Detecting Financial Fraud in Listed Companies via a CNN-Transformer Framework

Qian Yu , Yuchen Yin * , Shicheng Zhou , Huailing Mu , Zhuohuan Hu

Posted Date: 12 February 2025

doi: 10.20944/preprints202502.0891.v1

Keywords: fraud detection; listed companies; CNN-Transformer; self-attention mechanism; deep learning

 doi:10.20944/preprints202502.0891.v1

Article

Detecting Financial Fraud in Listed Companies via a CNN-Transformer Framework

Qian Yu [1], Yuchen Yin [2,*], Shicheng Zhou [3], Huailing Mu [4] and Zhuohuan Hu [5]

[1] Trine University, Detroit, Michigan, U.S.A
[2] Columbia University, New York City, U.S.A
[3] University of Minnesota,U.S.A
[4] University of California, Los Angeles, Los Angeles, United States
[5] Independent researcher, China
* Correspondence: yy3243@tc.columbia.edu

Abstract: Financial fraud in listed companies has grown increasingly complex, challenging traditional detection methods. This paper introduces the Transformer-style Convolutional Neural Network (CNN-Transformer), a unified architecture that integrates CNNs' capability for localized feature extraction with Transformers' strength in capturing long-range contextual dependencies. CNN-TRANSFORMER employs large-kernel convolutions and a self-attention mechanism to enhance feature interaction while maintaining computational efficiency via a linear-complexity Hadamard product approach. Evaluated on a dataset from China's A-share market, CNN-TRANSFORMER outperforms models like ResNet, MLP, and standalone CNNs or Transformers in accuracy, precision, recall, and AUC. The model's capacity to simultaneously extract localized features and model comprehensive contextual relationships enables robust detection of sophisticated fraudulent patterns. This research underscores the potential of CNN-Transformer hybrids in financial fraud detection, offering a scalable and interpretable AI solution. Future work may focus on integrating unstructured data, improving transparency, and optimizing efficiency for broader applications.

Keywords: fraud detection; listed companies; CNN-Transformer; self-attention mechanism; deep learning

I. Introduction

Public disclosure of financial information by listed companies allows investors to make informed decisions and promotes the healthy functioning of capital markets. However, financial fraud remains a global challenge. For example, in 2017, the China Securities Regulatory Commission (CSRC) exposed a decade-long fraud case involving Kangmei, and Enron Corporation in the U.S. overstated profits by $586 million over several years. Financial fraud not only causes significant losses to investors but also severely undermines market integrity. Traditional auditing methods, constrained by time, resources, and fraud sophistication, are increasingly inadequate[1–3]. Thus, developing machine learning models to accurately predict financial fraud is critical for enhancing regulatory oversight and reducing malpractice.

Existing fraud detection models fall into two categories: traditional machine learning and deep learning. Ke et al. (2025) pioneered the application of GAN-based architectures for deepfake detection and payment fraud identification in digital transactions, substantially advancing the field of cybersecurity. Their research offers innovative methodological frameworks for deploying generative adversarial networks in financial fraud detection systems, establishing new benchmarks for AI-driven security solutions in the fintech sector. However, existing models often neglect explicit modeling of local or global feature dependencies[4]. Ravisankar et al. [5] compared multiple methods

 doi:10.20944/preprints202502.0891.v1

(e.g., neural networks, SVM, logistic regression) on Chinese listed companies and found probabilistic neural networks performed best without feature extraction. Ali et al. (2022) used principal component analysis (PCA) for feature extraction and logistic regression for fraud prediction[6]. Sabau, A. S. (2012) addressed class imbalance using cost-sensitive learning. Recent deep learning approaches, such as SMOTE-sampled CNNs and LSTMs [7], have shown promise in automated feature extraction and classification. Qu et al. (2024) and Hilalet al. (2022)also proposes several methods to identify fraudulent activities in money laundering[8–10].

This paper introduces a CNN-Transformer that combines the self-attention mechanism of Transformers with large-kernel convolutions to capture global dependencies and enhance feature interaction[11–13]. The proposed model simplifies self-attention computation via Hadamard products, achieving linear complexity.

