International Journal of Global Sustainable Research (IJGSR)
Vol.4 No.10, 2026: 1547-1564

Causal Intent-Disentangled Feature Engineering for Fraud Detection in Government Expenditure

Nuh Saneraro Fakdawer[1*], Misfah Muslimawati[2], Amiruddin[3], Asri Usman[4]
[1]Universitas Ottow Geissler
[2,3,4] Universitas Hasanuddin
Corresponding Author: Nuh Saneraro Fakdawer nuhfakdawer2@gmail.com

A R T I C L E I N F O

Keywords : Government Expenditure Fraud, Causal Representation Learning, Graph Neural Networks, Domain Generalization, Explainable AI

Received : 2 August
Revised : 20 September
Accepted: 20 October

©2026 Fakdawer, Muslimawati, Amiruddin, Usman : This is an open-access article distributed under the terms of the Creative Commons Atribusi 4.0 Internasional.

A B S T R A C T

Government expenditure fraud models often confuse misconduct signals with benign administrative style, which weakens transfer across agencies, regions, procurement categories, and time. This study develops Causal Intent-Disentangled Feature Engineering (CIDFE), a heterogeneous-graph representation framework that separates fraud intent from contextual style using invariant prediction, adversarial domain confusion, orthogonality, counterfactual style swapping, and confounder-adjusted residuals. Leave-domain-out experiments show that CIDFE improves average and worst-domain AUPRC, reduces false positives for high-value legitimate contracts, and improves calibration relative to tabular, graph, textual, multimodal, and domain-generalization baselines. Shapley-based explanations further separate genuine red flags from administrative anomalies. The findings support CIDFE as a transferable, auditor-oriented feature engineering approach for public expenditure fraud detection

DOI: https://doi.org/10.59890/ijgsr.v4i10.15
ISSN-E: 3026-4677
https://journalijgsr.my.id/index.php/ijgsr/index

INTRODUCTION

Fraud in government expenditure is difficult to detect not because suspicious transactions are absent from public records, but because the signals of misconduct are mixed with ordinary administrative variation. Procurement systems generate large volumes of structured and unstructured information: contract values, payment timing, budget codes, supplier histories, tender relationships, audit notes, and descriptive text. Conventional screening can identify unusual values, yet unusual does not necessarily mean fraudulent. A high-value purchase may be legitimate, a delayed payment may reflect procedural constraints, and a distinctive phrase in a contract may simply be regional boilerplate. When these contextual patterns are learned as if they were fraud signals, detection models become accurate within familiar data but fragile when deployed across agencies, regions, procurement categories, or fiscal years. This transfer problem is especially consequential in public financial management because audit resources are limited and false positives redirect attention away from cases that warrant investigation.

Machine learning has improved the capacity to process public financial records at scale. Classical models such as logistic regression, decision trees, random forests, and gradient-boosted ensembles remain useful when fraud indicators are expressed as structured attributes, while anomaly-detection systems can surface unusual expenditure patterns even when labels are incomplete (Schiavon et al., 2026; Shovon, 2025). Sequential models add sensitivity to temporal behavior, and adaptive time-series approaches can respond to evolving compliance environments (Odunaike, 2025). Related applications in government deposit processing show that learned classifiers can complement conventional audit procedures by screening transactions more consistently than purely manual review (Sany & Uddin, 2023). These developments demonstrate the operational value of data-driven audit support, but most conventional pipelines still represent expenditure as independent rows of features and therefore underuse the relational structure through which collusion, bid rigging, coordinated suppliers, and approval chains emerge.

Graph-based learning addresses part of this limitation by treating agencies, suppliers, officials, tenders, contracts, and payments as connected entities rather than isolated observations. Heterogeneous and relational graph neural networks can encode different node and edge types and can therefore capture procurement semantics that are lost in flat tables (Sha et al., 2025; Wang et al., 2025). Textual information adds another layer of evidence. Transformer models such as FinBERT can represent contextual meaning in financial language, while multimodal models can combine narrative and structured evidence when one source alone is incomplete (Yang et al., 2020; Sharma & Hassan, 2026). Yet richer representations introduce a second problem: the model can become better at learning spurious patterns. Procurement wording, agency templates, supplier naming conventions, or category-specific terminology may correlate with past audit findings without being causally related to fraudulent intent. Soares et al. (2025) show why explainable analysis of procurement language matters:

semantic associations can appear predictive while reflecting incidental regularities rather than misconduct.

Causal and invariant representation learning provides a way to distinguish transferable fraud evidence from context-dependent correlation. Domain-generalization methods seek representations for which the relationship with the outcome remains stable across environments, and invariant information approaches explicitly penalize predictors that depend on environment-specific shortcuts (Li et al., 2022). Counterfactual augmentation offers a complementary test by changing attributes that should not alter the label while preserving the underlying causal structure (Zhao et al., 2022). Recent work extends causal representation learning to mixed structured and textual information and shows that confounder-aware representations can improve robustness when multiple data modalities are involved (Feng et al., 2026). Similar ideas have been used to separate authorial style from content (Man & Nguyen, 2024), to support cross-style causal representations in language tasks (Zhao et al., 2026), and to investigate whether semantic representations remain stable across linguistic settings (Salhan et al., 2026). Surveys of deep causal learning further underline that representation, discovery, and inference are related but distinct problems (Deng et al., 2025). In financial fraud detection, the central implication is clear: robust models need to isolate signals that are plausibly tied to misconduct rather than merely correlated with the administrative environment (Gupta et al., 2026). A further requirement is interpretability. Fraud-screening systems do not simply classify data; they allocate scarce audit attention and can expose suppliers, officials, and agencies to additional scrutiny. Public-sector use therefore requires explanations that are intelligible to auditors and defensible to oversight bodies. Shapley-based attribution provides a principled way to estimate the marginal contribution of features to a prediction (Giudici & Raffinetti, 2021). However, explaining an entangled feature space does not solve the underlying causal problem: an attribution may faithfully report that a model relied on a regional phrase or agency-specific template even though that dependence is undesirable. The more useful objective is to make the representation itself less dependent on contextual style and then explain predictions in terms of the remaining fraud-intent features and legitimate operational residuals.

