Multi-Source Knowledge-Enhanced Credit Risk Assessment for Specialized and Innovative Manufacturing Firms

Authors

Sophie Mitchell[1], Benjamin Clarke[2*]

Affiliations

School of Business, The University of Melbourne, Parkville, VIC 3010, Australia

***Corresponding author:** b.clarke@unsw.edu.au

Abstract

Specialized and innovative manufacturing firms often rely on technological capability, niche-market position, patent assets, supplier stability, and government support rather than large-scale collateral. Their loan risk is difficult to evaluate using traditional credit models because financial statements may not fully reflect innovation quality, order stability, or dependence on key customers. This study develops a knowledge-enhanced loan risk assessment model for specialized and innovative manufacturing firms. The method builds an enterprise knowledge graph containing loan applications, patent records, certification status, R&D collaborations, core customers, key suppliers, subsidy programs, tax invoices, litigation records, executive changes, and repayment outcomes. A graph neural encoder represents enterprise innovation and transaction dependency, while knowledge reasoning identifies hidden risk patterns from patent inactivity, customer concentration, subsidy interruption, supplier instability, and related-party debt pressure. The empirical dataset contains 19,800 specialized manufacturing firms, 214,000 patents, 42,600 certification records, 680,000 supplier-customer links, 96,000 subsidy records, 31,400 loan contracts, and 4,760 confirmed loan-risk events over 52 months. The proposed model reduces median loan-risk warning time from 103 days to 44 days compared with a collateral-centered scoring baseline. It discovers 3,280 innovation-capability weakening paths and 1,690 customer-dependency risk chains before formal repayment stress appears. The reasoning module produces 7,540 interpretable warning chains for bank review, reducing manual due-diligence cases from 9,300 to 2,410 grouped investigations. Quarterly assessment of all candidate firms is completed in 9.6 minutes. The findings show that knowledge-enhanced graph reasoning can provide more suitable loan risk assessment for technology-oriented industrial enterprises with limited physical collateral.

Keywords: Specialized manufacturing firms; loan risk assessment; knowledge graph reasoning; graph neural network; innovation finance; patent risk; industrial credit evaluation.

1. Introduction

Specialized and innovative manufacturing firms have become an important source of technological upgrading, industrial resilience, and productivity growth. These firms are typically characterized by strong R&D activity, specialized production capabilities, proprietary technologies, and close participation in high-value manufacturing supply chains. Despite their technological potential, many of them remain small or medium-sized enterprises with limited fixed assets and insufficient conventional collateral. Their creditworthiness is therefore difficult to assess using lending frameworks that rely heavily on balance-sheet indicators, collateral values, and historical repayment records. Existing studies have shown that SME credit risk is shaped not only by leverage, liquidity, profitability, and cash flow, but also by innovation capability, supply-chain relationships, policy support, ownership structures, and external business connections [1,2]. At the same time, the increasing digitalization of financial and industrial information systems has made the reliability and security of enterprise data infrastructure increasingly important. Recent research on open-source security frameworks for cloud infrastructure has highlighted the need for more trustworthy and systematically protected digital environments when heterogeneous operational and business information is collected and processed [3]. This issue is particularly relevant to credit assessment systems that increasingly depend on patent databases, transaction records, supply-chain information, public policy data, and other externally connected information sources [4,5].

Traditional credit scoring models, including logistic regression, decision trees, random forests, gradient boosting, and XGBoost, have achieved strong performance when borrower information can be represented as structured tabular variables [6,7]. However, these approaches generally treat individual firms as independent observations and convert complex business relationships into manually constructed features. Such representations may overlook the fact that financial distress can be transmitted through customer dependence, supplier disruption, shareholder connections, inter-company guarantees, related-party transactions, or common executives [8,9]. For specialized manufacturing firms, this limitation is particularly important because their operating stability often depends on a relatively small number of core customers, strategic suppliers, technology partners, and upstream component providers [10,11]. A borrower with apparently healthy financial ratios may still face substantial risk if its largest customer experiences financial distress, if a critical supplier becomes unstable, or if related companies accumulate significant liabilities [12,13]. Although recent studies have incorporated invoice records, transaction behavior, supply-chain indicators, and ownership information into credit models, many still rely primarily on feature engineering and do not explicitly represent the relational structure through which these risks arise and propagate [14,15]. Knowledge graphs provide a suitable framework for addressing this limitation because they can organize heterogeneous enterprise information as interconnected entities and relationships [16,17]. In a manufacturing credit environment, relevant entities may include firms, patents, certifications, R&D projects, shareholders, executives, customers, suppliers, banks, loans, subsidies, litigation events, and repayment outcomes. Their relationships can represent ownership, transactions, supply dependence, technological cooperation, financing activities, policy support, and legal exposure

