Available Online: 29 July 2026 DOI: 10.54254/2977-5701/2026.35777

Automated feature engineering for multisource heterogeneous data

***Bo Xu*[1,2*], *Lifei Lu*[1,2], *Yulin Lu*[1,2], *Sitong Li*[1,2]**

[1]School of Big Data and Artificial Intelligence, Guangdong University of Finance & Economics, Guangzhou, China
[2]Key Laboratory of Collaborative Innovation in Digital Economy, Guangzhou, China

*Corresponding Author. Email: 20171080@gdufe.edu.cn

Abstract. To address the challenges of high-dimensional feature redundancy and nonlinear feature extraction in multisource heterogeneous financial data, in this study, key indicators are extracted from multidimensional structured data, including macroeconomic and stock market data, and unstructured news sentiment data are incorporated. Moreover, text features and time-series data are fused through a tensor fusion network to construct an automated feature engineering framework for risk indicators capable of integrating multisource heterogeneous data. Automated feature engineering can be used to automatically generate high-quality features without requiring domain knowledge or human intervention, thereby improving model performance and enhancing efficiency.

Keywords: multisource heterogeneous data, feature dimensionality reduction, automated feature engineering, extreme financial risk

1. Introduction

Financial risk characteristics represent the behaviours of financial risk cases. Only by selecting appropriate and significant feature indicators can the behavioural characteristics of financial risk cases be accurately represented [1, 2]. In terms of financial risk feature extraction, early studies mainly reported the manual construction of risk features using price data and financial indicators, supplemented by statistical testing methods to select highly correlated indicators. With the advancement of information technology, Artificial Intelligence (AI) methods were subsequently introduced to further explore the nonlinear characteristics of financial risk [3]. Despite the effectiveness of these methods, several limitations remain [4-6]. First, most studies have been conducted using statistical models [e.g., the Value at Risk (VaR) model, Principal Component Analysis (PCA), and Generalized Autoregressive Conditional Heteroskedasticity (GARCH)] or have reported on the construction of risk factor systems on the basis of expert knowledge for feature selection. These approaches are not only complex and time-consuming but also often assume linear relationships among variables, which makes effectively capturing the nonlinear dynamics and abrupt risks of financial markets difficult. Consequently, they have significant limitations in the processing of high-dimensional, nonlinear financial time-series data. Second, the selection of some indicators is highly subjective and lacks a data-driven automated screening mechanism. Third, financial market structures undergo substantial changes, whereas the data used for index construction in previous studies have been derived mainly from traditional financial statistics. The sources of risk feature indicators are therefore overly limited, and fusion analyses that incorporate multidimensional factors such as macroeconomic conditions, cross-market linkages, and sentiment are lacking. These limitations collectively affect the accuracy and timeliness of financial risk prediction.

Recent cutting-edge studies have reported that, from the perspective of behavioural finance, big data derived from web searches and news texts can be used to construct risk perception indicators for economic entities, thereby helping improve the accuracy of financial risk monitoring and identification [7-9]. Accordingly, in this study, the problem is approached from multiple dimensions, including macroeconomic indicators, financial market prices, and the sentiment features of financial news, and financial risk early warning indicators are automatically constructed by integrating structured and unstructured data. Deep learning is further introduced for modelling and analysis to capture the nonlinear interactions among different risk factors within the financial system, thereby increasing the forward-looking ability and accuracy of risk identification and providing a reference for financial regulatory authorities to more effectively prevent extreme risks that may threaten China's financial system.

2. Status of research

Early financial feature extraction methods were performed by primarily using capital market data and traditional statistical models to analyse financial market time-series data. For example, Karaca et al. applied the maximum Relevance Minimum Redundancy (mRMR) method to global stock index data for feature selection [10]. Liu conducted an empirical analysis of extreme risk transmission among China's financial markets, including the bond and stock markets, using the Multivariate Multiple-Quantile Conditional Autoregressive Value at Risk (MVMQ-CAViaR) method [11]. With the explosive growth of financial market data and the increasing diversity of data types, financial data have become characterized by high frequency, high dimensionality and heterogeneity. Extracting key and representative features from such complex data has become a critical step in the prediction and early warning of extreme financial risk. Machine learning models offer significant advantages in nonlinear and nonstationary data processing and feature extraction and have therefore been widely applied to financial modelling. For example, Yun et al. used eXtreme Gradient Boosting (XGBoost) to perform feature selection on 71 market data features and technical indicator features and then used the resulting feature subset to predict the direction of the Korea Composite Stock Price Index (KOSPI) [5].

