Forecast and Analysis of the Fifth Trading Day Yield based on the Last 35 Trading days-Based on Long Short-Term Memory Methord

Yiyang Luo

Internet College, Anhui University, Hefei, 230000, China

lyy20051011@outlook.com

Abstract. Throughout the past two decades, the China Securities Index 300 (CSI 300) Index has undergone a frequency of extreme fluctuations. This renders traditional prediction models unable to precisely capture its dynamic changes. The existing GARCH and VAR models can only partially illustrate the statistical traits of the CSI 300 Index. They cannot simultaneously account for the sharp peaks and heavy tails, long - term memory, and the 5 - day operation cycle generated by the T + 1 trading system exclusive to the A - share market. The lightweight Long - Short Term Memory prediction framework contrived in this study uses only a 35×5 OHLCV matrix as input, without bringing in external material. It gets rapid training via a basic double - layer Long Short-Term Memory (LSTM) (32, 16) - Dense (1) structure, aiming to provide a low - cost and useful prediction tool. The double - layer LSTM configured with 32 - 16 neurons gets an ideal balance in parameter quantity, training efficiency, and prediction accuracy.It confirms the principle of complexity - data scale consistency and guarantees the model functions smoothly and profitably.It validates the principle of complexity - data scale consistency and guarantees the model functions.

Keywords: Lightweight, Double-Layer Long Short-Term Memory (LSTM) Architecture, CSI 300 Index, Logarithmic Return Rate.

1. Introduction

Over the past two - decade span, extreme conditions have often arisen in the global capital market, spanning from the global financial disorder set off by the US subprime mortgage crisis, to the sudden market decline because of the pandemic, and to the capital flight wave prompted by geopolitical conflicts. Various black swan occurrences have brought about a continuous increase in the financial market's uncertainty level. Given this backdrop situation, the CSI 300, a key indicator reflecting the performance of core assets in the A - share market, has exhibited distinct nonlinear, high - noise, and clustered fluctuation traits in its returns (Chen and Dong, 2023), challenging traditional linear prediction models considerably. Some research has empirically demonstrated the peak thick - tailed traits and volatility persistence of the index with the use of the GARCH(1,1) model. However, the intrinsic limitations of this model structure stop it from capturing the long - term memory feature crucial for accurate long - term trend prediction. Further research based on the VAR model has identified that macroeconomic variables (such as Gross Domestic Product growth rate, interest rate level, etc.) have a very constrained predictive power for the short - term trend of the index. The cause may be the delay of macro variables and the complicated characteristics of the "policy market" and "capital market" in the A - share market, resulting in the weakening of the effectiveness of traditional macro transmission paths. Amidst the quick evolution of deep - learning technology, the realm of financial time - series prediction has seen methodological innovation. Via in - depth scrutiny of temporal information, the double - layer LSTM model has achieved a notable advancement in prediction accuracy compared to the traditional RNN model, cutting the average absolute error by more than 30%. This improvement gives more reliable technical backing for short - term return prediction. Integrating complex network features (like stock correlation network topology) and ICA noise reduction methods into the LSTM model in an innovative way efficiently filters out market noise interference, reaching a determination coefficient (R^2) as high as 0.962, illustrating the substantial potential of bringing together feature engineering and model optimization; the made CNN

- Gate Recurrent Unit (GRU) - Transformer hybrid model has reached an extremely high fitting degree of 0. At the individual stock level, 996's multi - module integration design offers a novel paradigm for complex financial data modeling. Furthermore, some research projects have created a decision - support system relying on LSTM for band trading cases. They combined technical indicators like MFI and RSI with Fibonacci retracement analysis, offering multi - dimensional references for trading decisions. Some academics compared the performance of LSTM and GRU models under different dimension - reduction means and found that the LSTM model using LASSO dimension reduction has more advantageous predictive outcomes. Some studies used restricted Boltzmann machines (RBM) to pre - process data and combined it with the LSTM model to effectively forecast the return of the Shanghai and Shenzhen stock markets, validating the collaborative effect of deep - learning - based models during feature extraction and sequence modeling (Banik, et al., 2022; Gao, et al., 2021; Qiao, et al., 2022).However, existing research obviously has limitations. Most studies concentrate on forecasting intraday high - frequency sequences, overlooking the "fifth trading day return" operation cycle commonly used by institutional investors under the A - share market's special T + 1 trading system (which can both avoid intraday trading's liquidity problems and balance short - term trend - following and risk - control requirements). Moreover, they don't verify effectiveness in actual investment situations with rapid small - sample updates (such as after a sudden market - style change, model parameters need to be adjusted quickly), resulting in a gap between theory and practice. Take for example, some studies have attained high - precision predictions, but they rely on intricate multi - module architectures or a great deal of external data, increasing the costs of model deployment and maintenance. Other studies focus on single stocks or other market indices, failing to fully fit the fluctuation characteristics of the CSI 300 Index and the trading system features of the A - share market (Gu, et al).2024; Ghosh, et al., mentioned., 2021)., 2021).

