International Scientific Technical and Economic Research **(ISSN: 2959-1309)**
Vol. 4, No. 1, (March 2026), pp. 47-69
Journal homepage: https://www.istaer.online

RESEARCH ARTICLE **DOI: 10.71451/ISTAER2603**

Cross-Border Trade Fraud Detection via Integrated Heterogeneous Graph Neural Network and XGBoost

Xi Zeng*

Sogang Business School, Sogang University, Seoul, Republic of Korea

Abstract: Because cross-border trade fraud involves multiple types of entities, multiple business relationships and complex interactive structures, it exhibits high heterogeneity and strong concealment, which has brought significant challenges to the traditional risk identification methods. Aiming at the problem that existing methods struggle to balance the ability of structural modeling and classification performance, this paper proposes a cross-border trade fraud detection framework based on heterogeneous graph neural network (HGNN) and gradient lifting tree model XGBoost. Firstly, the cross-border trade system is modeled as a heterogeneous graph of multi type entities and multi relationship interactions, and HGNN is used to learn the high-order structural semantic representation of entities in complex trade networks; Then, the graph embedding features and statistical features are input into XGBoost to achieve high-precision classification of fraud. The experimental results on the real cross-border trade data set show that the AUC of the proposed model on the test set reaches 0.966, which is 18.7% and 3.4% higher than using XGBoost and HGNN alone, and significantly improves the recall rate of fraud samples in a variety of typical fraud scenarios. Ablation experiments further verified the key role of heterogeneous relationship modeling, attention mechanism and integration strategy in performance improvement. The above results show that HGNN–XGBoost integration framework has good detection performance and engineering application potential in complex heterogeneous scenes.

Keywords: Cross-border trade fraud detection; Heterogeneous graph neural network; Integrated learning; XGBoost; Risk identification

How to Cite: Zeng, X. (2026). Cross-Border Trade Fraud Detection via Integrated Heterogeneous Graph Neural Network and XGBoost. *International Scientific Technical and Economic Research, 4*(1), 47–69. **https://doi.org/10.71451/ISTAER2603**

Article history: Received: 30 Nov 2025; **Revised:** Dec 18 2025; **Accepted:** 03 Jan 2025; **Published:** 19 Jan 2026

1. INTRODUCTION

With the continuous deepening of the global trading system and the continuous advancement of the digital process, cross-border trade activities not only promote economic development, but also provide a more hidden and complex breeding environment for all kinds of fraud. Cross-border trade fraud often involves multinational entities, a variety of business processes and a variety of transaction media [1],[2]. Its behavior characteristics are no longer

[*] **Corresponding author:** Xi Zeng, Sogang Business School, Sogang University, Seoul, Republic of Korea. Email: 13587820138@163.com

limited to a single transaction exception, but form a complex risk network through multiple relationships such as account Association, logistics sharing, agency intermediary and geographical risk [3]. This highly heterogeneous and highly correlated business feature makes it difficult for the traditional risk identification methods that rely on the assumption of independent samples to effectively capture potential fraud patterns, which seriously restricts the identification accuracy and response ability of the cross-border trade risk control system.

The cross-border trade system naturally has the characteristics of coexistence of multiple types of entities and relationships. Different entities play a differentiated role in the business chain, and their risks are not directly determined by a single attribute, but hidden in the complex interaction structure [4]. At the same time, cross-border fraud is usually characterized by low incidence, high concealment and strong antagonism, resulting in significant category imbalance and noise interference in real data [5]. These factors together constitute multiple challenges for cross-border trade fraud detection at the levels of structural modeling, feature expression and classification decision-making, making the problem far more complex than traditional financial fraud detection.

In recent years, machine learning and deep learning methods have been widely used in the field of fraud detection, but the existing methods still have obvious limitations in dealing with cross-border trade scenarios. On the one hand, although the method based on statistical features or tree model is robust in nonlinear classification and unbalanced data processing, it is difficult to effectively use the complex relationship structure between entities; On the other hand, although graph neural network method can model structure dependence, its classification ability and training stability are still limited in the face of strong class imbalance, weak supervision signal and engineering scale large-scale data [6],[7],[8],[9]. The imbalance between the ability of single model paradigm in structure expression and risk classification has become a key bottleneck restricting the improvement of cross-border trade fraud detection performance.

Based on the above background, this paper proposes the organic combination of HGNN and gradient lifting tree model XGBoost, in order to leverage the complementary strengths of the complementary advantages of the two types of models in structural modeling and discriminant decision-making [10],[11]. Specifically, HGNN is used to learn high-order structural semantic representation from multi type entities and multi relationship interactions, so as to characterize potential fraud patterns in cross-border trade networks; XGBoost makes use of its advantages in feature selection, nonlinear combination and unbalanced classification to make collaborative classification between graph embedding and statistical features [12],[13],[14],[15]. Through this integrated design with clear division of labor and complementary advantages, the model cannot only capture complex structural dependencies, but also maintain stable and efficient classification performance in actual risk decision-making.

Around the above research ideas, the main contributions of this paper are reflected in the following aspects: first, starting from the characteristics of cross-border trade business, a unified heterogeneous graph modeling framework is constructed, and the multi-agent and multi relationship trade behavior is incorporated into the structured representation; Secondly, a fraud detection oriented HGNN representation learning method is designed to enhance the ability of the model to depict high-order structure semantics and abnormal patterns; Thirdly, a graph embedding driven HGNN–XGBoost integrated learning mechanism is proposed to realize the effective collaboration between structure representation and strong discriminant model; Finally, through systematic experiments and ablation analysis, the significant advantages of the proposed method in detection performance, stability and generalization ability are verified on real cross-border trade data. The above work provides a solution with theoretical value and practical feasibility for cross-border trade fraud detection in complex heterogeneous scenarios.

2. PROBLEM DEFINITION AND HETEROGENEOUS GRAPH MODELING

The core goal of cross-border trade fraud detection is to accurately identify potential high-risk entities or transactions in a complex and multi-agent trade network. Different from the traditional financial fraud scenario, cross-border trade involves many entities, such as enterprises, accounts, logistics, countries and intermediary service agencies [16],[17]. The fraud behavior is often not directly reflected by the abnormal attributes of a single node, but hidden in the structural mode formed by the interaction of multiple entities. Therefore, it is necessary to make a strict formal definition of the problem, and build a heterogeneous graph representation that can describe the interaction between multi type entities and multi relationships.

From the perspective of supervised learning, cross-border trade fraud detection can be defined as a binary problem. Suppose there are N target entities (such as enterprises or accounts) in the dataset, and each entity corresponds to a label $y_i \in \{0,1\}$, where $y_i = 1$ indicates that the entity is at risk of fraud, and $y_i = 0$ indicates that it is normal. The goal of the model is to learn a mapping function [18]:

$$f: \mathcal{X} \rightarrow [0,1], \tag{1}$$

The prediction probability $\hat{y}_i = f(x_i)$ of its output can be as close as possible to the real label y_i. Among them, x_i not only contains the attribute characteristics of the entity itself, but also implies its structural context information in the cross-border trade network. The definition emphasizes that the essence of fraud detection task is not single point classification, but conditional probability estimation based on complex relationship structure.

In order to fully describe the multi-agent interaction in the cross-border trade system, this paper models the original business data as a heterogeneous graph, which is defined as [19],[20]:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \psi), \tag{2}$$

Where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the edge set, $\phi: \mathcal{V} \rightarrow \mathcal{T}_v$ is the node type mapping function, $\psi: \mathcal{E} \rightarrow \mathcal{T}_e$ is the relationship type mapping function. The node type set $\mathcal{T}_v$ can include enterprises, accounts, logistics units, countries, etc., while the relationship type set $\mathcal{T}_e$ corresponds to transaction relationships, capital flow relationships, logistics sharing relationships, geographical relationships, etc.

