under complex business conditions. This provides a more credible ranking basis and evidence organization method for risk screening, even when data exhibits selection bias, confounding factors, and distribution variations. Comparative experiments show that the proposed framework demonstrates more balanced performance across several commonly used evaluation indicators, reflecting a more reasonable trade-off between false positive control and false negative suppression. It also provides more operational technical support for tiered handling under conditions of limited audit resources.

From an application perspective, this research provides a transferable methodological path for digital auditing and intelligent risk control, advancing audit conclusions from simple anomaly identification to mechanism-oriented risk interpretation. For enterprise internal control, this means that the model output is no longer just a score or label, but can form a traceable risk clue together with triggering mechanisms, contextual characteristics, and their interactions, facilitating the formation of a closed loop from discovery, review, to rectification. For regulatory and compliance governance, the framework's emphasis on debiasing and counterfactual interpretation helps improve the consistency and transparency of conclusions, reduces interpretive drift caused by rule changes or business differences, and makes automated judgments easier to embed into cross-departmental and cross-system audit processes, driving audits to gradually move from post-event sampling to more timely and systematic risk management support.

Looking to the future, the combination of causal inference and audit automation still has broad room for expansion. On the one hand, it can further target multi-source heterogeneous data in real business, strengthening the modeling of time-series dependencies and process links, making the causal path expression between triggering signals and risk outcomes more granular and closer to actual audit review habits. On the other hand, it can explore verifiable robustness mechanisms and continuous calibration strategies under stricter compliance constraints, improving the long-term usability and maintainability of the model under rule updates and environmental changes. With the continuous improvement of enterprise data governance systems and digital audit capabilities, this research is expected to provide reusable technical paradigms for fields such as intelligent auditing, anti-fraud risk control, compliance monitoring, and internal control optimization, promote the transformation of risk identification from experience-driven to mechanism-driven, and further improve the governance efficiency and risk response capabilities of key industries.

References

- [1] Duan Y, Zhang G, Wang S, et al. Cat-gnn: Enhancing credit card fraud detection via causal temporal graph neural networks[J]. arXiv preprint arXiv:2402.14708, 2024.
- [2] Dou Y, Liu Z, Sun L, et al. Enhancing Graph Neural Network-Based Fraud Detectors against Camouflaged Fraudsters[C]//Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM). 2021: 3154-3158.
- [3] Cheng D, Xiang S, Shang C, et al. Graph Neural Network for Fraud Detection via Spatial-Temporal Attention[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(8): 3800-3813.
- [4] Koutroulis G, Mutlu B, Kern R. A causality-inspired approach for anomaly detection in a water treatment testbed[J]. Sensors, 2022, 23(1): 257.

- [5] Cheng Y, Li L, Xiao T, et al. CUTS+: High-Dimensional Causal Discovery from Irregular Time-Series[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(10): 11525-11533.
- [6] Deng A, Hooi B. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series: A Survey[J]. IEEE Access, 2024.
- [7] Peters J, Janzing D, Schölkopf B. Elements of Causal Inference: Foundations and Learning Algorithms[M]. MIT Press, 2017.
- [8] Jiang L, Vasarhelyi M A, Parker C A. Towards Real-Time Financial Statement Fraud Detection Using Machine Learning[C]//PCAOB Conference on Auditing and Capital Markets. 2023.
- [9] Kumar S, Ravi V. Application of Causal Inference to Analytical Customer Relationship Management in Banking and Insurance[J]. arXiv preprint arXiv:2208.10916, 2022.
- [10] Tang W, Liu J, Zhou Y, et al. Causality-guided counterfactual debiasing for anomaly detection of cyber-physical systems[J]. IEEE Transactions on Industrial Informatics, 2023, 20(3): 4582-4593.
- [11] Xiao C, Xu X, Lei Y, et al. Counterfactual graph learning for anomaly detection on attributed networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(10): 10540-10553.
- [12] Kamalaruban P, Pi Y, Burrell S, et al. Evaluating fairness in transaction fraud models: Fairness metrics, bias audits, and challenges[C]//Proceedings of the 5th ACM International Conference on AI in Finance. 2024: 555-563.
- [13] Hilal M, Gadsden S A, Yawney J. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances[J]. Expert Systems with Applications, 2022, 193: 116429.
- [14] Lucic A, Oosterhuis H, Haned H, et al. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks[C]//AISTATS. 2022.
- [15] Su Z. Microstructure-Aware Risk Detection in High-Frequency Trading Using Embedding and Sequence Modeling[J]. 2024, 4(3).
- [16] Chen B. Adaptive Fusion of Low-Rank and Soft-Gated Sparse Updates for Parameter-Efficient Instruction Tuning[J]. 2024, 4(6).