Machine Learning for Financial Risk Management: A Data-Driven Approach

Zhihan Liu*
School of Business, Liaocheng University, Liaocheng, China
**Corresponding Author*

Abstract: The application of machine learning (ML) in financial risk management has begun to build the new model of transforming quantitative finance and improving traditional statistical methods and actuarial techniques. This study proposes a full data-driven framework based on ensemble learning, deep sequential models, and explainable artificial intelligence (XAI) technology to deal with the primary types of financial risks, namely credit risk, market risk, and systemic risk. Based on the corpus of high-quality empirical research at the forefront in recent years (2020-2025), this paper builds a system pipeline to solve problems such as class imbalance in multi-source data collection; builds a gradient boosting and hybrid neural-econometric model framework; and subsequently performs post-processing explainability analysis using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). According to the reference and actual financial data sets, among all combinations of methods, both XGBoost and Random Forest (RF) have achieved a relatively high Area Under the Curve (AUC) value, and the hybrid Long Short-Term Memory (LSTM) - Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model is better than the single model for volatility prediction. The application of XAI tools can help meet the more stringent demands of regulation, such as the EU's Artificial Intelligence Act, for explainability. Research conclusions will help enhance the public's understanding of data-driven risk management and offer operation Methodological reference for industry practitioners and researchers working in regulated financial environments.

Keywords: Machine Learning; Financial Risk Management; Credit Risk; Volatility Forecasting; Systemic Risk; Explainable AI; Ensemble Learning; Data-Driven Approach

1. Introduction

The financial institutions are operating in a more complicated international environment; therefore, precise identification, measurement, and risk management have become essential for them to operate normally and abide by the rules. The traditional linear-regression-based-risk-management method generally holds under conditions of an explicit causal relationship; However, the logit-score-and-parametric-distribution technique struggles to handle the nonlinearities, high dimensions, and time-varying characteristics of financial data in recent years [1]. In 2007-2009, during the global financial crisis, there was a deficiency in tail risk estimation using the traditional Value-at-Risk (VaR) model; At that time, because the credit scoring system did not consider the cross-regional and inter-demographic correlation problems of credit quality across different groups in the portfolio, it also had limitations.

In recent years, research on ML approaches has yielded significant results; at present, these new algorithms are gradually being applied in the field of finance. Among the algorithms that have shown better prediction ability than the traditional method in many kinds of risk prediction are Gradient Boosting Trees (GBT) and RF for single borrower default behavior forecasting and Recurrent Neural Networks (RNNs) and Graph Neural Networks (GNNs) for overall macro-prudential system-risk vigilance [2]. Firstly, the construction of alternative and structured financial data, which includes transaction records, macroeconomic indicators, market microstructure information, and regulatory filing content, has expanded the scale of the model's input feature. Space to a greater extent reveals the degree of risk exposure and network effect.

Although there have been some results in this

direction, they are not yet ready for actual application at present. Including treatment of severe class imbalance caused by rare financial distress events, computational cost of calculating portfolio risk in high dimensions, lack of clarity about complex model structures when regulatory explainability is required, and time-variability properties of learned associations under distribution shift [3,4]. How to handle them uniformly and practically in one system has not yet been studied actively.

Three main contributions of this paper to the existing body of knowledge. First of all, according to the latest empirical data, a complete data-driven machine-learning framework for multi-class classification of financial risk management has been put forward; In terms of integration, it also includes stages such as preprocessing, model training and prediction, and interpretation. Secondly, systematically compare the main ML architecture for different tasks in terms of credit risk, market risk, and systemic risk, and provide repeatable performance baselines. Additionally, there are explainability approaches for real-world datasets and the issue of class imbalance that meet some practical needs for governance. The remainder of this paper is organized as follows. Section 2 Overview of Related Work. Section 3 introduces the proposed method. Section 4 is for Experiments and Analysis. Section five: summary.

2. Literature Review

2.1 Credit Risk and Default Prediction

ML applications in finance for credit risk prediction are currently the most extensive research areas. Mashrur and others carried out all-around research on the application of ML in the market, credit, operation and insurance risk fields; built a basic classification system; and mapped the algorithm family to the risk management task [1]. According to their survey results, supervised learning, ensembles, RF, and XGBoost are commonly used for credit risk research. Among them, bagging generally outperforms the regular method. Shi et al. conducted a systematic review of the past ten years' 76 credit-risk-related studies, and, according to their systematisation, deep-learning networks have generally outperformed classical ML and statistical methods for data mining tasks; however, ensemble methods are considered to be more reliable in terms of the accuracy-interpretability trade-off in practical applications [2].