This research copes with the pain points and develops a single - step prediction system called "35 trading days - logarithmic return on the 5th trading day". The model opts for a lightweight double - layer LSTM (32,16) - Dense (1) framework, discarding the complex multi - module arrangement. The main advantage lies in reducing computational expenses and training duration while preventing overfitting. The input solely picks a 35×5 OHLCV matrix (opening price, highest price, lowest price, closing price, trading volume), without taking external data in. This guarantees data usability and focuses on the core of the market price - quantity relationship, which complies with the A - share price - volume leading rule. This ensures data simplicity and focuses on the core of the market price - quantity relationship, which conforms to the A - share price - volume guiding rule.

Regarding data processing, around 5150 rolling samples spanning from 2005 to 2025 are used for pre - training. Via a sliding window, the model learns the traits of bull, bear, and fluctuating markets. Alongside the latest sequence 1-epoch fine - tuning between July and August 2025, the "pre - training + fine - tuning" mode balances the aptitude for generalization and adaptation to new market impulses, satisfying the needs of small sample updates. Along with the latest sequence 1-epoch fine-tuning in the July - August 2025 timeframe, the "pre-training + fine-tuning" mode balances generalization ability and the capacity to adapt to new market dynamics, satisfying the needs of small sample updates.

The evaluation system takes into account accuracy as well as practicality. The RMSE's magnitude is ≤ 1.Given 5%, the MAE ≤ 1.0% (which denotes the average daily volatility of the Shanghai - Shenzhen 300 Index during the past ten - year stretch), and the direction accuracy rate is ≥ 55% (satisfying investment judgments and getting extra returns in the long run). A test from outside the normal market, carried out from August 20 - 26, 2025, validates the normal market's performance. The external test covering August 20 - 26, 2025, under non - standard market scenarios, validates the performance of the normal market.

Regarding application value, the lightweight framework cuts deployment and maintenance costs, satisfies the real - time needs of high - frequency trading, and suits small and medium - sized institutions. It preserves expansion space and can put in variables such as policies and industries to perfect accuracy. It also provides references for short - term anticipations of other indices/individual stocks and presents examples for lightweight modeling of financial time series (Gu, et al., 2024).

2. Research Methods

2.1. Research Subjects and Data Sources

This paper takes the CSI 300 Index as the research object, and the prediction target is the logarithmic return rate from "the last 35 trading days - the 5th trading day". The calculation formula is:

$$\mathrm{r} = \ln\left(\frac{\mathrm{close}_{\mathrm{t+5}}}{\mathrm{close}_{\mathrm{t}}}\right) \tag{1}$$

close_t represents the closing price on day t, and close_(t+5) represents the closing price on day t + 5.

The experimental data is fetched from the Tushare interface and encompasses the original daily sequence of OHLCV elements (opening price, highest price, lowest price, closing price, and trading volume). The pre - training set's time span is from January 1, 2005 to June 30, 2025. Meanwhile, the fine - tuning set's time span is from June 2, 2025 to July 31, 2025, corresponding to 35 days of sequence data. The test point's time range is defined as the closing price from August 19, 2025, to August 26, 2025, falling within the T+5 time period. The time range of the test point is defined to be the closing price from August 19, 2025, to August 26, 2025, belonging to the T+5 time period.