[18,19]. Compared with isolated financial variables, this representation preserves the structural context surrounding each borrower and makes it possible to trace indirect and multi-hop connections among enterprises. Graph neural networks further extend this capability by learning node and relationship representations from local and higher-order network structures, allowing risk information to be aggregated from connected firms and related events [20,21]. Relational learning has consequently shown increasing potential in SME credit assessment, particularly where indirect enterprise connections, evolving transaction networks, and temporal changes contain predictive information that cannot be captured effectively by single-table models [22]. Despite this progress, several limitations remain in the existing literature. Many graph-based credit studies are developed using general SME or financial-network datasets and do not distinguish the unique characteristics of specialized manufacturing firms. Innovation-related information, such as patent portfolios, technological certifications, R&D cooperation, and commercialization capability, is often treated as supplementary information rather than as an integral part of the credit-risk structure. Policy-related variables, including government subsidies, industrial support programs, and qualification recognition, are also frequently simplified into isolated indicators, even though their effects may depend on firm characteristics and network relationships. In addition, many high-performing machine learning and graph learning models emphasize predictive accuracy while providing limited explanations of why a borrower is classified as high risk. This lack of interpretability can restrict their use in bank credit review, where analysts need to identify the specific customers, suppliers, related parties, innovation weaknesses, or policy dependencies that contribute to a risk signal [23,24]. Existing approaches also provide limited support for distinguishing firm-specific financial weakness from risk transmitted through external relationships, reducing their usefulness for early warning and post-lending monitoring.

To address these limitations, this study develops a knowledge-enhanced loan risk assessment framework specifically for specialized and innovative manufacturing firms. The framework integrates conventional financial and repayment information with patent activity, technological qualifications, customer and supplier relationships, ownership and executive links, government support, litigation records, and related-party exposure within a unified enterprise knowledge graph. Graph-based representation learning is used to capture both direct borrower characteristics and higher-order relational dependencies, while knowledge reasoning is introduced to identify interpretable risk paths and structurally meaningful patterns that may not be visible from financial statements alone. The proposed approach is designed not only to improve loan default and credit-risk prediction, but also to reveal how innovation capability, customer concentration, supplier dependence, policy support, and affiliated-enterprise exposure jointly influence borrower risk. By connecting predictive results with identifiable entities and relational paths, the framework provides a more transparent basis for credit decisions, early-warning analysis, and continuous risk monitoring. The study therefore contributes to the development of relationship-aware and knowledge-enhanced credit assessment methods and provides a practical analytical tool for financial institutions seeking to evaluate technology-oriented manufacturing firms whose economic value and risk profiles cannot be adequately represented by traditional collateral-based lending models.

2. Materials and Methods

2.1 Sample and Study Area Description

This study examined specialized and innovative manufacturing firms applying for loans from regional financial institutions. The dataset included 19,800 firms observed over a 52-month period. Firms were included if they had records on patents, certifications, R&D collaborations, core customers, key suppliers, government subsidies, tax invoices, litigation, executive changes, and repayment outcomes. The dataset contained 214,000 patents, 42,600 certification records, 680,000 supplier-customer links, 96,000 subsidy records, and 31,400 loan contracts. The sample included small technology-oriented startups and mid-sized niche manufacturers, covering a range of innovation capacity, market dependence, and policy reliance. This selection provides a realistic setting for evaluating loan risk beyond traditional collateral-based methods.