This study addresses that gap through Causal Intent-Disentangled Feature Engineering (CIDFE). The novelty lies in treating feature engineering for government expenditure fraud detection as a causal separation problem rather than a conventional aggregation problem. CIDFE combines heterogeneous graph representation, transformer-based text encoding, explicit decomposition of intent and style subspaces, cross-domain prediction consistency, adversarial domain confusion, subspace orthogonality, counterfactual style swapping, confounder-adjusted residuals, and Shapley-based explanations in one auditor-oriented architecture. The study has three objectives: first, to determine whether intent-disentangled features improve fraud detection when regions, agencies, procurement categories, and fiscal years are held out from training; second, to test whether residual adjustment reduces false positives among legitimate high-value transactions while improving calibration; and third, to evaluate whether

the learned representation produces explanations that distinguish genuine red flags from benign administrative anomalies. These objectives move the analysis from in-sample accuracy toward transferability, operational burden, and accountable use in public financial management.

LITERATURE RIVIEW

Machine Learning and Relational Fraud Signals

Public-sector fraud analytics has developed from rule-based screening toward increasingly expressive predictive models. Structured learners remain effective when known risk indicators are measurable, and anomaly-detection systems can identify unusual expenditure behavior without requiring a complete set of fraud labels (Schiavon et al., 2026; Shovon, 2025). Time-sensitive approaches add value when the meaning of a transaction depends on its sequence or regulatory context (Odunaike, 2025), while machine learning in government deposit processing illustrates the practical complementarity between automated screening and traditional audit work (Sany & Uddin, 2023). The limitation is representational: a flat record does not naturally encode recurring supplier–agency relationships, common officials, co-bidding structures, or payment chains.

Graph neural networks provide a more suitable representation because procurement is inherently relational. Relational and heterogeneous graph models allow message passing to depend on relation type, enabling the model to distinguish, for example, an agency awarding a contract from an official approving a payment. Existing graph-based fraud research demonstrates that such models can detect coordinated patterns that tabular learners miss (Sha et al., 2025; Wang et al., 2025). The present study builds on this insight but does not assume that a richer graph representation is automatically more robust. If domain-specific wording or administrative routines are encoded into the graph representation, the model can still overfit to the environment in which fraud was previously observed.

Textual Evidence, Multimodality, and Spurious Correlation

Procurement and audit systems contain substantial narrative evidence. Contract descriptions, tender notices, audit observations, and inspection notes often contain context that is unavailable in numerical fields. Transformer language models provide contextual embeddings that are more sensitive to meaning than bag-of-words features, and domain-adapted financial language models such as FinBERT demonstrate the value of specialized textual representation (Yang et al., 2020). Multimodal systems extend this logic by combining text with structured financial evidence and can remain useful when parts of the evidence are missing or corrupted (Sharma & Hassan, 2026).

The same flexibility that makes language models powerful also makes them vulnerable to shortcut learning. Agency templates, local wording, supplier naming conventions, and procurement-category terminology can become proxies for the outcome even when they have no stable causal relation to fraud. Evidence from public bidding documents shows that explainable and unsupervised approaches can uncover such spurious semantic relationships

(Soares et al., 2025). Accordingly, the theoretical problem is not simply how to fuse more information, but how to prevent context-specific style from contaminating the representation that drives the fraud prediction.

Causal Invariance and Counterfactual Robustness

Domain generalization reframes robustness as the ability to preserve predictive relationships when the data-generating environment changes. Invariant information methods seek features that retain target-relevant information while discarding environment-specific variation (Li et al., 2022). This principle is particularly relevant to government procurement because the same misconduct can appear under different administrative conventions, while the same administrative pattern can be benign in one jurisdiction and suspicious in another. A model that depends on the latter type of association may perform well under random train-test splits but deteriorate sharply when transferred to an unseen agency or fiscal year.

Counterfactual reasoning strengthens this idea by asking whether the learned intent representation remains stable when a non-causal attribute is deliberately changed. Counterfactual link learning demonstrates how alternative but structurally coherent examples can be used to distinguish causal from incidental relations (Zhao et al., 2022). In representation learning, analogous strategies have improved robustness to authorship style (Man & Nguyen, 2024), cross-style language variation (Zhao et al., 2026), and cross-lingual conceptual variation (Salhan et al., 2026). Mixed structured-text causal inference further shows the value of representing confounders explicitly rather than assuming that a fused embedding is causally clean (Feng et al., 2026). CIDFE adapts these ideas to public expenditure by swapping style attributes while preserving transactional facts and red-flag semantics, then requiring the intent representation to remain stable.

The distinction between representation and inference is important. Deep causal learning research emphasizes that causal representation, causal discovery, and causal effect estimation solve different problems (Deng et al., 2025). CIDFE is primarily a representation-learning intervention: it does not claim that every latent dimension identifies a fully verified causal mechanism. Instead, it uses invariance, adversarial confusion, orthogonality, confounder adjustment, and counterfactual consistency to reduce reliance on features that behave like domain-specific nuisance factors. This is consistent with the emerging call for fraud analytics to separate stable drivers from incidental association and to make governance and transparency part of model design rather than post hoc considerations (Gupta et al., 2026).