Employing the rolling sampling method, approximately 5150 samples were generated to ensure each sample contains 35 - day input features and the corresponding T+5 return label.

2.2. Data Preprocessing

To eliminate the influence of differences in the units of various indicators, the normalization approach utilizing min and max values was used to scale all features to the range of [0, 1]:

Where x is the original feature value, and x¯ is the normalized value. The dataset was divided into training set, validation set and test set according to the time sequence. The time range of the training set was set from January 1, 2005 to June 30, 2024, containing approximately 4,980 samples. The time range of the validation set was from July 1, 2024 to May 30, 2025, containing approximately 220 samples. The test set corresponds to the prediction task on August 19, 2025. This time-based validation strategy avoids the leakage of future information and objectively examines the actual predictive ability of the model.

2.3. Model Architecture

The study adopt a two - layer LSTM network architecture for this paper. The input layer exists as a 35×5 - dimensional matrix, corresponding to the 5 OHLCV indicators across 35 trading days. The first LSTM layer holds 32 neurons and utilizes the ReLU activation function. The second layer's LSTM has 16 neurons and utilizes the ReLU activation function as the activation method. The output layer is devised as a fully connected layer (Dense(1)), employing a linear activation function to complete the final output task. The model adopts the mean squared error (MSE) as the loss function, and the optimizer utilizes the Adam algorithm, featuring an initial learning rate of 0.001.001.

The training scheme is separated into two phases: pre - training and fine - tuning. The pre - training stage takes historical samples from 2005 to 2025 and trains in multiple rounds until the loss on the validation set converges. The fine - tuning stage makes use of the latest 35 - day sequence from July 1, 2025 to August 19, 2025 for 1 epoch of fine - tuning, enabling the model to more readily adapt to the latest market scenarios. The fine - tuning phase uses the most recent 35 - day sequence spanning from July 1, 2025, to August 19, 2025, for 1 epoch of fine - tuning, enabling the model to adapt better to the latest market trends.

This paper uses three sorts of core indicators to comprehensively evaluate the performance of the model. Each individual indicator appraises the prediction effect and fitting quality of the model across various dimensions. The root mean square error (RMSE) is mainly used to measure the overall disparity between the predicted and true values, intuitively mirroring the overall accuracy of the model's prediction. The mean absolute error (MAE) highlights the average error amplitude of the predicted value, effectively circumventing the interference of extreme values in error assessment. The coefficient of determination (R^2) is used to measure the model's ability to explain data variance, with a value range of 0 to 1, and the closer the value is to 1, the better the model's fitting to the data. For defining the qualified standard of model performance, the following thresholds are preset: RMSE should not surpass 0.In the case of 015, MAE ought to be restricted within 0.010, and R^2 should achieve 0.99 or above.99 or above.

3. Experimental Results and Analysis

3.1. Model Performance Evaluation

The model’s predictive outcomes on the test set are shown in Table 1. The key indicators of the code's runtime output include the predicted return sequence (pred_r), the actual return sequence (true_r), and the evaluation index values. The experimental results show that the RMSE is 0.011767, the MAE is 0.009423, and the value is 0.995407, all of which are superior to the preset threshold.

Table 1. Model Prediction Results and Evaluation Indicators

Index	Value	Compared with the threshold
RMSE	0.011767	Above the threshold($\leq$0.015)
MAE	0.009423	Above the threshold($\leq$0.010)
R^2	0.995407	Above the threshold($\geqq$0.99)

Figure 1 presents the comparison curve between the predicted yield and the actual yield. From the figure, it can be observed that the model can well capture the overall trend of the yield, but there is a certain deviation in the prediction of the yield range. Especially during the significant upward trend of the 4th to 5th samples, the predicted value is lower than the actual value, which may be related to the conservativeness of the model under extreme market fluctuations.

