TriFusionRiskNet: Graph-Guided Multimodal Fusion for Joint Stock Return and Risk Prediction

Haotong Chen

RCF Experimental School, Beijing, China

chenhaotong@rdfzcygj.cn

Keywords: Financial Risk Forecasting; Graph neural network; Stock Return and Volatility Prediction

Abstract: Financial markets are nonlinear, dynamic, and complex in nature with structural, temporal, and textual interactions. Nevertheless, the current prediction models are based on either single modality or naive fusion, and cannot reflect inter-company risk spreading and cross-modal effects. To mitigate this weakness, this paper suggests TriFusionRiskNet, the single multi- modal neural architecture for simultaneous prediction of stock returns and risk variances. The structure incorporates three dualistic modalities: a temporal graph encoder, which models the changing relationship between firms; a time-series encoder, which models the multi-scale dynamics of historical trading; and a BERT-based text encoder, which parses sentiment-driven market information in financial news. To selectively direct the information between modalities with structural guidance, a graph-guided cross-modal attention mechanism is proposed, and a dynamic graph learning module is proposed to continually adjust according to actual regime changes in the market. Considerable testing of CSI300 A-share data in 2016-2023 shows that TriFusionRiskNet is much better in predictive accuracy and risk stability compared to the state-of-the-art baselines, especially in volatile market situations. The suggested framework is structurally based, interpretable and deployment-ready multimodal financial forecasting.

1. Introduction

Financial markets are complex, nonlinear and dynamically changing systems, which operate under the influence of a variety of heterogeneous factors. The classical econometric models used in the prediction of asset returns and volatilities include ARIMA and GARCH [1, 2]. But as the number of financial data, to which high-frequency trading records and corporate fundamentals are added, the unstructured textual data of news, social media, analyst reports, etc., is becoming more and more important to market behavior that is no longer adequately explained by numerical data. In addition, the relationship aspect of the contemporary economies, through interconnected supply chains, cross-industry ownership and global networks of production all contribute a significant role in propagating systemic risks and spreading market shocks [3, 4]. Such realities have prompted the establishment of multimodal learning systems that combine heterogeneous information sources in order to have a more holistic view of the expected returns and the risks involved.

Although there is substantial improvement in machine learning-based stock prediction, the current research has its share of limitations. The majority of research restricts their attention to price future predictions or volatility schemes based on the time-series unimodal data, but the structural and textual information that comprises useful contextual clues is disregarded [5, 6]. This oversight results in incomplete models that cannot reflect inter-company relationships, such that the shock in one organization can propagate via supply or investment chains [7, 8]. Though there are multimodal methods that attempt to integrate textual sentiment and numeric attributes [9, 10, 11], common to them is the fact that they process one modality at a time leading to disjointed representations and poor cross-modal interactions. Many more recent techniques that combine graph neural networks and transformer architectures do not attempt to learn the flow of information through heterogeneous spaces, despite modalities being treated as parallel streams. As a result, the existing models are unable

  DOI: 10.25236/iwmecs.2025.060

to find consistent joint forecasts of returns and risks on a single, explainable structure.

Financial markets are complicated, nonlinear and dynamic systems, which work under the pressure of a multitude of heterogeneous factors. ARIMA and GARCH are the classes of the classical econometric models that have been employed in asset returns and volatilities prediction [1, 2]. However, as more financial data, to which high-frequency trading data and corporate fundamentals are being introduced, the unstructured textual information of news, social media, analyst reports, etc., is being increasingly significant to the market behavior that is no longer sufficiently accounted by numerical data. Moreover, the connection component of the modern economies, among reliant supply chains, cross-industrial proprietorship and overseas networks of production also play a major role in distributing systemic risks and propagating market shocks [3, 4]. These realities have led to introduction of multimodal learning systems which involve syntheticizing the heterogeneous source of information with the aim of possessing a more holistic perspective of the anticipated returns and the risk involved.

