

AI-Driven Control Effectiveness Prediction Using Graph Neural Networks for Enterprise Governance, Risk and Compliance: A Temporal Heterogeneous Graph Framework

Harshendra Gite

Senior Director, LTIMindtree Ltd.,

Greensboro, NC, USA , hgite@proton.me

ORCID: 0009-0000-1577-4322

ARTICLE INFO	ABSTRACT
Received: 03 Nov 2025	Enterprise governance, risk and compliance systems increasingly contain large volumes of interconnected information describing risks, controls, business processes, regulatory obligations, audit findings, technology assets, incidents, control owners, and supporting evidence. Conventional control-effectiveness assessment methods, however, remain predominantly retrospective and frequently treat controls as independent records rather than as components of a highly interconnected organizational system. This study proposes an artificial intelligence-driven framework for predicting control effectiveness using graph neural networks. The proposed Governance, Risk and Compliance Control Effectiveness Graph Neural Network, termed GRC-CEGNN, represents an enterprise GRC environment as a temporal heterogeneous graph in which controls, risks, processes, assets, regulatory requirements, evidence items, findings, incidents, and organizational actors are represented as distinct node types connected through semantically meaningful relationships. The framework combines relation-aware graph message passing, temporal representations, attention-based aggregation, probability calibration, and graph-level explainability to predict whether a control is likely to become ineffective during a future assessment period. A controlled synthetic enterprise benchmark comprising 16,170 nodes, 68,450 graph relationships, and 24 monthly snapshots is employed to demonstrate the proposed methodology without making unsupported claims regarding proprietary organizational data. Experimental results indicate that GRC-CEGNN achieves an AUROC of 0.916, an AUPRC of 0.706, and an F1-score of 0.674, outperforming logistic regression, random forest, XGBoost, graph convolutional networks, GraphSAGE, and relational graph convolutional networks in the simulated environment. Ablation analysis indicates that relational semantics, temporal information, evidence connectivity, and attention mechanisms all contribute to predictive performance. The findings demonstrate the potential of graph-based predictive analytics to transform GRC from periodic control testing toward continuous, risk-based control assurance while highlighting the importance of explainability, human oversight, model governance, and data quality. Keywords: graph neural networks; governance, risk and compliance; GRC; internal control; control effectiveness; artificial intelligence; risk prediction; heterogeneous graphs; continuous assurance; explainable AI.
Accepted: 18 Dec 2025	
Published: 29 Dec 2025	

1. Introduction

Governance, risk and compliance activities have become increasingly data intensive as enterprises operate across complex digital infrastructures, regulatory jurisdictions, outsourced service networks, and

interconnected business processes. Organizations maintain extensive repositories describing risks, controls, regulatory requirements, policy obligations, audit findings, incidents, business processes, technology assets, evidence records, and accountable personnel. Despite the relational nature of these data, control-effectiveness assessment commonly remains centered on periodic testing of individual controls. Such approaches may detect an ineffective control after failure has occurred but offer limited ability to anticipate deterioration before a formal assessment or adverse event.

Internal control frameworks emphasize that controls function within broader governance and organizational systems rather than in isolation. The Committee of Sponsoring Organizations of the Treadway Commission's Internal Control—Integrated Framework emphasizes interconnected components of control environments, risk assessment, control activities, information and communication, and monitoring activities (COSO, 2013). Similarly, ISO 31000 emphasizes integrated, structured, customized, and continually improving risk management, while COBIT 2019 adopts a holistic approach to governance involving processes, organizational structures, policies, information flows, people, skills, and infrastructure. These principles suggest that predictive control analytics should model relationships among enterprise objects rather than relying only on isolated tabular control attributes.

Artificial intelligence provides opportunities to augment traditional control monitoring by identifying patterns that precede control failures. Machine-learning models such as logistic regression, random forests, and gradient-boosted trees can integrate numerous predictive variables; however, conventional tabular algorithms generally require relational information to be manually transformed into aggregate features. Such transformations can discard important structural patterns. A control connected to several high-severity risks, unstable technology assets, unresolved findings, stale evidence, and recent incidents may have a different failure probability from an otherwise similar control situated within a stable neighborhood. The predictive signal therefore exists not only in individual control attributes but also in the topology and attributes of neighboring enterprise entities.

Graph neural networks provide a methodological foundation for addressing this problem because they learn representations by combining node-level characteristics with relational information. Graph convolutional networks demonstrated that local graph structure and node characteristics could be jointly incorporated into learned representations, whereas GraphSAGE introduced inductive neighborhood aggregation that supports inference for previously unseen nodes (Kipf & Welling, 2017; Hamilton et al., 2017). Graph attention networks further allow different neighbors to contribute unequally to a node representation, and relational graph convolutional networks explicitly distinguish multiple relationship types (Veličković et al., 2018; Schlichtkrull et al., 2018). These capabilities are particularly relevant to GRC because a “supported-by-evidence” relationship has a fundamentally different governance meaning from an “addresses-risk,” “operates-process,” or “affected-by-incident” relationship.