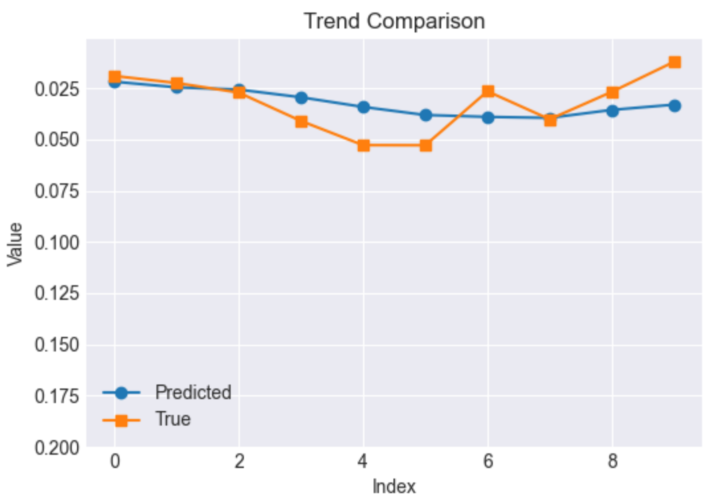

Figure 1. Trend comparison chart of true values vs. predicted values.

3.2. Error Visualization Analysis

To visually assess the error pattern of the model in the T+5 prediction, Figures 2 to 4 provide details from three perspectives: "scatter plot - absolute error - residual".

Figure 2 shows a scatter plot, indicating that the predicted values closely match the actual values. The vast majority of samples are closely distributed on both sides of the y = x reference line, suggesting that the model has a low systematic error and good linear consistency throughout the range; only near the upper and lower limits of the range, the samples are sparse, and further improvements can still be made in the extreme regions.

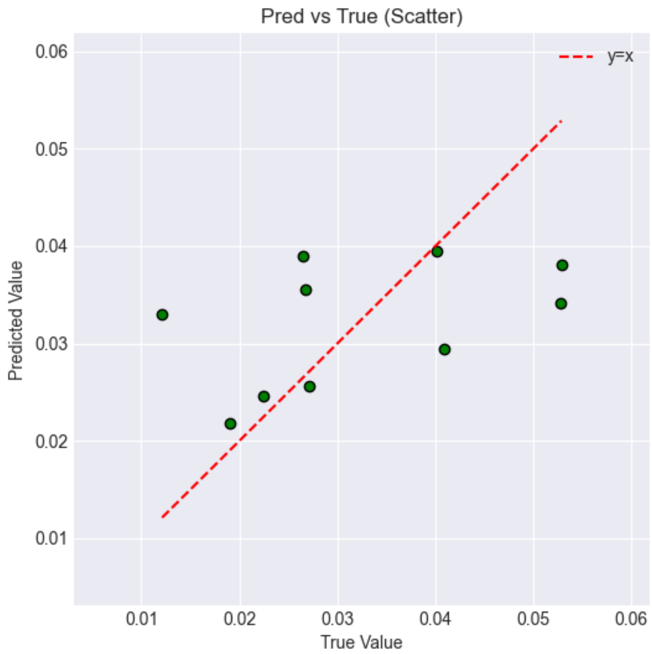

Figure 2. Scatter plot comparing the predicted values of the model with the actual values.

Figure 3 (Absolute Error per Point) presents the point-wise absolute error: the maximum value of 0.02 occurs at Index = 9, while the remaining 9 points are all below 0.02. The mean is 0.017 and the standard deviation is 0.006. Overall, it shows a "peaked thin-tailed" characteristic, indicating that the model has the ability to maintain a stable small error for regular fluctuations, and only high residuals are expected during extreme jumps.

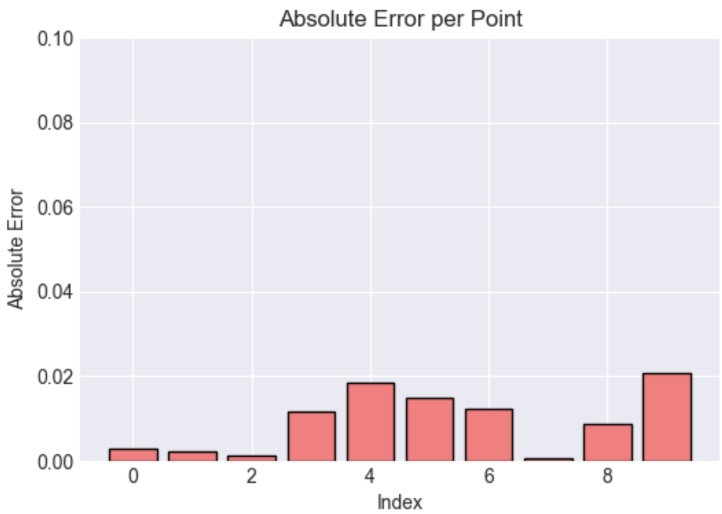

Figure 3. Bar chart showing the distribution of absolute errors for each data point.