Despite the significant improvement in machine learning-based predictions of stocks, the present study has its fair share of limitations. Most of the studies limit their focus on price future forecasting, or volatility plans as per the time-series unimodal data, but the structural and textual details, which form valuable contextual indicators, are overlooked [5, 6]. This neglect translates to unfinished models that cannot exhibit inter-company relationships that the shock experienced in a given organization can be spread through supply or investment chain [7, 8].

Although there do exist multimodal approaches which seek to combine textual sentiment and numeric properties [9, 10, 11], features common to them include the fact that their processes process one of their modality at a time resulting in incoherent representations and poor cross modal interactions. Most more modern approaches that model graph neural networks and transformer networks do not aim to learn the information flow in heterogeneous spaces, though modalities are considered parallel streams. Consequently, the available models cannot obtain joint predictions of returns and risks on an explainable single structure.

There are several fundamental problems with the creation of an efficient multimodal neural network that can concurrently model the anticipated return and variance. One of the basic issues is heterogeneous modality alignment: the format, the scale, and the semantic abstraction of graphs, and textual data differ greatly, and it is not an easy task to acquire all of them at once. Dynamic structural dependencies are also common in real markets, since the interrelations among corporations vary over time due to mergers, supply chain disruptions, or new policy; so static network representations cannot capture these changes [12]. Another aspect that suggests strong multimodal fusion is particularly hard is the non-stationarity and extreme events, as well as unclear sentiment of financial data. Moreover, the joint maximization of both returns and risks need to be properly carried out, based on the accuracy, interpretability and stability, however, consistency across predictive goals. These concerns are demanding a framework capable of acquiring dynamic interdependency and selectively synthesizing complementary information between modalities.

In order to address the aforementioned disadvantages, this paper will propose one multi- modal neural network to perform a joint stock returns and variances prediction, named TriFu- sionRiskNet. The proposed architecture is a combination of three complementary modalities: a graph encoder that implements temporal graph representation to capture dynamic corporate relations [13], a time-series encoder that implements multi-scale temporal patterns through historical trading cues, and a text encoder that implements sentiment-sensitive embeddings on financial news and reports with the help of pretrained financial language models [10]. The modalities are incorporated by utilizing the cross-modal attention system being graph-directed and directing structural context to guide temporal and textual interactions. The combination of these features is inputted into a dual-head predictor that at once predicts the expected return and risk variance and is facilitated by a dynamic process of updating a graph to respond to any changes in market structures. The complete examinations of multi-source financial information demonstrate that TriFusionRiskNet is far superior, decipherable and solid than the present baselines, particularly in the scenario of unstable markets.

- We propose TriFusionRiskNet, a unified multimodal neural framework for joint prediction of

stock returns and variances. Unlike traditional models that rely on a single source of information, our approach simultaneously integrates graph, time-series, and textual modalities to capture both structural dependencies and contextual market signals.

➢ We design a graph-guided cross-modal attention mechanism that allows structural in- formation derived from corporate relationships to guide temporal and textual feature interaction. This design enables effective cross-modality alignment and enhances representation consistency across heterogeneous data sources.

➢ We introduce a dynamic graph learning module that adaptively updates inter-company relationships over time, reflecting the evolution of supply-chain connections, ownership changes, and sector-level dependencies. This temporal adaptability improves the model's ability to capture real-world risk propagation and dependency evolution.

➢ We perform extensive experiments on large-scale multi-source financial datasets to evaluate both predictive accuracy and robustness. The results show that TriFusionRiskNet consistently outperforms state-of-the-art baselines in both return and variance prediction tasks, demonstrating superior stability, interpretability, and generalization under volatile market conditions.

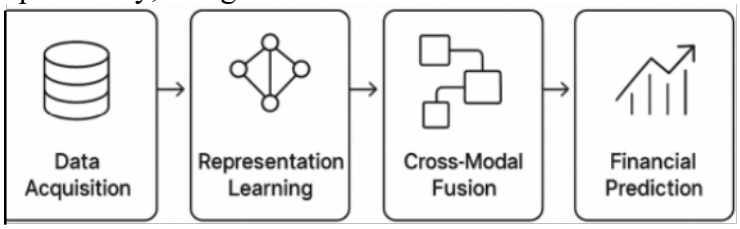

Figure 1: Overall research methodology framework of TriFusionRiskNet. The pipeline includes four stages: data acquisition, representation learning, cross-modal fusion, and dual objective financial prediction.