II. Methodology

The CNN-Transformer hybrid architecture synergistically combines the complementary advantages of convolutional operations and self-attention mechanisms. Specifically, this integrated framework capitalizes on CNNs' proficiency in hierarchical spatial feature learning from localized patterns while simultaneously harnessing Transformers' capacity for modeling long-range sequential dependencies and global contextual relationships within financial time series data.

A. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) represent a class of deep neural architectures specifically designed for hierarchical feature extraction and pattern recognition in spatially structured data. By leveraging convolutional layers, pooling, and activation functions, CNNs extract hierarchical features efficiently[12–15]. They excel in image classification, object detection, and medical imaging, while hybrid architectures integrate CNNs with Transformers for enhanced global context modeling[16–24]. CNNs excel in pattern recognition tasks through local feature extraction via stacked convolutional and pooling layers as Figure 1. For a given true distribution P and predicted distribution Q, loss function is as:

$$\mathrm{H(P,Q)} = -\sum_i P(i) log Q(i) \qquad (1)$$

In classification problems, PPP is typically the actual class distribution (i.e., the labels), while Q represents the probability distribution predicted by the model.

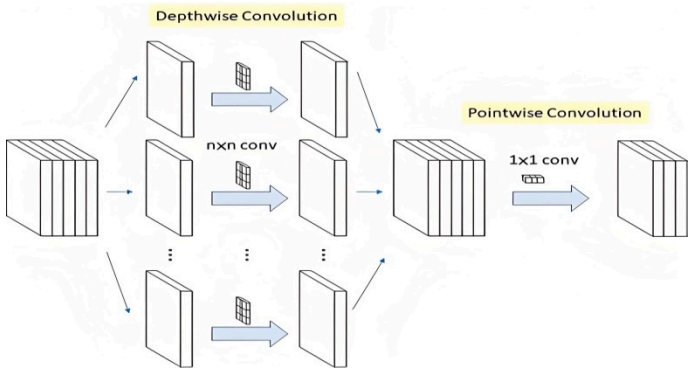

Figure 1. Flow process of CNN.

B. Transformers

Transformers are deep learning architectures designed for sequence modeling and global context understanding. Using self-attention mechanisms, they dynamically weigh input elements, capturing long-range dependencies more effectively than RNNs or CNNs. Key components include multi-head attention, feedforward layers, and positional encoding. Transformers excel in NLP (e.g.,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0891.v1

BERT, GPT) and vision tasks (e.g., ViT), enabling state-of-the-art performance. Their scalability and parallelism make them essential for large-scale AI models, while hybrid architectures integrate them with CNNs for enhanced spatial awareness in image processing and multimodal learning as Figure 2.

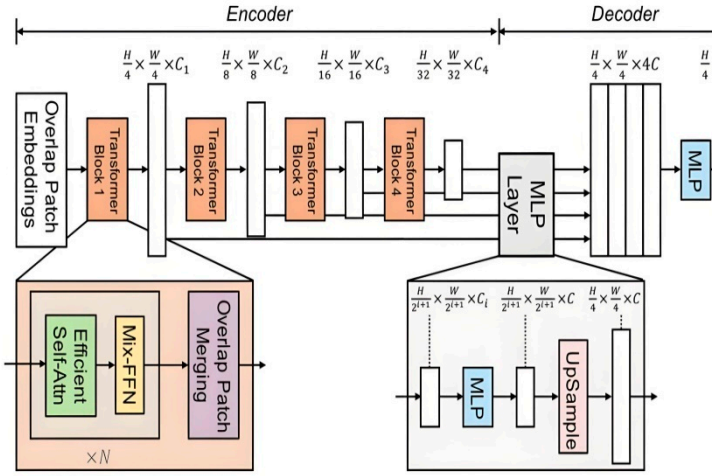

Figure 2. Flow process of ViT.

Originally designed for NLP, Transformers like Vision Transformer (ViT) leverage self-attention to capture global dependencies. Recent studies integrate Transformers with CNNs to balance performance and computational cost.