In recent years, studies have begun to incorporate nontraditional data sources on a large scale. Alam et al. [12] proposed integrating macroeconomic indicators with alternative datasets and using natural language processing techniques to extract investor sentiment from regulatory reports and news streams, thereby substantially improving the ability of the model to detect systemic financial risk at an early stage. To address the challenge of extracting patterns from massive and complex financial data, Alhomayani and Alruwaitee [13] proposed a novel financial prediction model, Quasi-Oppositional Coot algorithm with a Deep Learning-based Predictive Method on the Financial Sector (QOCODL-PMFC), that significantly improves prediction accuracy and robustness when processing integrated datasets comprising investor sentiment, market indicators, and economic variables.

The focus of research in China has primarily been on monitoring online public opinion in the Chinese context. Fan et al. [14] used a text co-occurrence mining approach to extract features from more than three million news articles published by major Chinese newspapers and measured the level of systemic financial risk on the basis of the sentiment of bank-related co-occurrence news. Ouyang et al. [15] mined investor sentiment from online communities such as Xueqiu and Eastmoney to construct a China-specific Financial Stress Index (FSI) and reported that fluctuations in investor sentiment often precede the outbreak of financial crises.

In summary, most studies have focused primarily on the analysis of structured data, whereas the mining of high-dimensional unstructured data such as textual sentiment remains inadequate. Moreover, the fusion of multisource heterogeneous data is still limited, and a unified modelling framework has yet to be established for jointly incorporating macroeconomic variables, financial market prices, and unstructured information such as financial news text and audio/video recordings of financial conferences.

3. Cross-modal data fusion on the basis of a tensor fusion network

To enable unified modelling of multisource heterogeneous data, data with different sampling frequencies were aligned to a daily time scale during the preprocessing stage to construct a unified temporal index. Specifically, low- and medium-frequency data were expanded using interpolation or forward filling, and high-frequency data were denoised through time-window aggregation to form a unified multivariate time-series matrix, which can be represented as $X \in \mathbb{R}^{T\times(d_m+d_s)}$.

Directly concatenating features from different modalities before they are fed into the prediction model may introduce distortions because of differences in feature scale and inconsistent noise characteristics; hence, capturing interactions among modalities is difficult. Therefore, a Tensor Fusion Network (TFN) was used to model high-order interactions between structured financial indicators and news sentiment indicators in financial market data, thereby achieving cross-modal data fusion. The core principle of this method is to jointly represent features from different modalities through tensor operations to capture cross-modal interactive information. Compared with traditional feature concatenation methods, the TFN integrates features from different data sources into a unified representation, which not only preserves the original information of each modality but also captures high-order nonlinear relationships; hence, it is particularly effective for multimodal data processing and complex feature integration. This integrated representation, combined with nonlinear activation functions, increases the ability of the model to capture nonlinear relationships, thereby generating more accurate and comprehensive feature representations for subsequent risk prediction. The TFN-based fusion architecture is shown in Figure 1.