Explainability and Auditor Accountability

Explainability is not optional in a government audit setting. A fraud score may determine which supplier is investigated, which agency receives additional scrutiny, and how audit resources are allocated. Shapley-based attribution is attractive because it decomposes a prediction into feature-level contributions and can be applied consistently across model families (Giudici & Raffinetti, 2021). Yet attribution quality depends on the quality of the feature space. If the representation itself is entangled, a perfectly faithful explanation can still reveal dependence on an undesirable proxy. CIDFE therefore links explainability to feature disentanglement: the auditor receives separate contributions from the intent block and the confounder-adjusted residual block, making it easier to distinguish a substantive red flag from a legitimate contextual anomaly.

METHODOLOGY

Research Design and Analytical Logic

The study adopts a comparative machine-learning experimental design centered on cross-domain generalization. Rather than relying on a conventional random split, the evaluation withholds complete domains so that the test data represent contexts not observed during optimization. Four domain axes are used: administrative region, procuring agency, procurement category, and fiscal year. This design directly tests the claim that fraud-intent features should transfer when administrative style and operational context change. The unit of analysis is an expenditure record linked to the entities and relations that generated it. Confirmed audit outcomes and corruption-risk indicators provide the supervision signal, while transaction attributes and narrative text provide the observable evidence.

CIDFE is positioned as a feature-engineering module between data ingestion and downstream classification. It does not replace standard classifiers; instead, it changes the representation they consume. The architecture first builds a heterogeneous graph, then encodes structured and textual evidence into a joint semantic representation, decomposes that representation into intent and style components, applies invariance and counterfactual constraints, and outputs two feature blocks: causal-intent features and confounder-adjusted residuals. This design allows the same engineered features to be evaluated with gradient-boosted trees and a calibrated neural network, separating the contribution of representation learning from the choice of final classifier.

Data Sources and Heterogeneous Expenditure Graph

The empirical corpus combines three complementary data sources. A national procurement disclosure portal supplies tender notices, award decisions, contract values, supplier identifiers, and awarding-agency codes. A regional audit repository contributes audit notes, inspection reports, and confirmed irregularity findings that provide labels and unstructured textual evidence. A municipal payment ledger contributes disbursement timestamps, budget-line codes, and payment approval chains. Records are linked through supplier tax identifiers, contract reference numbers, and agency codes, producing a corpus of several hundred thousand contracts across multiple fiscal years. A publicly

released procurement benchmark with corruption-risk indicators is additionally used to evaluate transferability with labels independent of the audit-office annotations (Gara et al., 2024).

The heterogeneous graph contains typed nodes for agencies, suppliers, officials, tenders, contracts, and payments. Typed edges represent awarding, approving, disbursing, and co-bidding relations. Structured attributes include contract amount, budget code, payment delay, supplier history, and related transaction characteristics. Textual attributes include contract descriptions and audit notes. Text is encoded with a pretrained transformer, while graph structure is encoded with a relational graph neural network so that message passing respects relation type (Sha et al., 2025; Yang et al., 2020; Wang et al., 2025). A multi-head attention readout pools the subgraph associated with each expenditure record into a joint embedding that becomes the input to the disentanglement stage. Monetary attributes are log-transformed and standardized within fiscal year to reduce inflationary drift, and text fields are truncated to 512 tokens before transformer encoding.