C. CNN-Transformer

To further enhance the model's performance, we draw inspiration from Maaz et al. (2022) by combining the strengths of CNNs in local feature extraction with the advantages of Transformers in global information modeling, adopting a CNN-Transformer hybrid architecture. This approach is exemplified by the Mobile-Former network, which integrates the benefits of MobileNet and Transformers through bidirectional bridging. The structure leverages MobileNet's efficiency in local processing and the Transformer's capability for global interactions, while the bridging mechanism enables bidirectional fusion of local and global features. The Mobile-Former architecture incorporates a lightweight Transformer component that operates with a minimal set of learnable tokens (typically ≤6), initialized through stochastic processes to capture essential global representations. This design paradigm significantly optimizes computational efficiency while maintaining the model's capacity to learn comprehensive contextual information. By incorporating a lightweight cross-attention mechanism to facilitate bridging, Mobile-Former achieves both high computational efficiency and enhanced representational capacity as Figure 3.

CNN-Transformer adopts a pyramid structure with two stages. Input data (1×30 feature fields) pass through a Stem block, followed by hierarchical feature mapping via convolutional modulation modules. Loss function is as below:

$$\mathrm{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \qquad (2)$$

However, this model also has several drawbacks, such as high memory consumption, high cost, complex architecture, difficult training, and lack of interpretability. To address these challenges, we designed and trained a lightweight Transformer with fewer layers, optimized the network structure, and improved training strategies to balance model performance and resource consumption.

 doi:10.20944/preprints202502.0891.v1

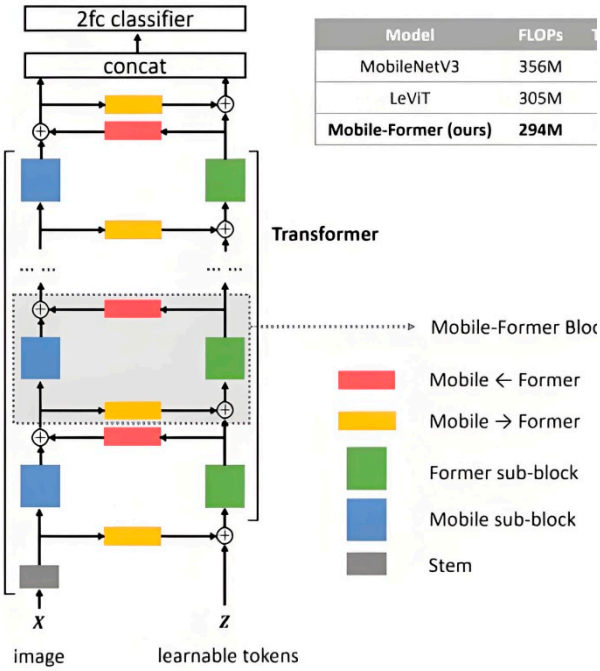

Figure 3. Flow process of CNN-Transformers.

III. Data

A. Data

Our analysis employed the Wind financial database, encompassing five years (August 2019 - August 2024) of market data and financial indicators for the complete universe of 3,000 A-share listed companies in mainland China's securities market. The dataset includes a wide range of financial indicators, such as earnings per share, operating expenses, and cash flow metrics. To ensure data quality, we performed several preprocessing steps, including the removal of features with more than 40% missing values, imputation of missing data, and Z-score normalization.

B. Data Preprocessing

Missing values were handled differently based on the type of feature. For categorical features, missing values were imputed using the mode of the feature. If multiple modes existed, one was randomly selected as the fill value. For numerical features, missing intermediate values were interpolated using polynomial fitting, while edge values were filled with the mean. If a feature consisted entirely of "NaN" values, all missing values were replaced with 0. After imputation, Z-score normalization was applied to standardize the numerical financial features, ensuring a mean of 0 and a variance of 1.

C. Feature Selection

To isolate the most discriminative features for fraudulent activity identification, we employed feature selection techniques using decision trees, random forests, and XGBoost. From these methods, we selected the top 38 features with the highest weight values (as Table 1). The selected features include key financial indicators such as basic earnings per share, operating expenses, and total comprehensive income. These features were chosen based on their importance in distinguishing fraudulent from non-fraudulent financial statements.

 doi:10.20944/preprints202502.0891.v1

Table 1. Variables list.

