

Research on Stock Price Prediction and Its Application Based on an Optimized LSTM Model

Dingyi Guan

School of Economics, Hangzhou Dianzi University, Hangzhou 310000, P.R. China

gdy573464659@163.com

Abstract

To address high dimensionality, strong noise and critical signal attenuation in stock price time series, as well as the limited accuracy and generalization performance of vanilla LSTM models, this paper constructs a hybrid FPCA-Attention-LSTM prediction model. The model first applies Functional Principal Component Analysis (FPCA) to reduce the dimensionality of high-dimensional technical indicators while preserving the dynamic morphological characteristics of time series. Then, a temporal attention mechanism is introduced to assign weights across time steps, enhancing the ability to capture critical market signals. Combined with a two-layer LSTM network, the model realizes the prediction of index rise-fall directions. Empirical tests are conducted using weekly data of the CSI Consumer Index from January 2019 to July 2024. The results show that the model achieves a rise-fall prediction accuracy of 65.38% and an F1-score of 0.4375, with comprehensive performance superior to benchmark models such as the vanilla LSTM and PCA-LSTM. The quantitative market timing strategy derived from the model output demonstrates better risk control performance and capital preservation capacity during market downturns. This research can provide methodological references for index-oriented quantitative timing investment.

Keywords

Functional Principal Component Analysis; Attention Mechanism; LSTM; Stock Price Prediction; Quantitative Trading.

1. Introduction

Price forecasting in financial markets is a core research topic in the field of quantitative investment. Stock prices, as a typical class of nonlinear and non-stationary time series, are influenced by intertwined factors such as macroeconomics, industry cycles, market sentiment, and extreme risks in commodity markets^[1], featuring complex fluctuation patterns and pervasive noise. Traditional linear statistical models struggle to effectively capture their dynamic characteristics. In recent years, Long Short-Term Memory (LSTM) networks have become a mainstream tool for stock price prediction by virtue of their strength in capturing long-term temporal dependencies. However, when processing high-dimensional technical indicators and long-period trading sequences, the vanilla LSTM model still suffers from bottlenecks such as dimensional redundancy and attenuation of critical information, leaving room for further improvement in prediction accuracy and generalization performance.

From the perspective of research evolution, stock price prediction methods have undergone gradual evolution from traditional statistics and machine learning to deep learning. Early linear time series models such as ARIMA are constrained by strict linear assumptions and fail to adapt to the nonlinear fluctuations and abnormal patterns of financial markets^[2]. Although traditional machine learning methods such as Support Vector Machines and gradient boosting trees improve the characterization of nonlinear relationships, they are insensitive to the sequential

order of time series, and their feature construction relies heavily on manual feature engineering, resulting in obvious limitations in generalization under high-dimensional noise environments [3].

LSTM alleviates the vanishing gradient problem of Recurrent Neural Networks (RNNs) through its gating structure and can better capture long-term dependencies in time series, gradually becoming the mainstream method for financial time series prediction^[4]. Existing studies on LSTM performance optimization are mainly carried out from four dimensions:

First, signal denoising and decomposition integration. In view of the prominent non-stationarity and complex noise components of financial time series, studies in this stream generally adopt a multi-scale hierarchical modeling framework. Original sequences are decomposed and denoised via various modal decomposition methods to separate the trend component and fluctuation components of different frequencies, which are then fed into LSTM networks for separate modeling and prediction. The final overall prediction result is obtained by synthesizing outputs of all components. Jinlu Guo (2020)^[5] constructed a two-layer decomposition framework combining variational mode decomposition and ensemble empirical mode decomposition, and empirical evidence showed that its prediction error for the CSI 300 Index is significantly lower than that of single decomposition models. Yiyue He et al. (2020)^[6,7] introduced complete ensemble empirical mode decomposition with adaptive noise to further reduce the lag of prediction results, and the index timing strategy built upon this method also significantly outperformed benchmark strategies such as traditional moving averages and support vector machines. Other studies have further improved prediction performance in terms of inflection point fitting accuracy and model stability by combining complementary ensemble empirical mode decomposition with broad learning systems^[8] and adaptive Fourier decomposition^[9] respectively.