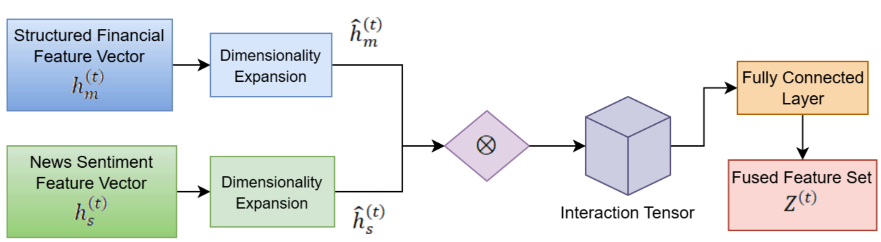

Figure 1. TFN-based multisource heterogeneous data fusion architecture

Specifically, at time t, after preprocessing and feature extraction, the structured financial feature vector is denoted as $h_m^{(t)} \in \mathbb{R}^{d_m}$, and the news sentiment feature vector is denoted as $h_s^{(t)} \in \mathbb{R}^{d_s}$. As shown in Equations (1) and (2), the TFN first concatenates the feature vector of each modality with a constant vector of ones, thereby preserving both unimodal features and bimodal interaction features. The Cartesian outer product is subsequently applied to the expanded feature vectors to generate a higher-order tensor $Z^{(t)}$ that contains all interaction information.

$$\widehat{h}_m^{(t)} = \begin{bmatrix} h_m^{(t)} \\ 1 \end{bmatrix}, \widehat{h}_s^{(t)} = \begin{bmatrix} h_s^{(t)} \\ 1 \end{bmatrix} \tag{1}$$

$$Z^{(t)} = \widehat{h}_m^{(t)} \otimes \widehat{h}_s^{(t)} \tag{2}$$

However, the outer product operation substantially increases the feature dimensionality, thereby increasing the computational complexity and potentially leading to overfitting. Therefore, a low-rank decomposition strategy is used to compress the fusion tensor and project it into a low-dimensional latent space. In the experimental setup, the features of each modality were mapped to a unified 32-dimensional embedding space, and the rank of the low-rank decomposition was set to r = 4. These parameter settings were based on the original TFN model and further optimized through cross-validation to balance model complexity and representational capacity.

Subsequently, the compressed fusion tensor is flattened and fed into a fully connected layer for nonlinear mapping. As shown in Equation (3), this layer maps the original or preprocessed feature vectors of each modality into a common embedding space with the same dimensionality, thereby yielding the final fused feature representation $H_{fusion}^{(t)}$.

$$H_{fusion}^{(t)} = \sigma\left(W \cdot flatten\left(Z^{(t)}\right) + b\right) \tag{3}$$

where W and b denote the weight matrix and the bias term, respectively, and σ denotes the ReLU activation function. Finally, $H_{fusion}^{(t)}$ is used as the input feature variable for the systemic financial risk early warning model to enable accurate prediction of the risk state.

4. Experimental results and analysis

The multisource heterogeneous data used in this study comprised three primary categories: macroeconomic data, financial market data and news sentiment data [16]. The macroeconomic data were collected primarily at monthly or quarterly frequencies and were of relatively low dimensionality with stable characteristics; the financial market data were collected predominantly at a daily frequency and were of moderate dimensionality with high volatility; and the news sentiment data were extracted from textual information and exhibited high-frequency characteristics and relatively high levels of noise.

Compared with lagging financial indicators, news sentiment indicators capture the expectation biases and psychological fluctuations of market participants, thereby enabling the identification of early warning signals of systemic risk. When financial news is dominated by negative sentiment, it often reflects heightened market uncertainty and increasing pessimism, which may trigger herding behaviour among investors, thereby leading to violent fluctuations in asset prices. Therefore, timely monitoring of news sentiment provides an important basis for predicting future financial risk and facilitates the development of more adaptive risk management strategies.

In this study, Python-based web crawling and a Convolutional Neural Network (CNN) were used to perform the sentiment analysis of financial news texts related to China's stock and securities markets, extract quantitative indicators that reflected changes in market sentiment, and incorporate these indicators into the risk early warning indicator system. The news data were obtained from the China Financial News Database (CFND) and covered the period from January 2005 to December 2024. The dataset consisted of articles from major financial media platforms, including *Sina Finance*, *RoyalFlush*, and *Eastmoney*, as well as financial news published in well-known domestic financial and general newspapers, such as *Shanghai Securities News*, *China Securities Journal*, *China Business News*, and *Economic Daily*, with a total of approximately 1.25 million valid news samples. The trained CNN model was then used to perform inference on the financial news across the entire sample period to obtain a sentiment label $L_{i,t} \in [-1,1]$ for each news item. The news sentiment labels were subsequently aggregated on a daily basis to construct the News Sentiment Indicator (NSI), which is calculated using Equation (4):