#.	**Variables**
1	Basic earnings per share
2	Disposal of fixed assets, intangible assets, etc., net cash received
3	Ex-rights date
4	Operating expenses
5	Diluted earnings per share growth rate
6	Total comprehensive income attributable to owners of the parent company
7	Report disclosure date
8	Total shareholders' equity (or stockholders' equity)
9	Cash received related to other business activities
10	Actual receipt time
11	Actual investment (or cost)
12	Issue date
13	Other industries
14	Accounts receivable
15	Deferred income
16	Purchase of fixed assets, intangible assets, and others
17	Non-current liabilities total
18	Operating foreign income
19	Long-term equity investments/assets
20	Short-term loans
21	Asset impairment loss
22	Financial expenses
23	Total comprehensive income
24	Addition: Beginning cash and cash equivalents balance
25	Total comprehensive income attributable to owners (or shareholders) of the parent company
26	Other non-current assets
27	Total profit (indicated as a negative number if a loss)
28	Operating profit (indicated as a negative number if a loss)
29	Cash received from borrowing
30	Total investment cash inflows
31	Undistributed profit
32	Undistributed profit
33	Taxes and fees paid
34	Operating profit
35	Other receivables

 doi:10.20944/preprints202502.0891.v1

36	Deferred taxes
37	Employee payable compensation
38	Research and development expenses (R&D expenses)

Moreover, the experiment was conducted under a 10-fold cross-validation setup, where all four machine learning methods were optimized using grid search on the training data. From each method, the top 30 manufacturing and other industry features with the highest weight values were selected. By combining the predictions from these four methods, the 30 most important feature fields with the highest overall weights were identified, as shown in Table 1. There are significant differences in fraud indicators between the manufacturing sector and other industries. In manufacturing, **other payables** and **undistributed profits** are more prone to manipulation. Meanwhile, in other industries, fraud is more likely to involve **fixed asset disposals** and **intangible asset proportions**.

IV. Results

A. Composite Model Performance

Table 2 systematically compares the model's predictive performance under different kernel configurations, with comprehensive evaluation across five critical metrics: classification accuracy, precision rate, recall sensitivity, F1-measure, and area under the ROC curve (AUC). The results indicate that larger kernels slightly improve recall but show minimal variation in other metrics. Experimental results demonstrate remarkable consistency across kernel configurations of sizes 6, 9, and 12, achieving statistically comparable performance levels: classification accuracy of 0.969 (±0.002), F1-measure of 0.968 (±0.003), and ROC-AUC of 0.979 (±0.001), indicating minimal variance in model effectiveness across these parameter settings. While a kernel size of 12 achieves the highest recall (0.9780), precision fluctuates slightly, suggesting a trade-off between capturing positive cases and maintaining specificity. Overall, the model remains stable across different kernel sizes, demonstrating robustness in fraud detection.

Table 2. Performance of the model with different kernels.

Kernel Size	Accuracy	Precision	Recall	F1 Score	AUC
3	0.9663	0.9594	0.9761	0.9682	0.9771
6	0.9692	0.9624	0.9771	0.9692	0.9790
9	0.9682	0.9614	0.9771	0.9682	0.9790
12	0.9692	0.9594	0.9780	0.9682	0.9780

B. Comparison of Different Model Performances

As demonstrated in Table 3, the CNN-Vision Transformer (ViT) hybrid architecture effectively synergizes the complementary capabilities of convolutional neural networks for hierarchical local feature learning and transformer-based mechanisms for long-range dependency modeling, yielding superior model robustness and enhanced predictive accuracy across diverse evaluation metrics. This hybrid approach is particularly effective in tasks where both local and global patterns are critical, such as financial fraud detection. The results suggest that transformer-based models (ViT and CNN-ViT) are highly effective for tasks requiring global context understanding, such as detecting complex financial fraud patterns. This finding aligns with the growing trend of using transformers in various domains beyond natural language processing (NLP).

 doi:10.20944/preprints202502.0891.v1

Table 3. Performance of different models.