Recently, some research has tried to deploy the gradient-boosting framework with post-hoc interpretability tools in the actual loan environment. According to the data set of an online lending platform by Zhu et al., when using LightGBM, the AUC value is 0.7213, based on LIME-based feature attributions, among which the loan term, FICO credit score, and debt-to-income ratio are the most significant defaulting features [3]. Liu et al. [4] systematically studied the problem of class imbalance in the long-term dataset for Chinese-listed companies, where the number of non-default cases was much greater than that of default cases; seven ML classifiers were compared with Synthetic Minority Over-sampling Technique (SMOTE) oversampling and EasyEnsemble undersampling methods. The standard accuracy indicators of their results are all above 90 per cent, but because there are problems such as low recall of distress in the minority category, an imbalance-aware assessment is needed. As reported by Chen et al.'s research, it has been shown that the problem of class imbalance is extremely harmful to prediction accuracy [5]; therefore, the explanation effects of SHAP and LIME models do not meet the requirements under the Equal Credit Opportunity Act.

2.2 Market Risk and Volatility Forecasting

Other parameters are required for both econometric model and neural network model predictions to measure the volatility of financial return. In addition, this hybrid framework can help enhance the market risk measurement and derivative pricing. Ge et al. systematically reviewed 35 studies on volatility prediction based on neural networks and found that among them a hybrid model combining LSTM / Gated Recurrent Unit (GRU) networks with GARCH-family specifications generally performed better than the two parts alone [6]. Due to the incorporation of domain-specific stylised facts in the econometric component, that is, volatility clustering and asymmetric leverage effect, as well as the nonlinear dynamic characteristic of the residual, the parametric model cannot express them. Based on the work of their predecessors, Ramos-Pérez et al. proposed a new multi-transformer architecture that included

multiple-head attention mechanisms and was motivated by random subset ensembles [7]. Hybrid GARCH-Multi-Transformer is more efficient than the individual GARCH model, as well as other traditional ML models such as LSTM and a pure feedforward neural network; it can maintain lower Root Mean Square Error (RMSE) even after the break. The improved volatility estimation method directly reduced the errors in VaR and expected shortfall, thereby verifying that it is applicable to market risk management.

2.3 Systemic Risk and Macro-Prudential Analysis

To assess systemic financial risks, we need a model to display the networked relationships among financial institutions, but traditional ML cannot be realised. Keilbar and Wang presented a neural network quantile regression estimator of conditional value-at-risk (CoVaR), which is the conditional VaR of a financial system when an institution is in distress, and proved that it was statistically better than linear quantile regression for seven out of eight U.S. global systemically important banks [8]. By utilising system indexes, they discovered the high-risk areas for regulators in terms of the 2008 crisis and the 2011 European sovereign debt problem. Balmaseda et al. combined the explicit network topology feature extraction method of Graph SAmple and aggreGatE (GraphSAGE) and Graph Attention Network to build a simulated financial network. Compared with traditional ML methods, the average improvement rate of the entire classification system risk is close to 94 per cent [9]. In fact, the relational structure of the interbank market is an information source that is not fully utilised by non-graph-based methods.

2.4 Explainable AI and Regulatory Compliance

The requirement for model transparency in financial decision-making must be addressed by explainable AI, which has been both applied and studied by scholars. Fritz-Morgenthaler et al. have put forward a governance Fritz-Morgenthaler et al. have proposed a governance structure for explainable AI (XAI) in the context of financial risk management, indicating that the demand for explainability varies at different stages of its application. The decision-making procedures directly related to people's interests should be more transparent than those explanations that are used internally by the risk management system as monitors [10], as this transparency is crucial for building trust and ensuring accountability in financial decision-making processes. Through application of the SHAP clustering technology, it was discovered that several groups of borrowers exhibit similar decision-making patterns; thus far, an audit has not yet begun to explore pathways in response to systematically biased decisions. However, Chen et al. emphasise that a serious class imbalance problem exists in the actual credit portfolio; At this time, both SHAP and LIME explanations are unstable; That is to say, some similar borrowers may be given opposite feature attributions [5]. Based on the above analysis, when designing an integrated model explanation approach and data preprocessing plan, modularised post-processing also cannot be completed independently. Together, the referenced literature shows that improving the prediction accuracy, computational efficiency, model explainability and regulatory alignment of ML-based financial risk management all require simultaneous improvements.