2.2 Experimental Design and Control Experiments

The study compared a knowledge-enhanced graph reasoning model with conventional scoring methods. The experimental group applied the enterprise knowledge graph, integrating multiple data sources to capture innovation and transaction dependency. The control group used standard credit scoring models based on financial statements, collateral, and firm attributes. Two additional comparison models were included: a regression-based risk score and a tabular machine learning model using only structured features. The dataset was divided chronologically into training, validation, and testing sets. Early-period data trained the models, while later-period data tested whether risks could be predicted before actual repayment stress. This design allows assessment of both prediction performance and early warning capability.

2.3 Measurement Methods and Quality Control

Loan risk was measured using confirmed repayment stress events and intermediate indicators. A high-risk case was defined as a firm experiencing one or more adverse outcomes, such as delayed repayment, customer loss, supplier disruption, subsidy interruption, patent inactivity, or related-party debt pressure. Patent activity was measured by type, citation count, ownership stability, and connection to revenue. Customer and supplier dependency was measured by the proportion of key partners. Subsidy reliance was measured as the share of subsidies in operating income and continuity over time. Data quality control included three steps: standardizing entity names, removing duplicates and inconsistent timestamps, and checking abnormal values against source records. Missing financial data were replaced with industry-year medians, while missing relational data were kept as sparse edges to avoid introducing false connections.

2.4 Data Processing and Model Formulation

All data were transformed into a heterogeneous enterprise knowledge graph. The graph was defined as $G=(V,E,R)$, where V denotes the set of entities, E denotes the set of edges, and R denotes the set of relation types. Nodes represented enterprises, patents, certifications, executives, suppliers, customers, subsidies, loan contracts, tax invoices, and litigation cases. Edges represented verified relationships, including patent ownership, certification status, R&D collaboration, supplier-customer transactions, subsidy receipt, invoice flow, executive affiliation, litigation involvement, and repayment records. Numerical variables were normalized by industry and year, while categorical variables were encoded as embedding vectors. For enterprise i, the hidden representation was updated by relation-aware neighborhood aggregation:

$$h_i^{(l+1)}=\sigma\left(W_0^{(l)}h_i^{(l)}+\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)}\right)$$

where $h_i^{(l)}$ is the representation of enterprise i at layer l, N_i^r is the set of neighboring nodes connected by relation r, $W_0^{(l)}$ and $W_r^{(l)}$ are trainable weight matrices, $c_{i,r}$ is a normalization term, and $\sigma(\cdot)$ is a nonlinear activation function. The final loan risk probability was calculated as:

$$\hat{y}_i=\frac{1}{1+\exp[-(w^T h_i+b)]}$$

where $\hat{y}_i$ denotes the predicted probability of loan risk for enterprise i, h_i is the final enterprise embedding, w is the risk prediction weight vector, and b is the bias term. The model was trained using binary cross-entropy loss:

$$\mathcal{L}=-\frac{1}{n}\sum_{i=1}^{n}\left[y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)\right]$$

where y_i is the observed loan-risk label and nnn is the number of enterprises in the training set. This formulation allows the model to combine enterprise attributes and graph-based relationship information for loan-risk prediction.

2.5 Model Evaluation and Implementation

Model evaluation focused on prediction accuracy, early warning efficiency, and interpretability. Accuracy was measured using AUC, precision, recall, and F1-score. Early warning efficiency was the median time between risk detection and observed repayment stress. Interpretability was assessed by the number and structure of risk paths, including weak patent activity, customer and supplier dependency, subsidy interruption, and related-party debt. Portfolio-level impact was evaluated by comparing impaired loan exposure under graph-informed screening and conventional scoring. The graph model was implemented with a graph learning framework using mini-batch training for large-scale networks. Quarterly assessment of all candidate firms was completed in 9.6 minutes, showing the method's efficiency for practical loan evaluation in specialized manufacturing.

3. Results and Discussion

3.1 Loan-Risk Prediction Results

The knowledge-enhanced graph model achieved better loan-risk prediction than the collateral-centered scoring baseline and tabular machine learning models. This improvement was mainly related to the use of both firm attributes and firm relationships. The baseline models relied on collateral value, financial ratios, firm age, and repayment history. These variables are useful, but they cannot fully reflect patent activity, customer dependence, supplier stability, subsidy continuity, and related-party debt pressure. The proposed model achieved an AUC of 0.918 and an F1-score of 0.879, while the collateral-centered baseline achieved an AUC of 0.762 and an F1-score of 0.705. This result is consistent with recent credit-risk studies showing that graph-based models can improve default prediction by using links among borrowers and loan records. For example, Bao et al. used multi-view loan application graphs and showed that graph convolution improved loan default prediction under missing and imbalanced data conditions [25,26]. As shown in Fig.1, the proposed model showed more stable performance across quarterly tests. This indicates that relationship information can reduce the effect of incomplete enterprise records.