In this heterogeneous graph, each node $v_i \in \mathcal{V}$ is associated with an attribute vector $\mathrm{x}_i \in \mathbb{R}^d$, which is used to describe the static or statistical characteristics of the entity; Each edge $(v_i, v_j) \in \mathcal{E}$ characterizes the interaction behavior between entities under specific relational semantics. Through this modeling method, different types of entities and their relationships in cross-border trade are unified into the same graph structure, providing structured input for subsequent graph based representation learning.

It should be noted that fraud tags are usually directly associated with some core entities (such as enterprise nodes), while most of the other nodes are in unlabeled or weak tag status. Therefore, cross-border trade fraud detection is essentially a semi supervised node classification problem on heterogeneous graphs at the graph modeling level. The model needs to use the information of labeled nodes to infer unlabeled or potentially high-risk nodes through structural propagation and relationship modeling.

At the label modeling level, cross-border trade fraud data generally show serious category imbalance. Let the number of positive (fraud) samples be N_1, and the number of negative (normal) samples be N_0. There are usually $N_1 \ll N_0$. The available scale factor for this imbalance

$$\rho = \frac{N_1}{N_0 + N_1} \tag{3}$$

The value is often far less than 0.2, or even less than 0.1. This feature makes the model prone to bias to most classes in the training process, resulting in insufficient recognition ability of fraud samples. Therefore, in the subsequent model design and training process, the imbalance of label distribution must be explicitly considered to avoid the high risk of fraud and missed detection in the case of high overall accuracy of the model.

3. OVERALL METHODOLOGY FRAMEWORK

This study proposes a HGNN–XGBoost integrated learning framework for cross-border trade fraud detection, which aims to fully integrate the complex heterogeneous relationship structure information in the cross-border trade scene and the gradient lifting model with strong classification ability. The overall framework follows the design idea of "structure representation learning → representation enhancement → classification integration". The HGNN is used to model the high-order structure of the multi-agent and multi relationship trade network, and then the learned graph embedded representation and traditional statistical features are input into XGBoost to achieve accurate recognition of complex fraud patterns.

Formally, the cross-border trade system is modeled as a heterogeneous graph:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T}_v, \mathcal{T}_e) \tag{4}$$

Where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the edge set, $\mathcal{T}_v$ and $\mathcal{T}_e$ are the node type and relationship type mapping functions respectively. The goal of HGNN is to learn the low dimensional embedded representation of nodes on this heterogeneous structure:

$$\mathrm{Z} = f_{\mathrm{HGNN}}(\mathcal{G}, \mathrm{X}) \tag{5}$$

Where X is the original attribute characteristic matrix, and $\mathrm{Z} \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the node representation of structure perception.

In the overall architecture, HGNN does not directly output the final fraud detection results, but acts as a structure representation learning module to provide high-quality, low redundancy and fraud semantic graph embedding features for the subsequent XGBoost detection model. This decoupling design effectively avoids the degradation of the classification performance of the graph model in the scenarios of strong class imbalance and noise labels.

In order to clearly show the module composition and function division of the overall model, table 1 gives the core modules of HGNN–XGBoost framework and their input and output definitions.

Table 1. Module division and function description of HGNN–XGBoost integrated model

Module number	Module name	Main input	Main output	Function description
M1	Data preprocessing module	Raw trade data	Standardization features	Cleaning, coding and normalization
M2	Heterogeneous graph building module	Entities and transactions	Heterogeneous graph $\mathcal{G}$	Building multi type nodes and relationships
M3	HGNN stands for learning module	$\mathcal{G}$, X	Graph embedding Z	Semantic representation of learning structure
M4	Feature fusion module	Z, X_{stat}	Enhanced feature H	Splicing and aligning multi-source features
M5	XGBoost classification module	H	Fraud probability	Gradient lifting classification

In the design of information flow, the model adopts a unidirectional progressive and phased decoupling structure. The original cross-border trade data is first mapped into a heterogeneous graph structure and input into HGNN. The node embedded representation is obtained through the multi-layer relationship aware messaging mechanism. Then, the graph embedding Z and the statistical feature vector X_{stat} are spliced in the feature space to form the enhanced feature for final classification [21]:

$$H_i = \left[z_i \parallel x_i^{stat} \right] \tag{6}$$

Where ∥ indicates vector splicing operation. This design makes XGBoost use the structural pattern and numerical distribution characteristics at the same time, so as to improve the recognition ability of covert and combined fraud [22].

To verify the stability of the information flow design under different feature compositions, table 2 shows the impact of different feature combinations on the model input dimension and information coverage.

Table 2. composition of input features under different feature fusion strategies

Policy number	Graph embedding dimension	Statistical feature dimension	Total dimension	Information type override
S1	64	32	96	Structure+basic attribute
S2	128	32	160	Structure+attribute
S3	128	64	192	Structure+attribute+timing
S4	256	64	320	High order structure+multimode
S5	256	128	384	Total feature fusion

In the training and reasoning process, the model adopts the end-to-end but non joint optimization strategy. Specifically, HGNN first independently optimizes its representation learning objectives on the training map, and its loss function is defined as [23],[24]:

$$\mathcal{L}_{\text{HGNN}} = \sum_{i \in \mathcal{V}_l} \ell(\mathbf{z}_i, y_i) \tag{7}$$

Where $\mathcal{V}_l$ is the set of labeled nodes. After the HGNN training, its parameters are frozen and only participate in the subsequent stages as a feature extractor. Then, XGBoost minimizes the objective function of the additive tree model with the enhanced feature H as the input:

$$\mathcal{L}_{\text{XGB}} = \sum_{i} \ell(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{8}$$

The phased training strategy not only ensures the quality of representation, but also significantly improves the training stability and engineering scalability of the overall model under large-scale cross-border trade data. In the reasoning stage, the model can complete the fraud risk assessment only by one time of forward graph embedded calculation and one time of tree model prediction, meeting the requirements of real-time and deployability in the actual business scenario.

4. DESIGN OF HETEROGENEOUS GRAPH NEURAL NETWORK FOR CROSS-BORDER TRADE

Cross-border trade fraud is often hidden in the complex network structure formed by multiple subjects and multiple relationships. Its key characteristics are not only determined by the attributes of a single entity, but also shaped by the interaction mode, relationship combination and path structure between entities. In order to effectively depict such complex semantics, this study designed a relationship aware HGNN structure for customization of cross-border trade scenarios, and achieved fine modeling of high-order structure semantics by introducing the message passing mechanism of relationship type constraints and heterogeneous attention aggregation strategy.

In this HGNN, different types of edges are regarded as information channels carrying differentiated semantics. Let the type of node v_i be $\tau(v_i)$ and the relationship type of edge (v_j, v_i) be $r \in \mathcal{R}$, then in the layer l network, the messages received by nodev_i from its neighbors are defined as:

$$\mathrm{m}_{i,r}^{(l)} = \sum_{j \in \mathcal{N}_r(i)} \mathrm{W}_r^{(l)} \mathrm{h}_j^{(l)} \tag{9}$$

Where $\mathcal{N}_r(i)$ represents the set of neighbor nodes connected with v_i through the relationship r, and $\mathrm{W}_r^{(l)}$ is the relationship specific learnable transformation matrix. The design avoids semantic confusion of different trade relationships (such as "transaction", "guarantee" and "common account") in the process of feature propagation, and enables the model to clearly distinguish the roles of different relationships in fraud propagation.