Second, network structure improvement. Hybrid architectures are adopted to integrate the feature extraction advantages of different models. Jingjing Geng et al. (2021)^[10] constructed a CNN-LSTM hybrid model, which improves temporal modeling efficiency by extracting local fluctuation features through convolutional layers. Subsequent studies further introduced volume-weighted average price to characterize investor behavior, integrated both technical and fundamental indicators^[11], and optimized parameters via orthogonal design, achieving significant accuracy advantages in medium- and long-term prediction.

Third, parameter optimization and generalization expansion. Gülmez (2025)^[12] and Ritu et al. (2025)^[13] respectively introduced intelligent optimization algorithms to achieve global parameter optimization. Cross-country empirical evidence shows that LSTM maintains stable generalization across different markets and time horizons, and outperforms traditional benchmark models in both prediction accuracy and stability^[14]. In addition, LSTM has been extended to scenarios such as pairs trading to improve investment returns through dynamic parameter adjustment^[15].

Fourth, multi-variable information focusing. To address long-sequence information attenuation and multi-variable interaction, attention mechanisms have been introduced into financial time series modeling. The Transformer-LSTM quantile regression model processes multi-variable interactions in parallel through multi-head attention, and captures the evolution of extreme risks in global stock markets more accurately^[16]. However, temporal attention optimization specifically for index rise-fall prediction still needs further in-depth exploration.

In summary, most existing studies improve LSTM from a single dimension, and systematic verification of the integrated application and synergistic effect of functional feature dimensionality reduction and attention mechanisms remains scarce. Meanwhile, specialized empirical research targeting consumer industry indices is insufficient, and applied research on constructing supporting quantitative trading strategies based on model output is also limited.

Against this background, this paper constructs an FPCA-Attention-LSTM hybrid prediction model. It realizes feature dimensionality reduction and noise filtering through functional principal component analysis, introduces an attention mechanism to strengthen the weight of key temporal signals, and completes time series prediction modeling in combination with the LSTM network. The CSI Consumer Index is selected as the research sample for empirical testing, and a quantitative market timing strategy is designed based on model prediction signals to verify its practical application effect, providing methodological references for quantitative investment practice of industry indices.

2. Model Construction and Trading Strategy Design

This chapter systematically elaborates the core methods and model architecture of the study. It first explains the core logic and adaptability of three basic methods, then constructs the hybrid prediction model, and finally designs the supporting quantitative trading strategy, providing a methodological framework for subsequent empirical testing.

2.1. Theoretical Foundations

2.1.1. Long Short-Term Memory Network

Long Short-Term Memory (LSTM) is a classic improved architecture of Recurrent Neural Networks (RNNs). It realizes the screening, updating and output of temporal information through three gating units—the forget gate, input gate and output gate—effectively solving the vanishing gradient problem that plagues traditional RNNs when processing long sequences, and can accurately capture long-term dependencies in time series. The update logic of its core cell state is as follows:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (1)$$

where f_t and i_t are the output values of the forget gate and input gate respectively; C_{t-1} denotes the cell state at the previous time step; x_t is the input vector at the current time step; h_{t-1} represents the hidden state at the previous time step; W_c and b_c are the weight matrix and bias term of the candidate cell state respectively; $\odot$ denotes element-wise multiplication (Hadamard product). The forget gate filters redundant historical noise, while the input gate controls the proportion of new information written into memory cells. The two jointly realize the dynamic update of memory cells. The output gate regulates the hidden state information transmitted externally based on the updated cell state.

2.1.2. Functional Principal Component Analysis

Functional Principal Component Analysis (FPCA) is an extension of traditional Principal Component Analysis (PCA) in function space. Its core idea is to treat discrete time series observations as sampling points of continuous functions, and perform orthogonal decomposition in function space to achieve data dimensionality reduction. The core calculation process is as follows: first, P-spline smoothing is applied to eliminate random noise and generate continuous function curves; then the covariance function is estimated and eigen decomposition is performed; finally, the original function is projected onto the orthogonal basis to obtain the low-dimensional principal component score sequence. In this paper, the number of principal components is selected with the criterion that the cumulative explained variance ratio reaches 90%, so as to retain most valid information while compressing dimensions.

