# Comparing LSTM and Transformer Architectures for Daily Close Price Prediction on Chinese A-Share Stocks

Zhiwei Zhang

College of International Education, Hong Kong Baptist University, Hong Kong, China
25634119@life.hkbu.edu.hk

**Abstract.** Recent studies have reported strong performance of Transformer-based architectures for financial time-series prediction mainly evaluated on conventional error metrics. However, improvements in predictive accuracy do not necessarily translate into superior trading performance. This paper conducts a cross-sectional study to evaluate the performance of different financial forecasting models in the high-noise Chinese A-share market. In the study, a random-walk model, an Extreme Gradient Boosting (XGBoost) regressor, a two-layer Long Short Term Memory (LSTM), and a small encoder-only Transformer are trained in parallel to predict the next-day residual close prices using daily OHLCV data from 284 CSI 300 constituents over the period of 2015-2024. Approximately 2 million forecasts out of sample are generated under a strict time-series split. Besides, the models are then evaluated on various error metrics, and a long-only backtest with transaction costs taken into account. Results show that while all the models show limited predictive accuracy, their trading outcomes are significantly divergent. The Transformer achieves the best overall predictive performance, although its advantage over the random-walk benchmark is not statistically significant based on the Diebold-Mariano test. While the two-layer LSTM acts as a long-biased filter, achieving the highest market exposure and buy-and-hold returns. The findings suggest that the differences in trading performance are mainly attributed to different exposure behaviours induced by distinct models, rather than merely predictive accuracy. Therefore, it is significant to further evaluate financial forecasting models with both statistical criteria and multiple economic factors

**Keywords:** Quantitative Trading, Deep Learning, Time-series Forecasting, Stock Prediction

## 1 Introduction

Stock price forecasting has recently become a challenging sequence prediction problem. In the deep learning era, the family of recurrent neural networks (RNNs), especially the Long Short-term Memory (LSTM) networks, has long been used for financial forecasting modelling in applied finance studies [1]. While the Transformer model, originally introduced for natural language processing (NLP), get extensively used for complicated time-series forecasting [2][3][4]. More recently, a transformer-

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_62

based inverted attention variant, iTransformer, has been introduced as a superior method to capture market dynamics [5]. Consequently, attention-based architectures are increasingly applied to financial forecasting [6][7].

In numerous studies, accuracy metrics are compared between attention-based models and their recurrent counterparts for stock price forecasting, where the former are found to be superior in terms of accuracy [8][9][10]. However, those accuracy metrics, especially based on point predictions do not necessarily translate into better trading performance. This is because in real trading, not only do statistical errors matter, but also other errors regarding transaction costs, market exposure, and drawdown risks that are not captured by accuracy metrics. This creates a gap between forecasting accuracy and economic utility, which reveals that evaluating forecasting models merely through statistical error metrics may overestimate their practical value.

The Chinese A-share market provides an excellent testing environment because of its high noise levels, fixed T+1 settlement rules, 10% up-and-down daily price limits, a large number of retail investors, and substantial trading frictions. These features offer a unique test bed for this investigation

This paper compares a two-layer LSTM, a small Transformer, and an XGBoost regressor against a random walk benchmark for the next-day close prediction on 284 CSI 300 constituents. Using daily OHLCV data of 284 CSI 300 constituents from 2015 to 2024, approximately 2 million forecasts out of sample are strictly generated and evaluated using both statistical prediction metrics and a costed long-only backtest, where Sharpe ratios, drawdowns, and market exposures are reported along with the typical error metrics as opposed to overfitting error scores. This design aims to provide new evidence on the practical value of distinct forecasting architectures in a high-noise market environment.

## 2 Methods

### 2.1 Dataset

This study is based on the daily open, high, low, close, and volume (OHLCV) data of 284 CSI 300 constituents in forward-adjusted form. The study covered a time span from January 2, 2015, to December 31, 2024. Each stock had around 2,430 trading days. During this period, the market has experienced extreme fluctuations such as the 2015 bubble and crash, the 2018 trade-war drawdown, the 2020 COVID-19 rebound, and the 2022-2024 consolidation. The study did not generate any additional features on top of the five raw OHLCV columns. Instead, it tested the inductive learning ability of several models. In the CSI 300 constituents, 16 stocks were excluded from the cross-section due to insufficient historical data or excessive trading suspensions.

