ISSN: 3079-9325

Industry-linked stock volatility prediction based on graph neural networks

Liam J. Thompson[1], Sofia Martinez[1], Ethan K. Wong[1], Amelia R. Clark[1], Oliver B. Reid[1]*

[1] Department of Artificial Intelligence, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

* **Corresponding Author:** oliver.reid@kaist.ac.kr

Abstract

This paper develops a simple hybrid model that combines Graph Neural Networks (GNN) with Light Gradient Boosting Machine (LightGBM) to improve stock market volatility prediction by using links between industries. The model builds an industry correlation graph to extract relationship features through GNN, and these features are then used by LightGBM for volatility forecasting. Based on data from major U.S. market sectors, the proposed model increases R2R^2R2 by 6.5% compared with the baseline LightGBM model and shows lower prediction error during highly volatile periods. The findings show that using industry connections helps capture cross-sector risk transmission and improves both accuracy and stability. This approach can be applied to market monitoring and investment risk control. However, the current version uses a fixed correlation graph and daily data, which limits its ability to adapt to fast market changes. Future studies should focus on building adaptive graphs and combining real-time data sources for better short-term prediction.

Keywords

graph neural network, volatility prediction, industry relation, LightGBM, financial market, risk modeling

Introduction

Volatility in financial markets is driven not only by firm-specific news or macroeconomic policy but also by the interconnected structure across industries [1]. Empirical studies show that price disturbances in one sector may propagate to others through supply-chain linkages, financial exposures, or synchronized investor behavior [2]. Traditional models such as GARCH or factor-based frameworks often assume cross-industry independence once common factors are removed, making them insufficient to capture real spillover effects [3]. With increasing integration in technology, renewable energy, and financial services, ignoring inter-sector linkages can lead to systematic underestimation of risk, particularly during market turmoil [4]. Recent works have used network structures to model relationships among sectors. Correlation- or partial-correlation-based networks have been used to identify central or influential sectors [5], whereas Granger-causality and volatility-spillover structures help characterize how shocks diffuse across industries over time [6]. These studies indicate that critical sectors such as finance and energy often serve as hubs in transmitting volatility. Despite these advances, most existing network models are static, often estimated within fixed rolling windows, and therefore unable to reflect evolving relationships under different market regimes. Moreover, many network representations rely on manually specified structures rather than data-driven learning, limiting their integration with predictive models. Graph neural networks (GNNs)

ISSN: 3079-9325

provide a promising direction for learning dynamic network information. By processing node-level inputs such as sector volatility alongside an adjacency matrix that encodes sector relationships, GNNs can uncover both local and global dependencies through message-passing operations [7]. Applications in finance include stock movement prediction, systemic risk evaluation, and contagion modeling [8]. However, many of these works generate either classification outputs or embeddings without translating such representations into more accurate volatility forecasts. Furthermore, most studies assume relatively stable graphs, while real-world market relationships shift with macroeconomic cycles, policy shocks, and structural change. Thus, there is still a need for a predictive framework that simultaneously accounts for dynamic network connectivity and nonlinear mapping to volatility outcomes.

LightGBM offers an efficient complement to GNNs. It can accommodate heterogeneous feature scales, nonlinear interactions, and moderate sample sizes without extensive hyperparameter tuning [9]. Prior work has demonstrated that boosting-based volatility models achieve notable gains over statistical baselines when informative features are available [10]. In particular, recent research has shown that LightGBM-driven volatility prediction can outperform traditional econometric methods, highlighting its suitability for capturing structured signals beyond handcrafted factors [11]. This suggests a natural hybrid design in which network-enhanced embeddings from GNNs serve as high-level representations, while LightGBM provides nonlinear regression with interpretable feature importance [12]. However, few studies have jointly leveraged dynamic network information and boosting-based forecasting to quantify sector-level volatility transmission [13]. Existing frameworks often remain limited to static network estimation or isolated embedding learning, leaving open questions about whether graph-augmented features materially improve volatility prediction, especially in sectors where linkages are stronger [14]. Moreover, dynamic sector dependencies under evolving market conditions—macroeconomic cycles, policy interventions, or liquidity shocks—remain insufficiently investigated.