In the cross-border trade network, there are significant differences in the importance of different relationship types to fraud identification. In order to quantitatively analyze the structural characteristics of various relationships in heterogeneous graphs, table 3 shows the statistical distribution of the main relationship types in the experimental data set.

Table 3. Structural statistical characteristics of different relationship types in cross-border trade heterogeneous diagram

Relationship type	Relationship meaning	Side quantity (10000)	Average degree	Proportion associated with fraud node
R1	Enterprise–enterprise transaction	182.4	7.6	21.3%
R2	Enterprise–account association	96.7	4.1	34.8%
R3	Enterprise–logistics sharing	74.2	3.5	18.6%
R4	Enterprise–customs broker	58.9	2.9	27.1%
R5	Enterprise–high-risk country	41.3	1.8	39.5%

It can be seen from table 3 that there are significant differences in the number, scale and fraud relevance of different relationships, which further verifies the necessity of introducing the relationship awareness mechanism in the messaging phase.

After completing the relationship level message aggregation, the model weights the information from different relationship sources through heterogeneous attention mechanism to realize the adaptive selection of key structure semantics. Specifically, the update of layer l node v_i is expressed as:

$$\mathrm{h}_i^{(l+1)} = \sigma\left(\sum\nolimits_{r \in \mathcal{R}} \alpha_{i,r}^{(l)} \cdot \mathrm{m}_{i,r}^{(l)}\right) \tag{10}$$

Where $\sigma(\cdot)$ is the nonlinear activation function, $\alpha_{i,r}^{(l)}$ is the attention weight of node v_i to relationship r, and its calculation method is defined as:

$$\alpha_{i,r}^{(l)} = \frac{\exp\left(\mathrm{a}^{\top}\left[\mathrm{h}_i^{(l)} \parallel \mathrm{m}_{i,r}^{(l)}\right]\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\mathrm{a}^{\top}[\mathrm{h}_i^{(l)} \parallel \mathrm{m}_{i,r'}^{(l)}]\right)} \tag{11}$$

This mechanism enables the model to dynamically adjust the influence weights of various trade relations at different nodes and levels, so as to highlight the structural model that has more discriminative power against fraud.

To further analyze the learning effect of heterogeneous attention mechanism on relationship weight distribution, table 4 shows the average attention weights of different relationship types on the fraud node after the training.

Table 4. Average attention weight distribution of fraud nodes in different relationship types

Relationship type	Average attention weight	standard deviation	Weight ranking
Enterprise–high risk country	0.312	0.041	1
Enterprise–account Association	0.247	0.036	2
Enterprise–customs broker	0.193	0.029	3
Enterprise–enterprise transaction	0.156	0.022	4
Enterprise–logistics sharing	0.092	0.018	5

The results show that the model can automatically give higher weight to the relationship highly related to fraud, and the weight distribution has good stability, which shows that heterogeneous attention mechanism has clear semantic selection ability in cross-border trade scenarios.

In terms of high-order structural semantic modeling, this study makes the node representation gradually integrate the information from different hops' neighborhood by stacking multi-layer HGNN, so as to capture the common hidden patterns such as "multi hop collaboration" and "indirect association" in cross-border fraud. The nodes of the L-th layer output represent:

$$\mathrm{z}_i = \mathrm{h}_i^{(L)} \tag{12}$$

It not only contains the attribute information of the node itself, but also implicitly encodes its high-order structural role in the heterogeneous trade network. It is observed in the experiment that when $L = 3$, the model achieves the best balance between performance and over smoothing risk, indicating that high-order semantics has significant gain for fraud detection, but the propagation depth needs to be reasonably controlled.

Through the above collaborative design of relationship aware messaging, heterogeneous attention aggregation and multi-layer structure semantic modeling, the proposed HGNN can provide a representation basis with high fraud sensitivity for subsequent classification models, and lay a key structural support for the effectiveness of the overall HGNN–XGBoost integration framework.

5. IMPROVEMENT OF HGNN REPRESENTATION LEARNING ALGORITHM

Although the heterogeneous graph neural network can effectively capture the Structural Semantics in the cross-border trade network, in practical applications, it is often difficult to make full use of the high-dimensional attribute information carried by the entity itself, and it is vulnerable to the interference of sparse relationships and abnormal connections. To this end, this study introduces several algorithm improvements in the representation learning phase of HGNN to enhance the discriminant and stability of node representation from three aspects: Joint embedding, robust modeling and fraud sensitive constraints.

First, in the joint embedding of structure and attribute, the model no longer simply takes the node attributes as the initial input, but constructs the final embedding by explicitly modeling the complementary relationship between structure representation and attribute representation. Let the structure of node v_i be expressed as z_i^s, and the attribute code be expressed as z_i^a, then its joint embedding is defined as:

$$\mathrm{z}_i = \phi(\mathrm{W}_s\mathrm{z}_i^s + \mathrm{W}_a\mathrm{z}_i^a) \tag{13}$$

Where W_s and W_a are learnable parameter matrices, and $\phi(\cdot)$ is a nonlinear mapping function. This fusion method realizes the alignment of Structural Semantics and numerical attributes in the feature space, and avoids the dimension expansion and semantic mismatch problems caused by simple splicing.

In order to evaluate the impact of joint embedding strategy on representation quality, table 5 shows the separability statistical results of node representation on fraud and non fraud samples under different embedding methods.

Table 5. Separability analysis of node embedding under different representation learning strategies

Representation	Embedded dimension	Intra class distance	Distance between classes	Inter class/intra class ratio
Attribute embedding only	128	0.842	1.116	1.33
Structure embedding only	128	0.793	1.204	1.52
Splicing fusion	256	0.765	1.387	1.81
Weighted joint embedding	128	0.621	1.412	2.27
Constrained joint embedding	128	0.598	1.439	2.41

The results show that the joint embedding strategy significantly improves the separability of fraud samples and normal samples in the embedding space while keeping the embedding dimension controllable.

Secondly, aiming at the problem of sparse relationship and abnormal structure that are common in cross-border trade networks, this study introduces a robust modeling mechanism in the HGNN representation learning process. Specifically, the structural representation of nodes with extremely low degree or abnormal height is often unstable and vulnerable to noise relations. To alleviate this problem, the model introduces the structure confidence weight γ_i in the message aggregation stage to dynamically adjust the proportion of structure information in the final representation:

$$\mathrm{z}_i = \gamma_i\mathrm{z}_i^s + (1-\gamma_i)\mathrm{z}_i^a \tag{14}$$

Where γ_i is calculated adaptively according to the number of effective neighbors and relationship consistency of nodes. This mechanism makes the model more dependent on attribute information when facing sparse connected nodes, and strengthens the relational semantics when the structural information is sufficient, so as to improve the robustness of the overall representation.

Table 6 shows the changes of model representation stability before and after the introduction of robust modeling mechanism under different node sparsity.

Table 6. Stability comparison of representation learning in sparse relation scenarios

Node average degree	No robust modeling variance	Robust modeling variance	Stability improvement
≤2	0.184	0.097	+47.3%
3–5	0.142	0.081	+43.0%
6–10	0.108	0.067	+38.0%
11–20	0.083	0.056	+32.5%
>20	0.069	0.051	+26.1%

The robust modeling strategy shows particularly significant stability improvement in low degree nodes and sparse relationship scenarios, which is of great significance for a large number of small and medium-sized enterprise nodes in cross-border trade.