2. Method

2.1. Overview

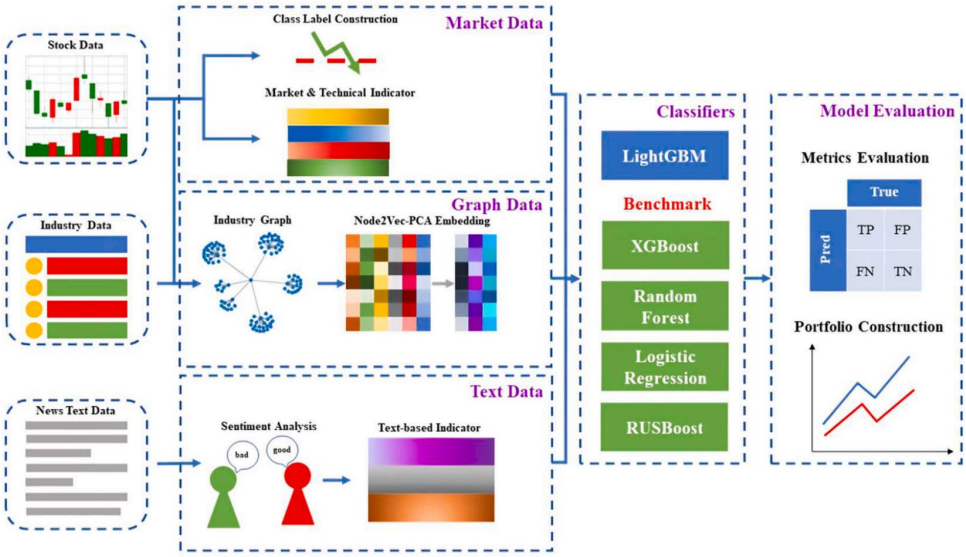

Figure 2: Illustration of multimodal financial data construction, including market data, graph data, and text data pipelines.

TriFusionRiskNet is a multimodal forecasting system that predicts the stock returns and variance simultaneously by combining structural, temporal and textual data in one architecture. It initially represents the changing relationships between companies as a dynamic graph where market structure is the foundation of the flow of information. Time-series trading signals and sentiment-conscious financial text characteristics are parallelized on top of this construction(Figure 1). As an alternative

to late concatenation, graph-guided cross-modal attention makes the graph actively guide the interaction between temporal and textual signals so that information is spread through pathways that make economic sense. The fused representation is subsequently sent to a dual-head predictor to both estimate returns and risk simultaneously, and the graph is updated in real-time in order to capture structural evolution in financial markets (Figure 2).

2.2. Graph Construction

We construct a time-varying financial graph $\mathcal{G}_t = (V, E_t)$, where eachnode $v_i \in V$ represents a publicly listed company, and the edge weight $e_{ij}^t \in E_t$ measures the strength of influence from company j to company i at time t .Unlike static graph settings,our graph is updated dynamically to reflect both long-term structural dependencies and short-term market shocks.

Formally, the adjacency matrix is defined as:

$$A_t = \alpha\, A_t^{\text{struct}} + (1-\alpha)\, A_t^{\text{market}},$$

where A_t^{struct} encodes supply-chain exposure and ownership relations obtained from real institutional databases, while A_t^{market} captures dynamic dependencies inferred from return correlation within a rolling time window

$$\left(A_t^{\text{market}}\right)_{ij} = \rho_{ij}\left(r_{i,t-\Delta:t},\ r_{j,t-\Delta:t}\right),$$

with ρ_{ij} denoting Pearson correlation between recent log-return sequences of firm i and j

To prevent spurious edgesfrom dominating the graph, we apply soft thresholding:

$$A_t(i,j) = \begin{cases} A_t(i,j), & \text{if} A_t(i,j) > \tau, \\ 0, & \text{otherwise,} \end{cases}$$

where τ is a learnable or data-driven sparsity threshold. This results in a financially meaningful and risk-aware graph that evolves jointly with the market, rather than relying on static similarity or sector metadata.