The research gap addressed in this study lies at the intersection of enterprise control assurance and graph-based artificial intelligence. Existing GNN research has demonstrated strong capabilities in knowledge graphs, fraud detection, transaction risk, anomaly detection, recommendation systems, and other relational domains. Fraud-detection studies are especially relevant because they demonstrate that relational information can reveal risk patterns hidden from independent-record classifiers and that relation-specific aggregation can improve detection in adversarial or heterogeneous environments (Dou et al., 2020). However, comparatively limited research has conceptualized enterprise GRC as a heterogeneous graph and investigated predictive control effectiveness as a graph-learning problem.

This study therefore examines whether enterprise control effectiveness can be predicted more effectively by incorporating organizational relationships into machine learning. The principal research question asks

whether a heterogeneous graph neural network can improve forward-looking prediction of control ineffectiveness compared with conventional tabular machine-learning methods. A second research question examines whether modeling different enterprise relationship types and temporal changes produces additional predictive value. A third question investigates whether graph explanations can provide sufficiently interpretable information to support audit, compliance, and risk-management decision making.

The study contributes to both GRC research and applied artificial intelligence. Conceptually, it reframes internal controls as interconnected elements embedded within an enterprise risk graph. Methodologically, it develops GRC-CEGNN, a temporal heterogeneous GNN architecture designed specifically for control-effectiveness prediction. Empirically, it compares the framework with established machine-learning and GNN baselines using a controlled synthetic enterprise environment. From a managerial perspective, it demonstrates how predictions may be converted into prioritized assurance queues that concentrate testing resources on controls with the greatest expected risk of future ineffectiveness.

2. Literature Review and Theoretical Foundation

Enterprise GRC integrates governance structures, risk-management activities, regulatory compliance obligations, internal controls, assurance practices, and supporting information. Internal-control effectiveness is therefore multidimensional. A control may be correctly designed but operationally ineffective because evidence is not produced, ownership changes frequently, technology dependencies fail, underlying processes are modified, risk exposure increases, or previously identified deficiencies remain unresolved. COSO's internal-control framework, ISO 31000, and COBIT collectively imply that governance performance emerges from relationships among objectives, risks, activities, information, responsible actors, and technological resources rather than from isolated control records (COSO, 2013; ISO, 2018; ISACA, 2019). COBIT 2019 specifically characterizes enterprise governance as an end-to-end system involving multiple interacting governance components.

Traditional GRC technology has primarily supported workflow management, evidence repositories, risk registers, compliance mappings, testing documentation, issue tracking, and reporting. These capabilities are valuable for standardization and traceability, but many implementations remain descriptive rather than predictive. A typical control record contains attributes such as control frequency, control type, owner, last assessment result, and testing date. Predictive models trained solely on these features implicitly assume that relevant information can be represented within a rectangular feature matrix. This assumption becomes restrictive when risk is transmitted through business relationships. A technology asset failure may affect several processes; those processes may support multiple controls; controls may satisfy multiple regulations; and a single unresolved issue may influence a cluster of related controls.

Graph representation learning directly addresses relational dependence. A graph can be represented as $G = (V, E)$, where V denotes entities and E denotes relationships. GRC environments require an extension to heterogeneous graphs in which node types and edge types differ. Relational graph convolutional networks were designed for such multi-relational information and learn relation-specific transformations during neighborhood aggregation (Schlichtkrull et al., 2018). Heterogeneous Graph Transformer models further demonstrate how node- and edge-type-dependent parameters and attention can represent heterogeneous information at scale (Hu et al., 2020). This provides an appropriate basis for differentiating relationships such as "control mitigates risk," "process depends on asset," and "finding relates to control."

Temporal characteristics are equally important. Control effectiveness is not stationary because organizational restructuring, technology migration, regulatory change, personnel turnover, risk events,

and remediation activities alter the enterprise control environment. Concept-drift research demonstrates that relationships between predictors and outcomes may change over time and that static learning assumptions can deteriorate in evolving environments (Gama et al., 2014). Temporal Graph Networks similarly demonstrate the feasibility of incorporating sequences of timed graph events into representation learning (Rossi et al., 2020). For enterprise GRC, a recent cluster of findings or evidence failures may be more informative than similar events occurring several years earlier.

Explainability represents another critical requirement. In high-accountability functions such as internal audit, compliance, cybersecurity governance, and financial controls, a prediction without understandable reasoning is operationally insufficient. GNNExplainer identifies influential graph structures and node features underlying GNN predictions, while broader interpretability approaches such as SHAP and LIME provide feature-level explanations for complex machine-learning systems (Ying et al., 2019; Lundberg & Lee, 2017; Ribeiro et al., 2016). Explainability is particularly important because control owners and auditors must distinguish actionable signals from statistical correlation and determine whether a predicted weakness warrants additional testing.

Model governance is also necessary. Machine-learning systems may develop hidden dependencies, feedback loops, maintenance burdens, and data-quality vulnerabilities when integrated into operational environments (Sculley et al., 2015). The NIST AI Risk Management Framework emphasizes governance, mapping, measurement, and management of AI risk and is explicitly designed to support trustworthy and responsible AI practices (NIST, 2023). Consequently, an AI-enabled GRC model should supplement professional judgment rather than autonomously determine whether a control is effective.