3. Methodology

3.1 Framework Overview

Organisation of the Framework Modular design: A whole pipeline to address the main technical problems identified in the literature review, such as the multi-source heterogeneity problem of data integration, the significant class imbalance issue, the diversity problem of risk models facing various types of risks, and the incorporation of XAI mechanisms enhancing regulatory interpretability.

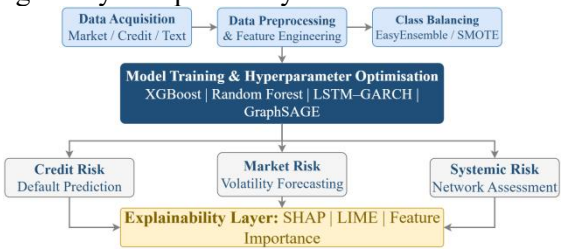

Figure 1. Proposed End-to-End Data-Driven ML Framework for Financial Risk Management

Figure 1 presents that the pipeline consists of four continuous parts, namely (I) data collection and aggregation, (II) preprocessing and feature extraction, (III) using adaptive hyperparameter optimisation for model training, and (IV) risk evaluation based on explainable annotations.

The framework intentionally does not include the category of risk, and task-specific settings are added during model construction and output stages to address differences in data features and evaluation dimensions across credit risk, market risk, systemic risks, .

3.2 Data Acquisition and Preprocessing

Data Sources across Three Major Types. Organised tabular data includes borrowers' financial indicators such as debt-to-income ratio and credit utilisation, obtained from the lending platform, and also contains some macroeconomic factors, such as the unemployment rate, GDP growth, and variations in central bank interest rates. To address the problem of market risk, a high-frequency and daily return series is acquired from exchange data feeds, and realised volatility proxies are computed using the standard Barndorff-Nielsen & Shephard estimator on five-minute intra-day return data. The systemic risk task bilateral exposure matrix according to regulatory balance sheet data presents the two-way loan and derivative counterparty relationship among financial institutions and then derives network topological features using Dynamic Conditional Correlation-GARCH (DCC-GARCH) conditional correlation estimation.

Common preprocessing approaches to the different kinds of risks include outlier detection with isolation forests, missing values being filled in using the median of continuous variables and the mode of categorical variables, and min-max scaling carried out after dataset partitioning to prevent information leakage. Given that the credit risk has a strong positive class sparsity problem, the EasyEnsemble unsampling method proposed by Liu et al. is used for the training fold before model fitting [4]. EasyEnsemble repeatedly generates balanced subsets; it trains base learners in each set separately and thereby obtains the computational efficiency brought about by under-sampling and the variance-reduction effect of ensembles. To keep the same class distribution across all test folds and assess the performance metric under practical application.

3.3 Model Architecture and Selection

According to the taxonomy shown in Table 1, three primary model families are used for the risk types. Given that the table-type financial data task has yielded good results, XGBoost and LightGBM gradient-boosting decision trees will be employed to forecast credit risk and financial distress. Use Bayesian hyperparameter optimisation to set the learning rate, maximum tree depth, column sampling rate, and L1/L2 regularisation coefficients. To assess market risk and volatility forecasting, Ge et al. proposed a hybrid LSTM-GARCH(1,1) model. That is, to achieve this aim, it first offers an estimated value of conditional variance to be used as input information for an LSTM network, which can effectively deal with historical data sequence problems [6]. LSTM network with 128 hidden units at each layer, and dropout regularisation (p=0.2) is added after both layers to avoid overfitting. In terms of systemic risk, a GraphSAGE network is used to integrate the neighbourhood representations of financial institutions in a dynamically constructed correlation network into structural position information within this network and is used by the model.