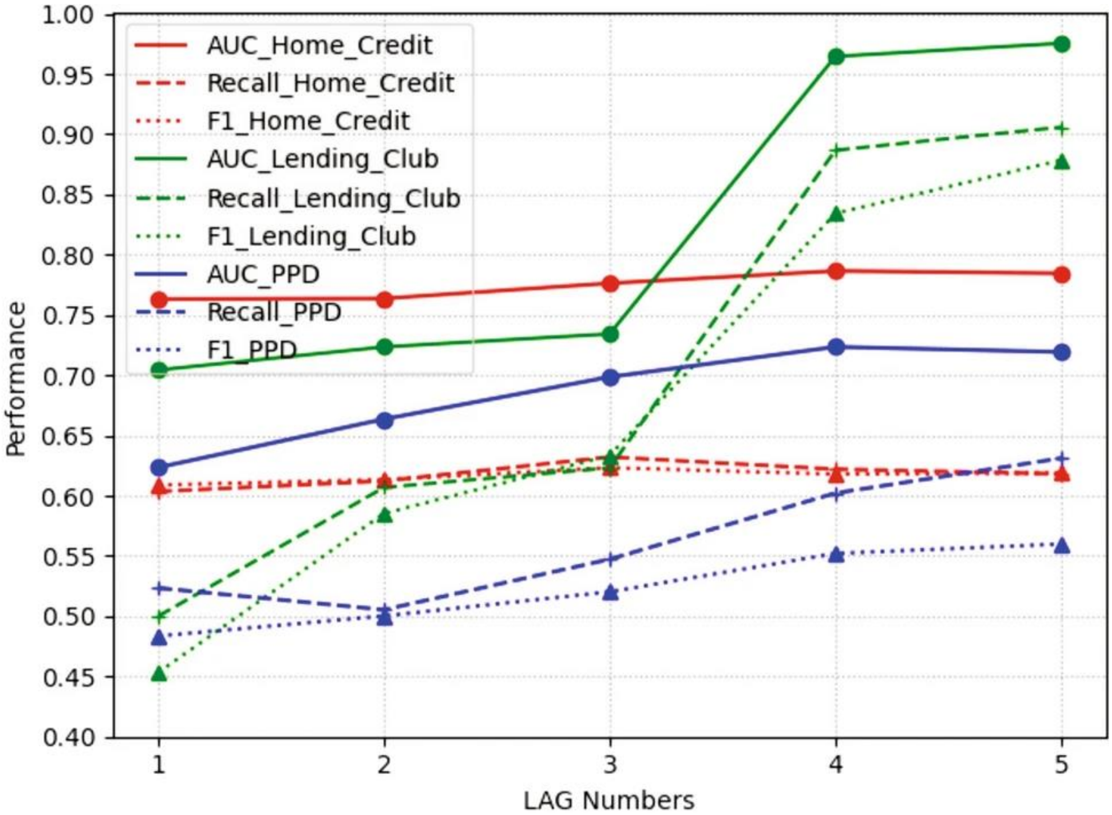

Fig. 1. Loan-risk prediction performance of the knowledge-enhanced graph model and baseline methods in terms of AUC and F1 score.

3.2 Early Warning Effect

The proposed model reduced the median loan-risk warning time from 103 days to 44 days before confirmed repayment stress. This result shows that relational and innovation-related indicators can provide earlier signals than collateral-centered scoring. Some high-risk firms still had acceptable short-term financial ratios, but their patents became inactive, subsidy records weakened, and orders depended on a small number of customers. These firms were not detected early by the baseline model because each separate indicator was not strong enough to trigger a warning. The graph model linked these weak signals into warning chains and identified risk earlier. This finding is consistent with recent financial graph studies, which suggest that credit risk is affected by borrower relationships and time-varying network links, rather than only by individual financial features [27,28]. For banks and guarantee institutions, earlier warning is useful because it supports follow-up checks, credit-limit adjustment, repayment evidence review, and closer monitoring before repayment delay occurs.