Compared with traditional PCA, FPCA has three notable advantages in processing financial time series. First, it adopts functional data representation, which fully retains the continuity and dynamic characteristics of time series, and avoids the destruction of key morphological information such as trend inflection points and fluctuation slopes caused by discrete dimensionality reduction. Second, with the P-spline smoothing mechanism, it effectively retains key dynamic features of financial data while performing denoising. Third, combined with the

explained variance ratio threshold, it realizes adaptive selection of the number of principal components, striking a balance between dimensionality reduction efficiency and information retention.

2.1.3. Temporal Attention Mechanism

The core idea of the temporal attention mechanism is to dynamically assign weights to hidden states at different time steps, guiding the model to focus on information intervals that contribute more to the prediction target. Its calculation logic is as follows: first, the importance score of the hidden state at each time step is calculated via a scoring function, and normalized by the Softmax function to generate corresponding weight coefficients. Finally, the hidden states are weighted and summed to obtain a context vector integrating global key information. The core formula is as follows:

$$c = \sum_{t=1}^T \frac{\exp(e_t)}{\sum_{i=1}^T \exp(e_i)} \cdot h_t \quad (2)$$

where e_t is the attention score at time step t , measuring the importance of information at time t to the prediction target; h_t is the hidden state at time t output by LSTM; c is the final output context vector; T is the time step length of the input sequence.

In stock price prediction scenarios, this mechanism can adaptively amplify signal weights corresponding to key events such as financial report disclosures and policy adjustments, and suppress interference from daily trading noise. It effectively alleviates the attenuation of key information in long sequences and improves the model's accuracy in identifying market trend inflection points.

2.2. FPCA-Attention-LSTM Hybrid Prediction Model

Aiming at the problems of high-dimensional feature redundancy and attenuation of temporal dynamic information in financial time series, existing studies have combined principal component dimensionality reduction and attention mechanism with LSTM to construct prediction models, and verified the effectiveness of this technical framework in consumer industry index prediction^[17]. However, traditional principal component analysis follows the linear dimensionality reduction logic of discrete vectors, which tends to damage the continuous dynamic characteristics of time series. To address this issue, this paper constructs an LSTM prediction model integrating functional principal component dimensionality reduction and temporal attention (FPCA-Attention-LSTM), which adopts a three-layer logical architecture as a whole:

The first layer is the feature dimensionality reduction module. The model input is the preprocessed weekly technical indicator sequence, with a time window length of 5—that is, feature data of 5 consecutive weeks are used to predict the rise-fall direction of the index in the next week. FPCA is applied to reduce the dimensionality of original high-dimensional indicators. P-spline smoothing preserves temporal dynamic characteristics, and principal component score sequences with a cumulative explained variance ratio exceeding 90% are extracted as input for subsequent networks, compressing data dimensions while optimizing input feature quality.

The second layer is the temporal feature extraction and attention module. A stacked two-layer LSTM structure extracts temporal dependency features layer by layer. The first LSTM layer is configured with 64 hidden units and outputs a complete temporal hidden state sequence to capture short-term price fluctuation patterns. An attention layer is embedded after the first LSTM layer to calculate dynamic weights for the output hidden state sequence, perform weighted aggregation of full temporal information, and strengthen the signal contribution of key time steps. The second LSTM layer is configured with 32 hidden units to extract deep-level information from the weighted feature sequence and characterize long-term trend patterns. L2 regularization (regularization coefficient $\lambda=0.001$) and recurrent Dropout (dropout rate = 0.1)

are applied to both LSTM layers. Batch normalization layers and Dropout layers (dropout rate = 0.2) are added between layers to suppress overfitting and accelerate model convergence.

The third layer is the output module. First, a fully connected layer with 16 neurons performs feature transformation on the context vector output by the attention layer, and the LeakyReLU activation function is adopted to alleviate the vanishing gradient problem. Finally, through an output layer with a single neuron combined with the Sigmoid activation function, the probability of index rise in the next week is output, completing the binary classification prediction of rise-fall directions.

2.3. Design of Quantitative Trading Strategy

To verify the practical application value of the model, this paper constructs a trend-following quantitative market timing strategy based on the rise-fall probability signals output by the model, combined with classic technical indicators. The strategy is designed from three dimensions: signal generation rules, risk control mechanism and performance evaluation system.