The data cleaning operation prevented destroying temporal dynamics, ensuring a continuous and 10-year aligned tensor for sequence models. Each stock's time series was then split into training, validation, and test sets at a 70/15/15 ratio without any shuffling. The resulting number of non-overlapping training, validation, and test windows at a cross-sectional level was 394,352 for training, 103,559 for validation, and

103,578 for testing. In addition, the test data covered the October 2024 stimulus rally and the following price fluctuations.

### 2.2 Models

Four predictive models were trained in parallel, all targeting the residual between the next-day close and the last value in the 30-day input window on the MinMax-scaled price series, to prevent models from generating "tomorrow equals today" prediction. Predictions were then converted back to the original scale of the price series, and the respective quality metrics were evaluated.

Firstly, the random-walk baseline predicts a residual of zero and has no parameters, where no model can extract predictable returns or residuals of any forecasts on one stock from its historical price sequence. Secondly, the XGBoost model is a gradient-boosted tree ensemble, which was trained on the 30-day window, yielding 150 flattened features per sample. The XGBoost model uses a maximum tree depth of 6 and 500 boosting rounds with early stopping on the validation error. It is a strong non-deep baseline, since it treats the 30 lags of a sample as independent features. But there is no sequence information that the model could use [11]. Thirdly, the LSTM model has 2 layers with a hidden size of 64 each including a dropout of 0.2 between the 2 layers, whose output is a linear function of the last step of the sequence (51,521 parameters in total). The LSTM mitigates the vanishing gradient problem of conventional RNNs by the gated memory of its hidden states [12]. Fourthly, the Transformer model used for stock price prediction is a small encoder-only model, consisting of 2 blocks and model dimension of 64. There are 4 attention heads, and the feed-forward width is 128. The activation function is a Gaussian Error Linear Unit, and the total number of parameters is 67,393.

While the deep learning models have comparable parameter counts, identical data splits, and shored residual target that exclude the impact of model capacity and data handling for raw error differences, they still cannot isolate the architectures from other confounds such as training dynamics or implicit regularization.

### 2.3 Training and Evaluation

In addition to the shallow models, two deep models are trained on residual targets using the Adam optimizer with a learning rate of 0.001 and gradient clipping with a maximum norm of 1.0 [13]. Both models are trained for up to 50 epochs in batches of size 64 with early stopping on the validation loss with a patience of 7. Random seeds are fixed for reproducibility. Early stopping is triggered well before the maximum number of epochs is reached: the LSTM after 7 epochs and the Transformer after 10 epochs. Min-max scaling is used for scaling the input of the deep models, which is fitted on the training part of a time series and then applied to the respective validation and test data of that time series. This procedure can prevent look-ahead bias by confining the scaling parameters to past information only, preserve the strict temporal causality needed for realistic forecasting, and normalize each stock independently to avoid cross-asset scale contamination. The four mentioned metrics— Mean Absolute Percentage Error

(MAPE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), and Direction Accuracy (DA)—are then calculated in original price units.

### 2.4 Backtest

Error metrics such as MAE and MAPE do not capture real-world trading friction, so they are not sufficient to evaluate time-series and financial forecasting models. The best MAE does not equal the best returns; a model with the lowest MAE can still lose money, specifically by making its largest errors on the most volatile days in the time series. A backtest of model residuals converted to market exposures and used in a long-only trading strategy can illustrate a number of key properties of models (i.e., Sharpe ratio, maximum drawdown, and average exposure). In this setup, each model's signals are used to trade independently daily with rebalancing. The portfolio is equally weighted, and a 10-basis-point transaction cost is applied to simulate friction in the Chinese A-share market. This approach can easily compare any number of models to a buy-and-hold benchmark and reveal the difficulties models have in translating strong statistical accuracy into strong out-of-sample trading returns.