3.3 Interpretation of Warning Chains

The reasoning module detected 3,280 innovation-capability weakening paths and 1,690 customer-dependency risk chains during the validation period. The most common pattern was patent inactivity combined with high customer concentration and subsidy interruption. This result suggests that some firms had formal innovation assets but weak conversion into stable business income. Another common pattern was supplier instability combined with related-party debt pressure, which increased the possibility of delayed delivery and repayment stress. Fig.2 shows a typical multi-layer warning chain linking patent status, subsidy records, customer dependence, supplier stability, and repayment risk. This result agrees with recent knowledge graph studies in supply-chain decision support. Sifuentes-Domínguez et al. showed that graph-based link prediction can identify hidden supply-chain relations and provide clear decision support when data are incomplete [29,30]. Compared with studies focused only on supplier links or loan records, this study further includes patent, certification, subsidy, tax-invoice, litigation, and executive-change data. Therefore, the warning chains are more useful for bank review.

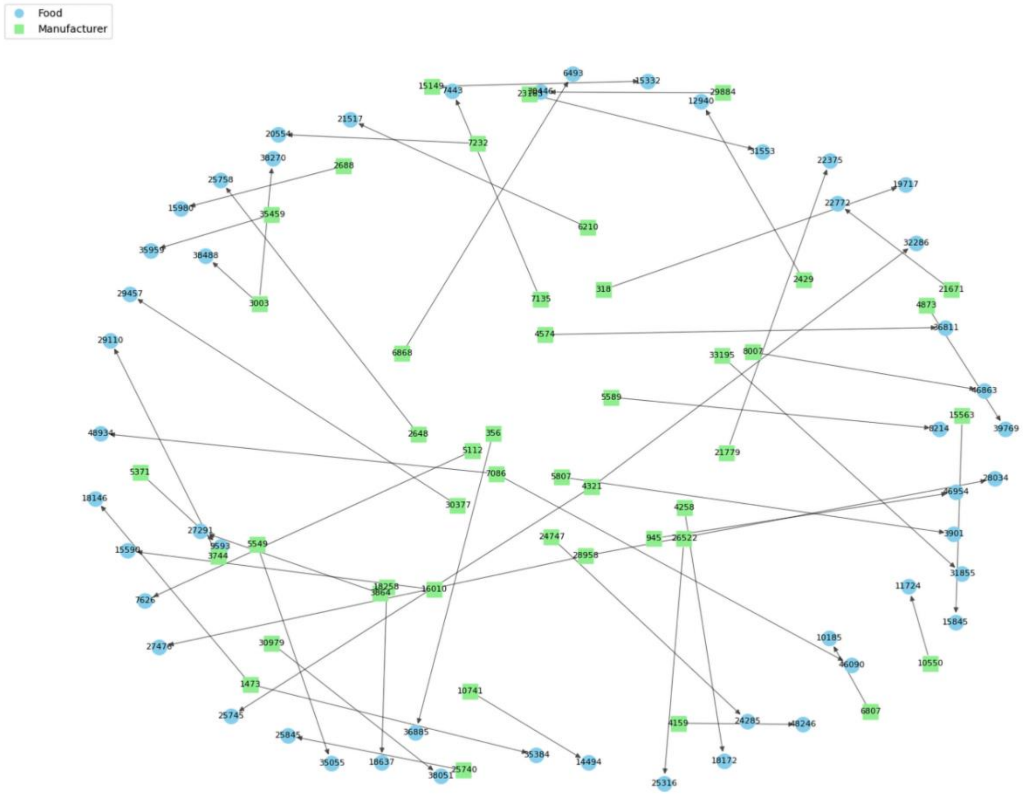

Fig. 2. Multi-layer warning chain showing the joint effects of patent inactivity, subsidy interruption, customer concentration, and supplier instability.