Finally, in order to further enhance the sensitivity of representation to fraud, this study introduces a fraud sensitive representation constraint strategy in the representation learning phase. The strategy enlarges the distance between the fraud node and the normal node in the embedded space, and compresses the distribution range of similar nodes. The constraint loss is defined as:

$$\mathcal{L}_{\text{fs}} = \sum_{(i,j)} \mathbb{I}(y_i = y_j) \parallel \mathrm{z}_i - \mathrm{z}_j \parallel^2 - \sum_{(i,k)} \mathbb{I}(y_i \neq y_k) \max\,(0, m - \parallel \mathrm{z}_i - \mathrm{z}_k \parallel)^2 \tag{15}$$

Where m is the interval superparameter. This constraint is added to the overall optimization goal of HGNN as a regular term, so that the model explicitly considers the fraud classification requirements while learning the structural semantics.

Combined with the above algorithm improvements, the proposed HGNN representation learning method has significantly enhanced the structure expression ability, noise robustness and fraud classification, and laid a high-quality representation foundation for the subsequent high-precision classification based on XGBoost.

6. INTEGRATED LEARNING MECHANISM OF HGNN AND XGBOOST

In the task of cross-border trade fraud detection, a single model is often difficult to take into account the expression ability and strong classification performance of complex structural patterns. Although HGNN has significant advantages in structural semantic modeling, its classification ability under highly unbalanced label distribution and heterogeneous noise characteristics is still limited; XGBoost is robust in dealing with nonlinear feature combination and unbalanced classification, but it is difficult to directly use the structure information of high-order graph. Based on this, this study proposes a graph embedding driven ensemble learning mechanism, which takes the structural representation obtained by HGNN learning as the core input feature and drives XGBoost to build a gradient lifting model with high classification ability.

At the feature construction level, the nodes embedded in $z_i \in \mathbb{R}^d$ output by HGNN are regarded as a low dimensional abstraction of the structural role of nodes in the cross-border trade network. In order to fully release the expression potential of graph embedding in tree model, this study does not directly use a single embedding vector, but constructs a variety of derived features based on embedding, including embedding component, statistical aggregation and local difference measure, forming an enhanced feature set:

$$f_i^{graph} = g(z_i) = \{z_i, mean(\mathcal{N}(i)), var(\mathcal{N}(i)), \parallel z_i - \bar{z}_{\mathcal{N}(i)} \parallel\} \tag{16}$$

Where $\mathcal{N}(i)$ represents the set of node neighborhoods. This construction method enables XGBoost to learn more complex discriminant rules based on structural similarity and structural deviation.

Table 7 compares the effects of different graph embedding feature construction methods on the scale and information coverage of XGBoost input features.

Table 7. Comparison of feature construction methods based on HGNN graph embedding

Construction strategy	Graph embedding dimension	Number of derived features	Total characteristic dimension	Structure information coverage
Original embedding	128	0	128	Node itself
Mean aggregation	128	128	256	First order neighborhood
Mean+variance	128	256	384	Local structure distribution
Deviation enhancement	128	192	320	Abnormal structure
Total quantity construction (in this paper)	128	320	448	Multiscale structure

After completing the construction of the graph embedding feature, the model integrates it with the traditional statistical features in cross-border trade (such as transaction frequency, amount distribution, country risk index, etc.) to form the final classification input:

$$h_i = [f_i^{graph} \parallel f_i^{stat}] \tag{17}$$

This collaborative classification mechanism enables XGBoost to use both structure induced features and business statistical features, so as to form a more robust decision boundary in complex fraud scenarios.

To verify the complementarity of structural representation and statistical features, table 8 shows the classification performance of XGBoost on the verification set under different feature combinations.

Table 8. Comparison of XGBoost classification performance under different feature combinations

Feature input type	AUC	F1-score	Recall	Precision
Statistical characteristics only	0.842	0.611	0.573	0.654

Graph embedding features only	0.876	0.648	0.621	0.679
Simple splicing	0.901	0.687	0.662	0.714
Weighted fusion	0.913	0.701	0.683	0.726
Collaborative classification (in this paper)	0.928	0.732	0.718	0.748

The results show that the collaborative input of structure representation and statistical features is significantly better than that of a single feature source, especially in the recall rate, which shows that the model can identify more covert fraud.

However, with the increase of feature dimension, the integrated model faces the risk of over fitting. Therefore, this study introduces multiple over fitting prevention strategies in the HGNN–XGBoost integration process, including graph embedding dimensional constraints, feature subsampling and sample weighting based on structural similarity. The objective function of XGBoost is defined as:

$$\mathcal{L} = \sum_{i} \ell(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k), \Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \parallel \mathbf{w}_k \parallel^2 \tag{18}$$

The regular term $\Omega(f_k)$ is used to limit the complexity of the tree model. In addition, in the training process, only part of the graph embedded derived features are selected randomly to participate in each round of tree growth, so as to reduce the risk of feature collinearity.

Table 9 shows the effects of different regularization and sampling strategies on the generalization performance of the model.

Table 9. Comparison of model generalization performance under different integration regularization strategies

Policy configuration	Training AUC	Test AUC	Generalization gap
No regularity	0.972	0.889	0.083
L2 regular	0.961	0.905	0.056
Characteristic subsampling	0.955	0.911	0.044
Sample weighting	0.949	0.917	0.032
Comprehensive strategy (in this paper)	0.946	0.924	0.022

It can be seen that the proposed integration strategy not only effectively inhibits over fitting, but also further improves the stable performance of the model on the test set.

7. MODEL TRAINING STRATEGY AND OPTIMIZATION

In the task of cross-border trade fraud detection, the core challenges faced in the model training phase mainly come from the extreme category imbalance, the complex distribution of heterogeneous features and the optimization stability problems brought about by the large-scale graph structure. Therefore, this study designed a set of phased, controllable and engineering scalable training and optimization strategies for HGNN–XGBoost integration framework to ensure the effective convergence and generalization ability of the model in the actual scene.

In terms of loss function design, considering that fraud samples usually account for only a small proportion in the real cross-border trade data, the direct use of standard cross entropy loss can easily lead to the model bias to most categories. In this study, the weighted cross entropy is introduced as the basic loss function in the HGNN representation learning stage, and its form is defined as:

$$\mathcal{L}_{\text{cls}} = -\sum_{i \in \mathcal{V}_l} (w_1 y_i \log \hat{y}_i + w_0 (1 - y_i) \log(1 - \hat{y}_i)) \tag{19}$$

Where w_1 and w_0 respectively represent the weight coefficients of fraud and normal classes, which are set adaptively according to the sample proportion. The loss function can explicitly amplify the influence of fraudulent samples on parameter update in the representation learning stage, so as to improve the sensitivity of embedding to minority classes.

To analyze the impact of different category weight settings on the training process, table 10 shows the convergence performance and stability index of the model on the validation set under different positive and negative sample weight ratios.

Table 10. Comparison of model training stability under different category weight configurations

Fraud weighting w_1	Normal weight w_0	Number of convergence rounds	Verify AUC	Loss fluctuation range
1	1	14	0.871	0.092
3	1	16	0.892	0.081
5	1	18	0.907	0.063
8	1	21	0.912	0.058
10	1	24	0.909	0.071

Moderately increasing the weight of fraud samples helps to improve the performance and convergence stability of the model, but too high weight will increase the number of training rounds and introduce additional fluctuations. Therefore, this paper uses the configuration of $w_1 : w_0 = 8 : 1$ in subsequent experiments.