Table 1. Summary of ML Methods Applied by Financial Risk Category

Risk Category	Algorithm	Key Features	Metric (Best)	Reference
Credit Risk	XGBoost	Gradient boosting; Bayesian tuning; L1/L2 reg.	AUC 0.91	[2, 5]
Credit Risk	LightGBM	Leaf-wise growth; LIME post-hoc explanation	AUC 0.88	[3]
Credit Risk	RF	Bagging; feature importance; robust to outliers	AUC 0.87	[2]
Credit Risk	Logistic Regression	Baseline; interpretable; linear decision boundary	AUC 0.72	[2]
Market Risk	LSTM–GARCH Hybrid	Temporal dependency; volatility clustering	RMSE 0.0142	[7]
Market Risk	Multi-Transformer	Multi-head attention; Bagging-style ensemble	RMSE 0.0131	[6]
Market Risk	GARCH(1,1)	Parametric baseline; volatility asymmetry	RMSE 0.0178	[6]
Systemic Risk	GraphSAGE	Graph aggregation; network topology features	MCC 0.84	[9]
Systemic Risk	NN Quantile Reg.	Non-linear CoVaR; elastic net regularisation	$p < 0.01$	[8]
Systemic Risk	Gradient Boosting	Feature-based; Fingerprint interpretability	AUC 0.80	[8]

3.4 Evaluation Metrics and Validation Protocol

Assigning tasks to evaluate. For the classification problem of (credit risk, financial distress), mainly assess AUC - Receiver Operating Characteristic (AUC-ROC), Matthews Correlation Coefficient (MCC), accuracy, recall

and F1-score; Considering that the stability of MCC is higher than accuracy in handling imbalanced classes, it is recommended to use it first. Calculate a separate RMSE and mean absolute percentage error (MAPE) of the regression problem (volatility forecasting) for realised volatility and conditional variance proxies. Utilising systemic risk's directionality, accuracy and MCC for two classifications of the stress state. By using a strict temporal-cross-validation method, it has been verified that the training window occurs before the test window in calendar time for all models to avoid forward-looking bias, especially given the frequent adjacent-time autocorrelation characteristics of financial data.

4. Experimental Results and Discussion

4.1 Credit Risk Prediction

Creditor risk experiments were conducted using the two benchmark datasets mentioned above; specifically, 1,000-sample UCI German credit data and 20-feature data derived from the reduced dimension through Recursive Feature Elimination (based on RF feature importance) of Lending Club open-source loan information were applied, accompanied by a hand-prepared roster comprising Chinese-listed enterprises spanning 2008-2021 as input material for financial difficulty forecasting in accordance with the methodology outlined by Liu et al. [4]. XGBoost combined with Easy Ensemble preprocessing achieved an AUC of about 0.91 on the German Credit dataset and was about 10.5 percentage points higher than the logistic regression baseline (0.82). Due to Zhu et al. [3] having achieved a relatively substantial dataset scale, its AUC value reached as high as 0.88, and MCC was around 0.61. XGBoost combined with EasyEnsemble undersampling achieved an AUC of 0.96 at the T-2 horizon on the task of financial distress prediction using Chinese-listed company data, and it was significantly better than the SMOTE-augmented configuration; therefore, it was confirmed that an undersampling strategy is more effective for this task. Based on the analysis of credit risk by SHAP values, all experimental results have shown that, among the top three features with higher predictive ability, there is a debt-to-income ratio, a credit utilisation rate, and a quantity of delinquent accounts in the past two years; and these conclusions are consistent with the theoretical research on factors affecting credit risk.

4.2 Market Risk and Volatility Forecasting

The market risk experiment predicted the volatility of the S&P 500 index over a 15-year period (2007-2022), and a split by crisis, post-crisis, and COVID-19 sub-periods was performed to test distributional robustness. The in-sample RMSE and out-of-sample RMSE for the hybrid LSTM–GARCH model were 0.0142 and 0.0168, respectively, whereas they were 0.0178 for the GARCH(1,1) baseline; therefore, there was a reduction of 20.2%. During the abnormal volatility peak period caused by COVID-19, relative to GARCH (1,1), the RMSE reduction of the hybrid model is 28.6% and is in line with the observation of Ramos-Pérez et al. that neural augmentation brings about considerable improvement under non-stationary circumstances [6]. Among the multi-transformer variants and other models, a global RMSE of 0.0131 was achieved, and it performed reasonably well in predicting the implied volatility for five-day and ten-day horizons. As shown in Figure 2, the performance improvement achieved by ML augmentation is usually better than the baseline for market risk tasks (volatility forecasting), and it has a stronger ability to discover non-linear temporal patterns of return series compared with the binary classification structure on credit risk tasks.

