https://doi.org/10.26599/AIR.2026.9150002 CAAI Artificial Intelligence Research

GraphCredit: Enhancing credit explainability via gated graph reasoning and causal LLM attribution

Bocheng Wang[1†], Di Han[2†], Yuchang Zou[2], Jianqing Li[1] ✉, Haochen Duan[2], and Canwei Dai[2]

ABSTRACT

Complex prediction models widely adopted in the field of financial risk control, despite their superior performance in credit default prediction accuracy, have long suffered from issues regarding the explainability of their predictive results. This makes it difficult to satisfy stringent requirements for regulatory compliance and algorithmic fairness. Current mainstream explainability techniques primarily focus on feature attribution and generally lack structured modeling and deep reasoning capabilities for complex community correlations between financial entities, thereby failing to reveal the transmission mechanisms of community-based risks. To this end, this paper proposes an explainability-enhanced framework, namely GraphCredit. Specifically, GraphCredit first extracts and quantifies the community risks of borrowers to construct a borrower-centric knowledge graph. Subsequently, GraphCredit employs a GraphSAGE model combined with a gating mechanism to achieve dynamic feature weighting and credit default prediction, achieving an average performance improvement of 8.70% compared with the SOTA models. Finally, leveraging large language models (LLMs), the framework transforms the extracted complex risk evidence chains into logically clear natural language narrative reports that comply with regulatory standards. Experimental results demonstrate that the explainability score under both human and LLM evaluations increased by 17.98% compared with shapley additive explanations (SHAP). GraphCredit elevates the explainability of credit default prediction from traditional static "feature attribution" to a dynamic "risk narrative" dimension, providing a new paradigm that balances high precision with robust trustworthiness for human−AI collaboration in high-risk financial decision-making.

KEYWORDS

credit default prediction; large language models (LLMs); explainability; risk narrative; knowledge graph

1 Introduction

With the continuous expansion of the global credit market, the issue of loan default risk has become prominent. Particularly during economic downturns or sudden crises (such as COVID-19), the continuous accumulation of non-performing loans often leads to a non-linear upward trend in loan default rates, posing challenges to the asset quality and financial stability of financial institutions [1]. In this context, achieving accurate prediction of loan defaults and control of credit risk holds significant practical importance for maintaining the stability of financial markets and the equilibrium of credit systems.

Existing paradigms for credit default prediction are shifting toward the application of multi-source heterogeneous data features, combining "hard information" with "soft information" to further enhance the precision of predictions. Hard information refers to objective, quantifiable, and verifiable factual data, typical features such as demographic information, balance sheet status, and credit history have been widely utilized to predict borrowers' default probabilities and facilitate credit decision-making [2, 3]. Soft information encompasses subjective, qualitative, and difficult-to-quantify data, such as calling activities, spatial dependence, and community influence, which are also being progressively validated and incorporated into credit default prediction research as key features [4–6]. While these tabular data have effectively driven improvements in prediction accuracy, ongoing research reveals that traditional tabular data, due to isolated features and missing values, make it difficult for models to establish effective data correlations. Thus, these models tend to overlook the implicit, unstructured relational risks (such as community-level default contagion) hidden within the data [7]. Graph-structured data, its inherent capability to digitize relationships, effectively compensates for the isolation of tabular data in describing associations and mitigates data sparsity issues, thereby making it possible to mine deep-seated, unstructured relational risks (such as community risks) [8].

The application of multi-source heterogeneous data features has achieved results in improving the accuracy and coverage of credit default prediction. However, financial risk control involves more than just precision; it relies heavily on the traceability of data sources, the logical consistency of feature construction, and the transparent presentation of decision rationales. These elements are to solidify the reliability of risk decisions and satisfy regulatory compliance requirements. Consequently, an in-depth exploration of the explainability and traceability of feature selection and risk decision-making has become important. Regarding feature selection, existing methods predominantly employ static approaches for feature selection and attribution. For instance, the leave-out covariates (LOCO) method indirectly reflects feature

1 School of Computer Science and Engineering, Macau University of Science and Technology, Macau 999078, China
2 School of National Finance, Guangdong University of Finance, Guangzhou 510521, China
†Bocheng Wang and Di Han contributed equally to this work.
Address correspondence to Jianqing Li, jqli@must.edu.mo

importance by measuring the degree of variation in evaluation metrics after removing target features [9]. Ying et al. [10] simulated and quantified the increase in prediction uncertainty after feature removal using learnable feature masks. Li and Wu [11] adopted shapley additive explanations (SHAP) based methods to rank feature importance, analyzing the direction and magnitude of the impact of various independent variables on default risk. While these methods provide quantitative importance scores, they fail to capture the relational dependencies among borrowers, limiting their ability to explain risk propagation effects in networked credit scenarios. In terms of risk decision-making, with the continuous development of large language models (LLMs), research has utilized methods such as prompt tuning, retrieval-augmented generation (RAG) and multi-agents to optimize LLMs, enhancing decision explainability in areas such as financial time-series forecasting and quantitative trading [12–15]. These LLMs-based approaches excel at generating semantically rich natural language explanations but primarily operate on textual inputs, lacking the capability to directly incorporate graph-structured relational information that is critical for credit risk analysis. Consequently, existing methods face a dichotomy: feature selection methods offer quantitative rigor but lack semantic explainability, while LLMs-based methods produce human-readable narratives but cannot integrate relational graph structures. Established graph neural networks (GNNs) explanation methods, such as GNNExplainer, focus on identifying subgraph structures or node features that contribute most to the prediction; nevertheless, a semantic gap remains, leaving the traceability and explainability of risk decisions insufficient [10]. Thus, there still lack a unified approach that combines the relational modeling capacity of GNNs with the semantic generation capability of LLMs to produce traceable, narrative-style risk explanations.

In summary, this paper studies how to transform the implicit relationships in tabular data—through objective graph construction and traceable graph reasoning—into structured, machine-readable risk evidence chains, thereby improving the accuracy of credit default prediction. Hence, this paper proposes a explainability-enhanced framework, GraphCredit. This framework achieves the explicit quantification and fusion of community risks and elevates traditional static "feature attribution" to a dynamic "risk narrative" dimension, providing a approach that balances high precision with robust explainability for high-risk credit decision-making.

The core contributions of this paper are summarized as follows:

- We design a community risk explicit modeling method based on Bayesian smoothing. This method transforms the implicit associations between borrowers into explicit risk features, enhancing the representation capability of the feature space regarding community risks.
- We develop a gated GNNs (Gated-SAGE) for heterogeneous feature fusion. Through a feature-level gating mechanism, the model achieves adaptive weight adjustment between individual borrower features and community features, providing a solution for credit default prediction that balances both precision and stability.
- We construct an explainability generation path, shifting from multimodal evidence chains to natural language explanations. By leveraging LLMs, the framework transforms SHAP-based feature attributions and complex evidence chains into natural language narrative reports, enhancing the understandability, trustworthiness, and completeness of risk reports.

2 Related work

In recent years, credit default prediction has gradually shifted toward three main areas: the construction of multi-source heterogeneous data features, the optimization of prediction models, and the enhancement of explainability in credit decision-making. Regarding the construction of multi-source heterogeneous data features, tabular data features have been widely incorporated into research to achieve more precise credit decisions. Furthermore, limited by the issues of isolated features and data sparsity in tabular data, graph-structured data features based on knowledge graphs have increasingly demonstrated their advantages. Consequently, relational risk features derived from graph-structured data have been progressively integrated into studies. In terms of prediction model optimization, the iterative development of models has effectively improved the accuracy of credit default prediction; however, the explainability of credit decision-making has gradually diminished. As for the enhancement of explainability in credit decision-making, existing work primarily focuses on credit feature selection, decision attribution, and LLMs-based explainability enhancement.

2.1 Data feature construction and prediction model optimization

Credit features have progressively evolved from historical credit features to multi-source heterogeneous data feature sets to capture the multi-dimensional factors of borrowers' credit risk. Driven by both the expansion of data dimensions and the increase in algorithmic complexity, credit prediction models have undergone an evolution from traditional statistical models to classic machine learning and then to deep learning, improving credit prediction accuracy. Typical studies were summarized in Table 1.