In terms of parameter optimization and convergence, HGNN and XGBoost adopt a phased optimization strategy. HGNN indicates that the learning phase uses the random gradient descent method based on Adam, and its parameter update form is:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon} \tag{20}$$

Where η is the learning rate. In order to prevent the gradient oscillation caused by the deep heterogeneous propagation, the gradient clipping and early stop mechanism are introduced in the training process, and the training is terminated when the loss of the verification set does not fall in several consecutive rounds.

Table 11 compares the convergence speed and final performance of HGNN training process under different optimal configurations.

Table 11. Influence of different optimization strategies on HGNN convergence behavior

Optimize configuration	Initial learning rate	Clip gradient	Number of convergence rounds	Final AUC

SGD	0.01	no	32	0.884
Adam	0.01	no	24	0.901
Adam	0.005	yes	19	0.913
Adam	0.003	yes	21	0.911

The results show that the combination of adaptive optimizer and gradient clipping can significantly accelerate the convergence speed and improve the performance of the model.

In terms of computational complexity and scalability, the main computational overhead of HGNN comes from message passing and attention computing, and its time complexity can be approximately expressed as:

$$O(\sum_{r\in\mathcal{R}} \mid \mathcal{E}_r \mid \cdot d) \tag{21}$$

Where $\mid \mathcal{E}_r \mid$ is the number of edges of relationship r, and d is the embedded dimension. By using small batch sampling and relational level parallel computing, this complexity shows good scalability in the actual large-scale cross-border trade map. The complexity of XGBoost stage is mainly determined by the number and depth of trees, and its growth rate is relatively controllable.

In summary, the proposed training and optimization strategy not only ensures the convergence stability of the model, but also effectively balances the performance improvement and computational cost, providing a feasible engineering basis for the deployment of HGNN–XGBoost integrated model in the real cross-border trade fraud detection system.

8. EXPERIMENTAL DESIGN AND RESULT ANALYSIS

In order to systematically evaluate the effectiveness and robustness of the proposed HGNN–XGBoost integrated model in the cross-border trade fraud detection task, this study built an experimental data set around the real business scenario, and conducted in-depth analysis from three aspects: overall performance, statistical significance and different fraud modes.

In terms of data set, the experimental data comes from the desensitization history of a cross-border trade risk control system, covering the enterprise subject, transaction behavior, account Association and country risk information. The data is uniformly modeled as a heterogeneous graph structure, including a variety of nodes and relationship types. To ensure the reliability of the experimental results, the data is divided into training set, verification set and test set in chronological order, with a ratio of 6:2:2, so as to avoid information leakage. The fraud detection task is modeled as a binary classification problem, and its prediction probability is recorded as $\hat{y}_i \in [0,1]$.

The overall assessment uses a variety of indicators, including AUC, F1 score, precision and recall, which are defined as [25],[26]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{22}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{23}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{24}$$

AUC is measured by the area under the ROC curve to measure the overall classification ability of the model under different thresholds.

Figure 1 shows the ROC curves and corresponding AUC values of different models on the test set, which is used to evaluate the overall classification ability of the models in the cross-border trade fraud detection task. The performance of each model shows a clear and consistent progressive relationship. The AUC of logistic regression is 0.693, indicating that only linear statistical features can capture some fraud signals, but the ability to depict complex nonlinear patterns is limited. The AUC of XGBoost increased to 0.814, which was about 17.5% higher than that of logistic regression, indicating that the gradient lifting model has obvious advantages in dealing with nonlinear feature combinations.

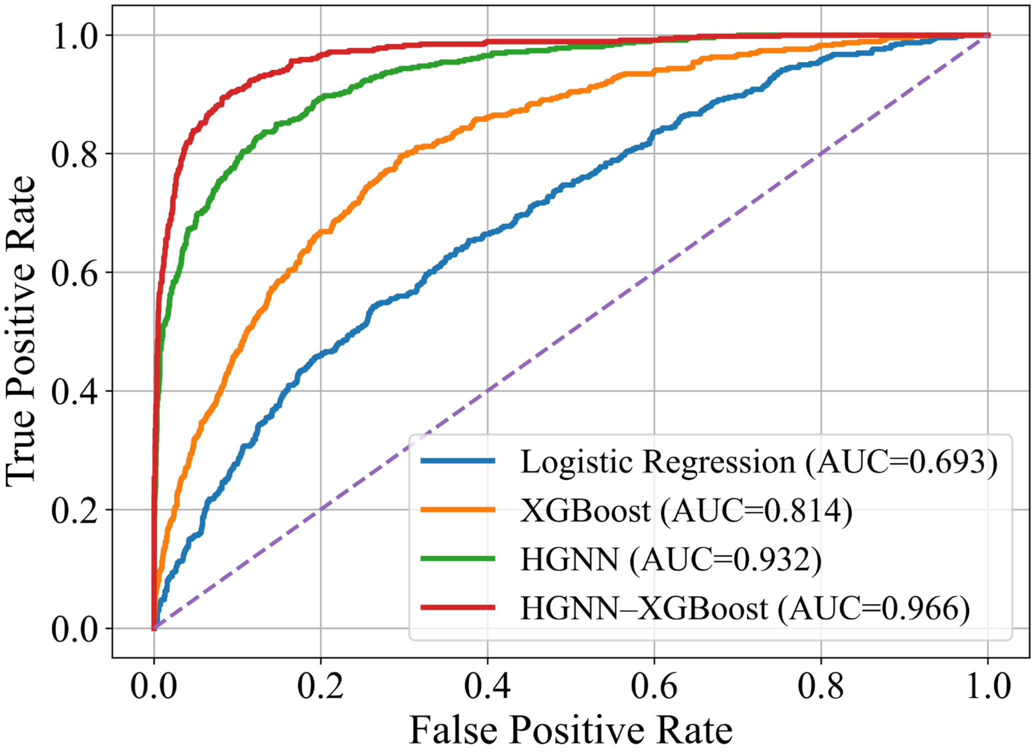

Figure 1. Comparison of ROC curves of different models on the test set

After further introducing heterogeneous graph structure modeling, the AUC of HGNN was significantly increased to 0.932, which was about 11.8% higher than that of XGBoost, indicating that the Structural Semantics of cross entity and multi relationship has key discriminative value in cross-border trade fraud detection. Finally, the HGNN–XGBoost integrated model achieved the highest AUC (0.966), which was about 3.6% higher than that of HGNN alone, showing that the learning and gradient lifting discriminant models are highly complementary in the overall ranking ability. The results show that the integrated framework can maintain a higher true positive rate in almost the entire false positive rate range, and provide reliable support for high-precision risk ranking.

Subsequently, in order to quantitatively compare the overall detection performance of different methods, table 12 summarizes the evaluation results of this model and various baseline methods on the test set. The comparison methods include traditional model based on statistical features, XGBoost alone, and graph neural network model without integration mechanism.

Table 12. Overall performance comparison of different methods in cross-border trade fraud detection task

Method	AUC	F1-score	Recall	Precision
Logistic Regression	0.693	0.512	0.471	0.561
Random Forest	0.756	0.574	0.533	0.621
XGBoost	0.814	0.664	0.631	0.700
HGNN	0.932	0.708	0.684	0.734
HGNN–XGBoost (in this paper)	0.966	0.748	0.718	0.758

HGNN–XGBoost is superior to the comparison model in all core indicators, especially in the recall index, which is of great significance to the business goal of "reducing missing detection fraud" in the actual cross-border trade risk control.

Figure 2 compares the recall performance of different models in a variety of typical cross-border trade fraud scenarios, focusing on the coverage of fraud samples. It can be seen that HGNN–XGBoost achieves the highest recall in all scenarios, especially in the scenarios of "multi account collaborative fraud" and "high-risk country association". Quantitatively, the recall of the integrated model in the multi account collaborative fraud scenario is about 0.75, which is about 23% higher than XGBoost (about 0.61) and about 10% higher than HGNN (about 0.68).