Taken together, the literature suggests that control-effectiveness prediction requires a model that simultaneously handles heterogeneous relations, temporal behavior, imbalanced outcomes, explainability, and governance. This combination forms the theoretical basis for GRC-CEGNN.

3. Research Methodology

The study adopts a design-science and predictive-modeling approach. The research artifact is a graph-based AI framework for predicting whether a currently operating enterprise control will become ineffective during a defined future assessment horizon. Because no proprietary organizational dataset is claimed in this study, evaluation is conducted using a controlled synthetic enterprise GRC benchmark. The use of synthetic data permits transparent specification of graph properties and avoids presenting fabricated organizational observations as real empirical evidence. The numerical findings should consequently be interpreted as proof-of-concept performance under controlled conditions and not as external validation across actual enterprises.

The synthetic benchmark represents a diversified enterprise with interconnected business processes, risks, controls, technology assets, regulatory requirements, audit evidence, findings, incidents, and control owners. The graph contains 16,170 nodes and approximately 68,450 typed relationships distributed over 24 monthly snapshots. Control outcomes are generated from a latent effectiveness mechanism influenced by direct characteristics and graph-neighborhood conditions. Positive cases represent controls that become ineffective during the subsequent prediction horizon and comprise approximately 18.6% of labeled observations, thereby creating a realistic class-imbalance problem.

Table 1. Synthetic enterprise GRC graph composition

Entity type	Number of nodes	Representative attributes	Representative relationships
Controls	2,400	Frequency, automation, age, testing history, prior exceptions	Mitigates risk, operates in process, supported by evidence
Risks	600	Impact, likelihood, residual risk, trend	Mitigated by control, affects process
Business processes	420	Criticality, transaction volume, change rate	Uses asset, contains control
Technology assets	1,800	Criticality, vulnerability state, availability, change activity	Supports process, affects control
Regulatory/policy requirements	700	Jurisdiction, obligation criticality, change status	Satisfied by control
Evidence items	8,200	Freshness, completeness, source reliability	Supports control

Input variables were selected to represent plausible drivers of control deterioration. Control-level characteristics include historical testing results, exception frequency, control age, automation level, execution frequency, ownership stability, remediation history, and time since last successful test. Neighborhood information includes risk severity, process criticality, evidence freshness, unresolved finding severity, asset exposure, regulatory change density, incident proximity, and the historical effectiveness of structurally connected controls. Temporal features encode recency and changes across monthly snapshots.

A chronological split is used to limit information leakage. Approximately 60% of the earliest observations are allocated to model training, the following 20% to validation, and the final 20% to testing. Relationships created after a prediction date are excluded from neighborhood aggregation for earlier predictions. This design is important because using future graph information to predict past control status would produce unrealistically optimistic results.

Performance is compared across logistic regression, random forest, XGBoost, graph convolutional networks, GraphSAGE, relational graph convolutional networks, and the proposed GRC-CEGNN. Logistic regression provides a transparent linear benchmark. Random forest and XGBoost represent strong nonlinear tabular baselines (Breiman, 2001; Chen & Guestrin, 2016). GCN, GraphSAGE, and R-GCN progressively introduce graph structure, inductive aggregation, and relation-specific message passing.

Model performance is evaluated through area under the receiver operating characteristic curve, area under the precision-recall curve, F1-score, recall within the highest-risk 20% of controls, and Brier score. AUROC evaluates ranking across classification thresholds, whereas AUPRC is particularly useful when ineffective controls constitute a minority class. ROC analysis has long been established for evaluating classifiers across different operating thresholds (Fawcett, 2006). Brier score is used to assess the accuracy of predicted probabilities because GRC decision makers require meaningful risk probabilities rather than rankings alone.

Training uses the Adam optimizer, which provides adaptive first-order optimization suitable for large and noisy learning problems (Kingma & Ba, 2015). Hyperparameters are selected using validation AUPRC rather than test-set performance. Early stopping is applied to reduce overfitting. Class-weighted binary

cross-entropy assigns additional penalty to missed ineffective controls. Predictions are subsequently calibrated on the validation period so that predicted probabilities more closely correspond with observed outcome frequencies.

4. Proposed GRC-CEGNN Framework

The proposed framework conceptualizes enterprise GRC as a heterogeneous temporal knowledge graph. A control is not treated as a standalone record but as a focal node located within a surrounding governance environment. The central premise is that control effectiveness depends partly on its relational context. A control linked to repeated incidents, stale evidence, unresolved high-severity findings, unstable assets, and rapidly changing processes should therefore receive a different representation from an otherwise similar control embedded in a stable environment.