## 3 Results

### 3.1 Predictive Accuracy

**Table 1.** Test-set predictive accuracy across 284 stocks in original price units; DA is undefined for the random walk.

| Model | MAE | RMSE | MAPE (%) | DA (%) |
|---|---|---|---|---|
| Random Walk | 0.6480 | 2.1309 | 1.5535 | — |
| XGBoost | 1.8634 | 7.8786 | 4.6471 | 50.17 |
| LSTM | 0.6797 | 2.1568 | 1.6575 | 48.37 |
| **Transformer** | **0.6457** | **2.1342** | **1.5507** | **52.19** |

According to Table 1, the cross-sectional point prediction by tested models performs poorly. The MAE of the Transformer is only marginally lower than that of the naive random walk (0.6457 vs. 0.6480, a gap of 0.0023). A Diebold–Mariano test of equal forecasts by the Transformer and the naive random walk yields a p-value of 0.1768. Hence, there is no statistically significant difference between the two models' forecasts, which challenges the predictive edge claimed by previous Transformer-related studies. Although the Transformer and the LSTM are the best models evaluated, they are not far superior to the naive random walk, whose performance "premium" is quite small. The worst model is the XGBoost model, with both MAE and RMSE about 3 times that of the naive random walk, directional accuracy no better than random chance, and a signal correlation between predictions and realized values of almost zero.

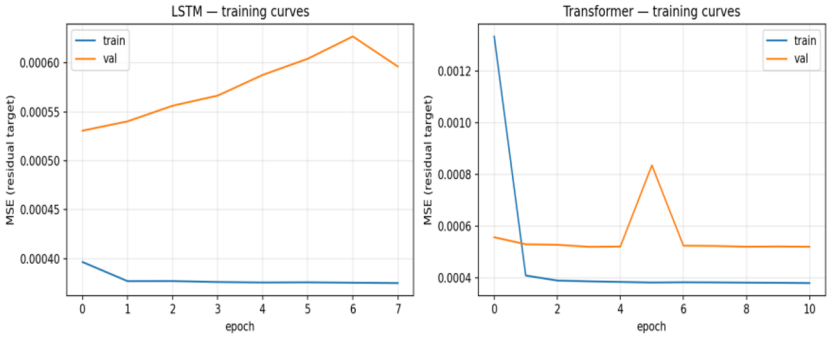



**Fig. 1.** Training and validation mean-squared error per epoch for the LSTM (left) and Transformer (right) (Photo/Picture credit: Original).

The training curves are shown in Fig. 1. The Transformer approaches its validation minimum by the end of the second epoch, after which it plateaus. In contrast, the LSTM approaches its training minimum with each additional epoch, but it never generalizes to new test cases given the task's extreme difficulty in extracting genuine signal. In the noisy market environment, the rational behaviour of a model is to converge quickly and then stop learning.

### 3.2 Backtest

These results are translated into trades using the costed long-only rule, and the outcomes are compared in terms of market exposure and subsequent economic performance. Although the statistical errors of the models are similar in magnitude, their resulting exposures and economic performances are different (see Table 2 and Fig. 2). "Mean Stk. Sh." averages per-stock annualized Sharpe ratios, "% Beat B&H" is the share of stocks beating their own buy-and-hold Sharpe, and "Exp." is the fraction of days invested. Concerning the portfolio performance, the Transformer achieves an annualized Sharpe ratio of 1.386, considerably more than that of a corresponding buy-and-hold portfolio with a Sharpe ratio of 0.618. The Sharpe ratio for the LSTM model is 0.676. In contrast, the exposure generated by the XGBoost model results in a Sharpe ratio of only 0.054, losing money as expected.

**Table 2.** Backtest of the equal-weighted long-only portfolio after a 10 bps round-trip cost.

| Model | Sharpe | Ann. Ret. (%) | MDD (%) | Tot. Ret. (%) | Mean Stk. Sh. | % Beat B&H | Exp. (%) |
|---|---|---|---|---|---|---|---|
| Random Walk | 0.000 | 0.00 | 0.00 | 0.00 | 0.000 | 27.11 | 0.00 |
| XGBoost | 0.054 | −0.00 | −13.91 | −0.00 | −0.004 | 28.87 | 51.03 |
| LSTM | 0.676 | 9.75 | −15.54 | 14.42 | 0.425 | 40.14 | 73.79 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Transformer** | **1.386** | 1.46 | **−0.68** | 2.13 | 0.120 | 33.45 | 4.15 |
| BuyHold (bench.) | 0.618 | **11.03** | −17.90 | **16.37** | — | — | 100.00 |