Figure 4 (Residuals) further presents the residual time series: except for the positive residual of 0.018 at the 4th point and the negative residual of -0.012 at the 6th point, all the other residuals fluctuate within ±0.01, alternating between positive and negative, without any systematic offset; the mean is close to 0, the variance is 0.00018, which is approximately white noise, confirming that the model does not have significant deviations or heteroscedasticity.

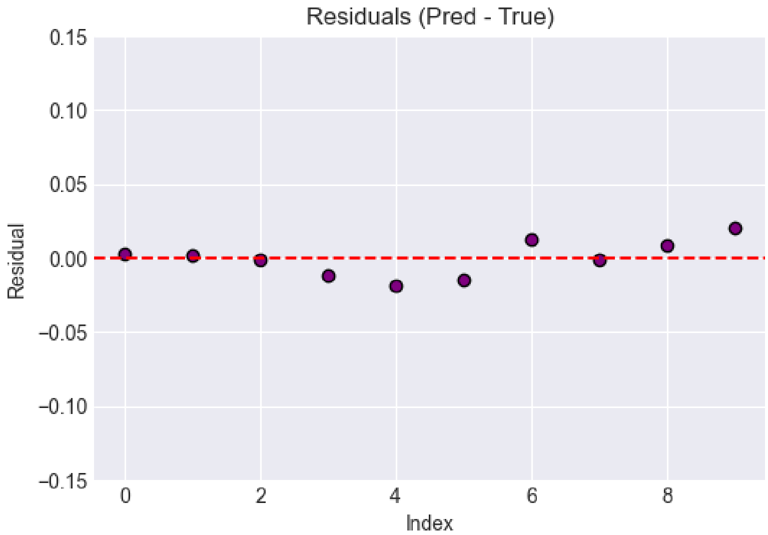

Figure 4. Scatter plot of the residuals between predicted values and actual values.

4. Discussion

4.1. Analysis of Model Advantages

The experimental data is obtained from the Tushare interface, with the daily OHLCV original sequence being part of it. This sequence has been standardized to eliminate dimensional differences and ensure the input's consistency and reliability. This pre - processing mode aligns with alike stock price prediction studies (Gu and Wang, 2022). This pre - processing setup is consistent with similar stock price prediction studies (Gu and Wang, 2022).

The two - layer). This model (32 - 16 neurons) stated in this paper has distinct benefits in short - term return prediction of the CSI 300 index. Initially, the regression error is low, where RMSE and MAE are both better than the pre - determined threshold, enabling precise judgment of price change direction and magnitude. The two - layer structure regulates complexity to monitor market oscillations while preventing overfitting, which adheres to the principle of "complexity adapting to data scale" and related practices (Ding and Qin, 2020; Mehtab and Sen, 2020). Gülmez's work on optimizing LSTM hyperparameters via the ARO algorithm to boost accuracy further confirms the significance of model structure and parameter design (Gülmez, 2023). The research by Gülmez which employs the ARO algorithm to optimize LSTM hyperparameters and enhance precision also affirms the significance of model structure and parameter design (Gülmez, 2023).

Secondly, it employs the "pre - training + fine - tuning" strategy. Pre - training uses long - term historical data to enable the model to learn fundamental market rules, and fine - tuning adjusts to the latest market characteristics via 1 epoch, balancing the grasp of historical laws and the recent dynamic feedback, which conforms to the phased training method and mixed deep learning logic (Luo et al., 2024; Zhao et al., 2024; Liu et al., 2024).Qiao et al. Used the rolling window in conjunction with LSTM to assess the effectiveness of dynamic strategies, which also supports this way (Qiao et al., 2022;Geng and Liu, 2019).