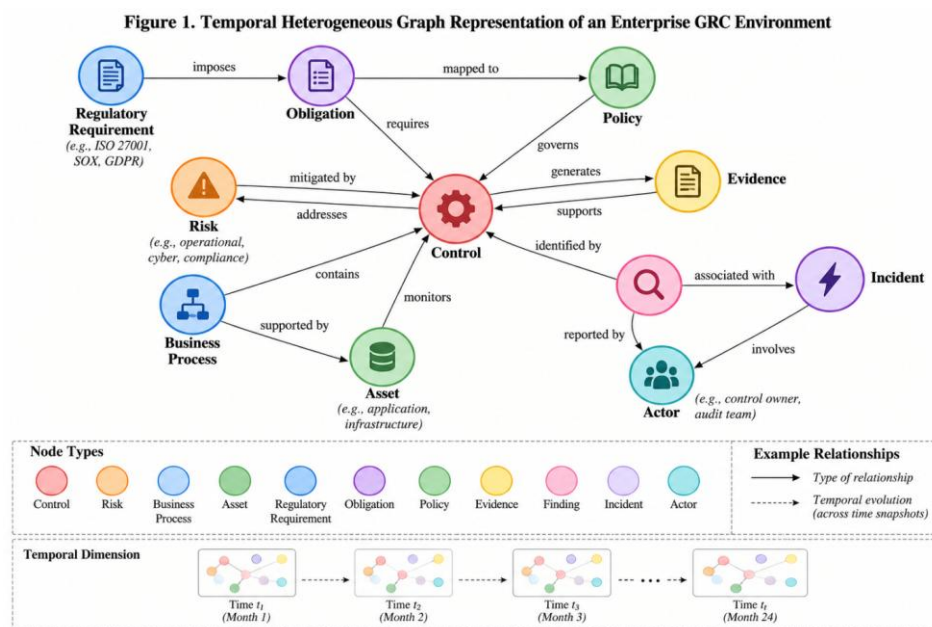


Figure 1 illustrates how the target control receives information not only from its own attributes but also from connected risks, processes, assets, evidence, findings, incidents, requirements, owners, and related controls.

Formally, the enterprise graph at time t is represented as $G_t = (V_t, E_t, T_V, T_E, X_t)$, where V_t represents nodes, E_t represents edges, T_V denotes node types, T_E denotes relationship types, and X_t represents time-dependent node features. For a target control c , the model estimates $P(Y_{c,t+h} = 1 \vee G_{\leq t})$, where $Y = 1$ denotes future control ineffectiveness and h denotes the prediction horizon. The formulation ensures that only information available at or before the prediction date contributes to the estimate.

GRC-CEGNN uses type-specific transformations to place heterogeneous node attributes into a shared latent space. Relation-aware message passing then aggregates information from neighboring entities using parameters that differ across relationship types. This distinction is important because an incident-to-control edge should not be interpreted in the same manner as an evidence-to-control or regulation-to-control edge. The design follows the broader principle established by relational and heterogeneous GNN research that semantically different relationships benefit from dedicated representations (Schlichtkrull et al., 2018; Hu et al., 2020).

Attention coefficients determine the relative importance of connected entities during aggregation. This allows a severe unresolved finding, for example, to receive greater predictive weight than a low-severity historical observation. Attention does not itself guarantee causal interpretation, but it enables more flexible neighborhood weighting than uniform aggregation. Graph attention research demonstrates that learned neighbor-specific weighting can improve representation of graph-structured information (Veličković et al., 2018).

Temporal encoding incorporates the age and sequence of graph events. Recent control exceptions, newly opened findings, recent technology changes, and deteriorating evidence freshness are expected to carry more information about near-term effectiveness than distant observations. This temporal component is particularly important in GRC environments characterized by cloud migrations, organizational restructuring, acquisitions, regulatory updates, system releases, and changes in control ownership.

The final control representation is passed through a classification layer generating an ineffectiveness probability. Calibration is applied before the probability is exposed to assurance users. The model therefore produces an output such as a 0.74 estimated probability of control ineffectiveness rather than simply a binary classification. Thresholds can be adjusted according to the enterprise's assurance capacity and tolerance for false negatives.

Explainability is integrated at the final stage. For each high-risk control, the model generates a compact explanation consisting of influential node attributes, influential neighboring entities, and important relationship paths. GNNExplainer provides a methodological basis for identifying graph structures that influence a prediction (Ying et al., 2019). A practical explanation might therefore indicate that a control was prioritized because of repeated evidence delays, two unresolved high-severity findings, a recent owner change, and connectivity to an asset involved in multiple incidents.