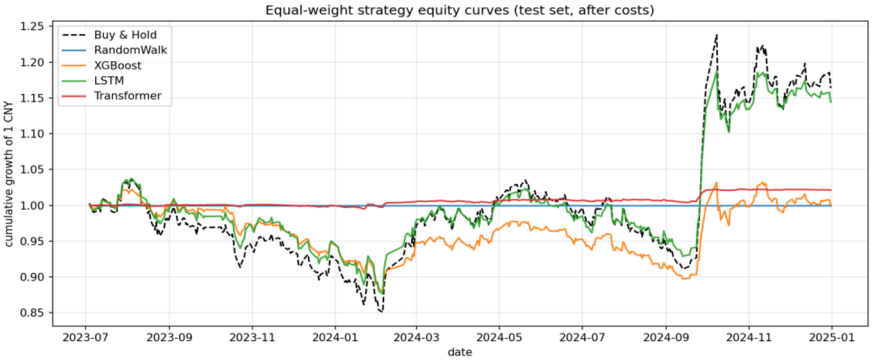



**Fig. 2.** Equal-weighted strategy equity curves on the test set, after costs (Photo/Picture credit: Original).

Fig. 2 shows how these differences evolve over time. As the market fell roughly 15% between August 2023 and early 2024, the LSTM rode the same drawdown to a 15.54% maximum loss, while XGBoost eroded capital due to transactional friction. The Transformer instead held a far flatter curve through the decline, with a maximum drawdown of only 0.68%, and gained mainly during the September 2024 rally.

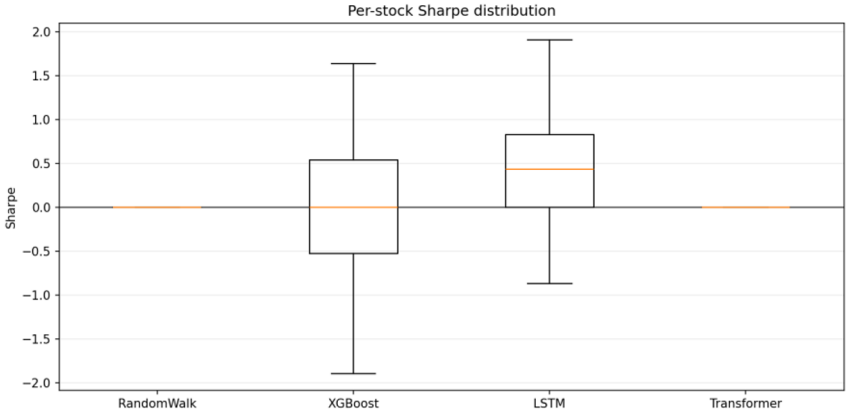



**Fig. 3.** Distribution of per-stock annualized Sharpe ratios across 284 stocks (Photo/Picture credit: Original).

The per-stock Sharpe distribution in Fig. 3 confirms that these results are not driven by outliers. The LSTM distribution sits in positive territory, with roughly 70% of stocks

achieving a positive Sharpe ratio, consistent with a robust trend-following filter. The Transformer distribution clusters near zero, whose expected shape of a highly selective strategy that stays out most of the time. However, XGBoost spreads symmetrically around zero, suggesting the trades are almost noisy.