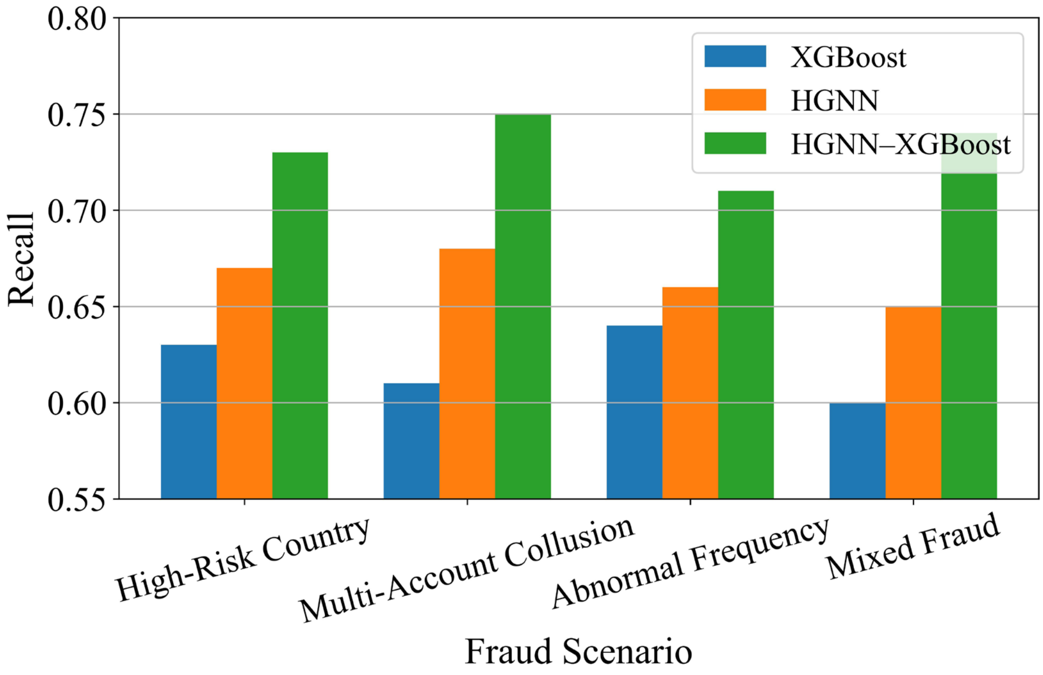

Figure 2. Comparison of recall models under different fraud scenarios

This result shows that the high-order structure representation learned by HGNN can effectively capture the cross agent collaborative behavior, and XGBoost further strengthens the ability to distinguish the key structural features. Even in the "abnormal transaction frequency" scenario with relatively weak structural dependence, HGNN–XGBoost still maintains a recall of about 0.71, indicating that the model is not only effective for complex structured fraud, but also has good scene generalization ability. Further, table 13 shows the AUC statistical results of each model under different fraud scenarios.

Table 13. Model AUC comparison under different fraud scenarios

Fraud scenario type	XGBoost	HGNN	HGNN–XGBoost
High risk country linkages	0.881	0.903	0.936
Multi account collaborative fraud	0.864	0.891	0.927
Abnormal transaction frequency	0.872	0.884	0.914
Hybrid fraud	0.858	0.889	0.921

The results show that the integration model maintains a consistent advantage in all fraud sub scenarios, and the improvement is greater in the scenarios with stronger structural dependence.

Figure 3 shows the precision recall curve of HGNN–XGBoost model on the test set, which is used to analyze the actual classification performance of the model under the condition of highly unbalanced categories. Within the range of recall from about 0.60 to 0.75, precision has always maintained above 0.70, which shows that the model can effectively control the false positive rate while expanding the coverage of fraud. This feature is particularly critical for the cross-border trade risk control system, because too many false positives will significantly increase the cost of manual review.

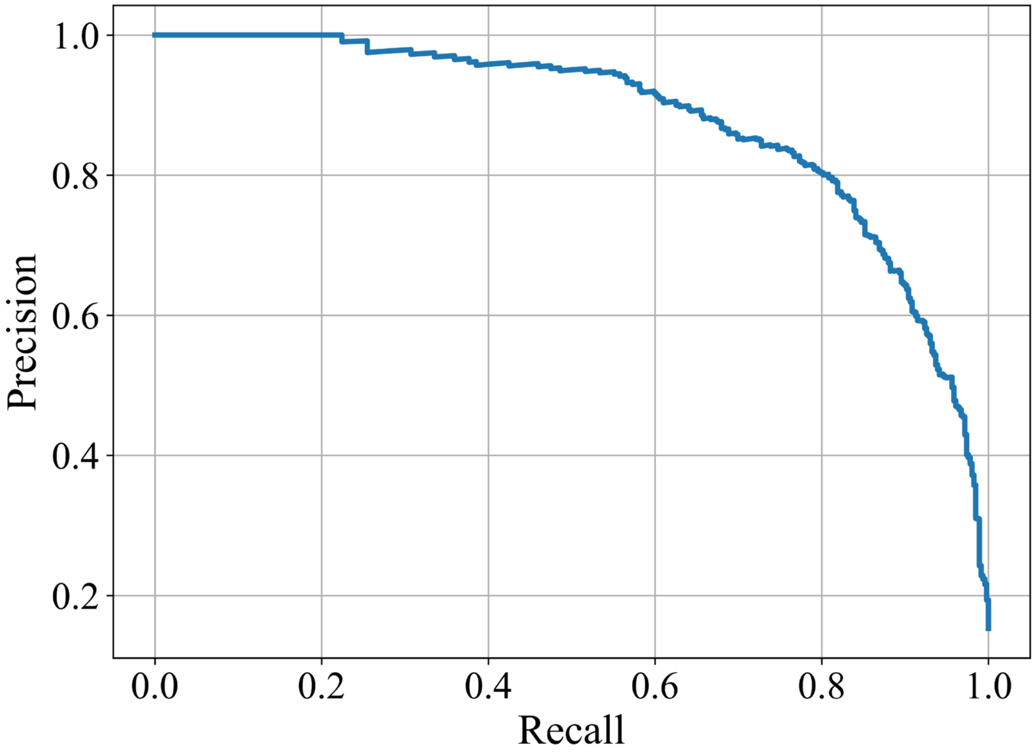

Figure 3. Precision recall curve of HGNN–XGBoost model

In addition, the precision – recall curve shows a smooth and monotonous change trend as a whole, without obvious oscillation, indicating that the model has high prediction stability under different threshold settings. Combining the results in Figure 1 and Figure 2, it can be seen that HGNN–XGBoost not only performs well in the global sorting capability (AUC), but also achieves a good balance between the high recall and controllable precision that the actual business is more concerned about, which verifies its application potential in the real cross-border trade fraud detection scenario.

9. ABLATION EXPERIMENT AND MODEL ANALYSIS

In order to systematically analyze the actual contributions of key components in the proposed HGNN–XGBoost integration model and verify the rationality of the source of model performance improvement, this study designed a series of ablation experiments to conduct quantitative and qualitative analysis of model behavior from the perspectives of heterogeneous relationship modeling, attention mechanism, integration strategy and key parameter settings.