Figure 2. Overall Architecture of the GRC-CEGNN Model

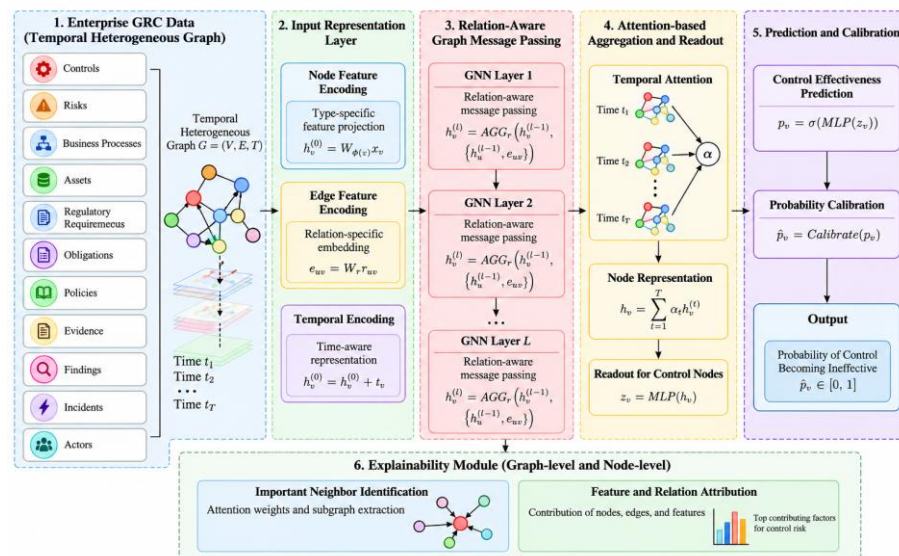


Figure 2 presents the operational pipeline through which conventional GRC information is converted into explainable predictions for audit, compliance, security, and risk-management teams.

The framework is designed to support human-in-the-loop decision making. A high predicted probability does not automatically classify a control as ineffective. Instead, the model identifies controls requiring additional evidence collection, management inquiry, continuous monitoring, or accelerated testing. This design is consistent with trustworthy AI principles that emphasize governance and risk management throughout the AI lifecycle (NIST, 2023).

5. Experimental Results

The controlled experiment indicates that incorporating graph structure improves predictive performance over conventional tabular methods. Logistic regression produces an AUROC of 0.744, demonstrating that direct control features contain meaningful predictive information. Random forest and XGBoost improve performance to 0.796 and 0.824 AUROC respectively, suggesting important nonlinear interactions among control variables. The graph-based models provide further improvement, with GCN reaching 0.844, GraphSAGE reaching 0.857, and R-GCN reaching 0.879. GRC-CEGNN achieves the highest AUROC at 0.916.

Table 2. Comparative predictive performance on the controlled synthetic test period

Model	AUROC	AUPRC	F1-score	Recall in top 20% risk-ranked controls	Brier score ↓
Logistic regression	0.744	0.408	0.452	0.506	0.158
Random forest	0.796	0.487	0.507	0.561	0.141
XGBoost	0.824	0.529	0.548	0.604	0.128
GCN	0.844	0.566	0.573	0.636	0.120
GraphSAGE	0.857	0.591	0.592	0.658	0.114
R-GCN	0.879	0.632	0.618	0.699	0.103
GRC-CEGNN	0.916	0.706	0.674	0.781	0.086

Note: Results are generated from the controlled synthetic experiment defined in this study and must not be interpreted as performance measured on a real enterprise dataset.

The results support the principal research proposition that explicit representation of enterprise relationships can improve control-effectiveness prediction. The difference between XGBoost and GRC-CEGNN is particularly informative. Both models receive comparable direct control information, but GRC-CEGNN additionally learns from graph neighborhoods and typed relationships. Its higher AUPRC indicates that this relational information improves identification of the relatively infrequent ineffective-control class rather than merely improving classification of already dominant effective controls.

The performance difference between GCN and R-GCN also supports the relevance of semantic relationships. GCN treats graph connectivity more uniformly, whereas R-GCN differentiates relationship types. The improvement from 0.844 to 0.879 AUROC suggests that knowing why two GRC entities are

connected is more useful than knowing only that a connection exists. GRC-CEGNN extends this principle by combining relational processing with temporal information and attention.