Signal generation takes model prediction as the core basis, combined with technical indicators such as MACD and 5-week moving average for secondary filtering. A buy signal is triggered when the model-predicted probability of index rise in the next week is higher than the preset threshold, MACD forms a golden cross, and the 5-week moving average turns upward. A sell signal is triggered when the predicted rise probability is lower than the threshold, MACD forms a death cross, and the 5-week moving average turns downward. When the model prediction falls within the neutral interval, or the two types of signals conflict, the original position is maintained or the investor stays on the sidelines.

Risk control rules are formulated in line with the characteristics of weekly trading. Trading signals are executed via full position rotation for a single underlying asset. A 1.5% take-profit threshold and a 0.5% stop-loss threshold are set, and positions are closed in advance when returns during the holding period hit the thresholds. The minimum holding period is set at 25 trading days (approximately 5 trading weeks) to avoid excessive trading caused by frequent signal switching and reduce transaction costs.

Four core indicators—cumulative return, compound annual growth rate, Sharpe ratio and maximum drawdown—are selected for performance evaluation, to comprehensively assess strategy performance from three dimensions: profitability, risk-adjusted return and downside risk.

3. Empirical Analysis

3.1. Indicator Selection and Data Sources

This paper selects the CSI Consumer Index (index code: 000932) as the research sample. Basic market data are obtained from the Baostock financial data API. The sample period spans from January 4, 2019 to July 15, 2024, with weekly sampling frequency, yielding a total of 282 trading weeks of observations. Raw indicators include 8 fundamental variables: opening price, highest price, lowest price, closing price, trading volume, turnover, turnover rate and price change rate. To fully characterize market volume-price characteristics, 25 technical indicators are calculated using the TA-Lib technical analysis library, which are mainly divided into five categories: trend-tracking indicators (e.g., MA, EMA, SAR, ADX), momentum indicators (e.g., MACD, CMO), volatility indicators (e.g., ATR, Bollinger Bands), volume-based indicators (e.g., OBV, AD), and price pattern indicators (e.g., Hilbert transform indicators, average price). The above 33 indicators reflect market conditions from different dimensions, providing a feature basis for subsequent functional principal component analysis.

3.2. Data Preprocessing

The time series data are preprocessed as follows: missing values are handled by combining forward filling and backward filling to ensure the continuity of the time series. The Interquartile Range (IQR) method is used to identify and truncate outliers, which suppresses the interference of abnormal fluctuations on the model while avoiding time series discontinuity caused by data deletion. Min-max normalization maps all indicators to the interval $[0,1]$ to eliminate the influence of dimensional differences. The extreme statistics required for normalization are calculated only based on the training set, and the test set uses the statistics from the training set for mapping to avoid look-ahead bias.

3.3. Results of Functional Principal Component Analysis

After P-spline smoothing preprocessing, FPCA is performed to reduce the dimensionality of the 33-dimensional technical indicators. The eigenvalues and explained variance ratios of each principal component are shown in Table 1.

Table 1 Statistics of Principal Component Explained variance ratios

Principal Component	Eigenvalue	Explained variance ratio (%)
PC1	1151.63	34.24
PC2	649.02	19.30
PC3	330.33	9.82
PC4	226.47	6.73
PC5	165.27	4.91
PC6	155.21	4.62
PC7	134.82	4.01
PC8	77.66	2.31
PC9	74.40	2.21
PC10	72.92	2.17
Cumulative	—	90.33

The results show that the first 10 principal components achieve a cumulative explained variance ratio of 90.33%, meeting the preset 90% threshold. While realizing data dimensionality reduction, it retains most of the effective information from the original indicators.

From the perspective of economic implications, the first principal component (PC1) mainly reflects the overall market trend. All technical indicators show consistent loading directions on this dimension, reflecting the dominant role of systematic risk factors. The second principal component (PC2) characterizes the deviation between short-term momentum and long-term trend, with differentiated loadings for short-cycle and long-cycle moving average indicators. The third principal component (PC3) mainly reflects market volatility characteristics, with high loadings for volatility indicators.

3.4. Model Training and Parameter Setting

Sample data are divided chronologically: the first 80% serves as the training set (sample period: January 2019 to May 2023), and the remaining 20% serves as the test set (sample period: June 2023 to July 2024), to avoid introducing future information. Aiming at the imbalanced distribution of rise and fall samples in financial time series, a dynamic class weight adjustment strategy is adopted to balance the model's learning effect on the two categories.