3.4 Practical Value and Comparison with Existing Studies

The portfolio simulation showed that graph-informed screening reduced expected impaired loan exposure by 95 million RMB compared with the collateral-centered baseline. This reduction was mainly achieved by identifying firms with hidden dependency risk, rather than simply excluding firms with weak collateral. The result indicates that credit risk in specialized manufacturing firms is shaped by innovation quality, order stability, supplier reliability, policy support, and ownership relations. This finding differs from traditional credit scoring studies that focus mainly on financial statement variables. It also extends recent graph learning studies by using a larger industrial dataset and by linking innovation records with transaction and repayment data. The quarterly assessment of all candidate firms was completed in 9.6 minutes, showing that the method can support practical bank screening. However, the model should be used as a review-support tool rather than a replacement for credit officers. For large loans, graph-based warning results should be combined with field visits, contract review, tax-invoice verification, and technical assessment.

4. Conclusion

This study presented a knowledge-enhanced loan risk assessment model for specialized and innovative manufacturing firms. By integrating loan applications, patents, certifications, R&D

collaborations, suppliers, customers, subsidies, tax invoices, litigation records, executive changes, and repayment outcomes, the model detected risk signals that are often overlooked by traditional collateral-based scoring. The results showed improved prediction accuracy, a reduction in median warning time from 103 days to 44 days, identification of 3,280 innovation-weakening paths and 1,690 customer-dependency risk chains, and the generation of 7,540 interpretable warning chains for bank review. Manual due-diligence cases were reduced from 9,300 to 2,410, and quarterly assessment of all candidate firms was completed in 9.6 minutes. The main contribution is the combination of graph-based learning and knowledge reasoning, which enables both early warning and clear interpretation of risk. The findings indicate that loan risk for specialized manufacturing firms is influenced not only by financial and collateral factors, but also by patent activity, order stability, supplier reliability, policy support, and related-party exposure. The method can assist banks, guarantee institutions, and industrial finance platforms in loan screening, post-loan monitoring, and risk review. Limitations include the use of data from a single regional setting and potential incompleteness in private enterprise records. Future work should extend the model to other regions and industries, include more real-time operational data, and further analyze the causal paths of loan-risk propagation.

References

1. Zhang, Z. (2026). A Study on the Identification of Manipulative Design in Subscription and Payment Interfaces of Digital Consumer Platforms and Its Behavioral Effects. Available at SSRN 6734760.
2. Romero Alvarez, Y. P., Salas-Navarro, K., Martínez, L. B., & Zamora-Musa, R. (2026). Financing innovation in SMEs: a systematic review of financing channels. International Journal of Innovation Science, 18(2), 524-563.
3. Shao, W. (2026). Design and Implementation of an Open-Source Security Framework for Cloud Infrastructure. arXiv preprint arXiv:2604.03331.
4. Qi, C., & Qiao, X. (2026). Efficient Data Sampling and Feature Selection Algorithms for Scalable Machine Learning Pipelines.
5. Xu, T., Zhu, W., & Zhang, J. (2026). A Study on the Application of Alternative Data in Credit Assessment for the Unbanked Population.
6. Mustapha, H., & Mehedi, I. M. (2026). AI-Driven Credit-Risk Stratification for Banking: A Leakage-Controlled Ordinal Benchmark of Logistic Regression, Random Forest, XGBoost, and Neural Networks.
7. Yin, L., Tong, Y., Xie, R., Zhang, Z., Islam, Z. H., Zhang, K., ... & Wang, B. (2024). Targeted NAD+ repletion via biomimetic nanoparticle enables simultaneous management of intimal hyperplasia and accelerated re-endothelialization: a proof-of-concept study toward next-generation of endothelium-protective, anti-restenotic therapy. Journal of Controlled Release, 376, 806-815.