Table 1 Review of typical studies related to credit default forecasting

Reference	Model	Hard	Soft	Group
Nie et al. [16]	LR-DT	√		
Noriega et al. [17]	LR, RF, SVM	√		
Lee et al. [7]	GCN	√	√	√
Zhou et al. [18]	GAT	√	√	√
Zandi et al. [19]	DML-GNN	√	√	√
Li et al. [8]	SAMTL-GNN	√	√	√
Wang et al. [20]	RGAT	√	√	√
Liu et al. [21]	LG-GNN	√	√	√

Early studies primarily relied on statistical models such as logistic regression (LR), linear discriminant analysis (LDA), and decision trees, utilizing "hard information" to facilitate credit decision-making. For instance, Nie et al. [16] developed a customer churn prediction model using logistic regression and decision trees based on four categories of variables—customer, card, risk, and transaction activities—utilizing credit card data from a Chinese bank. Although statistical models offer strong explainability and low computational cost, they are inherently limited in capturing non-linear relationships among features, which constrains their predictive performance in complex credit scenarios. Consequently, models capable of effectively capturing complex non-linear interactions between features, such as support vector machines (SVM), random forests (RF), and eXtreme gradient boosting (XGBoost), were introduced into the field of credit prediction. Concurrently, "soft information" began to be incorporated into research, resulting in a improvement in credit default prediction performance [17]. However, these models often rely heavily on manual feature engineering and lack the capacity to automatically learn hierarchical representations from raw data. With the advancement of deep learning, GNNs have gradually become the optimal solution for maximizing predictive accuracy. Meanwhile, unstructured data and graph-structured data have started to be applied, with various community features being integrated into credit default prediction research [7]. For example, Zhou et al. [18] designed a credit default risk prediction model based on graph attention networks (GAT), which improved the model's predictive capability by focusing on latent community correlations between users. Zandi et al. [19] constructed the DML-GNN-RNN network, enhancing default prediction accuracy by capturing the associations within credit data. Despite the superiority of GNN-based methods in default risk prediction, existing approaches still tend to define community features in a relatively narrow scope, which limits their ability to fully capture the multifaceted nature of borrower networks. In this context, this paper, based on GNNs and graph-structured credit data, further enriches community features from three dimensions—occupation, geography, and intention—thereby expanding the construction dimensions of community features to further improve credit default prediction accuracy.

2.2 Enhancement of credit decision explainability

The continuous evolution of data features and prediction models has effectively improved the accuracy of credit default prediction; however, credit decision-making has become a difficult-to-explain "black-box" problem. Existing research primarily seeks breakthroughs from two perspectives.

First, credit feature selection and decision attribution. Early studies typically employed feature ablation methods, which directly compare the degree of variation in evaluation metrics (such as AUC and accuracy) before and after removing target features to reflect factor importance [9]. This approach offers the advantages of being intuitive to implement and applicable across different models. However, it falls short in capturing the non-linear contributions of features within prediction models. With the increasing demand for explainability, post-hoc feature attribution methods represented by local interpretable model-agnostic explanations (LIME) and SHAP have been more widely applied in the field of credit default prediction. These methods provide feature contribution rankings and visual analyses for the individual or global prediction results of "black-box" models, thereby enhancing the explainability of credit decisions while facilitating key feature selection [22–24]. For instance, Li and Wu [11] utilized a SHAP-based method to rank feature importance, analyzing the direction and magnitude of the impact of various independent variables on default risk. Bussmann et al. [25] combined SHAP with correlation networks to analyze the similarity of applicable features between risky and non-risky borrowers. Nallakaruppan et al. [26] combined LIME, SHAP, and PDP to construct a credit risk assessment system, enhancing the transparency and explainability of decisions. Nevertheless, these methods typically provide only quantitative importance scores without generating semantically rich, human-readable explanations, leaving a gap between attribution results and interpretable decision narratives.

Second, LLMs-based explainability enhancement. Distinct from explanation methods that primarily handle structured data, the advantage of LLMs lies in their ability to analyze unstructured text data related to credit risk, thereby capturing broader risk signals and generating natural language-based decision reports [27]. However, due to the hallucination problem of LLMs, the explanation reports they generate rely on predictions, and often show significant discrepancies with empirical feature attribution results based on methods like SHAP [28, 29]. Therefore, existing research mostly employs methods such as chain of thought (CoT) and RAG to constrain the generation process of LLMs, thereby improving the credibility and reliability of their reports. For example, Chen et al. [30] used CoT to optimize the reasoning capability of LLMs, enhancing the transparency, fairness, and compliance of decisions. Li et al. [31] adopted RAG to bridge the gap of traditional CoT methods in modeling deep causal relationships of variables, effectively mitigating the logical inconsistency hallucination problem of LLMs. Despite these advances, current LLMs-based explanation approaches primarily operate on textual inputs and lack the capability to directly incorporate graph-structured data, which is critical for capturing complex borrower dependencies.

The aforementioned studies have improved the explainability of credit decisions, but limitations remain regarding the types of models focused upon. Although GNNs have been proven to effectively improve credit default prediction accuracy, research specifically targeting the explainability enhancement of GNN prediction results remains scarce [19]. In light of this, this paper proposes a generation path that integrates graph-structured data with natural language explanations. By integrating graph-structured evidence chains and SHAP attribution results into a knowledge base, and invoking LLMs via RAG technology, complex evidence chains are transformed into natural language risk reports that align with business scenarios, thereby addressing the aforementioned gap by enabling the fusion of relational graph information with semantically rich explanations to enhance the explainability of credit decisions while improving the trustworthiness of risk reports.

3 Problem definition and methodology

3.1 Problem definition

GNNs have been widely used in complex correlation scenarios such as financial risk control and credit default prediction, owing to their unified modelling of multi-source heterogeneous data and high-order correlation mining capabilities. The introduction of subgraph explainers (such as GNNExplainer [10]) further provides interpretable local subgraph structures for model decision-making, enhancing model explainability. Nevertheless, existing research still faces two key problems:

● Credit community risk mining: despite GNNs' high-order data mining capabilities, they remain inadequate for comprehensive relationship modelling. Credit default data exhibits highly correlated community risk features, hindering GNNs from constructing initial features that fully encapsulate semantic relationships, which prevents thorough mining of credit community risks.

● Prediction explainability: subgraph explanations visualise locally influential features but struggle to reveal their complex contextual dependencies within the broader graph structure. Furthermore, their inability to link individual feature importance to broader contextual factors complicates the generation of traceable graph-based evidence chains, thereby limiting explainability in credit default prediction.

Thus, the focus of this paper is to implement credit community risk mining and explainability enhancement by leveraging technologies such as knowledge graphs, machine learning, deep learning, LLMs, and SHAP. The goal is to facilitate credit risk decision-making that possesses both accuracy and explainability. Thus, this paper designs the GraphCredit framework (as shown in Fig. 1), which consists of three core modules: data processing module, graph prediction enhancement module, and explainability enhancement module.

3.2 Overview

GraphCredit first centers on the credit knowledge graph, employing machine learning algorithms to achieve the efficient fusion of tabular data and graph structures data, thereby laying the data foundation for mining community risks. Subsequently, based on the Bayesian smoothing algorithm, the framework realizes the mining and quantification of community risks. These constructed community features are then integrated into the knowledge graph to enhance the prediction performance of GNNs by leveraging community homophily. Finally, by integrating SHAP methods with LLMs, the framework constructs a credit knowledge base grounded in SHAP values and graph evidence chains. It utilizes CoT and few-shot learning techniques to design guiding prompts that drive the LLM, thereby generating dynamic risk narrative reports based on SHAP feature attribution and graph-structured evidence chains. To articulate the problem addressed, Section 3.3 introduces and defines the concepts relevant to the research question.