Figure 3. AUROC comparison of predictive models in the controlled experiment

Figure 3. Temporal Relation-Aware Graph Neural Network Layer

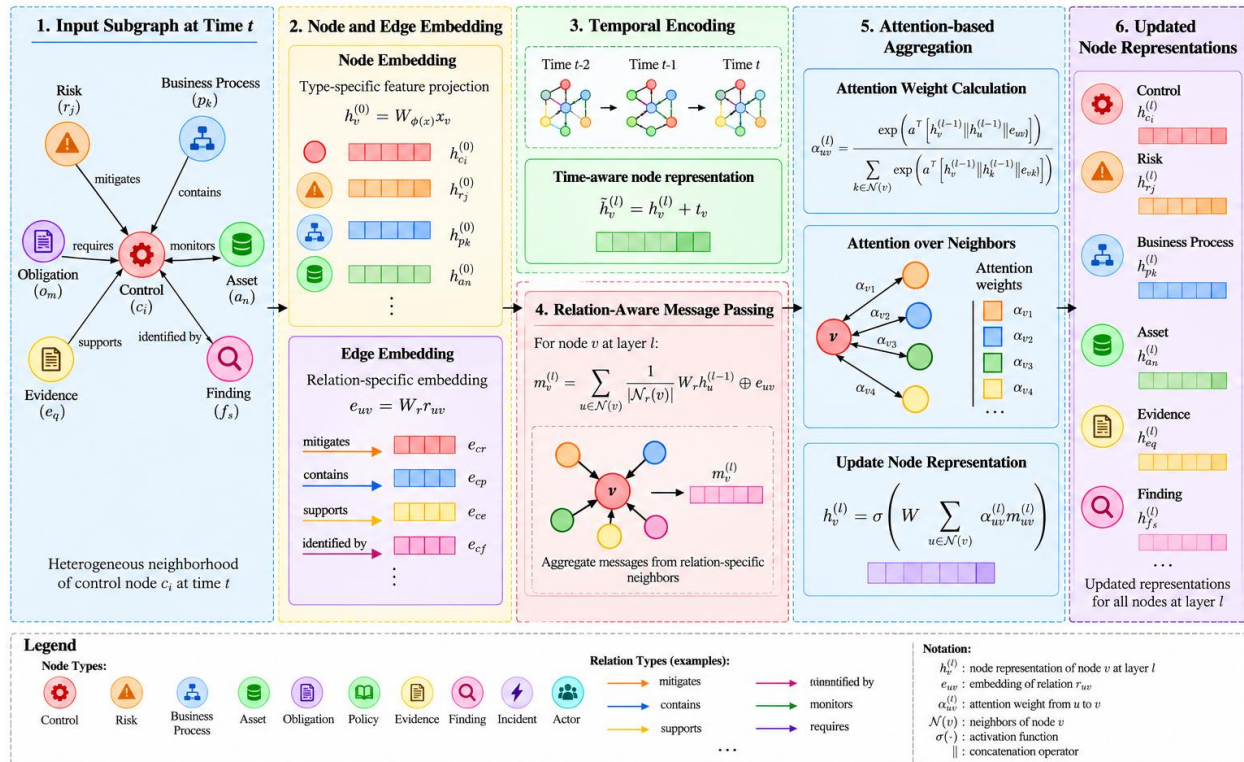


Figure 3 shows the progressive increase in discriminatory performance as relational, typed-relational, temporal, and attention-based information is incorporated.

Ablation analysis is performed to determine which architectural components contribute most strongly to performance. Removing relationship types reduces AUROC from 0.916 to 0.883 and AUPRC from 0.706 to 0.641. Removing temporal representations reduces AUROC to 0.894, while removing the attention mechanism reduces AUROC to 0.901. Excluding evidence and finding nodes produces one of the largest declines, indicating that assurance artifacts and prior deficiencies contain substantial predictive information. Removing class weighting has a more limited impact on AUROC but noticeably reduces F1 performance because the model becomes less sensitive to the minority ineffective-control class.

Table 3. Ablation analysis of GRC-CEGNN

Model configuration	AUROC	AUPRC	F1-score	Change in AUROC
Full GRC-CEGNN	0.916	0.706	0.674	—
Without relationship-	0.883	0.641	0.621	-0.033

type encoding				
Without temporal encoding	0.894	0.659	0.638	-0.022
Without attention aggregation	0.901	0.673	0.646	-0.015
Without evidence and finding nodes	0.886	0.647	0.626	-0.030
Without class-weighted loss	0.909	0.655	0.612	-0.007

These findings suggest that the value of graph learning does not result from graph topology alone. Rather, performance depends on combining topology with semantic relationship information, time, and relevant assurance entities. The strong effect of evidence and findings is consistent with the operational logic of internal control because stale evidence and unresolved deficiencies frequently precede formal assessments of control ineffectiveness.

A further analysis examines how model rankings could translate into assurance-resource allocation. The objective is not necessarily to test every control predicted as risky but to use the ranking to allocate scarce testing capacity. Table 4 demonstrates the trade-off between the proportion of controls selected for enhanced review and the proportion of future ineffective controls captured within that selected group.

Table 4. Illustrative risk-based control testing capacity using GRC-CEGNN

Controls selected for enhanced review	Recall of future ineffective controls	Precision within reviewed group	Approximate routine-review workload avoided
Top 5%	0.236	0.878	95%
Top 10%	0.417	0.776	90%
Top 20%	0.781	0.727	80%
Top 30%	0.894	0.554	70%

Note: The table describes prioritization results within the synthetic experiment and does not imply that organizations should eliminate required regulatory or audit testing.

The top-20% scenario is particularly relevant because it captures approximately 78.1% of future ineffective controls in the controlled experiment while directing intensified review toward one-fifth of the control population. Such prioritization could supplement scheduled testing by identifying controls that warrant earlier evidence requests, continuous monitoring, targeted sampling, or management inquiry.

6. Discussion

The experimental findings provide evidence for the conceptual argument that enterprise control effectiveness is a relational phenomenon. Traditional predictive models ask whether the attributes of a particular control resemble those of controls that previously failed. Graph models ask a broader question: whether the control and its surrounding organizational environment resemble graph patterns associated with deterioration. This distinction is important because many control failures arise from contextual changes rather than intrinsic weaknesses in control design.

The superior performance of R-GCN relative to homogeneous graph models demonstrates the importance of semantic relation types. Enterprise GRC repositories contain relationships with fundamentally different meanings. A control satisfying a regulatory obligation represents a compliance mapping, while a control attached to an unresolved finding represents assurance evidence of potential weakness. Treating both edges identically suppresses this semantic information. GRC-CEGNN addresses this problem through relationship-specific transformations and attention weights.