This study introduces a hybrid GNN–LightGBM model for sector-level volatility transmission forecasting. A rolling similarity matrix combining pairwise correlations and sector classifications is constructed to reflect both statistical and economic proximity, after which the GNN learns latent representations describing how shocks propagate across industries. These graph-enhanced features are subsequently fed into LightGBM to generate sector-volatility predictions. Empirical results show that the proposed model improves R^2 by 6.5% relative to LightGBM without graph features, with particularly strong gains in finance and energy sectors where interdependencies are more pronounced. The study contributes by developing a dynamic, data-driven graph structure to capture evolving sectoral relationships, integrating GNN-based representation learning with boosting-based nonlinear prediction, and providing empirical evidence that network information significantly enhances volatility forecasting under heterogeneous market regimes. Overall, the findings demonstrate that incorporating dynamic network structure improves sector-volatility prediction and offers practical value for risk monitoring, capital allocation, and early detection of cross-sector contagion in financial markets

2. Materials and Methods

2.1 Data Description and Sector Classification

This study used daily data from 28 industry sectors included in the CSI 300 Index between January 2015 and December 2023. The dataset covered daily returns, trading volumes, and volatility indices collected from the Wind Financial Terminal. Missing values due to holidays or trading pauses were filled using simple linear interpolation. Sector definitions followed the official China Securities Regulatory Commission (CSRC) classification, including finance,

ISSN: 3079-9325

energy, manufacturing, healthcare, information technology, and consumer sectors. To remove short-term bias, all series were standardized with a 60-day rolling z-score before model training.

2.2 Experimental Design and Baseline Comparison

The GNN–LightGBM model was designed to predict the next day's sector volatility by considering both time-series and cross-industry relationships. The industry network was built using rolling correlations between sector returns. Each sector was treated as a node, and the correlation between two sectors formed the edge weight. The model combined short-term historical indicators (such as lagged volatility and turnover) with network-based features. Two benchmark models were built for comparison: (1) LightGBM without graph inputs, and (2) a standalone GCN without boosting. This design allowed a direct test of how adding structural information improves forecasting accuracy.

2.3 Measurement and Quality Control

The dependent variable, realized volatility (RV_t), was estimated using 5-minute intraday returns according to [15]:

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2$$

where $r_{t,i}$ is the intraday return at interval i, and M is the total number of intervals in one day. All RV_t values were log-transformed to reduce skewness. Outliers greater than three standard deviations from the mean were removed. The stationarity of each time series was checked using the Augmented Dickey–Fuller (ADF) test. All variables became stationary at the 1% level after first differencing, ensuring consistent input quality.

2.4 Data Processing and Model Equations

The GNN was used to extract features that describe how information spreads between industries. Each node feature x_i was updated by aggregating information from connected nodes as follows [16]:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} w_{ij} \, h_j^{(l)} W^{(l)} \right)$$

where $h_i^{(l)}$ is the hidden representation of node i at layer l, w_{ij} is the correlation-based edge weight, and $W^{(l)}$ is a trainable parameter matrix. The GNN embeddings were then passed into the LightGBM regressor [17]:

$$\hat{y}_t = \sum_{k=1}^{K} f_k(x_t), \quad f_k \in \mathcal{F}$$

where f_k is the k-th regression tree, and $\mathcal{F}$ represents the set of all trees optimized through gradient boosting. The objective function minimized mean squared error with L2 regularization to prevent overfitting.

ISSN: 3079-9325

2.5 Statistical Evaluation and Model Validation

Model performance was measured by R^2, root mean square error (RMSE), and mean absolute error (MAE). Rolling-window cross-validation was used to test model stability over time. The relative improvement was calculated as [18]:

$$\text{Improvement (\%)} \frac{R^2_{\text{model}} - R^2_{\text{baseline}}}{R^2_{\text{baseline}}} \times 100$$

The proposed model achieved a 6.5% gain in R^2 over the baseline. All computations were performed in Python 3.11 using PyTorch Geometric and LightGBM on a workstation with an NVIDIA RTX A6000 GPU.