The construction and representation of high-quality knowledge graph serve as the foundation for the research focus of this paper. Accordingly, this paper first utilizes the hard and soft information used in traditional credit analysis as the initial node features for the base knowledge graph, denoted as G_{base}. Subsequently, based on G_{base}, nodes across three distinct dimensions (occupation, geography, and purpose) are defined, and edge relationships are established to construct a heterogeneous knowledge graph G_{relation}, grounded in community features. This facilitates the representation of real-world credit scenarios and the fusion of tabular data with graph structures. Furthermore, TransR is adopted to transform G_{base} and G_{relation} into low-dimensional dense vectors, constructing the graph-structured datasets D_{base} and D_{relation}, respectively. Finally, this paper apply a Bayesian smoothing algorithm to extract community risks from D_{relation} to construct the dataset $D_{\text{relation\_bayes}}$, which is then concatenated with D_{base} to form the final graph-structured dataset D_{final}.

$$G_{\text{base}} = \{(h,r,t) \mid h,t \in E_1, r \in R\} \tag{1}$$

$$G_{\text{relation}} = \{(h',r,t') \mid h',t' \in E_2, r \in R\} \tag{2}$$

where (h,r,t) in the knowledge graph G_{base} and G_{relation} describe the relation r between the head entity h and the tail entity t, E_1 and E_2 represent the entity sets of the base knowledge graph G_{base} and the community feature-based heterogeneous knowledge graph G_{relation}, respectively, while R denotes the relation set.

Regarding explainability enhancement, this paper integrates the feature attribution data $\text{Data}_{\text{shap}}$ generated by the SHAP method with the graph evidence chains $\text{Data}_{\text{graph}}$ constructed based on the knowledge graph. By utilizing risk report examples $\text{Data}_{\text{template}}$ to guide LLMs via prompt engineering, we construct dynamic risk narrative reports that combine both statistical and logical explanations. In summary, the inputs and outputs of the credit default prediction and explainability enhancement method based on the GraphCredit framework are defined as follows:

● Input: $x_i \in D_{\text{final}}$ represents the i-th feature vector in the dataset D_{final}.

● Output: $\dot{p}_i = \text{sigmoid}(y_i)$ represents the default probability of the i-th borrower, and y_i denotes the output of the Gated-SAGE.

● Output: $R^i_{\text{report}} = \text{LLM}\{\text{Data}_{\text{shap}}, \text{Data}_{\text{graph}}, \text{Data}_{\text{template}}\}$ is the natural language-based dynamic risk narrative report.

3.3 Data processing module

Data processing module aims to achieve refined processing of initial credit data. Through semantic modeling, knowledge graph construction, and embedding, this module extracts core community dimensions and enhances community risk correlations, thereby achieving the fusion of tabular data and graph-structured data. Meanwhile, the datasets Dbase and

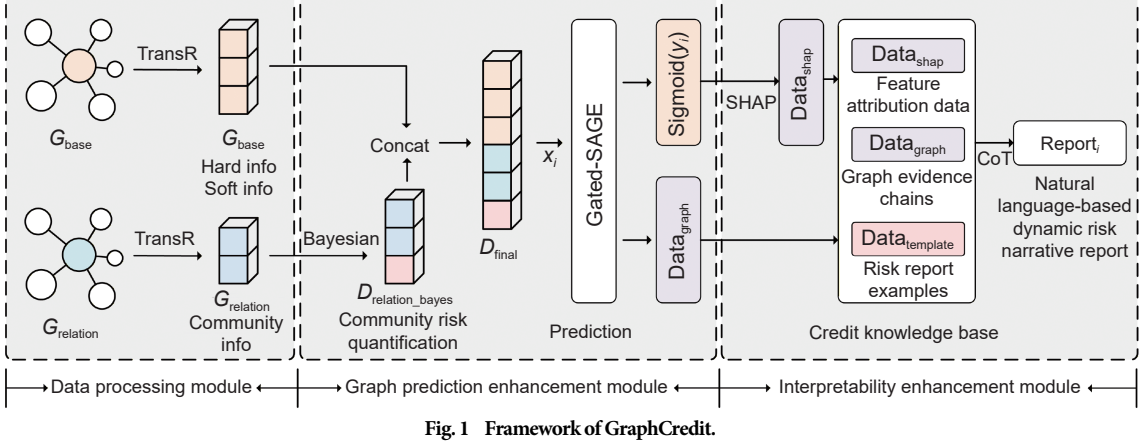

Fig. 1 Framework of GraphCredit.

Drelation constructed by this module provide data support for the graph prediction enhancement module.

3.3.1 Data cleaning and knowledge graph construction

Ontology, or the semantic data model, is a core element of knowledge graphs. It defines various concepts within the research domain and describes the attributes or features of each concept, being widely used for classifying and describing knowledge at the semantic level [32]. This paper designs a semantic data model based on credit business rules and financial industry business ontology (FIBO) standards. Centered on the borrower, this semantic data model connects information from different data sources. Nodes are connected via edges to form a semantic network, where nodes represent entities and the edges connecting them represent relationships between entities.

Based on the aforementioned semantic data model, this paper selects the default of credit card clients (DCCC) dataset as the experimental data source. Through the following three steps, while completing the construction of the credit knowledge graph G_{base}, we lay the foundation for acquiring community features.

First, community feature extraction. Features possessing community attributes—such as JobCategory, ZipCode, and Purpose—are extracted from the raw data. We define four types of nodes (Applicant, JobCategory, ZipCode, and Purpose) and establish their edge relationships, serving as three core community dimensions for community risk analysis.

Second, hierarchical strategy formulation. We establish a macroscopic rule repository for high-cardinality and noisy features. Through these rules, scattered features are aggregated into macroscopic categories, thereby reducing noise and enhancing community connectivity within the knowledge graph. Taking the JobCategory feature as an example, by formulating a macroscopic occupational domain rule repository (including Management, Medical, etc.), over 170,000 occupation categories are aggregated into 22 types. This reduces data noise and strengthens the occupational relevance of community risks.

Finally, knowledge graph construction. Based on the semantic data model, we construct the credit knowledge graph using a top–down approach and store it in a Neo4j database. Specifically, G_{base} contains 2,710,609 nodes and 31 types of relationships, while G_{relation} contains 349,756 nodes and 3 types of relationships. Table 2 shows some triplet samples of the final knowledge graph.

3.3.2 Knowledge graph embedding

Although credit knowledge graphs contain rich semantic relationships, their discrete symbolic representations are difficult for machine learning models to utilize directly. To this end, this paper adopts knowledge graph embedding (KGE) techniques to map entities and relations into a low-dimensional dense vector space, thereby transforming semantic information into computable continuous representations. In this vector space, the strength of semantic correlations between entities can be quantitatively measured via the distance and direction between vectors.

Considering that credit graphs contain multiple types of heterogeneous community risks (occupation, geography, and purpose), this paper selects the TransR model to perform embedding learning on the constructed credit knowledge graphs G_{base} and G_{relation}. TransR is capable of constructing distinct semantic spaces for different community risks. Through relation-specific projection matrices, it characterizes the intrinsic differences in how different community dimensions impact borrower risk, thereby achieving a more fine-grained semantic expression. Its energy score function is defined as

$$g(h,r,t) = \| W_r e_h + e_r - W_r e_t \|_2^2 \tag{3}$$

where $e_h, e_t \in \mathbb{R}^e$ and $e_r \in \mathbb{R}^r$ denote the entity embedding and W_r represents the projection matrix from the entity space to the relation space. Consequently, this paper constructs the graph-structured datasets D_{base} and D_{relation}, preserving the structured semantic information within the graphs. While providing feature inputs, this also serves as a reproducible and explainable data source for constructing graph evidence chains.

3.4 Graph prediction enhancement module

The graph prediction enhancement module integrates statistical and graph learning methods. This module employs Bayesian smoothing to extract community risks, constructing explainable risk scores. Subsequently, it incorporates a gating mechanism into GraphSAGE to construct the Gated-SAGE framework, achieving dynamic weight adjustment between traditional features and community features. Thereby, precise credit default prediction is achieved under the premise of fully accounting for community risks. Meanwhile, the feature set D_{final} constructed by this module and the prediction result y_i based on Gated-SAGE provide data support for the explainability enhancement module.