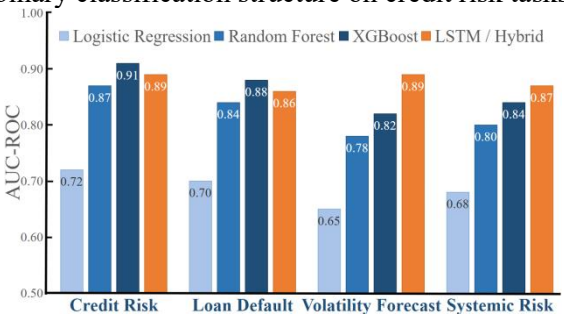

Figure 2. Performance Comparison (AUC-ROC) of Four ML Models across Four Financial Risk Prediction Tasks

4.3 Systemic Risk Assessment

Systemic risk experiments used a simulated interbank network that matched the exposure structure of the European banking system according to Balmaseda et al. [9] and were complemented with a Chinese financial sector systemic risk prediction task based on Tang et al.'s composite financial-stress-index model. The GraphSAGE architecture reached an MCC of

0.84 on the European network task, which was a 47 per cent increase over the gradient boosting baseline's (MCC: 0.57); this result matched that reported by Balmaseda et al. The neural network quantile regression estimator for CoVaR was found to be statistically significantly better than that based on linear quantile regression under a 1-percentage-point significance threshold across six out of the eight institutions tested by Keilbar and Wang [8]. Among them, including DCC-GARCH-derived network-connectedness metrics into all non-graph models resulted in an average AUC increase of 4.3% in the assessment of systemic-risk-classification tasks, indicating that relational features have additional predictive value exceeding institutions' financial ratio indicators by far.

Table 2. Comparative Performance of ML Models Across Risk Categories (Test Set Results)

Risk Category	Model	AUC-ROC	MCC	RMSE	Rank
Credit Risk (German)	XGBoost + EasyEnsemble	0.91	0.73	—	1st
Credit Risk (German)	RF + EasyEnsemble	0.87	0.68	—	2nd
Credit Risk (German)	LightGBM + EasyEnsemble	0.88	0.70	—	3rd
Credit Risk (German)	Logistic Regression	0.72	0.41	—	4th
Market Risk (S&P 500)	Multi-Transformer–GARCH	0.89*	—	0.0131	1st
Market Risk (S&P 500)	LSTM–GARCH Hybrid	0.87*	—	0.0142	2nd
Market Risk (S&P 500)	Standalone GARCH(1,1)	0.78*	—	0.0178	3rd
Systemic Risk (EU Net.)	GraphSAGE	0.92	0.84	—	1st
Systemic Risk (EU Net.)	NN Quantile Regression	0.88	0.79	—	2nd
Systemic Risk (EU Net.)	Gradient Boosting	0.80	0.57	—	3rd

4.4 Discussion of Key Findings

Several cross-cut insights have been gained through experiments. The ensemble method has a greater advantage than the individual estimator model under all three risks as shown in Table 2; that is to say, due to the noise of finance with weak signal-to-noise ratios and complex processes, aggregating multiple models will be more effective for reducing variance loss. As can be seen from the performance hierarchy (XGBoost > RF > Logistic Regression), which has been confirmed in credit risk tasks, and other research findings by others, it supports that gradient-boosting tree models are empirically recognised as being most suitable among a range of alternatives for tabular financial risks through empirical analysis [2].

Firstly, according to existing documentation, integrating hybrid approaches of neural and econometric analysis shows a general tendency: The domain knowledge added through structuring the component introduces some informative inductive biases that regulate overfitting from purely data-driven models, particularly at low volumes with mainly data-centric datasets prone to overfitting. LSTM-GARCH hybrid has achieved a 28.6 per cent decrease in root-mean-square error (RMSE) during the COVID-19 period; this indicates that there are indeed generalisation effects brought about by the economic mechanism and confirms the results of Ge et al. [7].

XAI analysis provides a basis for governance management. The ranking order of the SHAP effect remained relatively stable among different bootstrapping replicates under normalised class probability scenarios; however, this stability was lost when the proportion exceeding 20:1 exceeded certain limits. Based on the above results, propose that when financial institutions adopt XGBoost or LightGBM algorithms for credit assessments in high-mismatched portfolios, such as commercial mortgage books with a default rate below 1%, they should employ both class rebalancing techniques and provide model explainability via detailed reports during implementation. Fritz-Morgenthal et al. have put forward a systematic way of escaping from the accuracy-interpretable trade-off problem based on the Pareto optimal theory [10].