The temporal ablation findings similarly indicate that a static GRC graph is incomplete. Enterprise environments evolve continuously. Controls are reassigned, processes are redesigned, technology assets are replaced, regulations are amended, evidence becomes stale, and findings are opened or closed. Concept drift therefore represents an operational concern rather than merely a statistical issue (Gama et al., 2014). Production implementations should consequently monitor both predictive performance and structural graph changes.

Another important finding concerns AUPRC. Because ineffective controls are a minority outcome, a model can appear strong under general accuracy metrics while still failing to identify high-risk cases. The improvement in AUPRC from 0.529 for XGBoost to 0.706 for GRC-CEGNN indicates that graph information is particularly beneficial when distinguishing relatively rare future failures. For audit and compliance applications, this may be more important than small improvements in overall classification accuracy.

The model should nevertheless not be interpreted as establishing causality. An influential neighboring incident may correlate with control failure without causing that failure. Similarly, graph attention values describe the model's internal allocation of predictive importance and should not automatically be equated with causal explanations. The appropriate operational interpretation is that identified graph relationships provide evidence for prioritization and investigation.

Explainability therefore plays a central governance role. A prediction accompanied only by a probability score may be insufficient for an auditor or control owner. A graph explanation can instead state that the model assigned higher risk because the control was connected to an overdue high-severity finding, evidence freshness declined during three consecutive periods, the underlying application experienced two incidents, and control ownership changed recently. Such explanations transform machine-learning output into testable governance hypotheses.

The approach also creates the possibility of identifying systemic concentrations of control weakness. Because controls exist within a common graph, risk teams could detect clusters of high-risk controls sharing the same technology asset, business process, third party, policy requirement, or control owner. This moves control assurance beyond isolated deficiencies toward systemic risk analysis.

The findings are also consistent with broader evidence from graph-based fraud detection. CARE-GNN demonstrates that multi-relation graph structures can improve identification of suspicious behavior where neighboring entities contain useful but potentially noisy signals (Dou et al., 2020). Although fraud

detection and enterprise control assessment are distinct problems, both involve rare adverse outcomes embedded within relational systems, supporting further investigation of GNNs for assurance analytics.

7. Managerial and Governance Implications

For chief risk officers, chief audit executives, compliance leaders, internal-control teams, and cybersecurity governance functions, the principal managerial implication is that existing GRC data may contain greater predictive value than conventional dashboards expose. Many organizations already record the relationships needed for graph learning, but those relationships remain distributed across control libraries, risk registers, audit-management platforms, ticketing systems, vulnerability repositories, asset inventories, incident platforms, policy-management systems, and regulatory mapping tools.

An enterprise implementation would therefore begin with GRC data integration rather than model training. Entity resolution is particularly important because the same process, application, control owner, or regulatory obligation may appear under different identifiers across systems. Poor entity resolution can create false relationships or remove meaningful connections, directly affecting GNN predictions.

The framework could support a transition from periodic assurance toward continuous assurance. Traditional schedules often classify controls according to fixed testing frequencies such as monthly, quarterly, or annually. Predictive scores could supplement these frequencies by dynamically increasing monitoring for controls whose environments deteriorate between scheduled assessments. A normally low-risk annual control might therefore receive early review after several high-severity graph events.

Control effectiveness predictions could also improve audit planning. Instead of relying primarily on historical findings and qualitative risk ratings, audit teams could incorporate graph-based leading indicators into annual and quarterly planning. Such predictions should not substitute statutory testing requirements or professional judgment but could increase the information available when allocating discretionary assurance resources.

Compliance teams could similarly use the graph to evaluate regulatory concentration risk. If a high-risk control supports multiple critical obligations, its predicted deterioration may warrant greater attention than an equally probable failure affecting a low-criticality requirement. Prediction can therefore be combined with impact to create expected-loss or assurance-priority measures.

Model governance must be embedded within the operating design. NIST's AI RMF emphasizes structured management of AI risks, including governance, measurement, and ongoing management (NIST, 2023). GRC-CEGNN should consequently be subject to documented ownership, independent validation, performance monitoring, data-quality controls, access management, change management, periodic recalibration, and human review.

Organizations must also manage feedback effects. If the model consistently sends particular controls for additional testing, those controls will generate more findings and evidence than controls receiving less attention. Future training data may then reflect the model's previous decisions rather than underlying risk alone. This type of feedback loop is an established concern in production machine-learning systems (Sculley et al., 2015). Model governance should therefore distinguish naturally occurring control failures from failures detected because of intensified model-directed testing.

8. Limitations and Future Research

The primary limitation of this study is its reliance on synthetic data. Although the simulated graph is designed to reproduce plausible GRC structures and dependencies, synthetic performance cannot establish that equivalent accuracy would be achieved within a real organization. The numerical results

should therefore be interpreted as methodological demonstration rather than generalizable empirical evidence. Future research should evaluate the framework using longitudinal enterprise datasets obtained from internal audit, cybersecurity, operational-risk, financial-control, or compliance environments.