3.4.1 Community risk quantification based on bayesian smoothing

Credit decision scenarios typically require estimating the default probability of specific groups (e.g., JobCategory, ZipCode, Purpose). Taking JobCategory as an example, if Eq. (4) is directly used to estimate the default probability, the unreliability of small samples and data sparsity will amplify data noise and reduce statistical significance, consequently affecting the performance of the prediction model.

$$\mathrm{p}_{\text{observed\_job}} = \frac{\mathrm{Count}_{\text{default\_job}}}{\mathrm{Count}_{\text{total\_job}}} \tag{4}$$

This paper employs a Bayesian smoothing algorithm to calibrate community risk quantification by integrating prior default information. Specifically, the process begins by determining the global prior parameter p_{global} based on the D_{final}. The smoothing intensity m, which controls the trade-off between fidelity to the data and smoothness of the estimate, is then established via cross-validation. Subsequently, a weighting factor w_i is applied to balance observed data with prior experience. This allows for the calculation of smoothed default probability estimates $\hat{p}_i$, for three distinct community categories: JobCategory,

Table 2 Triplet samples of the final knowledge graph

Category	Triplet structure	Sample
Attribute	(Entity, Attribute, Value)	(app_220176, int_rate, 12.99)
Relation	(Entity, Relation, Entity)	(app_220176, job_category, Management)
Label	(Entity, Target, Label)	(app_220176, is_default, 0)

ZipCode, and Purpose. These estimates are finally integrated into the feature set $D_{\text{relation\_bayes}}$. The specific formulas are as Eqs. (5) and (6):

$$p_{\text{global}} = \frac{\text{Count}_{\text{default\_all}}}{\text{Count}_{\text{total\_all}}} \tag{5}$$

$$\hat{p}_i = w_i \cdot p_i + (1 - w_i) \cdot p_{\text{global}}, \quad p_i = \frac{k_i}{n_i}, \quad w_i = \frac{n_i}{n_i + m} \tag{6}$$

where p_{global} denotes the default probability of the entire training set, and i represents the community category, k_i, n_i, p_i represent the sample size, number of defaults, and raw default probability of community i respectively, while $\hat{p}_i$ represents the Bayesian smoothed default rate.

3.4.2 Credit default prediction based on gated-SAGE

In credit scenarios, each borrower node contains two distinct categories of features: personal features D_{base}, which encompass traditional hard and soft information, and community risk features $D_{\text{relation\_bayes}}$. This paper introduces a gating mechanism into the classic GraphSAGE model [33] to address the integration of heterogeneous features, therefore balancing the contributions of these two feature types toward the final prediction. The input features of this module are defined as follows:

- Personal feature vector: $x_{\text{base}} \in D_{\text{base}}$.
- Community risk feature vector: $x_{\text{relation\_bayes}} = [\hat{p}_{\text{JobCategory}}, \hat{p}_{\text{ZipCode}}, \hat{p}_{\text{Purpose}}] \in D_{\text{relation\_bayes}}$.
- Initial representation of the central node, obtained by projecting features into a unified semantic space via multilayer perceptron (MLP): $h_v^{(0)} = \text{MLP}(x_{\text{base}} \parallel x_{\text{relation\_bayes}})$.

The specific workflow of Gated-SAGE is as follows: First, the model implements neighbor information aggregation based on the GraphSAGE model, enabling the central node (the borrower) v to absorb the risk level of the community to which it belongs. Its aggregated neighbor representation at the k-th layer, $h_{\mathcal{N}(v)}^{(k)}$ is defined as Eq. (7):

$$h_{\mathcal{N}(v)}^{(k)} = \text{AGGREGATE}^{(k)}\left(\left\{h_u^{(k-1)}, \forall u \in \mathcal{N}(v)\right\}\right) \tag{7}$$

where $\mathcal{N}(v)$ is the set of neighbors of the central node v, meaning other borrowers sharing the same community (JobCategory, ZipCode, and Purpose) with v; $h_u^{(k-1)}$ is the feature vector of the neighbor node u at the $(k-1)$-th layer, and $h_{\mathcal{N}(v)}^{(k)}$ is the aggregated set of neighbor features. In this paper, we adopt a mean pooling function as the aggregation function AGGREGATE. Meanwhile, to avoid information loss during subsequent fusion, the central node features are included in the mean pooling. The specific formulas are defined as Eqs. (8) and (9):

$$h_{\mathcal{N}(v)}^{(k)} = \sigma\left(W_{\text{mean}}^{(k)} \cdot \text{MEAN}\left(\{h_v^{(k-1)}\} \cup \{h_u^{(k-1)}\}\right)\right) \tag{8}$$

$$\text{MEAN}(\cdot) = \frac{1}{|\mathcal{N}(v)| + 1}\left(h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right) \tag{9}$$

where $h_v^{(k-1)}$ represents the feature vector of the central node v at the $(k-1)$-th layer, $\{h_v^{(k-1)}\} \cup \{h_u^{(k-1)}\}$ represents the feature set of neighbors containing central node v, $\text{MEAN}(\cdot)$ represents the mean calculation function, $W_{\text{mean}}^{(k)}$ represents the learnable weight matrix at layer k, $\sigma(\cdot)$ represents the ReLU activation function, and $|\mathcal{N}(v)|$ represents the number of neighbor nodes.

Subsequently, by incorporating a gating mechanism, the model calculates a gating vector based on the central node features and the aggregated neighbor features, realizing the fusion update of the personal feature vector and the community risk feature vector. The gating vector $g^{(k)}$ and the gating fusion update formula are defined as Eqs. (10) and (11):

$$g^{(k)} = \sigma\left(W_g^{(k)} \cdot [h_v^{(k-1)} \parallel h_{\mathcal{N}(v)}^{(k)}] + b_g^{(k)}\right) \tag{10}$$

$$h_{\text{v\_final}}^{(k)} = \text{ReLU}\left(W^{(k)} \cdot [g^{(k)} \odot h_v^{(k-1)} \parallel (1 - g^{(k)}) \odot h_{\mathcal{N}(v)}^{(k)}]\right) \tag{11}$$

where $W_g^{(k)} \in \mathbb{R}^{d*2d}, b_g^{(k)} \in \mathbb{R}^d$ represents the gating weight matrix and the bias term respectively, $W^{(k)} \in \mathbb{R}^{d'*2d}$ represents the final weight matrix, $\odot$ represents the Hadamard product, and $h_{\text{v\_final}}^{(k)}$ represents the final node representation after the k-th layer gating fusion. Meanwhile, $h_{\text{v\_final}}^{(k)}$ will be integrated into the credit explainability knowledge base as the graph-structured evidence chain $\text{Data}_{\text{graph}}$.

Finally, based on the final node $h_{\text{v\_final}}^{(k)}$, credit default prediction is performed, and the formula for the default probability y_i is defined as Eqs. (12) and (13):

$$y_i = \text{Sigmoid}(z_v) = \frac{1}{1 + e^{-z_v}} \tag{12}$$

$$z_v = W_p \cdot h_{\text{v\_final}}^{(k)} + b_p \tag{13}$$

where $W_p \in \mathbb{R}^{1*d'}, b_p \in \mathbb{R}$ represent the weight and bias term respectively, and z_v represents the log-probability.

3.5 Explainability enhancement module

The explainability enhancement module aims to construct a dynamic, understandable, and trustworthy decision basis for the prediction results of the GraphCredit framework. This module first integrates SHAP-based feature attribution with knowledge graph-based graph evidence chains to construct a credit explainability knowledge base. Furthermore, referring to the method of Shen et al. [34], we design a report generation framework fusing CoT and few-shot learning to transform complex model outputs and graph-structured evidence chains into structured natural language narrative reports.