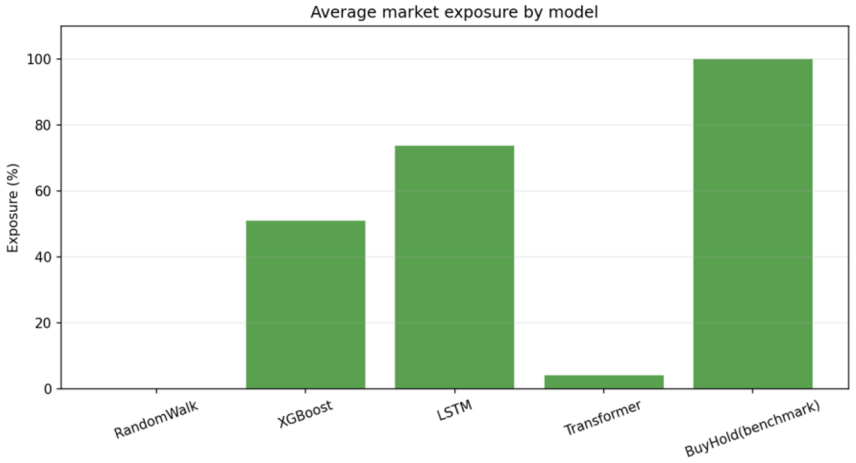



**Fig. 4.** Average market exposure by model; the Transformer functions effectively as a 96% stay-out filter (Photo/Picture credit: Original).

Average market exposure in Fig. 4 explains the structural contradiction. The Transformer behaves as a defensive stay-out signal that buys only occasionally, at 4.15% exposure; the LSTM acts as a stay-in signal that occasionally sells, at 73.79% exposure; and XGBoost trades at high frequency, averaging about 110 position changes per stock at 51.03% exposure. The true value of the attention mechanism therefore lies in autonomously building a defensive exposure logic rather than in precise price forecasting.

## 4 Discussion

### 4.1 The Point Prediction Edge Illusion

The predictive advantage of deep learning models is limited in out-of-sample cross-sectional prediction. The Transformer improves on the random walk by an MAE of only 0.0023 and a MAPE of 0.0028%, which are statistically insufficient to support a large deep-learning premium. As noted by Zeng et al., a deep-learning model can have an edge in stock price and return prediction, but the edge arises from the model's ability to make good multi-step forecasts rather than self-attention in the network [13]. The five OHLCV lag features used for training are inadequate to achieve better than a zero-parameter random walk. Stock prices and returns are affected by various economic factors including recent news, earnings announcements, large-scale macro events, and

sector-wide movements, none of which is included in the feature set. Therefore, simply optimizing a model for point prediction alone is insufficient and ineffective.

### 4.2 Forecasting Accuracy versus Economic Utility

Forecasting accuracy and economic utility are not equivalent fundamentally. A model with weak predictive accuracy can still achieve a superior Sharpe ratio by remaining in cash most of the time as a risk filter such as the Transformer performance. Because the self-attention mechanism is difficult to find consistent and high-correlation keys within a noisy 30-day window, the Transformer rarely crosses the long-only threshold, and whose output remains close to zero because signal quality is almost poor in high-noise market.

In contrast, the LSTM behaves as a long-biased filter, remaining invested for 73.79% of the testing period and capturing market movements. With a directional accuracy of 48.37%, it can produce positive residuals on most days. While its inherent constraint of sequential memory drives the LSTM model to accumulate the upward drift of the training data and retain a persistent long bias. Therefore, the LSTM model functions more as a brute-force filter than an intelligent forecaster.

XGBoost represents the worst-case scenario. It exhibits a near-random directional accuracy of 50.17% and a 51.03% market exposure, but its lack of temporal awareness drives approximately 110 position changes per stock, compared to only 1.6 for the Transformer. The excessive turnover leads to high transaction costs, underlining the necessity of a coherent exposure logic to improve risk-adjusted returns.

### 4.3 Limitations

Three caveats apply to the results reported in this study. First, by fixing the feature set to the five raw OHLCV columns, the models were decoupled from the broader space of possible features. Models could achieve greater gains by incorporating technical indicators, cross-sectional factor exposures, or text-derived financial sentiment indicators [14]. Second, the results are evaluated using a simple long-only, all-or-nothing trading rule. The models' predictive potential could be more effective by integrating the models' 'stay-out' periods into a volatility-adjusted or long-short framework. Third, the dataset inherently reflects a survivorship bias because of the excluded 16 companies, which implicitly avoids drawdown losses and delisting costs associated with failed firms, thereby overestimating the stability and net returns of distinct portfolio strategies. Fourth, the 10-basis-point (bps) round-trip cost used for the out-of-sample simulations is merely a baseline assumption. In reality, high-frequency strategies are fragile given the market restrictions. Therefore, the outcomes may turn negative when the round-trip cost is doubled to 20 bps, and the survivorship bias would be compounded. For more reliable evaluation, the dataset should adopt whole point-in-time index constituent universes and the explicit incorporation of forced-exit costs.