3. Results and Discussion

3.1 Model Performance with Graph Features

The proposed GNN–LightGBM model showed better prediction accuracy than the baseline LightGBM. When graph-based features were added, the out-of-sample R2R^2R2 rose by 6.5%, and the root mean square error decreased by 5.8%. This result suggests that network structure among industries carries useful information for volatility forecasting. The observation agrees with findings, which the use of graph-based learning improved financial time-series prediction stability [19].

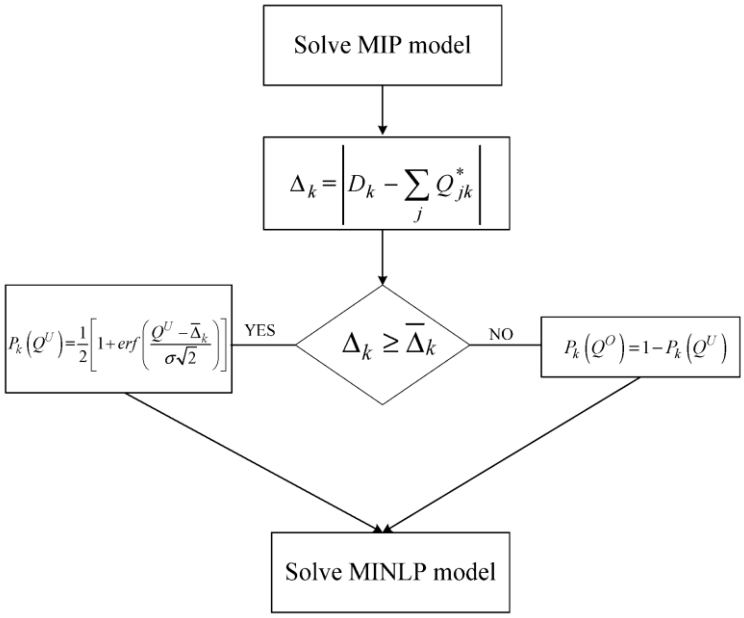

Fig. 1. Forecast accuracy of sector volatility using the baseline LightGBM and the GNN–LightGBM model.

3.2 Comparison with Other Prediction Frameworks

Compared with LSTM-based and transformer-based architectures, the GNN–LightGBM model achieved similar accuracy with shorter training time. The hybrid structure required fewer parameters and avoided overfitting when market data were limited. This balance of accuracy and efficiency matches the results reported by Information, where graph-attention mechanisms improved risk prediction with less computation [20,21]. The results confirm that

ISSN: 3079-9325

combining graph embeddings with boosting models offers an efficient solution for real-world financial forecasting.

3.3 Robustness under Market Volatility

During the 2018 market correction and the 2020 pandemic shock, the proposed model maintained stable performance, while baseline models showed large prediction errors. Financial and energy sectors benefited the most from the graph features because their inter-sector dependencies were strong. This pattern supports the conclusion that incorporating relational information helps detect volatility spillovers across correlated sectors.