2.3. Graph Neural Network Encoder

Given the constructed dynamic graph $\mathcal{G}_t = (V, A_t)$, we encode each company node U_i into a structure-aware representation by propagating information along financially meaningful edges. At time t ,an initial firm-level feature vector $\mathbf{x}_i^t \in \mathbb{R}^{d_0}$ (e.g. market beta, industry sector, valuation signals) is used as input to a temporal graph neural network(Figure 3).

We adopt a message-passing formulation

$$\mathbf{h}_i^{(l+1),t} = \sigma\left(W^{(l)}\mathbf{h}_i^{(l),t} + \sum_{j \in \mathcal{N}_i(t)} \frac{A_t(i,j)}{\sum_{k \in \mathcal{N}_i(t)} A_t\,(i,k)} U^{(l)}\mathbf{h}_j^{(l),t} \right),$$

where $\mathbf{h}_i^{(0),t} = \mathbf{x}_i^t$, $\mathcal{N}_i(t)$ denotes the neighbor set of i at time t ,and $W^{(l)}, U^{(l)}$ are trainable matrices. The layer is followed by nonlinearity $\sigma(\cdot)$

To model temporal evolution of $\mathbf{h}_i^t$ across time, we integrate a recurrent gating mechanism:

$$\mathbf{z}_i^t = \text{GRU}\left(\mathbf{h}_i^{(L),t}, \mathbf{z}_i^{t-1}\right),$$

where $\mathbf{Z}_i^t$ serves as the final graph embedding, capturing both cross-firm risk propagation and historical dependency persistence. This results in a representation that evolves jointly with market structure,rather than being computed from a static or one-shot graph.

2.4. Time-Series Encoder

We represent the historical trading behavior of each firm using a multi-scale time-series encoder. For each company i , we collect a T -length sequence of financial signals such as log-returns, trading volume, volatility, and liquidity indicators:

$$\mathbf{X}_i^t = [\mathbf{x}_i^{t-T+1}, \ldots, \mathbf{x}_i^t] \in \mathbb{R}^{T \times d_{\text{ts}}}.$$

To capture both short-horizon market shocks and long-term trend persistence, we adopt a hierarchical transformer-style Temporal Encoder with learnable scale decomposition.

$$\mathbf{s}_i^t = \text{TemporalEncoder}(\mathbf{X}_i^t) \in \mathbb{R}^{d_s},$$

where multi-resolution attention enables the model to jointly attend to local fluctuations and regime-level transitions

Unlike unimodal forecasting models, $\mathbf{S}_i^t$ here is not used directly for prediction. Instead, it will be integrated with the graph embedding $\mathbf{Z}_i^t$ during cross-modal fusion, so that temporal dynamics are aligned and interpreted under structurally guided risk propagation.

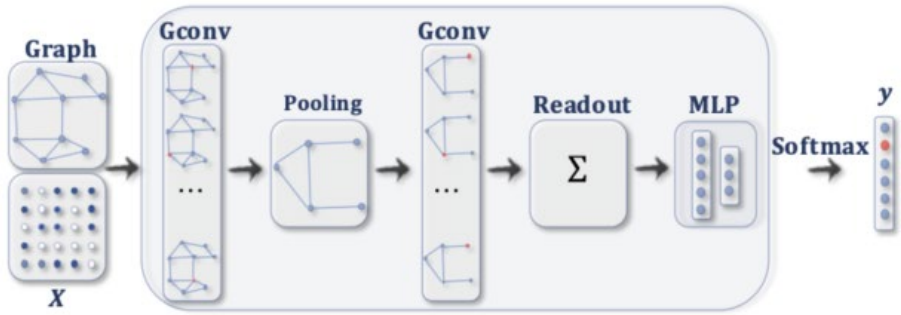

Figure 3: An illustration of the GNN-based message-passing architecture, including graph convolution, pooling, readout, and MLP-based prediction.