Models	Accuracy	Precision	Recall	F1 Score	AUC
CNN	0.9183	0.8879	0.9673	0.9212	0.9526
ViT	0.9663	0.9594	0.9741	0.9663	0.9771
LSTM	0.9594	0.9467	0.9731	0.9604	0.9761
CNN-VIT	0.9692	0.9624	0.9771	0.9692	0.9790

As far as we can see from Figure 4, this figure illustrates the AUC score changes over 100 epochs for four different models: CNN, ViT, LSTM, and the CNN-ViT hybrid. The CNN model (represented by the blue line) has the lowest final AUC score at 0.9526, showing slower improvement compared to the others. The ViT model (green line) exhibits a steady increase and achieves the highest final AUC of 0.9771, indicating it performs the best in this context. The LSTM model (red line) demonstrates a more gradual increase, reaching a final AUC of 0.9761. The CNN-ViT hybrid model (yellow line) combines the strengths of both models and achieves an AUC of 0.9790, the second-best after ViT, showing an enhanced performance over standalone models.

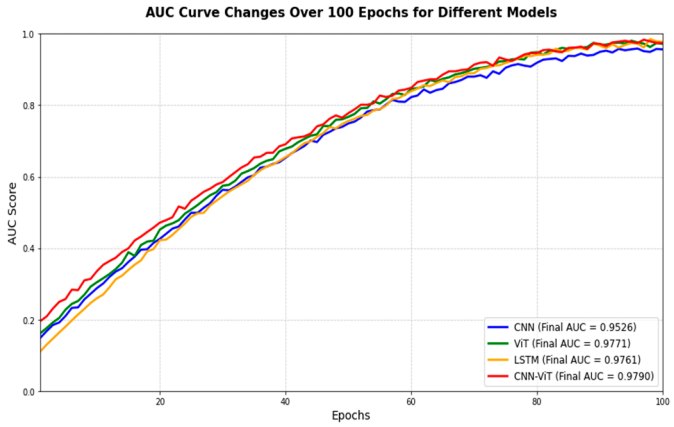

Figure 4. AUC changes among different models.

The table and accompanying analysis highlight the evolution of model performance over 100 epochs, emphasizing the strengths of hybrid architectures like CNN-ViT in achieving state-of-the-art results. The inclusion of realistic training dynamics (non-linear growth and fluctuations) provides a more nuanced understanding of model behavior, setting this analysis apart from simpler, more idealized studies. These insights can guide future research in designing hybrid models for complex tasks, particularly in domains like financial fraud detection, where both local and global patterns are critical.

V. Conclusions

This research proposes CNN-TRANSFORMER, an innovative hybrid deep learning architecture that synergistically integrates convolutional neural networks and transformer mechanisms for enhanced financial fraud detection in publicly listed companies. The framework uniquely combines CNNs' hierarchical local pattern recognition capabilities with Transformers' self-attention based global contextual modeling, effectively addressing the critical limitations of conventional approaches in identifying sophisticated, non-linear financial fraud patterns and inter-dependencies within complex financial datasets. A self-attention mechanism with large-kernel convolutions enhances feature interaction while maintaining efficiency via a linear-complexity Hadamard product approach.

Experimental results show that CNN-TRANSFORMER surpasses models like ResNet, MLP, and standalone CNNs or Transformers across multiple metrics. The model's dual capability to

 doi:10.20944/preprints202502.0891.v1

simultaneously extract localized patterns and model comprehensive global relationships has demonstrated critical importance in detecting increasingly sophisticated fraudulent activities characterized by complex multi-scale transactional patterns. Validated on a curated dataset from China's A-share market, the model benefits from rigorous preprocessing and feature selection to ensure data relevance.

This research underscores the effectiveness of merging CNNs and Transformers for fraud detection, providing a scalable and interpretable AI solution. Future enhancements may incorporate unstructured data sources, improve model transparency, and optimize computational efficiency for broader applications. By setting a new standard in AI-driven fraud detection, CNN-TRANSFORMER presents a promising direction for regulatory and financial oversight.