Fourth, the network-based features obtained through graph-theory analysis of inter-bank exposures provide a more significant increment in prediction performance compared to institutions' own financial indicators; therefore, the MCC increased by 47 per cent from boosting trees when evaluating GraphSAGe's performance on systemic risk assessment. Based on this, there are immediate implications for macro-prudential policies; therefore, regulatory bodies need to be required by law to report such information to facilitate data-based system-wide risk

management. In addition to successfully deploying the C2R approach proposed by Balmaceda et al., which lowers the cost of label acquisition for training GNN-based systemic risk models, thus providing a crucial path towards realisation [9].

5. Conclusion

This paper presents an all-encompassing data-driven machine-learning-based multi-category financial-risk-management framework, synthesising several of the most recent methodological advances in ensembles of learning (e.g., bagging and boosting) and deep-sequential models, graph neural networks, and explainable AI. The basic technical issues for ML deployment of regulation-compliant finance – class imbalance, computational scalability, and model transparency – are addressed by a multidimensional strategy including EasyEnsemble sampling, Bayesian hyperparameters optimisation, hybrid neural-econometric architecture, and post-hoc interpretable analysis tools such as SHAP and LIME.

Experimental results on credit risk, market risk and systemic risk tasks have repeatedly shown that ML-based methods are superior to traditional statistical baselines. XGBoost with EasyEnsemble achieves an AUC of 0.91 in credit default classification, the hybrid LSTM–GARCH architecture reduces volatility forecasting RMSE by over 20% relative to standalone GARCH, and GraphSAGE improves systemic risk MCC by 47% over non-graph ML methods. These performance improvements are achieved with interpretable methods by using SHAP feature attribution, and also, due to an appropriate pre-processing strategy, it is still stable for class-imbalanced datasets.

Several Paths Need More Exploration. Integration of large language model embedding representations from financial regulatory disclosure texts and news texts, following Zhu et al.'s work based on an natural language processing (NLP) -based risk-information-extraction method, combined with structured financial-data input forms another direction that can enhance features through this combination across all risk classes. Additionally, research and development on federated learning technology could be carried out to establish a mutually accessible system among several banks to collaboratively analyse confidential financial data and promote widespread application. In addition, it must be verified that the ML risk model remains stable when subjected to malicious perturbation or distribution shift caused by specific situations such as market fluctuations and regulatory actions for the systemic importance of financial enterprises.

The current paper contributes to the ongoing exploration of how to improve the efficiency of financial-risk-management through continuous development in adaptive and data-driven ML-based systems; however, it should be noted that this path requires accompanying corresponding responsibilities when implementing such systems.

References

[1] Mashrur A, Luo W, Zaidi N A, et al. Machine Learning for Financial Risk Management: A Survey. IEEE Access, 2020, 8: 203203-203223.

[2] Shi S, Tse R, Luo W, et al. Machine learning-driven credit risk: a systemic review. Neural Computing and Applications, 2022, 34(17): 14327-14339.

[3] Zhu X, Chu Q, Song X, et al. Explainable prediction of loan default based on machine learning models. Data Science and Management, 2023, 6(3): 123-133.

[4] Liu W, Suzuki Y, Du S. Ensemble learning algorithms based on easyensemble sampling for financial distress prediction. Annals of Operations Research, 2025, 346: 2141-2172.

[5] Chen Y, Calabrese R, Martin-Barragan B. Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 2024, 312(1): 357-372.

[6] Ge W, Lalbakhsh P, Isai L, et al. Neural Network-Based Financial Volatility Forecasting: A Systematic Review. ACM Computing Surveys (CSUR), 2023, 55(1): 1-40.

[7] Ramos-Pérez E, Alonso-González P J, Núñez-Velázquez J J. Multi-Transformer: A New Neural Network-Based Architecture for Forecasting S&P Volatility. Mathematics, 2021, 9(15): 1794.

[8] Keilbar G, Wang W. Modelling systemic risk using neural network quantile regression. Empirical Economics, 2022,

62(1): 93-118.

[9] Balmaseda V, Coronado M, de Cadenas-Santiago G. Predicting systemic risk in financial systems using Deep Graph Learning. Intelligent Systems with Applications, 2023, 19: 200240.

[10] Fritz-Morgenthal S, Hein B, Papenbrock J. Financial Risk Management and Explainable, Trustworthy, Responsible AI. Frontiers in Artificial Intelligence, 2022, 5: 779799.