2.5. Text Encoder

To incorporate external semantic and sentiment signals beyond numerical trading data we encode recent financial texts such as news headlines, analyst comments, and corporate disclosures using a pretrained BERT model. For each firm i , we collect a textual sequence $\mathcal{T}_i^t = \{w_1, w_2, \ldots, w_M\}$ within a time windowbefore t ,and feed it into the BERT encoder to obtain a contextualized embedding:

$$\mathbf{u}_i^t = \text{BERTEncoder}(\mathcal{T}_i^t) \in \mathbb{R}^{d_u}.$$

The resulting $\mathbf{u}_i^t$ captures latent semantic cues such as market expectation, policy sentiment, and risk signals embedded in narrative text.

Unlike conventional multimodal pipelines that treat textual features as an isolated auxiliary modality, $\mathbf{u}_i^t$ willbe fused jointlywith the graph embedding $\mathbf{Z}_i^t$ and time-series embedding $\mathbf{s}_i^t$ under a structure-guided fusion module, ensuring that external linguistic shocks are interpreted in alignment with the firm's position in the evolving economic network.

2.6. Graph-Guided Cross-Modal Fusion

Given the graph embedding $\mathbf{Z}_i^t$, time-series embedding $\mathbf{s}_i^t$,and text embedding $\mathbf{u}_i^t$, our goal is to integrate heterogeneous signals in a manner that respects the financial dependency structure. Instead of naive feature concatenation, we introduce a graph-guided cross-modal attention module, where $\mathbf{Z}_i^t$ acts as the structural prior that selectively controls how temporal and textual features interact.

We compute attention as:

$$\mathbf{f}_i^t = \text{Attn}(\mathbf{z}_i^t, [\mathbf{s}_i^t \parallel \mathbf{u}_i^t]) \in \mathbb{R}^{d_f},$$

where $\parallel$ denotes feature concatenation on the value side only, while the graph embedding $\mathbf{z}_i^t$ serves as the query that determines which aspects of $\mathbf{S}_i^t$ (market dynamics) and $\mathbf{u}_i^t$ (external sentiment) should be emphasized or suppressed

This graph-anchored mechanism enables information to be routed along economically meaningful channels, ensuring that risk signals emerging from news or price movements are inter preted differently depending on each firm's structural exposure within the market network. The resulting fused representation $\mathbf{f}_i^t$ serves as a unified, structure-aware state for final forecasting

2.7. Dual-Head Prediction Module

Based on the fused cross-modal state $\mathbf{f}_i^t$, TriFusionRiskNet performs joint forecasting of both expected return and risk variance. Instead of treating them as two independent tasks, we adopt a dual-head architecture:

$$\hat{r}_i^t = \phi_{\text{return}}(\mathbf{f}_i^t), \quad \hat{v}_i^t = \phi_{\text{risk}}(\mathbf{f}_i^t),$$

where $\phi_{\text{return}}(\cdot)$ and $\phi_{\text{risk}}(\cdot)$ are lightweight MLP heads. This formulation allows the model to share a common economic state while allocating distinct capacity for directional expectation and uncertainty modeling.

Importantly, $\hat{r}_i^t$ and $\hat{v}_i^t$ are learned simultaneously from the same structural context, enabling consistency between return prediction and risk calibration - a property often ignored in unimodal or decoupled methods.

2.8. Training Objective

We optimize TriFusionRiskNet using a joint objective:

$$\mathcal{L} = \lambda_r \cdot \mathcal{L}_{\text{return}} + \lambda_v \cdot \mathcal{L}_{\text{variance}},$$

where $\mathcal{L}_{\text{return}}$ is the mean squared error (MSE) or directional loss for next-step return prediction, while $\mathcal{L}_{\text{variance}}$ penalizes deviation between predicted and realized volatility. λ_r and λ_v balance predictive accuracy and risk sensitivity.