Consequently, any achieved outperformance should be strictly interpreted as an upper-bound effect relying on a survivorship-free sample set and realistic transaction costs estimates, which may disappear in a frictional real-world trading environment.

## 5 Conclusion

This paper compares the performance of different architectures on the high-noise Chinese A-share market by testing them on a dataset of 284 CSI 300 constituents traded from 2015 to 2024. In order to make a fair comparison, it keeps the amount of capacity and the dimensionality of the feature space of the models constant. Surprisingly, the learning models—the XGBoost regressor, the two-layer LSTM, and the Transformer—did not necessarily perform better than a naive random-walk in terms of point prediction errors.

However, the backtest results of the models showed a large gap, which is mainly due to the distinct exposure logic used by different architectures. For example, the Transformer, an attention-based model that learns risk aversion. As a result, it stays largely in cash and achieves the highest Sharpe ratio of 1.386 in this experiment against a benchmark Sharpe ratio of 0.618. On the other hand, the LSTM model acts as a long-biased filter and remains invested 73.79% of the time on average. These results indicate that in high-noise and frictional markets, even a weak directional edge combined with an effective exposure rule is able to achieve risk-adjusted returns higher than those of a benchmark.

In the future, further research plans should shift focus from minimizing point errors to directly optimizing risk-adjusted returns. The models can be improved by incorporating external information and macroeconomic sentiment. Furthermore, it would be effective to adopt deep reinforcement learning methods, such as Proximal Policy Optimization (PPO), to train the models so as to optimize their exposure robustly and adaptively.

## References

1. Sunny, M.A.I., Maswood, M.M.S., Alharbi, A.G.: Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. In: 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), pp. 87–92. IEEE, Giza (2020)
2. Vaswani, A.: Attention Is All You Need. In: Luxburg, U. von, Bengio, S., Fergus, R., Garnett, R., Guyon, I., Wallach, H., Vishwanathan, S.V.N. (eds.) 31st Annual Conference on Neural Information Processing Systems (NeurIPS 2017), Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008. Curran Associates, Inc., Red Hook (2017)
3. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In: AAAI Conference 2021, AAAI Conference Proceedings, vol. 35(12), pp. 11106–11115. AAAI Press, Palo Alto (2021)
4. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition Transformers With Auto-Correlation for Long-Term Series Forecasting. In: NeurIPS 2021, Advances in Neural Information Processing Systems, vol. 34, pp. 22419–22430. Curran Associates, Inc., Red Hook (2021)
5. Liu, Y., Hu, T.G., et al.: iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In: Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna (2024)

6. Hu, Z., Zhao, Y., Khushi, M.: A Survey of Forex and Stock Price Prediction Using Deep Learning. Applied System Innovation 4(1), Art.no. 9 (2021)
7. Wang, J. et al.: Asian Stock Markets Closing Index Forecast Based on Secondary Decomposition, Multi-Factor Analysis and Attention-Based LSTM Model. Engineering Applications of Artificial Intelligence 113, Art.no. 104908 (2022)
8. Yao, Y. et al.: Stock Index Forecasting Based on Multivariate Empirical Mode Decomposition and Temporal Convolutional Networks. Applied Soft Computing 142, Art.no. 110356 (2023)
9. Wang, D.: A Study on Stock Price Prediction Based on LSTM and GRU. E-Commerce Letters 13(2), 3203–3210 (2024)
10. Zeng, A. et al.: Are Transformers Effective for Time Series Forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence 37(9), 11121–11128 (2023).
11. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794. ACM, New York (2016)
12. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8), 1735–1780 (1997)
13. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR), pp. 1–15. San Diego (2015)
14. Gu, W., Li, S., Wang, Z.: Predicting Stock Prices With FinBERT-LSTM: Integrating News Sentiment Analysis. In: Proceedings of the 8th International Conference on Cloud and Big Data Computing (ICCBDC), pp. 1–6. ACM, Oxford (2024)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