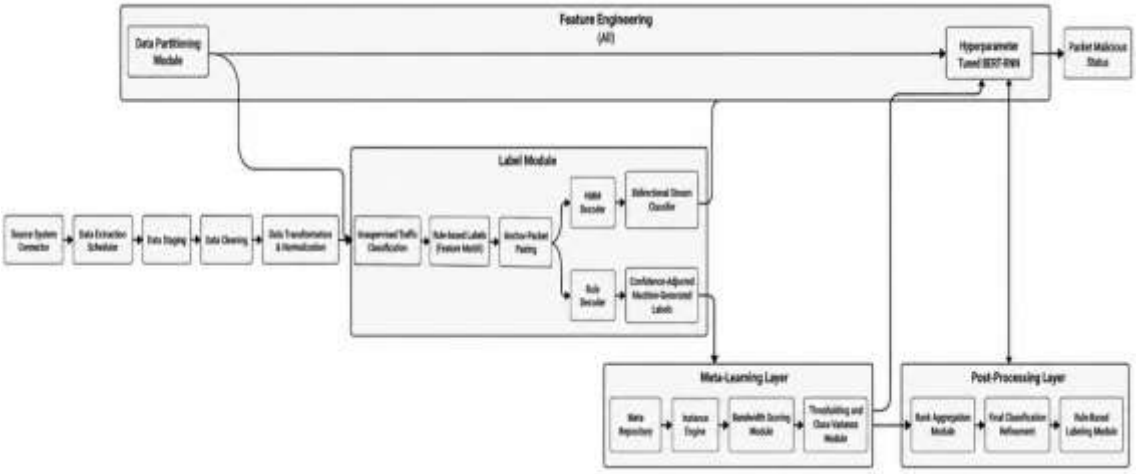


Figure 1. Overall GEFDS Architecture with CIDFE Module

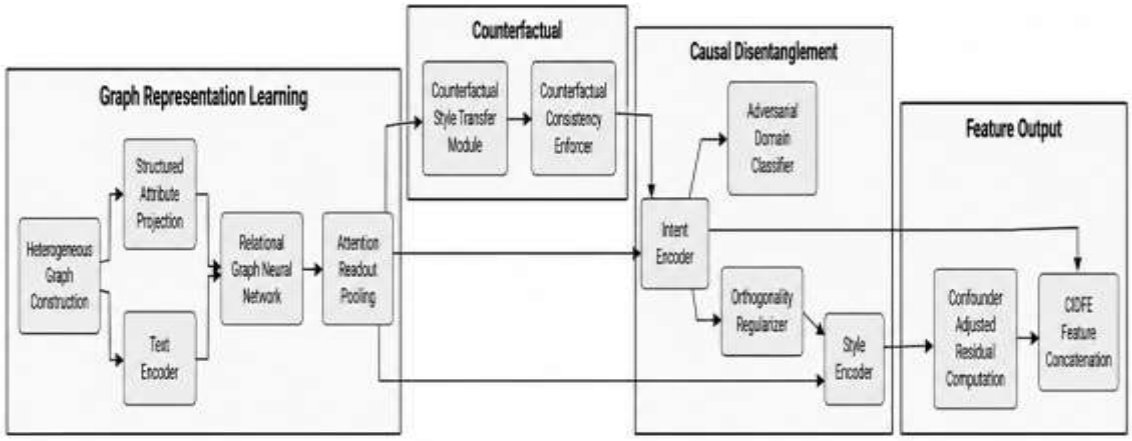


Figure 2. CIDFE Internal Architecture

Intent–Style Decomposition and Causal Invariance

CIDFE assumes that the joint representation contains two conceptually distinct sources of variation. The intent component captures information expected to remain relevant to misconduct across domains, including abnormal payment timing, suspicious supplier concentration, unusual relational structures, and red-flag semantics. The style component captures administrative variation that may be predictive in one environment but should not define fraud, such as regional boilerplate, agency-specific phrasing, supplier naming conventions, and category-dependent terminology. Two neural encoders map the joint representation into separate latent subspaces. The decomposition is deliberately operational rather than metaphysical: it is judged by whether the intent block predicts fraud consistently while carrying little recoverable information about the domain.

Three coupled constraints shape the intent subspace. First, a cross-domain prediction-consistency objective requires a common predictor to perform consistently across training domains. Second, an adversarial domain classifier attempts to recover the domain label from the intent representation while the intent encoder is trained through gradient reversal to make that prediction difficult. Third, an orthogonality penalty discourages the intent and style subspaces from collapsing into redundant representations. The combined objective follows the domain-generalization principle that stable target information should be retained while environment-specific shortcuts are suppressed (Li et al., 2022). An independent post hoc domain probe is later used to assess whether this separation has actually been achieved.

Counterfactual Style Swapping and Residual Adjustment

Counterfactual augmentation provides the principal stress test of the separation. A conditional style-transfer module changes supplier names, agency boilerplate, and category-specific phrasing using style information sampled from a different domain while preserving transactional facts and red-flag semantics. The counterfactual record is re-encoded through the same graph and text pipeline, and a latent consistency objective requires its intent representation to remain close to that of the original record. The mechanism adapts the logic of counterfactual augmentation used in robust representation learning (Man & Nguyen, 2024; Zhao et al., 2022; Zhao et al., 2026). If the intent representation changes materially when only style has been altered, the model has not achieved the desired disentanglement.

The style block is not simply discarded. CIDFE uses it to estimate legitimate operational variation in the structured attributes, and the residual component is retained as a second feature block. This adjustment is intended to prevent procurement opportunity, agency workload, supplier scale, budget size, and related operational factors from masquerading as fraud. The final feature vector therefore combines intent features with confounder-adjusted residuals. This dual representation is designed to support both sensitivity to genuine red flags and restraint when high-value but legitimate transactions differ from the average case.

Domain Partitioning, Baselines, and Metrics

For each domain axis, selected domains are held out completely from training and used only for testing. No supplier, official, or contract from a held-out domain is used during model optimization. This leave-domain-out procedure is stricter than random splitting because it tests the exact failure mode that motivates the study. Baselines span five families: classical tabular learners, including logistic regression, random forest, and XGBoost (Schiavon et al., 2026); sequential models represented by LSTM-style approaches (Odunaike, 2025); graph models including GCN, GAT, RGCN, and heterogeneous GNNs (Sha et al., 2025; Wang et al., 2025); text and multimodal models using transformer evidence (Yang et al., 2020; Sharma & Hassan, 2026); and domain-generalization or causal representation approaches, including invariant learning (Li et al., 2022), mixed-data confounder representation (Feng et al., 2026), and causality-guided counterfactual representation learning (Zhao et al., 2026). All baselines receive the same domain splits and comparable input information.

Because fraud detection is highly imbalanced, AUPRC is the principal performance measure. The analysis reports average AUPRC across held-out domains and worst-domain AUPRC to capture transfer failure in the most difficult context. Precision and recall support interpretation of classification behavior. The false-positive rate is calculated separately for contracts above the ninetieth percentile of value because expensive but legitimate transactions are especially vulnerable to anomaly-based overflagging. Expected calibration error measures whether predicted probabilities correspond to observed risk, while precision at the top five percent of scores reflects the reality that audit teams can review only a small fraction of alerts. Finally, post hoc domain-probe accuracy measures how much domain information remains in the intent representation; lower values indicate stronger invariance.