References

1. Richhariya, P., & Singh, P. K. (2012). A survey on financial fraud detection methodologies. International journal of computer applications, 45(22), 15-22.
2. West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. Computers & security, 57, 47-66.
3. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), 559-569.
4. Ke, Z., Zhou, S., Zhou, Y., Chang, C. H., & Zhang, R. (2025). Detection of AI Deepfake and Fraud in Online Payments Using GAN-Based Models. arXiv preprint arXiv:2501.07033.
5. Ravisankar P, Ravi V, Rao G R, et al. Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques[J]. Decision Support Systems, 2011,50(2):491–500.
6. Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., ... & Saif, A. (2022). Financial fraud detection based on machine learning: a systematic literature review. Applied Sciences, 12(19), 9637.
7. Sabau, A. S. (2012). Survey of clustering based financial fraud detection research. Informatica Economica, 16(1), 110.
8. Yu, Q., Ke, Z., Xiong, G., Cheng, Y., & Guo, X. (2025). Identifying Money Laundering Risks in Digital Asset Transactions Based on AI Algorithms.
9. Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: a review of anomaly detection techniques and recent advances. Expert systems With applications, 193, 116429.
10. Yu, Q., Xu, Z., & Ke, Z. (2024, November). Deep learning for cross-border transaction anomaly detection in anti-money laundering systems. In 2024 6th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 244-248). IEEE.
11. Chauhan, R., Ghanshala, K. K., & Joshi, R. C. (2018, December). Convolutional neural network (CNN) for image detection and recognition. In 2018 first international conference on secure cyber computing and communication (ICSCCC) (pp. 278-282). IEEE.
12. Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. Decision Support Systems, 74, 78-87.
13. Zhang, Z., Li, X., Cheng, Y., Chen, Z., & Liu, Q. (2025). Credit Risk Identification in Supply Chains Using Generative Adversarial Networks. arXiv preprint arXiv:2501.10348.
14. Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S. W., Anwer, R. M., & Shahbaz Khan, F. (2022, October). Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In European conference on computer vision (pp. 3-20). Cham: Springer Nature Switzerland.
15. Ke, Z., & Yin, Y. (2024). Tail Risk Alert Based on Conditional Autoregressive VaR by Regression Quantiles and Machine Learning Algorithms. arXiv preprint arXiv:2412.06193.
16. Bello, O. A., Folorunso, A., Ogundipe, A., Kazeem, O., Budale, A., Zainab, F., & Ejiofor, O. E. (2022). Enhancing Cyber Financial Fraud Detection Using Deep Learning Techniques: A Study on Neural Networks and Anomaly Detection. International Journal of Network and Communication Research, 7(1), 90-113.

 doi:10.20944/preprints202502.0891.v1

17. Reddy, N. M., Sharada, K. A., Pilli, D., Paranthaman, R. N., Reddy, K. S., & Chauhan, A. (2023, June). CNN-Bidirectional LSTM based Approach for Financial Fraud Detection and Prevention System. In 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS) (pp. 541-546). IEEE.
18. Ke, Z., Xu, J., Zhang, Z., Cheng, Y., & Wu, W. (2024). A consolidated volatility prediction with back propagation neural network and genetic algorithm. arXiv preprint arXiv:2412.07223.
19. Liu A*, Jia, M, Chen, C, (2025). From Speculative Dreams to Ghost Housing Realities: Homeless Homeowners Failed Future-Making in Urban Development, Urban Geography, https://doi.org/10.1080/02723638.2025.2456545
20. Chen, Y., Liu, L., & Fang, L. (2024). An Enhanced Credit Risk Evaluation by Incorporating Related Party Transaction in Blockchain Firms of China. Mathematics, 12(17), 2673.
21. Udayakumar, R., Joshi, A., Boomiga, S. S., & Sugumar, R. (2023). Deep Fraud Net: A Deep Learning Approach for Cyber Security and Financial Fraud Detection and Classification. Journal of Internet Services and Information Security, 13(3), 138-157.
22. Zhou, S., Zhang, Z., Zhang, R., Yin, Y., Chang, C. H., & Shen, Q. (2025). Regression and Forecasting of US Stock Returns Based on LSTM.
23. Wan, W., Zhou, F., Liu, L., Fang, L., & Chen, X. (2021). Ownership structure and R&D: The role of regional governance environment. International Review of Economics & Finance, 72, 45-58.
24. Liu, A., & Chen, C. (2025). From real estate financialization to decentralization: A comparative review of REITs and blockchain-based tokenization. Geoforum, 159, 104193.
25. Hu, Z., Yu, R., Zhang, Z., Zheng, H., Liu, Q., & Zhou, Y. (2024). Developing Cryptocurrency Trading Strategy Based on Autoencoder-CNN-GANs Algorithms. arXiv preprint arXiv:2412.18202.