3.5.1 Feature attribution based on the SHAP method

SHAP is a model interpretation method based on game theory [24]. This paper calculates the SHAP values φ_j of all features for each prediction sample x_i, quantifying the contribution degree and direction of each feature to the prediction result. For a given sample x and prediction model f, its predicted value $f(x)$ can be defined as the linear sum of the contributions of all features:

$$f(x) = y_i = \varphi_0 + \sum_{j=1}^{M} \varphi_j \tag{14}$$

where $\varphi_0 = E[f(x)]$ is the predicted value of all samples, and M is the number of features. The SHAP value φ_j for feature j is defined as Eq. (15):

$$\varphi_j(f, x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!}[f_x(S \cup \{j\}) - f_x(S)] \tag{15}$$

where F, S are the feature set and feature subset, respectively, and $f_x(S) = E[f(x)|x_S]$ represents the expected output of the model when the feature values belonging to subset S in sample x are known.

Since the Gated-SAGE model is a complex graph neural network, its predictions rely upon both node features and graph structure, thereby failing to satisfy the feature independence

assumption underpinning traditional SHAP methods. Consequently, this paper further employs an enhanced KernelSHAP approach to efficiently estimate SHAP values for individual features via linearly weighted regression. Specifically, for target node x_i:

First, a subgraph dataset is constructed. Samples similar to target node x_i in community features are selected to build local subgraphs for joint conditional distribution estimation.

Second, synthetic samples are generated. For target node x_i, original values within feature subset S_k are retained, while other features not included in the subset are filled using corresponding feature values from central nodes sharing the same community type as x_i within the subgraph dataset. Construct the synthetic sample z_k by assigning the hybrid feature vector to x_i.

Finally, generate K sets of synthetic samples z_k using the above method and input them into Gated-SAGE to obtain the predicted value $f(z_k)$ for the target node. Employ the weighted least squares formula identical to KernelSHAP to solve for the estimated SHAP values φ_j of each feature:

$$\min_{\varphi_0,\ldots,\varphi_M} \sum_{k=1}^{K} \pi(S_k) \left[f(z_k) - \left(\varphi_0 + \sum_{j \in S_k} \varphi_j \right) \right]^2 \tag{16}$$

where $\varphi_1, \ldots, \varphi_M$ is the approximate estimate of the SHAP value φ_j for each feature. Meanwhile, the aforementioned approximate SHAP values will be incorporated into the credit explainability knowledge base as feature attribution data $\text{Data}_{\text{shap}}$.

3.5.2 Natural language report generation based on LLMs

Natural language report generation based on LLMs is the core output component of the explainability enhancement module, aiming to transform structured data into fluent, professional, and insightful natural language risk narrative reports. Therefore, this paper first constructs a credit explainability knowledge base and subsequently designs a hierarchical report generation framework that fuses CoT reasoning and few-shot learning.

The credit explainability knowledge base $K = \{\text{Data}_{\text{shap}}, \text{Data}_{\text{graph}}, \text{Data}_{\text{template}}\}$ constructed in this paper integrates three types of core information:

- Feature attribution dataset $\text{Data}_{\text{shap}}$: stores the feature contribution matrix calculated by the SHAP method. This dataset transforms complex model decisions into explicit marginal contributions of each feature, providing quantitative statistical evidence for the prediction results.
- Graph-structured evidence chain $\text{Data}_{\text{graph}}$: stores structured risk information mined by GraphCredit. It includes key community associations of borrower nodes, high-risk neighbor states, and risk transmission paths (e.g., "Applicant → Belongs to → High-risk Professional Community"), providing traceable evidence chain.
- Credit Risk Report Paradigm Library $\text{Data}_{\text{template}}$: by collecting and annotating nearly 100 credit risk assessment reports written by domain experts, standardized narrative logic, linguistic styles, and report structures are extracted to form high-quality prior knowledge, which is used to guide and constrain the output of the LLMs.

Subsequently, this paper constructs a report generation framework comprising three steps: information aggregation, logical reasoning, and style alignment. Specifically, first, multi-view information aggregation and context construction. Key information is extracted from the input borrower i's feature attribution $\text{Data}^i_{\text{shap}}$ and graph-structured evidence chain $\text{Data}^i_{\text{graph}}$ to achieve structural alignment. To achieve deep alignment between model interpretation and classic business logic, enabling LLMs to output natural language narrative reports with clear business logic, this paper systematically maps and fuses the quantified community risk features with the 5C principles of traditional credit risk management (capacity, character, capital, collateral, conditions). Furthermore, prompt engineering converts the fused features into structured natural language descriptions to constitute the initial context for generation. Specific feature selection is shown in Table 3.

Secondly, reasoning and report generation based on CoT. The risk report COT based on $\text{Data}_{\text{template}}$ is embedded into the prompt. For each borrower i, $\text{Data}^i_{\text{shap}}$ and $\text{Data}^i_{\text{graph}}$ are combined to synthesize feature attribution and graph-structured evidence chains, thereby constructing decision recommendations.

Finally, report style alignment based on Few-shot Learning. To ensure output stability, this paper adopts an In-context Learning strategy, embedding credit risk reports from $\text{Data}_{\text{template}}$ that are similar to the current borrower's risk pattern into the prompt as Few-shot examples. Thereby, a dynamic risk narrative report

Table 3 Main features for credit risk reporting

Dimension	Design logic	Selected feature	Feature source
Capacity	Core repayment basis	Dti AnnualInc LoanAmnt	$\text{Data}_{\text{shap}}$
Character	Repayment willingness and stability	Credit History length PubRec EmpLength	$\text{Data}_{\text{shap}}$
Capital	Financial foundation and leverage	RevolUtil TotalAcc OpenAcc	$\text{Data}_{\text{shap}}$
Collateral	Asset buffer	MortAcc HomeRisk score	$\text{Data}_{\text{shap}}$
Conditions	Community risk	JobCategory ZipCode Purpose	$\text{Data}_{\text{graph}}$

Report_i is constructed for each borrower i without updating the model parameters.

4 Experiments

To systematically evaluate the effectiveness and performance of GraphCredit framework, this section designs experiments for three core research questions. First, we verify the competitiveness of the framework in terms of prediction accuracy through comparisons with traditional and state-of-the-art baselines. Second, we conduct ablation studies to analyze the enhancement provided by the explicit quantification of community risk features via Bayesian smoothing. Finally, we construct a multi-dimensional quantitative evaluation system to measure the practical efficacy of the explainability reports generated by the framework across three dimensions: informativeness, causal reasoning, and decision support. The research focuses on the following four questions.

RQ1: Can the GraphCredit framework maintain prediction accuracy comparable to or exceeding that of baseline models while introducing complex explainability mechanisms? Can community risk features constructed improve model prediction performance?

RQ2: Compared to using community risk features directly, does the risk score explicitly quantified via Bayesian smoothing yield performance improvements?

RQ3: Compared with traditional feature fusion methods, is it necessary to incorporate a gating mechanism into GraphCredit framework?

RQ4: Compared to baseline methods, do the dynamic risk narrative reports generated by GraphCredit demonstrate superior performance?

4.1 Experimental setup

The dataset utilized in our experiments comprises a total of 314,544 samples. To ensure robust model training and evaluation, we randomly partitioned the data into training and test sets using a 7:3 split ratio. All baseline models and our proposed GraphCredit framework were trained for 50 iterations to ensure convergence and fair comparison. The experiments were conducted under identical hardware and software configurations to maintain consistency across all methods.

To determine the optimal hyperparameters for each model, we employed Optuna as the primary parameter optimization method. Due to space constraints, this paper provides a representative summary using XGBoost as an illustrative example. Through systematic grid search, the optimal hyperparameters for XGBoost were identified as follows: learning rate = 0.1, maximum tree depth = 4, and number of trees = 80. The same rigorous tuning process was applied to all comparative models to ensure fair and optimal performance.

4.2 Basic performance comparison

To examine the accuracy of GraphCredit framework in the field of credit default prediction, we selected RF, MLP, XGBoost, GAT, GraphSAGE, GCN(Hetero), and LG-GNN models as baseline. Meanwhile, addressing the imbalanced sample distribution in credit risk control scenarios and the specific features of credit business, we constructed a multi-dimensional evaluation system covering discrimination, class balance, and accuracy.