$$NSI_t = \frac{\sum_{i=1}^{N} (L_{i,t} \times w_{i,t})}{N_t} \tag{4}$$

where N_t denotes the total number of news reports on trading day t and $L_{i,t}$ denotes the CNN classification label of the i-th news report on trading day t. Considering the varying influence of individual news reports, a weight $w_{i,t}$ is assigned to each news report. The resulting sentiment series captures the overall trend of market sentiment: Lower values indicate stronger market panic and a higher level of potential systemic financial risk, whereas higher values indicate more optimistic market sentiment. This indicator, together with the macroeconomic indicators, constitutes the input variables for the subsequent risk early warning model.

Using financial news data that were collected between April and December 2023 as an example, the calculation procedure for the above sentiment indicator is illustrated, and the temporal variation in the daily sentiment indicator is examined. The sentiment values obtained for the sample period were standardized according to the above procedure to produce a time series of the sentiment indicator, as shown in Figure 2. The news sentiment indicator fluctuated substantially in May, remained relatively stable from June to August with only slight fluctuations around 0, and fluctuated substantially again after August. A possible explanation is that, following the successive closures of Silicon Valley Bank, Signature Bank and First Republic Bank in March, the U.S. banking crisis continued to escalate, which further undermined the stability of the banking system.

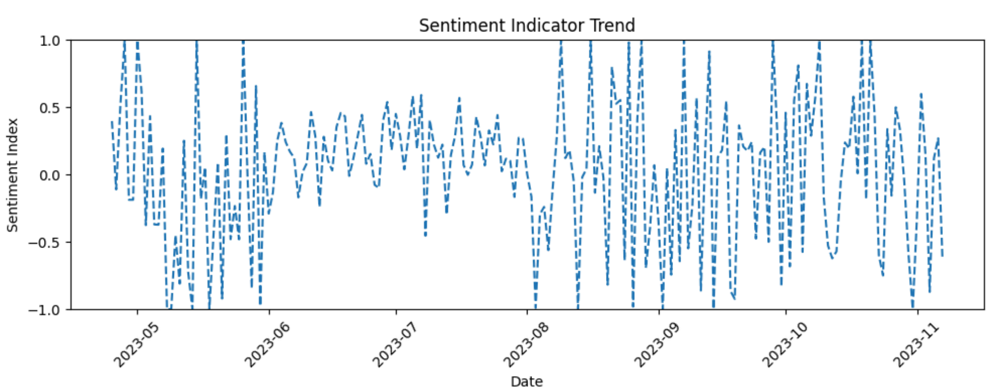

Figure 2. Time series of changes in sentiment indicators

To verify the effectiveness of the multisource data fusion strategy, comparative experiments were conducted to perform a sensitivity analysis of model performance from the perspective of input feature construction. Specifically, three input configurations were constructed, namely, (1) numerical financial indicators only, (2) direct concatenation of numerical indicators and text features, and (3) TFN-based cross-modal feature fusion, and comparative experiments were conducted using the same transformer-based prediction model architecture and training parameter settings. The experimental results reveal that the TFN-based fusion method outperformed the other two methods in terms of evaluation metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), thus indicating that it can more effectively capture the higher-order interactions between different modalities.

As shown in Table 1, the incorporation of sentiment text features increased the overall predictive performance of the model, thus indicating that sentiment information extracted from text provides valuable complementary information for early warning of extreme financial risk. In terms of the fusion strategy, compared with direct feature concatenation, the TFN-based cross-modal fusion method achieved the best performance in terms of RMSE, MAE and R^2, which indicates that the TFN can more effectively capture the higher-order interactions between different modalities, thereby increasing the predictive accuracy and stability of the model.