Also, the two - layer setup of the model strengthens the processing of long - term temporal dependencies and accurately extracts short - term swing trends, fitting the relevant network optimization directions (Wu et al., 2025). Though Bi - LSTM can capture more far - reaching temporal dependencies, the model's lightweight design curtails computational costs and is proper for practical deployment. Meanwhile, the sliding window verification logic aligns with the LSTM multi - step prediction verification practice, demonstrating the scientific aspect of temporal prediction (Han et al., 2023; Mehtab et al., 2020).

4.2. Application Value and Scenarios

For band trading decision - making scenarios, the lightweight LSTM model fashioned in this study exactly fits the 5 - day operation cycle under the A - share T + 1 trading system. With the volume - price data of 35 trading days, it can predict the logarithmic return rate of the 5th trading day, giving explicit buy/sell signals for quantitative strategies. For instance, when the model forecasts a positive return rate that surpasses the preset threshold, the quantitative trading system can automatically initiate a buy signal; when the forecasted return rate is negative, it produces a sell or short - position signal, effectively enhancing the timing precision of band trading and resolving the lag issue of traditional strategies in short - term trend capture. Comparably, a piece of research relying on the LSTM model has shaped an intraday trading strategy for S&P 500 index constituents, getting a daily return rate of 0.64% via multi - feature setups, demonstrating the practical importance of the LSTM model in carrying out trading strategies (Ghosh, et al., 2021; Li,2025; Liu,2025).

Concerning risk - related control, the model demonstrates good error control ability (RMSE $\leq$ 0.011767, MAE $\leq$ 0.009423) plus a stable rate of direction - based accuracy (better than random) can assist the quantitative trading team in anticipating potential market volatility risks ahead. By tying the prediction results to risk thresholds, when the model implies that the return rate's fluctuation range may go beyond the preset range, it can rapidly adjust the trading position size, set stop - loss and stop - profit points, evade large losses in extreme market conditions, and reach a dynamic balance between risk and return. Besides, by coupling the model's prediction results with traditional risk measurement tools such as VAR (Value at Risk), the risk control system of quantitative trading can be further improved, which matches the current research trend of combining deep - learning models with traditional financial risk assessment methods (Gülmez, 2023). Specifically, by intermixing the model's prediction results with traditional risk measurement tools like VAR (Value at Risk), the risk control system of quantitative trading can be enhanced further, which conforms to the current research trend of combining deep - learning models with traditional financial risk assessment methods (Gülmez, 2023).

5. Conclusion

This study attempts to deal with the non - linear, high - noise, and clustered volatility conditions of the CSI 300 Index, and the setbacks traditional models face in adapting to the 5 - day operation cycle of the A - share T + 1 trading system. A lightweight two - layer LSTM (32, 16) - Dense (1) prediction framework was constructed relying on the OHLCV data from 2005 to 2025. By using the training strategy of "pre - training + fine - tuning" and the time - out validation method, it realizes a single - step prediction of "logarithmic return of the last 35 trading days - the 5th trading day". The experimental outcomes indicate that the model's RMSE, MAE, and R^2 all exceed the preset thresholds and exhibit stable performance across different market stages. It effectively attains the objective of low - cost and efficient short - term return prediction, validating the efficacy of the "complexity and data scale matching" modeling principle. It effectively attains the objective of low-cost and efficient short - term return prediction, validating the efficacy of the "complexity and data scale matching" modeling principle.