Implementation and Reproducibility

The relational encoder uses two message-passing layers with 256-dimensional hidden states and four attention heads in the readout. Intent and style encoders are two-layer multilayer perceptrons with 512 hidden units. The style-transfer component is a six-layer conditional transformer decoder. The unified objective is optimized with Adam in mini-batches of 256; loss weights are selected on a validation domain rather than on the test domains. Downstream models include an XGBoost ensemble with 500 trees, maximum depth six, and learning rate 0.05, alongside a three-layer multilayer perceptron with 256, 128, and 64 units, batch normalization, calibration through temperature scaling, and Shapley-based explanation. Experiments are repeated over five random seeds and summarized using means and standard deviations. A self-diagnosing invariance mechanism monitors domain-probe accuracy and triggers retraining when domain predictability exceeds its validation-time reference by more than two percentage points. These settings define the experimental configuration used for all reported comparisons.

Implementation Governance and Human-in-the-Loop Protocol

Model deployment is evaluated as an audit-support process rather than a stand-alone prediction exercise. The operating workflow therefore separates automated scoring from investigative judgment. Incoming transactions are encoded using the most recent graph snapshot and ranked by calibrated risk. Auditors receive the risk score together with the dominant Shapley contributions from the intent and residual blocks, the relevant graph neighborhood, and provenance information linking each feature back to the underlying procurement or payment record. A flagged transaction is not treated as proof of misconduct; it enters a triage queue in which reviewers can confirm, reject, or defer the alert. Confirmed outcomes are retained as new supervision for subsequent retraining, while rejected alerts are analyzed to determine whether the error arose from legitimate operational context, missing graph relations, or drift in administrative language. This human-in-the-loop structure is consistent with the role of machine learning as a complement to conventional audit rather than a substitute for professional evidence assessment (Sany & Uddin, 2023).

The monitoring protocol also recognizes that public expenditure systems evolve. Cross-departmental budget monitoring research shows that automated anomaly detection becomes more useful when it is integrated with correction and management processes rather than operated as an isolated classifier (Y. Li, 2024). CIDFE therefore tracks four deployment diagnostics in addition to predictive performance: domain-probe accuracy, expected calibration error, the high-value false-positive rate, and the proportion of top-ranked alerts overridden by auditors. A sustained rise in domain-probe accuracy signals that the intent representation is becoming more context-specific; worsening calibration indicates that risk probabilities no longer match observed outcomes; an increase in high-value false positives suggests that residual adjustment is no longer absorbing legitimate opportunity effects; and a high override rate may indicate that the representation is missing operational information available to human reviewers. These diagnostics provide a practical basis for deciding whether to refresh graph features, expand the set of domains, revise style-swapping rules, or retrain the model.

Data governance is treated as part of model validity. Linkage across procurement, audit, and payment systems depends on stable identifiers and consistent entity resolution, yet real government databases often contain renamed suppliers, duplicated identifiers, incomplete approval chains, and delayed audit labels. Before training, records should therefore pass deterministic linkage checks, duplicate resolution, missingness profiling, and temporal consistency tests. Text used for counterfactual style swapping should be versioned so that synthetic transformations remain distinguishable from original evidence. Because the model creates latent representations that may encode sensitive relational information, access to embeddings and graph neighborhoods should follow the same control principles applied to case-management data. These safeguards do not improve AUPRC directly, but they reduce the risk that apparent model robustness is produced by leakage, inconsistent entity resolution, or poorly documented data transformations.

RESULT

Cross-Domain Fraud Detection Performance

CIDFE consistently outperforms conventional and recent baselines under leave-domain-out evaluation. Regional average AUPRC reaches 0.847, compared with 0.796 for both the heterogeneous GNN and multimodal fusion baselines. The gap remains visible for agency transfer, where CIDFE records 0.831, and becomes more pronounced for procurement-category transfer, where CIDFE reaches 0.812 compared with 0.714 for the heterogeneous GNN and 0.718 for multimodal fusion. Temporal transfer is the most difficult setting because changes in fiscal year can alter monetary scales, regulations, and administrative conventions simultaneously. Even there, CIDFE records a temporal average AUPRC of 0.748 and a worst-domain AUPRC of 0.688, compared with 0.693 and 0.556 for the heterogeneous GNN. The strongest counterfactual baseline reaches 0.701 temporal average and 0.563 worst-domain AUPRC, indicating that counterfactual reasoning alone does not match the full disentanglement architecture.

Table 1. Cross-Domain Fraud Detection Performance Measured by Average and Worst-Domain AUPRC.

Method	Regional Avg.	Regional Worst	Agency Avg.	Category Avg.	Temporal Avg.	Temporal Worst
Logistic Regression	0.612	0.471	0.598	0.554	0.531	0.402
Random Forest	0.668	0.523	0.651	0.607	0.584	0.448
XGBoost	0.701	0.556	0.689	0.641	0.617	0.479
LSTM	0.694	0.541	0.677	0.628	0.609	0.463
GCN	0.723	0.572	0.708	0.661	0.638	0.497
GAT	0.741	0.589	0.724	0.679	0.654	0.512
RGCN	0.785	0.631	0.757	0.702	0.681	0.541
Heterogeneous GNN	0.796	0.644	0.769	0.714	0.693	0.556
BERT Classifier	0.712	0.548	0.697	0.663	0.641	0.488
Multimodal Fusion	0.796	0.639	0.771	0.718	0.696	0.559
IRM	0.774	0.612	0.752	0.694	0.672	0.528
Causal Confounder Learner	0.789	0.628	0.764	0.707	0.685	0.547
Causality-Guided Counterfactual	0.801	0.647	0.776	0.721	0.701	0.563
CIDFE (Ours)	0.847	0.702	0.831	0.812	0.748	0.688

The domain-probe analysis supports the same interpretation. An independent classifier trained on the intent features predicts the domain with 0.583 accuracy, only slightly above the 0.561 majority-class rate. Applied to the multimodal fused embedding, the same probe reaches 0.812. The difference

indicates that CIDFE suppresses much of the region-, agency-, category-, and time-specific information that remains available in an entangled representation. Figure 3 provides a visual counterpart: the learned intent space separates records primarily by fraud status, while points from different domains remain mixed within the fraud and benign clusters. This pattern is consistent with the stated objective of preserving outcome-relevant structure without organizing the latent space around administrative context.