This system aims to measure model performance: AUC and KS evaluate the discriminative ability regarding customer risk levels, F1-Score and Recall evaluate the ability to identify defaulting customers, reflecting the effectiveness of risk prevention and control, and Accuracy measures overall prediction correctness. Higher values for these indicators signify superior model performance in the corresponding dimensions. Table 3 and Table 4 present the prediction results of each model based on personal features alone and after the introduction of explicit community risk features, respectively.

In terms of model prediction accuracy, GraphCredit achieves a multi-dimensional performance balance. It attains optimal values in metrics such as AUC, KS, F1-Score, and Recall, while maintaining the second-best accuracy, demonstrating a evaluation advantage. Specifically, compared to the random foreset, MLP and XGBoost models widely applied in the credit domain, graphcredit's advantage lies in its GNN architecture. By transforming discrete raw features and relationships into low-dimensional dense embeddings, it captures complex long-range dependencies between entities, thereby learning deep risk representations that transcend tabular data. Compared to the GAT, GraphSAGE, GCN (Hetero), and LG-GNN models, which are also based on graph structure. Although designed mechanisms such as attention and local-to-global approaches can adaptively integrate local and global features, GraphCredit's gating mechanism dynamically adjusts the weights of neighbor information, enabling more robust and accurate credit default prediction.

Regarding the effectiveness of community risk features, the GraphCredit framework constructs community risk patterns by quantifying the three-dimensional community risk features of JobCategory, ZipCode, and Purpose. The prediction accuracy of all models improved after the introduction of explicit community risk features, further validating the necessity of these features. This enhancement demonstrates the framework's capability to capture

Table 4 Prediction results based on personal features

Model	AUC	F1	Recall	KS	Accuracy
RF	0.5846	0.3378	0.5932	0.1280	**0.8041**
MLP	0.5921	0.3423	0.6426	0.1334	0.7476
XGBoost	0.5837	0.3392	0.6635	0.1256	0.7448
GAT	0.5577	0.3327	0.7404	0.1101	0.7589
GraphSAGE	<u>0.6026</u>	<u>0.3503</u>	0.7325	<u>0.1483</u>	0.4677
GCN(Hetero)	0.5976	0.3480	0.7106	0.1423	0.4784
LG-GNN	0.5975	0.3491	<u>0.7414</u>	0.1430	0.4583
Ours (GraphCredit)	**0.7734**	**0.4506**	**0.7473**	**0.3666**	<u>0.7627</u>

*The experimental data represent the average of the results from 30 independent runs. The optimal values are shown in bold, sub-optimal values are underlined. The same below.

localized risk dynamics and systemic patterns that are often missed by traditional individual-centric models. By integrating these community-level insights, GraphCredit provides a more holistic and accurate representation of the borrower's risk profile, bridging the gap between isolated node features and collective network behaviors.

4.3 Analysis of community risk construction mechanism

To conduct an investigation into the effectiveness of the core step of quantifying community risk within the GraphCredit framework, this section designs an ablation experiment. While keeping borrower personal features constant, the input features of the model are replaced from explicit community risk features (continuous) to conventional community risk features (discrete), and the differences in prediction performance are compared. The experimental results are shown in Table 6, forming a direct comparison with the performance using explicit community risk features as detailed in Table 5. Meanwhile, Table 7 presents the degree of performance improvement achieved by the proposed framework. This comparative analysis serves to validate that the transition from discrete representations to continuous, quantified community risk features is important for capturing the nuanced risk dynamics within a graph-structured credit environment. By isolating the impact of feature representation, the paper underscores how the explicit quantification of community risk provides the model with more granular and information, leading to the superior predictive accuracy observed in the GraphCredit framework.

Results show that, for both traditional machine learning models (RF, MLP, and XGBoost) and GNN based models (GAT, GraphSAGE, GCN(Hetero), LG-GNN and GraphCredit), the prediction performance when using continuous features, as measured by key metrics (AUC, F1-score, KS, and accuracy), consistently surpasses that achieved with discrete features, with

Table 5 Prediction results based on personal and explicit community risk features (continuous type)

Model	AUC	F1	Recall	KS	Accuracy
RF	0.7691(↑)	0.4745(↑)	0.6103(↑)	0.3791(↑)	**0.8319**(↑)
MLP	0.7981(↑)	0.4675(↑)	0.7150(↑)	0.3910(↑)	0.8179(↑)
XGBoost	0.7552(↑)	0.4634(↑)	0.7526(↑)	0.3918(↑)	0.8178(↑)
GAT	0.8054(↑)	0.4515(↑)	0.7624(↑)	0.3944(↑)	0.8232(↑)
GraphSAGE	0.7960(↑)	0.4838(↑)	0.5921	0.3875(↑)	0.7692(↑)
GCN(Hetero)	0.7950(↑)	0.4859(↑)	0.5614	0.3890(↑)	0.8093(↑)
LG-GNN	0.7959(↑)	0.4845(↑)	0.4996	0.3875(↑)	0.7918(↑)
Ours (GraphCredit)	**0.8432**(↑)	**0.4861**(↑)	**0.7865**(↑)	**0.4009**(↑)	0.8306(↑)

*(↑) represents that the discrimination, category balance, and accuracy of the model have been enhanced.

Table 6 Prediction results based on personal and conventional community risk features (discrete type)

Model	AUC	F1	Recall	KS	Accuracy
RF	0.6349	0.3700	0.6092	0.2040	0.8095
MLP	0.7854	0.4439	0.6506	0.3694	0.8085
XGBoost	0.7460	0.4604	0.6751	0.3031	0.7953
GAT	0.7732	0.4426	0.7473	0.3420	0.8034
GraphSAGE	0.6591	0.3816	0.6127	0.2254	0.6110
GCN(Hetero)	0.6557	0.3768	0.6046	0.2205	0.6082
LG-GNN	0.6573	0.3788	0.5995	0.2217	0.6148
Ours (GraphCredit)	**0.7989**	**0.4707**	**0.7517**	**0.3954**	**0.8289**

Table 7 Performance improvement (Tables 5 and 6)

Model	AUC	F1	Recall	KS	Accuracy
RF	0.1342	0.1045	0.0011	0.1751	0.0224
MLP	0.0127	0.0236	0.0644	0.0216	0.0094
XGBoost	0.0092	0.0030	0.0775	0.0887	0.0225
GAT	0.0322	0.0089	0.0151	0.0524	0.0198
GraphSAGE	0.1369	0.1022	−0.0206	0.1621	0.1582
GCN(Hetero)	0.1393	0.1091	−0.0432	0.1685	0.2011
LG-GNN	0.1386	0.1057	−0.0999	0.1658	0.1770
Ours (GraphCredit)	0.0443	0.0154	0.0348	0.0055	0.0017
Average	0.0635	0.0591	0.0037	0.1050	0.0765

only a slight decline in Recall observed for some models. For instance, when using discrete features, the AUC of GraphCredit framework decreased from 0.8432 to 0.7989, and the F1-Score dropped from 0.4861 to 0.4707. This confirms that the Bayesian smoothing algorithm within the GraphCredit framework enhances feature discriminability and model learnability by transforming discrete symbolic information into continuous signals with explicit business meaning (default probability).

4.4 Ablation experiment

To evaluate the impact of different feature fusion strategies on the model's credit default prediction performance and to further validate the effectiveness of the gating mechanism within the GraphCredit framework, this study employs the traditional feature concatenation strategy (Concat) and the attention-based feature fusion strategy (Attention) as comparative baselines, as shown in Table 8.

The experimental results show that the proposed gating mechanism achieves the highest predictive performance, ranking first among the three evaluated strategies. In comparison, while the attention-based strategy effectively integrates multi-source information, its performance is lower than that of the gating mechanism. The concatenation strategy exhibits the weakest performance, ranking third.

Thus, results indicate the necessity of incorporating a gating mechanism into the GraphCredit framework. By enabling adaptive weight allocation, the gating mechanism facilitates feature selection and integration. Consequently, this contributes to improved risk identification capability in complex credit scenarios.