The main parameters of the model are set as follows: the number of hidden units of the two LSTM layers is 64 and 32 respectively, with an additive attention layer embedded between them. The Adam optimizer is adopted, combined with an early stopping strategy, Dropout regularization and batch normalization to suppress model overfitting. After training, the

optimal classification threshold is determined based on the precision-recall curve of the test set, with the criterion of maximizing the F1-score.

3.5. Analysis of Model Prediction Results

3.5.1. Benchmark Model Comparison

To verify the performance advantages of the proposed hybrid model and analyze the independent effect of each functional component, four benchmark models are set for comparative analysis: vanilla LSTM, PCA-LSTM, FPCA-LSTM and Attention-LSTM. The prediction performance of each model on the test set is shown in Table 2.

Table 2 Performance Comparison of Different Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Vanilla LSTM	51.92	43.24	80.00	0.5614
PCA-LSTM	36.54	33.33	65.00	0.4407
FPCA-LSTM	59.62	47.37	45.00	0.4615
Attention-LSTM	50.00	33.33	30.00	0.3158
FPCA-Attention-LSTM	65.38	58.33	35.00	0.4375

The following conclusions can be drawn from the comparison results in Table 2. First, the FPCA-Attention-LSTM model constructed in this paper delivers the best comprehensive performance, with a test set accuracy of 65.38%, 5.76 percentage points higher than the suboptimal FPCA-LSTM model, and the improvement in precision is even more pronounced, verifying the effectiveness of the fusion strategy of FPCA dimensionality reduction and attention mechanism. Second, FPCA significantly outperforms traditional PCA in dimensionality reduction effect. The accuracy of the FPCA-LSTM model is 23.08 percentage points higher than that of the PCA-LSTM model. The reason is that traditional PCA adopts a linear dimensionality reduction method based on discrete vectors, which damages the continuity of time series, while the functional analysis framework of FPCA better fits the data characteristics of financial time series. Third, FPCA and attention mechanism exert a synergistic effect. Introducing the attention mechanism on the basis of FPCA-reduced features improves model accuracy by 5.76 percentage points. However, adding the attention mechanism directly to original features instead reduces model accuracy. This result indicates that the effectiveness of the attention mechanism is closely related to the quality of input features. Functional features purified by FPCA dimensionality reduction are more compatible with the information filtering logic of the attention mechanism, and their combination can strengthen the weight of key information while filtering out noise.

3.5.2. Confusion Matrix Analysis

The confusion matrix of the model's prediction results on the test set is shown in Table 3. The test set contains 52 samples in total, and the overall prediction accuracy of the model is 65.38%. Among 32 actual falling samples, 27 are correctly identified, corresponding to a fall recall rate of 84.38%. Among 20 actual rising samples, 7 are correctly identified, corresponding to a rise recall rate of 35.00%.

Table 3 Confusion Matrix

True Value \ Predicted Value	Fall	Rise
Fall	27	5
Rise	13	7

During the test period from June 2023 to July 2024, the CSI Consumer Index was dominated by a downward trend, with falling samples accounting for 61.5% of the total. The prediction results show that the model has strong ability to identify falling markets, but is relatively

cautious in judging rising markets, with more false negatives than false positives. From the perspective of strategy application, this prediction characteristic helps control investment drawdown risk during market downturns, which is consistent with the risk control objective of timing trading strategies.

3.6. Backtesting of Quantitative Trading Strategy

3.6.1. Backtesting Setup

The trading strategy constructed based on model output is backtested via the Backtrader quantitative trading framework in the Python environment. The backtesting period runs from March 9, 2021 to July 15, 2024, with initial capital set at CNY 1,000,000 and transaction commission at 0.03%. A buy-and-hold strategy for the CSI Consumer Index is adopted as the benchmark. By comparing the performance of the two strategies, the effectiveness of the proposed model strategy is verified.

3.6.2. Analysis of Backtesting Results

The historical backtesting performance comparison of the two strategies is shown in Table 4.