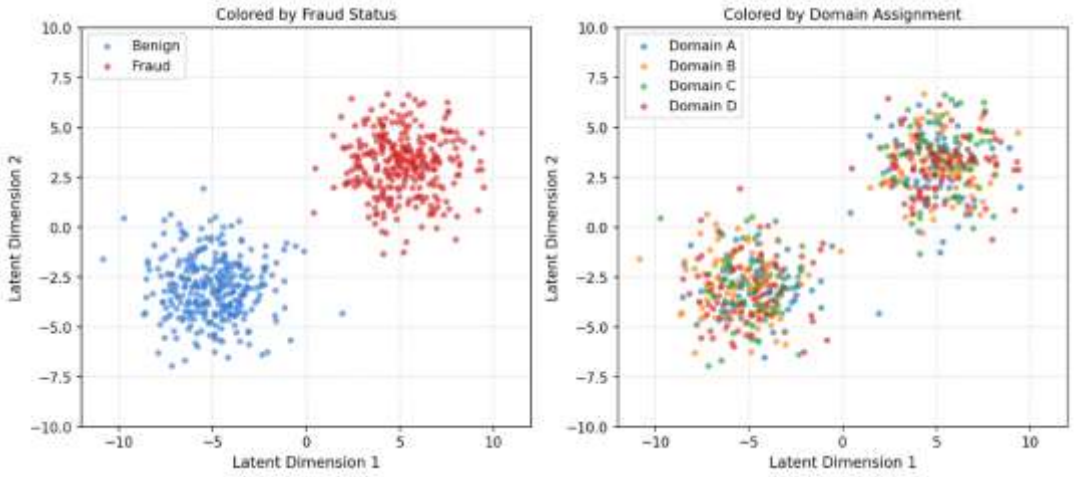


Figure 3. Latent Space Projection of Expenditure Records Colored by Fraud Status and Domain Assignment

False-Positive Burden and Calibration

CIDFE also reduces the operational burden associated with high-value legitimate transactions. For contracts above the ninetieth percentile of contract amount, the regional false-positive rate falls to 0.118, compared with 0.187 for the heterogeneous GNN and 0.174 for the causality-guided counterfactual baseline. The result is consistent with the purpose of the residual block: variation related to procurement opportunity and budget scale is absorbed rather than treated automatically as evidence of misconduct. Calibration improves at the same time. The expected calibration error of the calibrated neural classifier falls to 0.031 when it consumes CIDFE features, compared with 0.058 when the same downstream architecture uses entangled features. Precision among the top five percent of risk scores reaches 0.791, which is operationally important because high-risk ranking, rather than exhaustive classification, is often the practical basis for audit triage.

Ablation Results

The ablation experiment clarifies why the full architecture is needed. Removing adversarial domain confusion produces the largest deterioration in worst-domain performance, reducing temporal worst-domain AUPRC from 0.688 to 0.597 and increasing domain-probe accuracy to 0.671. Without the orthogonality penalty, regional average AUPRC falls to 0.813 and domain-probe accuracy rises further to 0.704, indicating partial collapse of the intended separation. Removing counterfactual consistency has a particularly visible effect on high-value false positives, which increase from 0.118 to 0.163. Omitting

confounder-adjusted residuals raises the high-value false-positive rate to 0.157 and reduces regional average AUPRC to 0.826. No single component reproduces the full configuration, suggesting that transferability results from the interaction among domain confusion, subspace separation, counterfactual stress testing, and residual adjustment rather than from one isolated regularizer.

Table 2. Ablation Results for CIDFE Under Regional and Temporal Domain Shifts

Configuration	**Regional Avg.**	**Temporal Worst**	**High-Value FPR**	**Domain Probe Acc.**
Full CIDFE	0.847	0.688	0.118	0.583
w/o adversarial loss	0.812	0.597	0.141	0.671
w/o orthogonality penalty	0.813	0.641	0.134	0.704
w/o counterfactual consistency	0.826	0.652	0.163	0.612
w/o confounder-adjusted residuals	0.826	0.663	0.157	0.591

Attribution Faithfulness

Shapley analysis provides a final test of whether the learned features are useful for auditor-facing interpretation. Confirmed fraud cases concentrate their largest contributions in intent dimensions associated with red-flag concepts such as bid rigging, split purchasing, shell suppliers, conflicts of interest, and invoice inflation. Benign cases show weaker contributions on those intent dimensions and relatively greater influence from residual features associated with contract amount, payment timing, budget code, supplier size, and agency workload. The separation does not prove wrongdoing; rather, it shows that the model can distinguish the feature block that carries fraud-relevant evidence from the block that captures legitimate operational variation. This distinction is more informative than a conventional explanation of an entangled embedding because auditors can see whether a high score is being driven by a substantive red flag or by contextual deviation.