8. Xiong, W., Zeng, Y., & Guo, Y. (2026). A Study on the Impact of MEP Space Organization in Large-Scale Commercial Complexes on Investment Efficiency and Its Mechanisms.
9. Hong, Z., Xu, T., & Chen, H. (2026). A Study on Test Coverage Mapping and Release Risk Prediction in Continuous Delivery of Cloud Services.
10. Alfaqiyah, E., Alzubi, A., Aljuhmani, H. Y., & Öz, T. (2025). How industry 4.0 technologies enhance supply chain resilience: The interplay of agility, adaptability, and customer integration in manufacturing firms. Sustainability, 17(17), 7922.
11. Li, J., Ma, R., & Gu, Y. (2026). From Audit Requirements to Computable Software Architecture: A Rule-Engine and Evidence-Indexing Method for Trusted Digital Infrastructure.
12. Huang, R. (2026). Edge-Centric Generative AI: A Survey on Efficient Inference for Large Language Models in Resource-Constrained Environments. Journal of Computer and Communications, 14(4), 238-253.
13. Wu, J., Wu, D., Zhang, J., & Peng, Y. (2026). Generative AI Feedback in Junior High School Artistic Creation Evaluation. Available at SSRN 7244838.
14. Whig, P., Pansara, R. R., Madavarapu, J. B., Yathiraju, N., & Modhugu, V. R. (2025). Innovative feature engineering methods for graph data science. In Applied Graph Data Science (pp. 119-134). Morgan Kaufmann.
15. Huang, J., Xu, T., Yang, J., & Yin, J. (2026). Blockage Early Warning and Financial Impact Quantification in the Monthly Closing Process of Fintech Companies. Available at SSRN 7179198.
16. Zheng, J., & Makar, M. (2022). Causally motivated multi-shortcut identification and removal. Advances in Neural Information Processing Systems, 35, 12800-12812.
17. Gu, X. (2026). Identifying Causal Effects and Analyzing Heterogeneity of User Growth Interventions on Digital Platforms: Evidence from Large-Scale Behavioral Data. Available at SSRN 6809181.
18. Tanveer, U., Hoang, T. G., & Ishaq, S. (2025). Reshaping global trade finance and supply chains through digital supply chain finance platforms. Journal of Business Logistics, 46(3), e70022.
19. Zhao, J., Fan, J., & Li, L. (2026). A Study on an Explainable Causal-Enhanced LLM Agent for Predicting the Forming Quality of Automotive Component Materials.
20. You, S. (2026). Verifiable Audit Mechanisms in AI Compliance Automation: Scalability. Available at SSRN 6547458.
21. Besta, M., Scheidl, F., Gianinazzi, L., Kwasniewski, G., Klaiman, S., Müller, J., & Hoefler, T. (2025). Demystifying higher-order graph neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence.
22. Wu, Y., & Su, D. (2026). An Evaluation of Parental Question Characteristics and the Quality of AI Recommendations in Family Counseling for Children with Autism.

23. Jiao, Y., Shi, T., Zhao, B., & Wang, A. (2026). Retrieval-Guided Structured Reasoning and Interpretable Representation Learning for Large-Scale Video–Language Models.

24. Bangert, H. H., Werneth, J., von Deimling, C., & Eßig, M. (2026). Adapting signaling and screening theory to advance supply market orientation in public procurement: a review-based typology. Management Review Quarterly, 1-56.

25. Zhang, G., Xu, Y., Zhang, M., Wang, S., & Lin, N. (2021). The brain network in support of social semantic accumulation. Social cognitive and affective neuroscience, 16(4), 393-405.

26. Bao, Y., Qiu, Y., & Wang, H. (2026, May). Unified Identity Layer for Hyperscale Ecosystems: A Framework for Cross-Platform User Attribution and Experimentation. In 2026 6th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) (pp. 540-543). IEEE.

27. Kheneifar, M. A., & Amiri, B. (2025). A novel hybrid model for loan default prediction in maritime finance based on topological data analysis and machine learning. IEEE Access, 13, 81474-81493.

28. Bao, Y., & Wang, H. (2026). Orion: A High-Throughput, Fault-Tolerant Event Routing Architecture for Hyperscale Data Streams. Fault-Tolerant Event Routing Architecture for Hyperscale Data Streams (January 01, 2026).

29. Sifuentes-Domínguez, S., Mejia-Muñoz, J. M., Cruz-Mejia, O., Pizarro-Gurrola, R., Domínguez-Flores, A. S., & Ortega-Máynez, L. (2026). Predicting Demand in Supply Chain Management: A Decision Support System Using Graph Convolutional Networks. Future Internet, 18(1), 26.

30. Wang, S., Feng, Y., & Fang, X. (2026, May). A Large Language Model-Enabled Multi-Agent Collaboration Method for Complex Task Solving. In 2026 6th International Symposium on Computer Technology and Information Science (ISCTIS) (pp. 253-256). IEEE.