4.5 Efficacy evaluation of explainability reports

To verify the decision support value of the GraphCredit framework in actual credit approval processes, this section systematically evaluates the generated dynamic risk narrative reports from three dimensions: explanatory semantic expression, risk auditing capability, and specific operational actionability. Furthermore, it demonstrates the distinctions between using solely SHAP feature attribution and the proposed GraphCredit framework.

To overcome the subjectivity in traditional manual evaluation, this paper constructs a standardized evaluation framework based on "LLM-as-a-Judge". This framework encompasses three dimensions: Informativeness, Logical & Causal Reasoning, and Decision Support. Specific scoring details are presented in Table 9.

Subsequently, we integrated the evaluation framework into GPT-4o as a prompt to conduct a double-blind quantitative audit on 100 randomly selected test samples. The specific audit results are compared in Table 10, and the core evaluation results are contrasted in Table 11.

The efficacy evaluation results indicate that the dynamic risk narrative reports generated by the GraphCredit framework outperform reports relying solely on SHAP feature attribution across all dimensions, with an overall score improvement of approximately 13.67 points (increasing from 76.01 to 89.68). Specifically:

For the informativeness: reports based on SHAP feature attribution can only list abstract information such as "JobCategory feature contribution +0.12." In contrast, relying on graph-structured evidence chains, the GraphCredit framework provides factual bases such as "The overall default rate of this community is only 1.42%." By fusing concrete entities, quantitative statistics, and business semantics, it enhances information density and comprehensibility.

For the Logical & Causal Reasoning: SHAP values can only indicate the statistical correlation between features and prediction results, but cannot explain the underlying causal mechanisms. The GraphCredit framework constructs a complete evidence chain through COT reasoning, associating and fusing community default probabilities with borrower individual feature contributions. Even if the model determines a borrower to be low risk, the GraphCredit framework can still issue early warnings through neighbor node information (e.g., high-default-rate professional communities).

For the Decision Support: SHAP reports stop at merely indicating high or low risk, whereas the GraphCredit framework can directly generate reasonable decision recommendations by combining community risk.

In conclusion, by fusing feature attribution (SHAP) with logical evidence (graph-structured evidence chains) and utilizing LLMs for semantic integration and narrative generation, the GraphCredit framework transforms explainability outputs from traditional, discrete feature importance into structured, traceable, and dynamic risk narrative reports. This not only enhances the readability and credibility of the reports but also provides a novel paradigm for human–AI collaboration in the field of credit default prediction that balances high accuracy with strong trustworthiness.

5 Conclusions

Addressing the core issues of insufficient explainability in complex models and the difficulty in quantifying and tracing community risk in credit risk prediction, this paper proposes an innovative explainability enhancement framework—GraphCredit. By

Table 8 Ablation experiment results

Strategy	AUC	F1	Recall	KS	Accuracy
Concat	0.8304	0.4513	0.7265	0.4631	0.7167
Attention	0.8347	0.4680	0.7582	0.4699	0.8104
Ours (Gated)	**0.8432**	**0.4861**	**0.7865**	**0.4009**	**0.8306**

Table 9 Detailed table of the standardized assessment framework

Dimension	Weight	Definition
D1: informativeness	30%	Does it have a specific business entity information been provided? Is the report substantial?
D2: logical & causal reasoning	30%	Does it have a reasonable causal chain been constructed?
D3: decision support	40%	Does it have practical guiding significance? Can manual review be achieved based on the report?

Table 10 Analysis of explainability report content

Method	Report$_i$
GraphCredit	Credit risk attribution audit report: Case ID: app_268101 Prediction conclusion: The risk control system predicts this user (app_268101) as a low-risk (normal) client. Core attribution logic: The core logic for the system's low-risk judgment lies in the fact that the user's high-quality personal credit record and the low-risk environment of their community effectively offset the potential risks associated with their occupation and housing status. As indicated by SHAP data, the risk score of the user's region (zip_risk_score: −0.01) makes a clear negative contribution to the overall risk. Combined with the knowledge graph, it is found that the user resides in the "Zip_295" area, where the community's overall default rate is only 1.42%, significantly reducing the user's environmental risk exposure. Furthermore, as shown by SHAP data, longer employment length (emp_length: −0.02) and credit history length (credit_history_length: −0.01) both contribute negatively to the risk. Integrating graph insights reveals the user has a stable 4-year employment record and a credit history spanning 133 months, explaining why the model judges them to possess strong repayment stability and credit awareness. Contradictions and implicit risks: Although the user's personal financial indicators, such as annual income (100,000) and debt-to-income ratio (3.47%), perform excellently, the system still identifies certain risk points. While SHAP data shows that their professional community risk (job_risk_score: 0.01) makes a slight positive contribution to risk, graph analysis reveals the user belongs to the "Finance and Insurance" industry community with a moderate default rate, and their housing status is "Rent" (community default rate 0.55). These are the primary risk sources identified by the system. However, the weight of these negative factors is completely offset by strong positive signals, including the extremely low-risk residential community, stable personal compliance history, and healthy financial status. The risk control system determines that the community contagion risk is at a controllable level in this case. Final recommendation: Based on the system evaluation, the user's credit fundamentals are solid, and the overall risk is controllable. Action: Approve credit application and proceed with loan disbursement according to standard procedures.
SHAP	Judged as a normal user. According to model analysis, the applicant's emp_length is the strongest single feature reducing their predicted risk (contribution value −0.02). Meanwhile, features such as zip_risk_score, credit_history_length, and open_acc (number of open accounts) also collectively reduce the risk. Although annual_inc (annual income) is the primary factor driving up their risk prediction value (contribution value +0.02), its negative impact is offset by multiple positive features. Overall, the sum of risk-reducing feature contributions exceeds that of risk-increasing features; thus, the model determines the user as low risk.

Table 11 Explainability performance scoring table

Method	Dimension	Score	Score_total	AVERAGE
GraphCredit	D1	27	88	89.68
	D2	26		
	D3	35		
SHAP	D1	22	76	76.01
	D2	23		
	D3	31		

*AVERAGE $= (\sum_{i=1}^{100} \text{Score\_total}(\text{Report}_i))/100$

integrating graph structural reasoning, statistics, and LLMs technology, this framework provides a solution for high-risk credit decision-making that combines high accuracy with strong trustworthiness.

The core work and contributions of this paper can be summarized in the following three aspects: First, it proposes a method for the explicit quantification and fusion of community risk based on Bayesian smoothing, transforming community risk into explicit risk features from the three dimensions of occupation, geography, and purpose, thereby enhancing the feature space's capability to represent community risk. Second, it designs a gated graph neural network, Gated-SAGE, tailored for the dynamic fusion of heterogeneous features, which adaptively balances individual features and community risk features through a learnable gating mechanism, effectively improving prediction accuracy. Third, it constructs a generation path from multi-modal evidence to natural language-based dynamic risk narrative reports. It incorporates SHAP feature attribution, graph evidence chains, and the classic 5C credit principles as prior knowledge into LLMs to generate dynamic risk narrative reports that possess comprehensibility, credibility, and completeness. Experimental results indicate that the GraphCredit framework achieves optimal comprehensive performance across multiple credit default prediction benchmarks, surpassing traditional tabular models (such as XGBoost) and state-of-the-art graph models (such as GAT). Ablation analysis further confirms that explicit quantified community risk features are the key source of performance improvement. In the explainability evaluation, the explainability reports generated based on LLMs significantly outperform SHAP feature attribution methods across the three dimensions of Informativeness, Causal Reasoning, and Decision Support, enhancing decision-making efficiency and transparency.

In summary, by integrating the relational advantages of graph structural reasoning, statistical feature attribution, and the semantic generation capabilities of LLMs, the GraphCredit framework provides a feasible technical path and theoretical reference for building trustworthy, reliable, and compliant intelligent credit risk control systems.

Acknowledgement

This work was supported by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011633 and the Educational Commission Key Program of GuangDong Province of China under Grant 2024ZDZX1036.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Article History

Received: 26 December 2025; Revised: 29 March 2026; Accepted: 20 April 2026

References

[1] Nigmonov, A.; Shams, S., COVID-19 pandemic risk and probability of loan default: Evidence from marketplace lending market, *Financial Innovation*, vol. 7, no. 1, pp. Article No. 83, 2021.