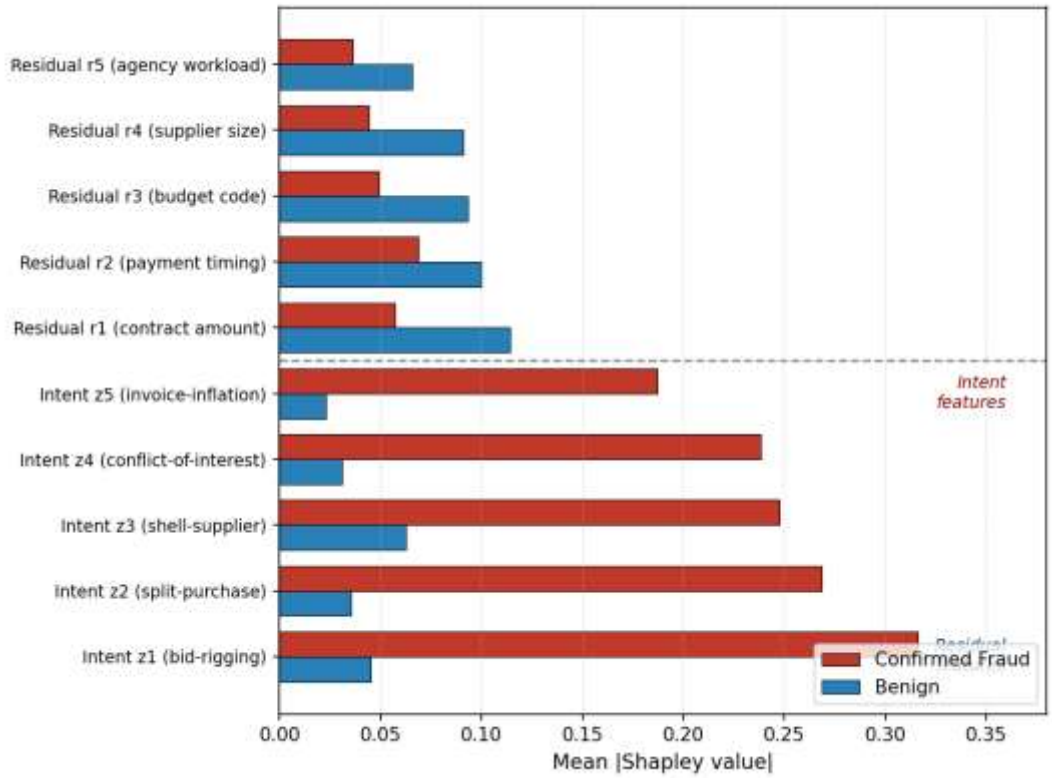


Figure 4. Shapley Value Contributions of Causal Intent and Confounder-Adjusted Residual Features

DISCUSSION

The findings support the central claim that transferability in public expenditure fraud detection is primarily a representation problem. Graph and multimodal baselines are strong within the same information environment because they can exploit rich relational and textual structure. Their weakness emerges when the environment changes. CIDFE improves worst-domain performance not by discarding rich information, but by reorganizing it: fraud-intent evidence is encouraged to remain stable, while administrative style is placed in a separate subspace and operational variation is retained through residual features. This explains why the method outperforms both conventional graph models and invariant-learning baselines applied to a single fused representation. Invariance appears more effective when the model is given an explicit place to put nuisance variation rather than being asked to suppress it implicitly.

The ablation results sharpen this interpretation. Adversarial domain confusion contributes most strongly to worst-domain transfer, whereas counterfactual consistency and residual adjustment are especially important for limiting false positives. These functions are related but not interchangeable. Domain confusion removes easily recoverable contextual identity from the intent representation; orthogonality prevents the two latent blocks from becoming redundant; counterfactual style swapping tests whether intent survives plausible changes in presentation; and residual adjustment preserves legitimate operational context instead of erasing it. Together, they create a more balanced

form of robustness than a single regularization objective. This finding is consistent with broader causal-representation research that treats invariance as something that must be designed and tested under intervention-like changes rather than inferred from in-sample accuracy alone (Deng et al., 2025; Feng et al., 2026; Li et al., 2022).

For public-sector auditing, the reduction in high-value false positives may be as important as the improvement in AUPRC. An audit model that flags every large contract as suspicious would have little operational value, even if its aggregate discrimination score were acceptable. By retaining confounder-adjusted residuals, CIDFE recognizes that size, workload, and procurement opportunity contain legitimate information that should influence interpretation without becoming direct proxies for fraud. This approach complements the broader move toward explainable fraud analytics (Giudici & Raffinetti, 2021; Soares et al., 2025). It also reinforces the principle that automated alerts should support rather than replace professional judgment. A Shapley explanation describes the model's reasoning; it does not establish the legal or factual guilt of a supplier, official, or agency.

Responsible deployment nevertheless requires caution. The current domain axes cover region, agency, procurement category, and time, but fairness risks may arise along dimensions that were not explicitly disentangled, including supplier size, ownership structure, locality, or the socioeconomic characteristics of regions. A low domain-probe score therefore should not be treated as a universal fairness certificate. Operational integration is another constraint. Government audit institutions often depend on relational databases, spreadsheets, and legacy case-management systems that are not designed to ingest dual-block latent features or real-time graph updates. CIDFE is more realistic as a scheduled analytical layer with cached features and human review than as an autonomous enforcement engine.

Finally, fraud is adaptive. Once detection rules become known, strategic actors may imitate legitimate styles or shift collusion into relationship patterns not represented in the training graph. Concept drift may also arise abruptly when procurement regulations, budget classifications, or disclosure rules change. The self-diagnosing invariance loop provides an early warning when domain predictability increases, but it cannot determine automatically whether the correct response is loss reweighting, domain expansion, model redesign, or retraining from scratch. The practical contribution of CIDFE is therefore not a claim of permanent invariance. It is a structured way to monitor and reduce known forms of spurious dependence while preserving auditor oversight and making the representation more transferable than conventional feature engineering.