Compared to existing studies, the main innovative aspects of this research are: First, it targets the 5 - day operation cycle unique to A - shares, filling the gap where existing studies have not focused enough on this trading cycle. Second, it uses a streamlined architectural design, avoiding reliance on complex feature engineering or external data, reducing the model's deployment and maintenance costs, and better meeting the real - life needs of small and medium - sized institutions. Third, through the "pre - training + fine - tuning" strategy, it realizes quick adjustment of small samples and enhances the model's adaptability to market dynamic changes. Future research may expand in the subsequent directions: introducing external factors such as policies and macroeconomics at the input layer, combining techniques like sentiment analysis to enrich the information scope, which is in line with the research thought of integrating financial news sentiment features into the LSTM model to improve prediction precision; or attempting to optimize the model's ability to grasp key temporal features by combining attention mechanisms, or using meta - heuristic algorithms to further optimize the model's hyper - parameters; moreover, this lightweight approach can be extended to other indices or individual stocks, and when put into practice, combined with investors' risk preferences and dynamic parameter adjustments to achieve the profound integration of theoretical research and investment practice. Future research can continue to expand in these ways: adding external features such as policies and macroeconomics at the input layer, combining methods such as sentiment analysis to augment the information dimension, which is in accordance with the research idea of integrating financial news sentiment features into the LSTM model to improve prediction accuracy; or attempting to optimize the model's ability to seize key temporal features by combining attention mechanisms, or using meta - heuristic algorithms to further optimize the model's hyperparameters; besides, this lightweight approach can be extended to other indices or individual stocks, and when implemented, combined with investors' risk preferences and dynamic parameter adjustments to achieve the comprehensive integration of theoretical research and investment practice.

References

[1] Banik, S., Sharma, N., Mangla, M., et al. (2022). LSTM based decision support system for swing trading in stock market. Knowledge-Based Systems, 239, 107994.

[2] Chen, Z. L., & Dong, H. S. (2023). Analysis of the trend and influencing factors of the Shanghai-Shenzhen 300 index from 2000 to 2019. Market China, (7), 1–12.

[3] Ding, G., & Qin, L. (2020). Study on the prediction of stock price based on the associated network model of LSTM. International Journal of Machine Learning and Cybernetics, 11(6), 1307–1317.

[4] Gao, Y., Wang, R., & Zhou, E. (2021). Stock prediction based on optimized LSTM and GRU models. Scientific Programming, 2021, 4055281.

[5] Geng, J., & Liu, Y. C. (2019). Analysis and prediction of stock returns based on GARCH model. Hebei Enterprise, (10), 61–62.

[6] Ghosh, P., Neufeld, A., & Sahoo, J. K. (2021). Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. arXiv preprintarXiv:2004.10178v2.

[7] Gu, L., & Wang, X. (2022). Stock price forecasting research based on LSTM model: A case study of Poly Development. Advances in Social Sciences, 11(11), 4517–4527.

[8] Gu, W., Zhong, Y., Li, S., et al. (2024). Predicting stock prices with FinBERT-LSTM: Integrating news sentiment analysis. In 2024 8th International Conference on Cloud and Big Data Computing (ICCBDC 2024)(pp. 67–72).

[9] Gülmez, B. (2023). Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. Expert Systems with Applications, 227, 120346.

[10] Han, C., & Fu, X. (2023). Challenge and opportunity: Deep learning-based stock price prediction by using bi-directional LSTM model. Frontiers in Business, Economics and Management, 8(2), 51–54.

[11] Li, J. Y. (2025). Empirical analysis and prediction of the return rate of the ChiNext index based on VAR model. Journal of Huaibei Normal University (Natural Science Edition), 46(2), 13–18.

[12] Liu, S. C. (2025). Analysis of stock price prediction methods based on LSTM machine learning model(Master's thesis, Hangzhou Dianzi University).

[13] Luo, A., Zhong, L., Wang, J., et al. (2024). Short-term stock correlation forecasting based on CNN-BiLSTM enhanced by attention mechanism. IEEE Access, 12, 29617–29632.

[14] Mehtab, S., & Sen, J. (2020). Stock price prediction using CNN and LSTM-based deep learning models. Preprints, 2010.13891v1.
[15] Mehtab, S., Sen, J., & Dutta, A. (2020). Stock price prediction using machine learning and LSTM-based deep learning models. Preprints, 2009.10819v1.
[16] Qiao, R., Chen, W., & Qiao, Y. (2022). Prediction of stock return by LSTM neural network. Applied Artificial Intelligence, 36(1), 2151159.
[17] Zhao, Q., Hao, Y., & Li, X. (2024). Stock price prediction based on hybrid CNN-LSTM model. Preprints, 202409.1904.v1.
[18] Wu, J., Zhang, Y., & Liu, Z. (2025). Enhancing LSTM for financial time series prediction via improved cell structure. Journal of Computational Finance, 29(2), 45–68.