Firstly, in order to evaluate the role of heterogeneous relationship modeling and attention mechanism in cross-border trade fraud detection, the experiment constructed a control model by gradually removing or simplifying the relevant modules. Specifically, it includes: reducing heterogeneous relationships to isomorphic graphs, removing relationship level attention weights, and replacing attention aggregation with uniform aggregation. The AUC change of the model on the test set is used to measure the contribution of each component, and its optimization objectives are consistent [27]:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{reg}} \tag{25}$$

Table 14 shows the change trend of the overall performance of the model under different structural configurations. When the heterogeneous relationship is simplified to the homogeneous structure, the AUC of the model decreases significantly, indicating that the semantic information carried by different relationship types in cross-border trade cannot be ignored; After further removing the attention mechanism, the performance degradation is more significant, indicating that the adaptive learning of relationship importance plays a key role in distinguishing complex fraud patterns.

Table 14. Comparison of heterogeneous relationship and ablation results of attention mechanism

Model configuration	Heterogeneous or not	Attention or not	AUC	F1-score
Isomorphic GNN	no	no	0.861	0.654
Heterogeneous inattention	yes	no	0.891	0.681
Heterogeneous+homogeneous polymerization	yes	no	0.903	0.694
Complete HGNN	yes	yes	0.932	0.721

From a quantitative point of view, the AUC of the complete HGNN is about 0.932, while after removing relational attention, it decreases to about 0.891, a decrease of nearly 4.4%, which verifies the importance of attention mechanism in the selection of Structural Semantics in multi relational scenes. The results show that heterogeneous relationship modeling and attention mechanism have obvious superposition effect in performance improvement, rather than simple redundancy.

Secondly, in order to verify the effectiveness of HGNN and XGBoost integration strategy, the experiments compared a variety of classification methods, including using only HGNN built-in classification header, using only XGBoost (with statistical features as input), and XGBoost model under different feature fusion methods. The objective function form of the integrated model remains as follows:

$$\mathcal{L}_{\text{XGB}} = \sum_{i} \ell(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{26}$$

Table 15 shows the AUC comparison results of different classification strategies on the test set. It can be seen that HGNN alone has achieved high performance, but after the introduction of XGBoost, the ranking ability of the model has been further improved, especially in the high-risk sample interval.

Table 15. Experimental results of effectiveness verification of integration strategy

Classification mode	AUC	Recall	Precision
XGBoost (statistical characteristics)	0.814	0.631	0.700
HGNN+linear classification	0.932	0.684	0.732
HGNN + MLP	0.944	0.702	0.741
HGNN–XGBoost	0.966	0.748	0.758

Further quantitative results are shown in table 15. Compared with HGNN alone, HGNN–XGBoost has increased the AUC by about 3.4%, and the improvement in recall index is more obvious, indicating that the integrated model is more conducive to capture covert fraud samples.

Finally, in order to analyze the sensitivity of the model to key parameters, this study conducted a systematic experiment on the embedding dimension d and network layer L of HGNN. The model output is expressed as:

$$\mathrm{z}_i \in \mathbb{R}^d, \mathrm{z}_i = \mathrm{h}_i^{(L)} \tag{27}$$

Table 16 shows the change trend of model performance under different embedded dimension settings. When d increases from 64 to 128, the performance of the model is significantly improved; After further increasing to 256, the AUC increase tends to be saturated or even slightly decreased, indicating that too high dimensions may introduce redundant information and increase the risk of over fitting.

Table 16. Comparison of model performance under different key parameter settings

Embedded dimension d	Number of layers L	AUC	Training time (relative)
64	2	0.901	1.0×
128	2	0.932	1.4×
128	3	0.966	1.9×
256	3	0.961	2.8×

Based on the above ablation experiments and parameter analysis, it can be concluded that the improvement of model performance comes from the synergy of several key designs. Heterogeneous relationship modeling and attention mechanism provide high-quality structural representation. The integration strategy of HGNN–XGBoost further enlarges the classification value of structural features, while reasonable parameter configuration achieves a good balance between performance and computational cost. These results fully verify the rationality and stability of the proposed method from the experimental level.

10. CONCLUSIONS AND FUTURE WORK

Focusing on the research problem of cross-border trade fraud detection, which is highly complex and practical, this paper proposes an integrated learning framework combining heterogeneous graph neural network and XGBoost. By modeling the cross-border trade system as a heterogeneous graph structure with multi type entities and multi relationship interactions, the model can learn the high-order structural semantic representation from the complex network, and on this basis, it can achieve efficient and robust fraud classification with the help of the gradient lifting tree model. The experimental results show that the proposed method is significantly better than many comparison methods in the overall ranking ability, fraud sample coverage and stability under different fraud scenarios, which verifies the effectiveness of structural representation learning and strong discriminant model collaborative design in cross-border trade risk control scenarios.

The research in this paper shows that it is difficult to meet the requirements of complex structure modeling and high-precision risk classification only relying on a single model paradigm. By using HGNN as the structural semantic extractor and XGBoost as the classification decision module for decoupling integration, the training stability and engineering deployability can be significantly improved while ensuring the model expression ability. This result provides a scalable modeling idea for fraud detection tasks in complex heterogeneous networks, and also provides a reference for subsequent research on collaborative optimization between structural learning and discriminant learning.

Although the experimental results are satisfactory, the method in this paper still has some limitations. Firstly, the current model is mainly based on static heterogeneous graph, and has not explicitly described the dynamic characteristics of cross-border trade behavior over time, which may affect the response ability of the model in the face of rapidly evolving fraud strategies. Secondly, the training process of HGNN has a certain dependence on the quality of graph structure and relationship integrity. When some relationships are missing or there is systematic noise, the reliability of structure representation may be affected. In addition, although the integration framework has good scalability in engineering, it still needs to further optimize the computational efficiency and storage overhead in large-scale real-time systems.

Facing the future work, this method can be expanded and deepened from many directions. On the one hand, the dynamic graph neural network can be introduced into the existing framework to explicitly model the temporal evolution of cross-border trade relations, so as to improve the model's ability to perceive new and sudden fraud; On the other hand, online learning and incremental update mechanism can be combined to enable the model to continuously adapt to changes in data distribution in real business systems. In addition, at the actual deployment level, the model is deeply integrated with the rule system and the manual audit process, and the interpretability and causal analysis methods are introduced to help improve the comprehensibility and business acceptability of model decisions. Through the above expansion, the proposed HGNN–XGBoost integration framework is expected to play a more lasting and stable role in the real cross-border trade risk control system.

Abbreviations

HGNN, Heterogeneous Graph Neural Network;
XGBoost, Extreme Gradient Boosting;
AUC, Area Under the ROC Curve;
ROC, Receiver Operating Characteristic curve;
TP, True Positive;
FP, False Positive;
FN, False Negative;
TN, True Negative;
GNN, Graph Neural Network;
MLP, Multi-Layer Perceptron;
SGD, Stochastic Gradient Descent;
Adam, Adaptive Moment Estimation;

AI, Artificial Intelligence;
ARIMA, Autoregressive Integrated Moving Average;
SARIMA, Seasonal Autoregressive Integrated Moving Average;
LSTM, Long Short-Term Memory;
GRU, Gated Recurrent Unit;
RNN, Recurrent Neural Network;
MAE, Mean Absolute Error;
RMSE, Root Mean Square Error;
MAPE, Mean Absolute Percentage Error;
AIC, Akaike Information Criterion;
ACF, Autocorrelation Function;
PACF, Partial Autocorrelation Function;
GPU, Graphics Processing Unit;
CUDA, Compute Unified Device Architecture.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The individual contributions are specified as follows: X.Z.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, X.Z.