[2] Han, D.; Guo, W.; Chen, Y.; Wang, B.; Li, W., Personal credit default prediction fusion framework based on self-attention and cross-network algorithms, *Engineering Applications of Artificial Intelligence*, vol. 133, pp. Article No. 107977, 2024.

[3] Sharma, N.; Ranjan, V. Credit card fraud detection: A hybrid of PSO and K-means clustering unsupervised approach. In: Proceedings of the 13th International Conference on Cloud Computing, Data Science & Engineering, 445–450, 2023.

[4] Gao, W.; Liu, Y.; Yin, H.; Zhang, Y., Social capital, phone call activities and borrower default in mobile micro-lending, *Decision Support Systems*, vol. 159, pp. Article No. 113802, 2022.

[5] Medina-Olivares, V.; Calabrese, R.; Dong, Y.; Shi, B., Spatial dependence in microfinance credit default, *International Journal of Forecasting*, vol. 38, no. 3, pp. 1071–1085, 2022.

[6] Adbi, A.; Lee, M.; Singh, J., Community influence on microfinance loan defaults under crisis conditions: Evidence from Indian demonetization, *Strategic Management Journal*, vol. 45, no. 3, pp. 535–563, 2024.

[7] Lee, J. W.; Lee, W. K.; Sohn, S. Y., Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers, *Expert Systems with Applications*, vol. 168, pp. Article No. 114411, 2021.

[8] Li, Z.; Wang, X.; Yao, L.; Chen, Y.; Xu G.; Lim. E.-P. Graph neural network with self-attention and multitask learning for credit default risk prediction. In: *Web Information Systems Engineering–WISE 2022. Lecture Notes in Computer Science, Vol. 13724*. Chbeir, R.; Huang, H.; Silerstri, F.; Mano opoulos, Y.; Zhang, Y., Eds Springer Cham, 616–629, 2022.

[9] Verdinelli, I.; Wasserman, L., Feature importance: A closer look at shapley values and LOCO, *Statistical Science*, vol. 39, no. 4, pp. 623–636, 2024.

[10] Ying, R.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. *arXiv preprint* arXiv: 1903.03894, 2019.

[11] Li, H.; Wu, W., Loan default predictability with explainable machine learning, *Finance Research Letters*, vol. 60, pp. Article No. 104867, 2024.

[12] Jung, H. S.; Lee, H., Explainable zero-shot trading using multi-agent LLM architecture: A backtested approach for Bitcoin price, *Information Processing & Management*, vol. 63, no. 2, pp. Article No. 104466, 2026.

[13] Han, D.; Guo, W.; Chen, H.; Wang, B.; Guo, Z., LEST: Large language models and spatio-temporal data analysis for enhanced Sino-US exchange rate forecasting, *International Review of Economics & Finance*, vol. 96, pp. Article No. 103508, 2024.

[14] Huang, H.; Li, J.; Zheng, C.; Chen, S.; Wang, X.; Chen, X., Advanced default risk prediction in small and medium-sized enterprises using large language models, *Applied Sciences*, vol. 15, no. 5, pp. Article No. 2733, 2025.

[15] Han, D.; Chen, J.; Guo, Z.; Li, Q.; Wang, B. A novel exchange rate forecasting paradigm based on multi-agent collaboration and multimodal big data-driven methods. *Big Data Mining and Analytics* 2025.

[16] Nie, G.; Rowe, W.; Zhang, L.; Tian, Y.; Shi, Y., Credit card churn forecasting by logistic regression and decision tree, *Expert Systems with Applications*, vol. 38, no. 12, pp. 15273–15285, 2011.

[17] Noriega, J. P.; Rivera, L. A.; Herrera, J. A., Machine learning for credit risk prediction: A systematic literature review, *Data*, vol. 8, no. 11, pp. 169–169, 2023.

[18] Zhou, B.; Jin, J.; Zhou, H.; Zhou, X.; Shi, L.; Ma, J.; Zheng, Z., Forecasting credit default risk with graph attention networks, *Electronic Commerce Research and Applications*, vol. 62, no. November–December, pp. Article No. 101332, 2023.

[19] Zandi, S.; Korangi, K.; Óskarsdóttir, M.; Mues, C.; Bravo, C., Attention-based dynamic multilayer graph neural networks for loan default prediction, *European Journal of Operational Research*, vol. 321, no. 2, pp. 586–599, 2025.

[20] Wang, J.; Liu, G.; Xu, X.; Xing, X., Credit risk prediction for small and medium enterprises utilizing adjacent enterprise data and a relational graph attention network, *Journal of Management Science and Engineering*, vol. 9, no. 2, pp. 177–192, 2024.

[21] Liu, Y.; Wang, X.; Meng, T.; Ai, W.; Li, K., LG-GNN: Local and Global Information-aware Graph Neural Network for default detection, *Computers & Operations Research*, vol. 169, pp. Article No. 106738, 2024.

[22] Ribeiro, M.; Singh, S.; Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 97–101, 2016.

[23] Cil, A. E.; Yildiz, K., A systematic literature review on applications of explainable artificial intelligence in the financial sector, *Internet of Things*, vol. 33, pp. Article No. 101696, 2025.

[24] Lundberg, S. M.; Lee, S. I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 4768–4777, 2017.

[25] Bussmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J., Explainable machine learning in credit risk management, *Computational Economics*, vol. 57, no. 1, pp. 203–216, 2021.

[26] Nallakaruppan, M. K.; Balusamy, B.; Shri, M. L.; Malathi, V.; Bhattacharyya, S., An Explainable AI framework for credit evaluation and analysis, *Applied Soft Computing*, vol. 153, pp. Article No. 111307, 2024.

[27] Golec, M.; AlabdulJalil, M. Interpretable LLMs for credit risk: A systematic review and taxonomy. *arXiv preprint* arXiv: 2506.04290, 2025.

[28] AlMarri, S.; Juhasz, K.; Ravaut, M.; Marti, G.; Al Ahbabi, H.; Elfadel, I. Interpreting LLMs as credit risk classifiers: Do their feature explanations align with classical ML? *arXiv preprint* arXiv: 2510.25701, 2025.

[29] Li, G., Big data and computing models, *Big Data Research*, vol. 10, no. 1, pp. 9–16, 2024.

[30] Chen, X.; Zhou, S.; Liang, K.; Yuan, D.; Chen, H.; Sun, X.; Meng, L.; Liu, X. Putting on the thinking hats: A survey on chain of thought fine-tuning from the perspective of human reasoning mechanism. *arXiv preprint* arXiv: 2510.13170, 2025.

[31] Li, Y.; Shen, Y.; Nian, Y.; Gao, J.; Wang, Z.; Yu, C.; Li, S.; Wang, J.; Hu, X.; Zhao, Y. Mitigating hallucinations in large language models via causal reasoning. *arXiv preprint* arXiv: 2508.12495, 2025.

[32] Zhou, D.; Zhou, B.; Zheng, Z.; Soylu, A.; Cheng, G.; Jimenez-Ruiz, E.; Kostylev, E. V.; Kharlamov, E. Ontology reshaping for knowledge graph construction: applied on Bosch welding case. In: *The Semantic Web –ISWC 2022. Lecture Notes in Computer Science, Vol. 13489*. Sattler, U.; Hogan, A.; Keet, M.; Presutti, V.; Almeida, J. P. A.; Takeda, H.; Monnin, P.; Pirrò, G.; d'Amato, C. Eds. Springer Cham, 770–790, 2022.

[33] Hamilton, W. L.; Ying, R.; Leskovec, J. Inductive representation learning on large graphs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 1025–1035, 2017.

[34] Shen, T.; Wang, H.; Qin, C.; Sun, R.; Song, Y.; Lian, D.; Zhu, H.; Chen, E., Prompting is not enough: Exploring knowledge integration and controllable generation on large language models, *Big Data Mining and Analytics*, vol. 9, no. 2, pp. 563–579, 2026.