CONCLUSIONS AND RECOMMENDATIONS

This study introduces CIDFE as a causally motivated feature-engineering architecture for fraud detection in government expenditure. The results show that explicitly separating fraud intent from administrative style improves generalization across regions, agencies, procurement categories, and fiscal years. CIDFE records the strongest average and worst-domain AUPRC across the reported comparisons, lowers the false-positive burden among legitimate high-value contracts, improves probability calibration, and produces auditor-facing explanations in which substantive red flags are more clearly separated from operational anomalies. The gains are not attributable to a single component: adversarial domain confusion, orthogonality, counterfactual consistency, and confounder-adjusted residuals each contribute to the final behavior of the model. For practice, public audit institutions should treat CIDFE-like systems as decision-support infrastructure rather than automated adjudication. Deployment should combine periodic graph reconstruction, explicit monitoring of domain-probe accuracy and calibration, human validation of high-risk alerts, documentation of counterfactual data generation, access controls, and fairness audits beyond the domains used during training. Where audit labels arrive slowly, retraining thresholds should reflect label latency rather than trigger automatically on short-term fluctuation. For model developers, the main recommendation is to evaluate fraud systems under genuine domain shifts, because random train-test splits can conceal dependence on administrative style. For public managers, the priority is to preserve data linkage quality and audit provenance so that relational and textual signals remain traceable during review.

FURTHER STUDY

Several extensions are necessary before the evidence can be generalized broadly. First, future work should evaluate CIDFE in additional public procurement systems with different legal frameworks, languages, and data architectures, including settings where confirmed fraud labels are sparse. Second, prospective evaluation is needed to determine whether gains persist when the model is deployed over time rather than tested retrospectively. Third, fairness assessment should be expanded to supplier size, ownership type, geographic disadvantage, and other protected or policy-relevant dimensions. Fourth, more rigorous counterfactual validation is needed because the boundary between style and substance is not always sharp; an attribute treated as stylistic in one procurement regime may carry substantive risk information in another. Finally, future studies should compare alternative disentanglement objectives, graph-update strategies, and explanation interfaces with auditor behavior, measuring not only predictive performance but also review time, override patterns, investigation yield, and user trust.

ACKNOWLEDGMENT

The authors acknowledge the contribution of the public expenditure, procurement, and audit-data sources described in the study and the broader research community whose methods informed the comparative evaluation.

REFERENCES

Deng, Z., Tian, H., Zheng, X., & Zeng, D. (2025). Deep causal learning: Representation, discovery and inference. ACM Computing Surveys.

Feng, L., Lin, A., & Luo, Z. (2026). MixingInsights: A framework for causal inference with confounder representation learning from mixed structured and textual data. Cognitive Computation.

Gara, M., Iezzi, S., & Siino, M. (2024). Corruption risk indicators in public procurement: A proposal using italian open data. Bahamasamlconference ….

Giudici, P., & Raffinetti, E. (2021). Shapley-lorenz eXplainable artificial intelligence. Expert Systems with Applications.

Gupta, D., Chugh, P., Sood, K., & Grima, S. (2026). … intelligence in financial fraud detection: A systematic review and FinTech-oriented ADO-TCCM meta-framework for trust, governance, and transparency. FinTech.

Li, B., Shen, Y., Wang, Y., Zhu, W., Li, D., Keutzer, K., et al. (2022). Invariant information bottleneck for domain generalization. Proceedings of the AAAI Conference on Artificial Intelligence.

Li, Y. (2024). Application of artificial intelligence in cross-departmental budget execution monitoring and deviation correction for enterprise management. Journal of Artificial Intelligence Review.

Man, H., & Nguyen, T. H. (2024). Counterfactual augmentation for robust authorship representation learning. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Odunaike, A. (2025). Designing adaptive compliance frameworks using time series fraud detection models for dynamic regulatory and risk management environments. Comput. Eng. Appl.

Salhan, S., Zhou, E., & Buttery, P. (2026). Do monolingual language models learn cross-lingual universal conceptual representations? Conference on Concept Representation Learning.

Sany, S. A., & Uddin, H. (2023). Machine learning-based fraud detection and conventional audit approaches in government deposit processing. American Journal of Interdisciplinary Studies.

Schiavon, J., Castro, M., & Carvalho, J. (2026). Ensemble-based anomaly detection in brazilian parliamentary expenses using unlabeled public data. International Journal of Data Science and Analytics.

Sha, Q., Tang, T., Du, X., Liu, J., Wang, Y., et al. (2025). Detecting credit card fraud via heterogeneous graph neural networks with graph attention. 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT).

Sharma, A., & Hassan, M. (2026). MissingMod-audit: Reliability-aware multimodal fraud detection under missing and corrupted financial evidence. International Journal of Science, Engineering and Information Technology.

Shovon, M. (2025). Application of machine learning and anomaly-detection algorithms for fraud detection in US government financial operations. American Journal of Interdisciplinary Studies.

Soares, H., Moura, R., Machado, V., Paiva, A., Lima, W., et al. (2025). The detection of spurious correlations in public bidding and contract descriptions using explainable artificial intelligence and unsupervised learning. Electronics.

Wang, J., Liu, J., Zheng, W., & Ge, Y. (2025). Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions. IEEE Access.

Yang, Y., Uy, M., & Huang, A. (2020). Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.

Zhao, C., Wan, S., Sheth, P., Patwa, K., et al. (2026). Causality guided representation learning for cross-style hate speech detection. Proceedings of the ACM Web Conference.

Zhao, T., Liu, G., Wang, D., Yu, W., et al. (2022). Learning from counterfactual links for link prediction. International Conference on Machine Learning.