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used DeepSeek in order to check spell and grammar. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

REFERENCES

[1] Liang, Y. (2025). Financial Legal Risks and Prevention Mechanisms in Cross-Border Mergers and Acquisitions: A Systemic Analysis. *Law and Economy*, *4*(4), 18-27. DOI: **https://doi.org/10.63593/LE.2788-7049.2025.05.003**

[2] Yu, L., Cong, Q., & Li, S. (2024). Study on international cooperation to address cross-border telecommunication network fraud offence. *Journal of Politics and Law.*, *17*, 51. DOI: **https://doi.org/10.5539/jpl.v17n2p51**

[3] Howson, K., Ferrari, F., Ustek-Spilda, F., Salem, N., Johnston, H., Katta, S., ... & Graham, M. (2022). Driving the digital value network: Economic geographies of global platform capitalism. *Global Networks*, 22(4), 631-648. DOI: **https://doi.org/10.1111/glob.12358**

[4] Bokrantz, J., Shurrab, H., Johansson, B., & Skoogh, A. (2025). Unravelling supply chain complexity in maintenance operations of battery production. *Production Planning & Control*, *36*(13), 1752-1773. DOI: **https://doi.org/10.1080/09537287.2024.2414334**

[5] Siqi, C., Rajamanickam, R., Manap, N. A., & Zahir, Z. M. (2024). Application of Blockchain Technology in Cross-Border Telecommunications Network Fraud to Ensure China's Judicial Justice. *Jurnal IUS Kajian Hukum dan Keadilan*, *12*(3), 472-486. DOI: **https://doi.org/10.29303/ius.v12i3.1554**

[6] Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of classification methods on unbalanced data sets. *Ieee Access*, *9*, 64606-64628. DOI: **https://doi.org/10.1109/ACCESS.2021.3074243**

[7] Yan, S., Liu, R., Zhang, Y., Yao, X., Yang, Y., Wang, Q., ... & Wang, S. (2024). Investigation and application of data balancing and combined discriminant model in rock burst severity prediction. *Scientific Reports*, *14*(1), 29657. DOI: **https://doi.org/10.1038/s41598-024-81307-z**

[8] Kyriazos, T., & Poga, M. (2024). Application of machine learning models in social sciences: managing nonlinear relationships. *Encyclopedia*, *4*(4), 1790-1805. DOI: **https://doi.org/10.3390/encyclopedia4040118**

[9] Shahbazi, M. A., & Azadeh-Fard, N. (2025). Hierarchical data modeling: A systematic comparison of statistical, tree-based, and neural network approaches. *Machine Learning with Applications*,

100688. DOI: **https://doi.org/10.1016/j.mlwa.2025.100688**

[10] Yan, L., & Xu, Y. (2024). XGBoost-Enhanced Graph Neural Networks: A New Architecture for Heterogeneous Tabular Data. *Applied Sciences (2076-3417)*, *14*(13). DOI: **https://doi.org/10.3390/app14135826**

[11] Deng, D., Chen, X., Zhang, R., Lei, Z., Wang, X., & Zhou, F. (2021). XGraphBoost: extracting graph neural network-based features for a better prediction of molecular properties. *Journal of chemical information and modeling*, *61*(6), 2697-2705. DOI: **https://doi.org/10.1021/acs.jcim.0c01489**

[12] Mosa, M. A. (2025). Optimizing text classification accuracy: a hybrid strategy incorporating enhanced NSGA-II and XGBoost techniques for feature selection. *Progress in Artificial Intelligence*, 1-25. DOI: **https://doi.org/10.1007/s13748-025-00365-0**

[13] Demir, S., & Sahin, E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications*, *35*(4), 3173-3190. DOI: **https://doi.org/10.1007/s00521-022-07856-4**

[14] Yaqoob, A., Verma, N. K., Aziz, R. M., & Shah, M. A. (2024). Optimizing cancer classification: a hybrid RDO-XGBoost approach for feature selection and predictive insights. *Cancer Immunology, Immunotherapy*, *73*(12), 261. DOI: **https://doi.org/10.1007/s00262-024-03843-x**

[15] Kumar, K., Samui, P., & Choudhary, S. S. (2026). Prediction and interpretation of liquefaction occurrences using explainable machine learning models. *Sādhanā*, *51*(1), 4. DOI: **https://doi.org/10.1007/s12046-025-03005-4**

[16] Chang, Y., Iakovou, E., & Shi, W. (2020). Blockchain in global supply chains and cross border trade: a critical synthesis of the state-of-the-art, challenges and opportunities. *International Journal of Production Research*, *58*(7), 2082-2099. DOI: **https://doi.org/10.1080/00207543.2019.1651946**

[17] Liu, Y. J., & Ha-Brookshire, J. E. (2025, January). Mapping Success: A Study on Firm Capabilities in Chinese Cross-Border E-Commerce. *In International Textile and Apparel Association Annual Conference Proceedings (Vol. 81, No. 1)*. Iowa State University Digital Press. DOI: **https://doi.org/10.31274/itaa.18539**

[18] Nguyen, H., Vu, T., Vo, T. P., & Thai, H. T. (2021). Efficient machine learning models for prediction of concrete strengths. *Construction and Building Materials*, *266*, 120950. DOI: **https://doi.org/10.1016/j.conbuildmat.2020.120950**

[19] Luo, Z., Li, Z., Dong, C., Dai, X., Shen, X., Li, J., & Bi, G. (2024). Multi-participants trading mode in Cross-Border electricity Market: A non-cooperative game approach. *International Journal of Electrical Power & Energy Systems*, *160*, 110093. DOI: **https://doi.org/10.1016/j.ijepes.2024.110093**

[20] Luo, Z., Dong, C., Dai, X., Wang, H., Bi, G., & Shen, X. (2024). Research on decision-making behavior of multi-agent alliance in cross-border electricity market environment: an evolutionary game. *Global Energy Interconnection*, *7*(6), 707-722. DOI: **https://doi.org/10.1016/j.gloei.2024.11.009**

[21] Rubin-Delanchy, P., Cape, J., Tang, M., & Priebe, C. E. (2022). A statistical interpretation of spectral embedding: the generalised random dot product graph. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4), 1446-1473. DOI: **https://doi.org/10.1111/rssb.12509**

[22] Gharagoz, M. M., Noureldin, M., & Kim, J. (2025). Explainable machine learning (XML) framework for seismic assessment of structures using Extreme Gradient Boosting (XGBoost). *Engineering Structures*, *327*, 119621. DOI: **https://doi.org/10.1016/j.engstruct.2025.119621**

[23] Bachiri, K., Yahyaouy, A., Malek, M., & Rogovschi, N. (2025). MM-HGNN: Multimodal Representation Learning Heterogeneous Graph Neural Network. *International Journal of*

Computational Intelligence Systems, *18*(1), 178. DOI: **https://doi.org/10.1007/s44196-025-00820-9**

[24] Yu, L., Sun, L., Du, B., Liu, C., Lv, W., & Xiong, H. (2022). Heterogeneous graph representation learning with relation awareness. *IEEE Transactions on Knowledge and Data Engineering*, *35*(6), 5935-5947. DOI: **https://doi.org/10.1109/TKDE.2022.3160208**

[25] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6. DOI: **https://doi.org/10.1186/s12864-019-6413-7**

[26] Melnykova, N., Patereha, Y., Skopivskyi, S., Farion, M., Fedushko, S., & Drohomyretska, K. (2025). Machine learning for stroke prediction using imbalanced data. *Scientific Reports*, *15*(1), 33773. DOI: **https://doi.org/10.1038/s41598-025-01855-w**

[27] Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., ... & Holzinger, A. (2022). Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 329-341. DOI: **https://doi.org/10.1109/TPAMI.2022.3145392**