Table 4 Backtesting Results of Two Strategies

Indicator	Proposed Model Strategy	Buy-and-Hold Strategy
Total Return	0.13%	-26.19%
Compound Annual Growth Rate	0.04%	-8.66%
Maximum Drawdown	0.09%	29.62%
Calmar Ratio	0.44	-0.29
Sharpe Ratio	-5.7470	-1.2650

Backtesting results show that the proposed model strategy outperforms the buy-and-hold strategy in the overall downward market during the test interval. In terms of returns, the proposed strategy achieves positive returns and generates significant excess returns compared with the buy-and-hold strategy, effectively avoiding most market downside risks through timing operations. In terms of risk, the maximum drawdown of the proposed strategy is far lower than that of the buy-and-hold strategy, reflecting strong downside risk control ability and capital preservation effect. In terms of risk-adjusted returns, the Calmar ratio of the proposed strategy is positive while that of the buy-and-hold strategy is negative, indicating better return per unit of drawdown.

Affected by the persistent market decline during the backtesting period, the Sharpe ratios of both strategies are negative, indicating that returns of both strategies are lower than the contemporaneous risk-free rate. In this scenario, the horizontal comparison reference value of this indicator is limited. Comprehensively considering absolute return, drawdown control and capital preservation, the proposed model strategy demonstrates more advantages in a downward market environment.

4. Research Conclusions and Implications

4.1. Research Conclusions

Aiming at the problems of high-dimensional technical feature redundancy, loss of temporal dynamic information and long-sequence dependency attenuation in financial index rise-fall prediction, this paper constructs an FPCA-Attention-LSTM hybrid prediction model and conducts empirical tests based on weekly data of the CSI Consumer Index. The main conclusions are as follows:

First, functional principal component analysis is suitable for dimensionality reduction of financial time series. Compared with traditional PCA, FPCA performs orthogonal decomposition

on smoothed indicator trajectories in function space, which can retain temporal dynamic information such as trend slopes and inflection points. Empirical results show that selecting the first 10 principal components achieves a cumulative explained variance ratio of 90.33%, compressing 33-dimensional original indicators to 10 dimensions while balancing dimensionality reduction efficiency and information retention.

Second, the fusion of FPCA and attention mechanism improves the prediction performance of LSTM models, and the two exert synergistic effects. FPCA denoises and purifies high-dimensional features, while the attention mechanism weights and strengthens key temporal information. The combined model achieves a test set accuracy of 65.38%, with comprehensive performance superior to benchmark models such as vanilla LSTM and PCA-LSTM.

Third, the quantitative market timing strategy constructed based on this model has practical application value. In the downward market backtesting interval, the strategy avoids most downside risks through timing operations, with a maximum drawdown lower than the buy-and-hold strategy. It effectively realizes capital preservation and delivers sound risk control effect, which can provide reference for index investment in bear markets.

4.2. Practical Implications

From the perspective of individual investors, the proposed model framework can serve as a reference tool for industry index timing. It simplifies high-dimensional technical indicator systems through functional principal component dimensionality reduction, and screens key market signals combined with the attention mechanism, alleviating information overload in investment decision-making. During market downturns, investors can leverage the model's ability to identify falling markets to control portfolio drawdown and assist in constructing defensive investment strategies.

From the perspective of quantitative institutions, on the one hand, this method can be extended to multi-industry indices and stock pools, and multi-source information such as macroeconomic factors and market sentiment indicators can be incorporated to further optimize prediction performance. On the other hand, the integration idea of FPCA dimensionality reduction and attention mechanism can be applied to other time series prediction scenarios. Improving model input quality helps enhance information utilization efficiency, providing technical references for the development of quantitative market timing strategies.

4.3. Research Limitations and Future Outlook

There are still some limitations in this study, which can be further deepened from three aspects: sample expansion, information supplementation and interpretability improvement. In terms of sample coverage, this paper only conducts empirical tests on weekly data of the CSI Consumer Index since 2019, and the sample period is dominated by downward market trends. The return performance of the model in bull markets remains unvalidated, and its generalization ability across industry indices and individual stocks needs further testing. In terms of information dimension, the current model only constructs the feature system based on price-volume technical indicators, without incorporating multi-dimensional information such as macroeconomic factors, market sentiment indicators and policy texts. Future research can enrich input features by fusing multi-source data to improve the model's comprehensive characterization of market conditions. In terms of model application, the black-box nature of deep learning models restricts their application in practical scenarios to a certain extent. Subsequent research can introduce interpretability analysis methods such as SHAP and LIME to clarify the transmission mechanism of various market factors on index prices, and strengthen the economic interpretability and practical reference value of model conclusions.