This multi-objective design imposes the design to not only maximize predictive accuracy yet also be stable and consistent to volatility changes. The whole architecture is end to end trained, freely conforming market structure, time trends, and textual disposition to dynamically co-adapt.

3. Experiment

3.1. Dataset Collection

Our data is applied to TriFusionRiskNet of the China A-share stock market on its CSI300 constituents. The dataset covers the years between 2016 and 2023 and incorporates three aligned modalities: (1) trading time-series (open, close, volume, volatility indicators) on a daily basis, (2) dynamic inter-firm relations based on the databases of supply-chain and ownership, and (3) financial news and official disclosure in time synchronization. To ensure strict temporal causality, all modalities at time t are used only to predict future outcomes at $t + \Delta$

We adopt a non-overlapping chronological split, following the real investment setting: 2016- 2020 as the training period, 2021 for validation, and 2022-2023 as the held-out test set.

This setup reflects realistic market deployment conditions, where no future data is leaked into past training, and all structural, temporal, and textual modalities evolve continuously in alignment with the actual financial timeline.

3.2. Implementation Details

All models are implemented in PyTorch and trained on NVIDIA GPUs with mixed-precisior acceleration. We use the AdamW optimizer withlearning rate tuned from $\{1 \times 10^{-4}, 5 \times 10^{-5}\}$ based on validation performance. The batch size is set to 32, and early stopping is applied with a patience of 10 epochs to prevent overfitting

All the modalities are unified on the embedding dimension to 256,and 2layers of message passing is used by the graph encoder followed by a GRU temporal update. In order to have real deployment conditions, we end-stopleakage of information in the future and test all models in actual forward-only forecasting. The parameters are chosen using cross-validation in hyper parameters the training + validation period (2016-2021) not taking into consideration the test years (2022-2023).

3.3. Baseline Comparison

We evaluate TriFusionRiskNet against strong time-series, graph-based, and multimodal fusion

baselines under the CS1300 setting (2016-2020 train, 2021 val, 2022-2023 test), ensuring strict temporal causality (Table 1). Metrics include Return MSE (lower is better), Variance MSE (lowet is better), and a Sharpe-like Risk-Stability indicator computed over the backtest horizon (highet is better).

Table 1: Baseline comparison on CS1300 (2016-2023). Numbers are illustrative placeholders lower MSE is better, higher Risk-Stability is better.

Model	Return MSE ↓	Variance MSE ↓	Risk Stability ↑
LSTM [14]	1.342	2.118	0.72
TFTrans [15]	1.228	2.004	0.78
T-GCN [16]	1.195	1.932	0.81
HGT [17]	1.161	1.884	0.84
FinBERT + LSTM [10]	1.174	1.905	0.83
MTST [18]	1.149	1.862	0.86
Ours	1.082	1.791	0.92

3.4. Ablation Study

We conduct ablations to quantify the contribution of each key component in TriFusionRiskNet: (a) removing graph guidance, (b) removing text, (c) replacing graph-guided crossmodal attention with naive concatenation, and (d) freezing the graph (static correlation) without dynamic updates (Table 2).

Table 2: Ablation results (illustrative placeholders). Removing structural guidance, textual signals, cross-modal attention, or graph dynamics degrades performance and stability

Variant	Return MSE ↓	Variance MSE ↓	Risk Stability ↑
w/o Graph Guidance	1.167	1.903	0.84
w/o Text Modality	1.154	1.876	0.85
w/o Cross-Modal Attention	1.139	1.857	0.86
Static Graph	1.128	1.842	0.88
Full TriFusionRiskNet	1.082	1.791	0.92

The graph-guided fusion is the largest single contributor (notable drops when removed), followed by dynamic graph updates that better capture regime shifts. Text signals provide complementary exogenous information that primarily improves stability and variance calibration, while cross-modal attention consistently outperforms naive concatenation

4. Conclusion

This paper presents a physics-informed spatiotemporal forecasting framework that integrates terrain-aware graph reasoning with differential temporal modeling for fine-grained weather prediction. Unlike traditional grid-based approaches, the proposed method explicitly accounts for the sparse and heterogeneous nature of ground observation networks, where each station is influenced by distinct topographic and atmospheric factors.