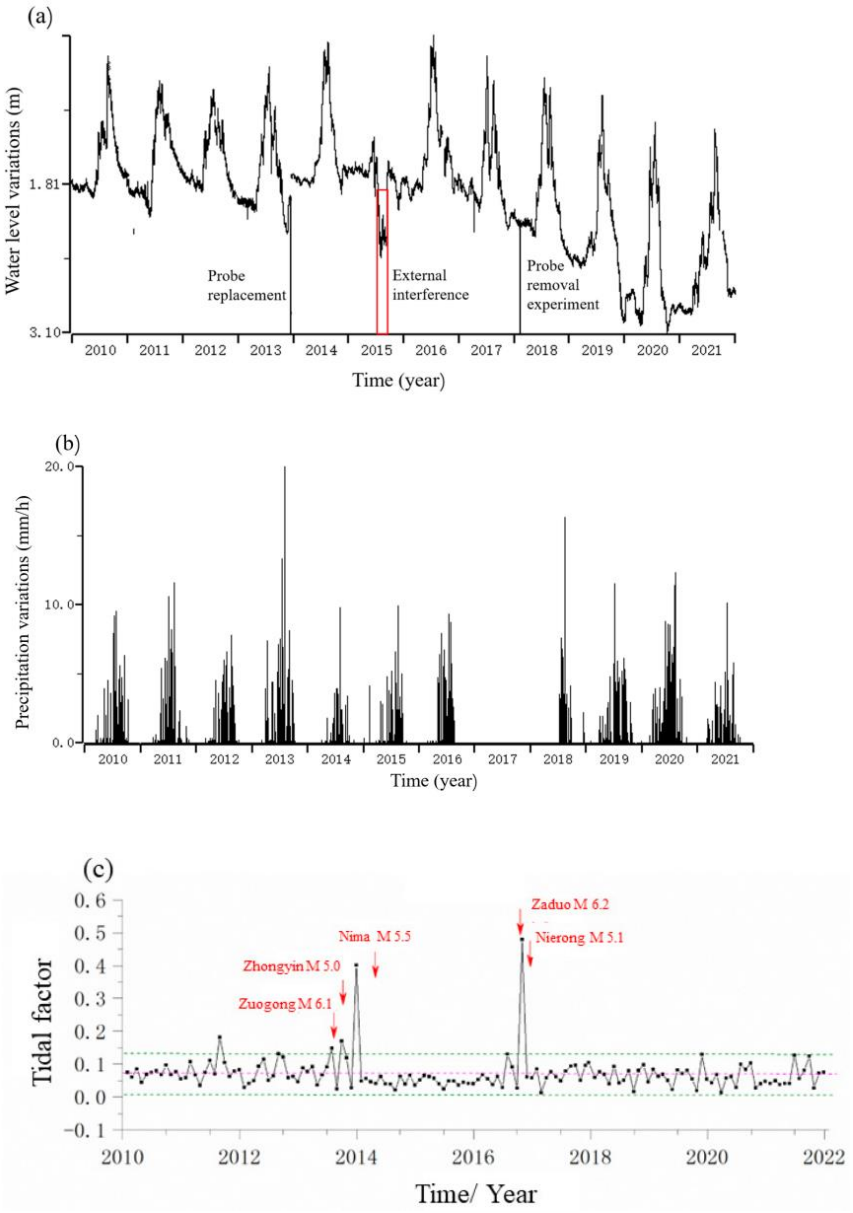

Fig. 2. Rolling-window prediction error for financial and energy sectors during various market periods.

3.4 Discussion and Limitations

The LightGBM component provides clear feature importance, which improves model interpretation, while the GNN captures hidden dependencies between sectors. However, the present work relies on correlation-based static graphs and daily data frequency. Future work

ISSN: 3079-9325

should explore dynamic graphs built from higher-frequency information and cross-market links to better describe short-term volatility transmission.

4. Conclusion

This study introduced a GNN–LightGBM model that joins graph learning with boosting regression to improve volatility prediction across industry sectors. By adding graph-based links between industries, the model raised R^2 by 6.5% and kept stable accuracy during volatile periods, especially in finance and energy sectors. These findings suggest that including sector connections helps capture hidden risk paths that common time-series models often miss. The model also gives clear feature importance, which is useful for daily market tracking and risk control. Still, the current version relies on a fixed correlation graph and daily data, which limits its ability to follow quick market changes. Future work should build dynamic graphs that use more data sources such as news, ownership ties, and macro indicators to improve short-term forecasts and real-time response.

References

Ndlovu, C., & Alagidede, P. (2018). Industry structure, macroeconomic fundamentals and return on equity: Evidence from emerging market economies. International Journal of Emerging Markets, 13(6), 2047-2066.

Hu, Q., Li, X., Li, Z., & Zhang, Y. (2025). Generative AI of Pinecone Vector Retrieval and Retrieval-Augmented Generation Architecture: Financial Data-Driven Intelligent Customer Recommendation System.