Table 1. Comparison of model performance under three feature fusion strategies

Input configuration	RMSE	MAE	R^2
Numerical financial indicators only	0.3142	0.4215	0.6613
Numerical indicators + text (direct concatenation)	0.2978	0.4086	0.6847
Numerical indicators + text (TFN-based feature fusion)	0.2689	0.3724	0.7215

5. Conclusion

In this study, a financial extreme risk early warning indicator system that comprises three dimensions, namely, macroeconomic conditions, market volatility, and news sentiment, was constructed. The traditional quantitative indicators were further integrated with text sentiment indicators through a tensor fusion model to construct a risk feature set that is better suited to the characteristics of the modern financial environment. The financial extreme risk indicator system with the fusion of multisource heterogeneous data can capture risk information from diverse sources, including changes in macroeconomic fundamentals, financial market instability, and biases in investor expectations, which increases its ability to identify potential risks.

Funding

The Project of Philosophy and Social Science Planning of Guangdong (GD24CGL08)

References

[1] Song, Y., Du, H., Piao, T., & Shi, H. (2024). Research on financial risk intelligent monitoring and early warning model based on LSTM, transformer, and deep learning. *Journal of Organizational and End User Computing (JOEUC), 36*(1), 1–24.

[2] Yang, H., Liu, L., Cui, J., Wu, W., & Gao, Y. (2025). Research on dynamic measurement and early warning of systemic financial risk in China based on TVP-FAVAR and deep learning model. *Systems, 13*(8), 720.

[3] Garcia-Jorcano, L., & Sanchis-Marco, L. (2021). Systemic-systematic risk in financial system: A dynamic ranking based on expectiles. *International Review of Economics & Finance, 75*, 330–365.

[4] Guo, Z., Wang, H., Liu, Q., & Li, X. (2014). A feature fusion based forecasting model for financial time series. *PLoS ONE, 9*(6), e101113.

[5] Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Systems with Applications, 186*, 115716.

[6] Tsai, C. F., & Hsiao, Y. C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems, 50*(1), 258–269.

[7] Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry, 33*(1), 3–12.

[8] Huang, T., & Yu, L. (2019). Application of SDAE-LSTM model on financial time series forecasting. *Computer Engineering and Applications, 55*(1), 142–148.

[9] Wang, G. J., Chen, Y., Zhu, Y., & Li, X. (2024). Systemic risk prediction using machine learning: Does network connectedness help prediction? *International Review of Financial Analysis, 93*, 103147.

[10] Karaca, Y., Zhang, Y. D., & Muhammad, K. (2020). Characterizing complexity and self-similarity based on fractal and entropy analyses for stock market forecast modelling. *Expert Systems with Applications, 144*, 113098.

[11] Liu, Y., Li, Z., & Liu, H. (2020). Interconnectedness, extreme risk spillover and China's financial systemic risk ——Empirical analysis based on the MVMQ-CAViaR method. *Research on Economics and Management*, (2), 19–29.

[12] Alam, M. S., Mrida, M. S. H., & Rahman, M. A. (2025). Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI). *American Journal of Scholarly Research and Innovation, 4*(1), 63–100.

[13] Alhomayani, F. M., & Alruwaitee, K. A. (2024). A new financial risk prediction model based on deep learning and quasioppositional coot algorithm. *Alexandria Engineering Journal, 108*, 60–69.

[14] Fan, X., Wang, Y., & Wang, D. (2022). A study of the systemic risk measurement and early warning mechanism of the banking industry in China based on news and machine learning. *The Journal of World Economy, 45*(4), 3–30.

[15] Ouyang, Z., Chen, S., Yang, X., & Li, X. (2023). Economic policy uncertainty, network public opinion and systemic risk of financial institutions. *Journal of Management Sciences in China, 26*(4), 62–86.

[16] Xu, B., Lu, L., & Tu, W. (2025). FRP-TNN: An intelligent framework using transformer for financial market risk prediction to enhance decision-making under uncertainty. *International Journal of Data Science and Analytics, 20*, 7071–7087.