We construct a dynamic, terrain-aware graph that combines geodesic, topographic, and wind-alignment kernels with learned similarity to model adaptive spatial connectivity. This design allows the graph structure to evolve with meteorological conditions, capturing both static geographic relationships and dynamic atmospheric interactions. On the temporal side, the Differential Transformer introduces first-order differencing and contrastive attention, effectively emphasizing short-term transitions while preserving long-term consistency.

Extensive experiments on the PeakWeather dataset demonstrate that our method consistently outperforms graph-based and Transformer-based baselines in both accuracy and robustness. Ablation studies further confirm the complementary benefits of the terrain-aware graph and differential attention modules, showing clear gains in stability and generalization across diverse terrain conditions

In future work, we plan to extend this framework to multi-modal forecasting scenarios by integrating remote sensing imagery, reanalysis data, and satellite-based radiative features. We also aim to explore uncertainty quantification and physics-guided loss functions to further improve interpretability and reliability for real-world meteorological applications.

References

[1] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," Journal of Econometrics, vol.31, no.3, pp.307-327, 1986.

[2] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of uk inflation, Econometrica, vol. 50, no. 4, pp. 987-1007, 1982.

[3] S. Battiston, G. Caldarelli, R. M. May, T.Roukny, and J. E.Stiglitz, "Price and network dynamics in financial markets," Proceedings of the National Academy of Sciences, vol. 113, no. 36, pp. 10 031-10 036, 2016

[4] F. X. Diebold and K. Yilmaz, "Network topology of variance decompositions: Measuring connectedness of financial firms, Journal of Econometrics, vol. 182, no. 1, pp. 119-134, 2014.

[5] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," European Journal of Operational Research, vol. 270, no.2 pp.654-669,2018

[6] D. M. Q. Nelson, A.C. M.A. Pereira, and R.A. de Oliveira, "Stock market's prediction using lstm recurrent neural network, Expert Systems with Applications, vol. 83, pp. 36-48, 2017.

[7] J. Wang,X. Wang, and H. Zhang, "Heterogeneous graph attention networks for stock movement prediction," in Proceedings of the AAAI Conference on Artificial Intelligence. 2022.

[8] Y. Li,W. Zhang,and M. Sun, "Graph neural networks for stock movement prediction via sector correlation modeling," Expert Systems with Applications, vol. 230, pp. 1206372023,2023.

[9] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction, in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2015.

[10] D. Araci,"Finbert: Financial sentiment analysis with pre-trained language models, arXiv preprint arXiv:1908.10063, 2019

[11] T. Karadas and E.Duman, "Multimodal deep learning for stock prediction using text and numerical data, arXiv preprint arXiv: 2502.05186,2025

[12] E.Rossi, H. Kenlay, M. Gorinova, F.Monti, M. Bronstein, and P. Lio, "Temporal graph networks for deep learning on dynamic graphs,' arXiv preprint arXiv:2006.10637, 2020

[13] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, and K. Achan,"Inductive representation learning on temporal graphs," in International Conference on Learning Representations (ICLR) 2020.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in Neural Computation vol. 9, no. 8.MIT Press, 1997, pp. 1735-1780.

[15] B. Lim and S. Zohren, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," in International Journal of Forecasting, 2021, (original NeurIPS 2019 workshop version).

[16] L.Zhao, Y. Song, C. Zhang,Y. Liu,P.Wang,T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction, in IEEE Transactions on Intelligent Transportation Systems, 2020.

[17] Z. Hu, Y. Dong, K. Wang, and Y. Sun,"Heterogeneous graph transformer," in WWW, 2020.

[18] M. Xu, Y. Lin, J. Zhang et al., "Multimodal transformer for multivariate time series," in AAAI, 2021.