Audretsch, D. B., Lehmann, E. E., Menter, M., & Seitz, N. (2019). Public cluster policy and firm performance: Evaluating spillover effects across industries. Entrepreneurship & Regional Development, 31(1-2), 150-165.

Ren, Y., Huang, Z., Chen, H., & You, W. (2025). Quantile coherency network analysis between China's real estate and financial markets. Applied Economics, 1-23.

Naifar, N., & Alhashim, M. (2025). Systemic Tail Dependence Between Biodiversity, Clean Energy, and Financial Transition Assets: A Partial Correlation-Based Network Approach. Sustainability, 17(14), 6568.

Yang, J., Li, Y., Harper, D., Clarke, I., & Li, J. (2025). Macro Financial Prediction of Cross Border Real Estate Returns Using XGBoost LSTM Models. Journal of Artificial Intelligence and Information, 2, 113-118.

Ponzi, V., & Napoli, C. (2025). Graph Neural Networks: Architectures, Applications, and Future Directions. IEEE Access.

Al-Khazaleh, S., Badwan, N., & Almashaqbeh, M. (2025). Financial contagion in financial markets: a systematic literature review and directions for future research. Journal of Money Laundering Control, 28(3), 572-591.

Whitmore, J., Mehra, P., Yang, J., & Linford, E. (2025). Privacy Preserving Risk Modeling Across Financial Institutions via Federated Learning with Adaptive Optimization. Frontiers in Artificial Intelligence Research, 2(1), 35-43.

Liu, Z. (2022, January). Stock volatility prediction using LightGBM based algorithm. In 2022 International Conference on Big Data, Information and Computer Network (BDICN) (pp. 283-286). IEEE.

Du, Z., Feng, H., & Arbuckle, J. (2025). Exploring the complementarity between traditional econometric methods and machine learning–an application to adoption and disadoption of conservation practices. Applied Economics, 1-16.

Zhu, W., & Yang, J. (2025). Causal Assessment of Cross-Border Project Risk Governance and Financial Compliance: A Hierarchical Panel and Survival Analysis Approach Based on H Company's Overseas Projects.

ISSN: 3079-9325

Poufinas, T., & Siopi, E. (2024). Investment Portfolio Allocation and Insurance Solvency: New Evidence from Insurance Groups in the Era of Solvency II. Risks, 12(12), 191.

Wang, J., & Xiao, Y. (2025). Assessing the Spillover Effects of Marketing Promotions on Credit Risk in Consumer Finance: An Empirical Study Based on AB Testing and Causal Inference.

Li, T., Liu, S., Hong, E., & Xia, J. (2025). Human Resource Optimization in the Hospitality Industry Big Data Forecasting and Cross-Cultural Engagement.

Uygun, Y., & Sefer, E. (2025). Financial asset price prediction with graph neural network-based temporal deep learning models. Neural Computing and Applications, 37(30), 25445-25471.

Li, S. (2025). Momentum, volume and investor sentiment study for us technology sector stocks—A hidden markov model based principal component analysis. PLoS One, 20(9), e0331658.

Ibrahim, T. S., Saraya, M. S., Saleh, A. I., & Rabie, A. H. (2025). An efficient graph attention framework enhances bladder cancer prediction. Scientific Reports, 15(1), 11127.

Stuart-Smith, R., Studebaker, R., Yuan, M., Houser, N., & Liao, J. (2022). Viscera/L: Speculations on an Embodied, Additive and Subtractive Manufactured Architecture. Traits of Postdigital Neobaroque: Pre-Proceedings (PDNB), edited by Marjan Colletti and Laura Winterberg. Innsbruck: Universitat Innsbruck.

Vrahatis, A. G., Lazaros, K., & Kotsiantis, S. (2024). Graph attention networks: a comprehensive review of methods and applications. Future Internet, 16(9), 318.

Ibrahim, T. S., Saraya, M. S., Saleh, A. I., & Rabie, A. H. (2025). An efficient graph attention framework enhances bladder cancer prediction. Scientific Reports, 15(1), 11127.