References

- [1] Zhang Q., Yan Z., Hu J. Extreme Risk of International Crude Oil and Stock Price Prediction of Chinese Overseas Listed Enterprises — An Empirical Study Based on Machine Learning Model. *Journal of Econometrics*, 2026: 1-24.
- [2] Ci B., Zhang P. Financial Time Series Prediction Based on ARIMA-LSTM Model. *Statistics & Decision*, 2022, 38(11): 145-149. DOI: 10.13546/j.cnki.tjyc.2022.11.029.
- [3] Zhang J., Gan Y. Shanghai Composite Index Prediction Based on Deep Learning Support Vector Machine. *Statistics & Decision*, 2019, 35(02): 176-178. DOI: 10.13546/j.cnki.tjyc.2019.02.041.
- [4] Zhang X. Research on Application of LSTM Model Stock Price Prediction Based on Deep Learning. *Science Technology and Innovation*, 2024(23): 23-26. DOI: 10.15913/j.cnki.kjycx.2024.23.006.
- [5] Guo J. Research on CSI 300 Index Prediction Based on VMD-EEMD-LSTM Model. *Modern Finance and Economics (Journal of Tianjin University of Finance and Economics)*, 2020, 40(08): 31-44. DOI: 10.19559/j.cnki.12-1387.2020.08.003.
- [6] He Y., Li P., Han J. Research on Stock Market Index Prediction Modeling Based on CEEMDAN-LSTM. *Statistics and Information Forum*, 2020, 35(06): 34-45.
- [7] He Y., Gao N., Han J., et al. Research on Quantitative Timing of Stock Index Based on Long Short-Term Memory Network. *Statistics & Decision*, 2020, 36(23): 128-133. DOI: 10.13546 /j .cn ki.tjyc. 2020.23.028.
- [8] Han Y., Zhang D., Sun K., et al. Research on a New Stock Prediction Model Combining Long Short-Term Memory Network and Broad Learning. *Operations Research and Management Science*, 2023, 32(08): 187-192.
- [9] Wang J., Mai J. Financial Signal Denoising and Prediction Based on AFD-LSTM Model. *Financial Theory and Teaching*, 2024, 42(04): 18. DOI: 10.13298/j.cnki.ftat.2024.04.007.
- [10] Geng J., Liu Y., Li Y., et al. Stock Index Prediction Model Based on CNN-LSTM. *Statistics & Decision*, 2021, 37(05): 134-138.
- [11] Li X., Tang P. Stock Price Prediction Based on Technical Analysis, Fundamental Analysis and Deep Learning. *Statistics & Decision*, 2022, 38(02): 146-150. DOI: 10.13546/j.cnki.tjyc.2022.02.029.
- [12] Gülmez B. A hybrid approach for stock market price forecasting using long short-term memory and seahorse optimization algorithm. *Annals of Data Science*, 2025.
- [13] Ritu A. N., Khandakar N. S., Bhuiyan M. M., et al. Optimizing stock market prediction using long short-term memory networks. *Journal of Computer and Communications*, 2025, 13(2): 207-222.
- [14] Yang Q., Wang C. Research on Global Stock Index Prediction Based on Deep Learning LSTM Neural Network. *Statistical Research*, 2019, 36(03): 65-77. DOI: 10.19343/j.cnki.11-1302/c.2019.03.006.
- [15] Diao H., Zhang Y. Research on Optimization of CSI 300 Stock Pairs Trading Strategy — Based on LSTM Dynamic Parameter Adjustment and Market Capitalization Screening Method. *Price: Theory & Practice*, 2024(03): 149-155. DOI: 10.19851/j.cnki.CN11-1010/F.2024.03.113.
- [16] Yao Y., Wang X., Chen W., et al. Research on Extreme Risk Spillover of Global Stock Markets Based on Transformer-LSTM Quantile Regression. *Chinese Journal of Management Science*, 2025: 1-17.
- [17] Zhou Z., He X. Stock Price Prediction Method Based on Optimized LSTM Model. *Statistics & Decision*, 2023, 39(06): 143-148.