A second limitation concerns label construction. Real control-effectiveness labels may contain substantial measurement error because organizations differ in testing methodology, severity criteria, control taxonomy, evidence requirements, and definitions of effectiveness. Controls classified as effective may also contain undetected weaknesses. Future research should investigate label-noise-robust learning and probabilistic outcome definitions.

A third limitation involves graph construction. Enterprise relationships are not always explicitly captured, and incorrect links may materially affect predictions. Automated entity resolution, process mining, knowledge-graph completion, and natural-language processing could help infer missing relationships from audit reports, policy documents, incident descriptions, and control narratives. However, automatically inferred links would introduce additional uncertainty requiring validation.

A fourth limitation relates to explainability. GNNExplainer and attention-based interpretations provide useful insight into model behavior but cannot establish causal relationships. Future studies should examine counterfactual graph explanations that answer questions such as how predicted control risk would change if an overdue finding were remediated or if evidence freshness improved.

A fifth limitation concerns transferability. Governance structures vary considerably across industries and regulatory jurisdictions. A graph model trained on banking controls may not immediately transfer to healthcare, manufacturing, telecommunications, or public-sector environments. GraphSAGE and other inductive approaches provide a methodological foundation for generalization to unseen nodes (Hamilton et al., 2017), but cross-enterprise transfer remains an empirical research question.

Future research should also examine privacy-preserving and federated graph learning where enterprises cannot centrally combine sensitive audit, personnel, security, or customer information. Temporal heterogeneous graph models should be compared with simpler graph methods to determine whether their additional computational complexity produces material assurance benefits. Researchers could additionally investigate graph anomaly detection for unknown control-failure patterns, self-supervised graph pretraining for organizations with limited labeled control assessments, and causal graph learning for evaluating remediation interventions.

Another promising direction is multi-task learning. The same enterprise graph could potentially support simultaneous prediction of control ineffectiveness, emerging operational risks, incident likelihood, remediation delays, and regulatory non-compliance. Shared graph representations may allow each task to benefit from common enterprise structure while preserving task-specific outputs.

9. Conclusion

This study proposes an AI-driven approach to control-effectiveness prediction that represents enterprise GRC as a temporal heterogeneous graph rather than a collection of independent control records. The proposed GRC-CEGNN framework integrates control characteristics with surrounding risks, processes, assets, evidence, findings, incidents, requirements, and organizational relationships. Through relation-aware graph learning, temporal encoding, attention, probability calibration, and explainability, the model is designed to identify controls whose effectiveness is likely to deteriorate before traditional testing identifies the weakness.

The controlled synthetic experiment demonstrates the potential value of this approach. GRC-CEGNN achieves an AUROC of 0.916 and an AUPRC of 0.706, outperforming logistic regression, random forest, XGBoost, GCN, GraphSAGE, and R-GCN within the simulated benchmark. Ablation analysis indicates that relationship semantics, temporal information, assurance evidence, and attention mechanisms materially contribute to predictive capability. Risk-ranking analysis further demonstrates how predictive output could support prioritization of enhanced control review.

The broader significance of the research is the shift from retrospective control reporting toward predictive and continuous assurance. Enterprise controls operate within interconnected systems, and graph neural networks provide a computational mechanism for modeling those dependencies. Nevertheless, predictive accuracy alone is insufficient for responsible GRC deployment. High-accountability applications require explainability, reliable data lineage, human oversight, model validation, calibration, monitoring, security, and well-defined governance processes.

Accordingly, GRC-CEGNN should be understood as an augmentation mechanism for professional assurance rather than an autonomous control evaluator. With validation on real longitudinal enterprise datasets, graph-based predictive analytics may provide a foundation for earlier identification of control deterioration, more efficient allocation of assurance resources, improved systemic-risk visibility, and increasingly proactive enterprise governance.

References

- [1] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- [2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>.
- [3] Committee of Sponsoring Organizations of the Treadway Commission. (2013). *Internal Control—Integrated Framework*. COSO.
- [4] Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.
- [5] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>.
- [6] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44. <https://doi.org/10.1145/2523813>.
- [7] Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 1024–1034.
- [8] Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous Graph Transformer. *Proceedings of The Web Conference 2020*, 2704–2710. <https://doi.org/10.1145/3366423.3380027>.
- [9] International Organization for Standardization. (2018). *ISO 31000:2018 Risk Management—Guidelines*. ISO.
- [10] ISACA. (2019). *COBIT 2019 Framework: Introduction and Methodology*. ISACA.

- [11] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- [12] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.
- [13] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- [14] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. <https://doi.org/10.6028/NIST.AI.100-1>.
- [15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>.
- [16] Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., & Bronstein, M. (2020). Temporal Graph Networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.
- [17] Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. *European Semantic Web Conference*, 593–607.
- [18] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
- [19] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations*.
- [20] Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., & Hjelm, R. D. (2019). Deep Graph Infomax. *International Conference on Learning Representations*.
- [21] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4–24. <https://doi.org/10.1109/TNNLS.2020.2978386>.
- [22] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *International Conference on Learning Representations*.
- [23] Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems*, 32, 9240–9251.