International Scientific Conference 20 UNITECH 26

INTERNATIONAL SCIENTIFIC CONFERENCE
19-21 November 2026
GABROVO

EVALUATION OF THE EFFECTIVENESS OF A BILSTM MODEL IN FORECASTING RETURNS OF MAJOR EUROPEAN STOCK INDICES

Mantas Bertulis[1], Grigorij Žilinskij[2], Rima Tamošiūnienė[3], Laura Gudelytė-Žilinskienė[4]

[1,2,3,4]Vilnius Gediminas Technical University, 11 Sauletekio Ave., Vilnius, Lithuania
**Mantas Bertulis: mantas.bertulis@vilniustech.lt*

Abstract
Forecasting stock index returns is challenging because financial markets exhibit nonlinear relationships and changing patterns. This study evaluates a Bidirectional Long Short-Term Memory (BiLSTM) model for forecasting next-trading-day logarithmic returns of DAX, CAC 40, and FTSE 100. Daily closing prices from October 2016 to September 2026 were obtained from Yahoo Finance. The same architecture and data preparation, training, and testing procedures were applied to each index. Performance was assessed across five training runs using MAE, RMSE, predictive R^2, and directional accuracy, with the training-sample mean return serving as a benchmark. FTSE 100 recorded the lowest absolute errors and the lowest variability of actual test returns. DAX achieved a predictive R^2 of 0.00352, indicating an approximately 0.35% reduction in mean squared error relative to the benchmark. Predictive R^2 was negative for CAC 40 and FTSE 100. Directional accuracy ranged from 48.30% to 50.50% and was below the benchmark for all three indices. The findings show that lower absolute errors do not necessarily indicate better benchmark-relative or directional performance. The selected specification offered only a small improvement in squared-error performance for DAX.

Keywords: BiLSTM, Stock Index Returns, Financial Forecasting, European Stock Markets, Forecasting Accuracy.

INTRODUCTION

Stock market forecasting is complicated by noise, non-linearity, and changing dependencies (Li et al., 2026). The efficient market and random walk hypotheses also raise questions regarding the forecasting of future changes based on historical data (Beniwal et al., 2024). Factors that influence prices include financial events, corporate information, and investor sentiment (D. Wang et al., 2026).

The importance of forecasting is closely linked to investment decision-making and risk management (Giantsidi & Tarantola, 2025). Putra & Nurmawati (2024) made forecasts to select stocks and construct suitable portfolios. Kehinde et al. (2026) discuss financial solutions involving the use of dynamic clustering methods and Data Envelopment Analysis (DEA) together with TabPFN model to evaluate corporate operational efficiency.

The limitations of classical methods encourage the adoption of deep learning architectures. LSTM gates help retain long-term information, while BiLSTM processes historical sequences in both directions (Giantsidi & Tarantola, 2025). However, neural networks can overfit the data, making proper preparation and testing important (An et al., 2024).

BiLSTM is also applied to forecast returns, (Ouyang et al., 2023) combined it with data decomposition and emissions trading data. Comparison regarding price forecasting yield inconsistent results, Putra & Nurmawati (2024) found BiLSTM to outperform LSTM, whereas Beniwal et al. (2024) did not find such an advantage in their overall assessment.

This is an open access article licensed under Creative Commons Attribution 4.0 International
Print ISSN: 1313-230X, Online ISSN: 3033-1404

European markets have already been examined in related studies. Mutinda & Geletu (2025) forecasted the DAX, while Mubarak (2026) analysed S&P 500 and FTSE 100 returns. These studies prove the relevance of the selected markets but employ different research approach, when it comes to model and data preparation.

The aim of this study is to evaluate and compare the effectiveness of a BiLSTM model in forecasting next trading day logarithmic returns of DAX, CAC 40, and FTSE 100. Each model is trained separately using the same architecture and data preparation, training, and testing procedures. Performance is assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), predictive R^2, and Directional Accuracy (DA), with the training-sample mean return serving as a benchmark. The contribution is an empirical comparison across three European markets that examines whether lower absolute forecast errors are accompanied by improvements over a simple benchmark and more accurate predictions of return direction.

LITERATURE REVIEW

Financial time series are characterised by non-stationarity, heavy distribution tails, and market-regime-dependent autocorrelation structures (T. Wang et al., 2026). To handle these changing structures, (Zhang et al., 2026) proposed a sliding-window EMA decomposition and a dual-stream network that separately models the trend and more complex fluctuations.

Historical price information is complemented by textual signals. Sathya et al. (2025) combined prices, technical indicators, and sentiment from news and social media, while J. Chen & Li (2025) applied BiLSTM for sentiment analysis and integrated the resulting information with market data. This shows different approaches to constructing inputs.

ARIMA models linear temporal dependencies, whereas GARCH is used to describe volatility clustering. Their structures may limit the modelling of more complex dynamics (Beniwal et al., 2024; T. Wang et al., 2026). However, in the table presented by J. Wang et al. (2024), the performance of the hybrid BiLSTM model relative to ARIMA-based model on RMSE, depended on the index.

Data decomposition is employed to separate components of different frequencies. Mutinda & Geletu (2025) combined CEEMDAN, LSTM, and BPNN for DAX forecasting. Dong et al. (2026) integrated SEVMD with LSTM, BiLSTM, and TCN across three indices, including FTSE 100. These studies evaluate forecasts of price and index levels.

Mubarak (2026), examining returns directly, found that applying ICA to the FTSE 100 dataset increased the directional accuracy of the LSTM model from 50.9% to 57.4%. Ouyang et al. (2023) employed GAVMD-Carbon-BiLSTM to forecast returns and volatility. In both instances, the entire enhanced process is evaluated, therefore the results cannot be attributed solely to the network architecture.

These studies show that model comparison is influenced by the variable being predicted, the choice of market, and data preparation. Giantsidi & Tarantola (2025) identify the lack of standardised evaluation procedures as an unresolved issue in their research. This highlights the need for common conditions when applying a single model across multiple markets.

LSTM gates help retain long-term information, while BiLSTM combines sequence representations and processed in both directions (Giantsidi & Tarantola, 2025). Zheng et al. (2021) proposed an evolutionary BiLSTM architecture with different objective functions. Rani et al. (2026) enhanced the architecture with residual connections and cosine learning rate scheduling.

The advantage of BiLSTM is not universal. Putra & Nurmawati (2024) found lower error rates compared to LSTM, however, Beniwal et al. (2024) reported an overall testing MAPE of 10.38% for LSTM and 10.90% for BiLSTM. Purnomo et al.

(2024) found that BiLSTM outperformed Random Forest in terms of accuracy, but the Random Forest was more efficient regarding training time and resources.

Architectural results are also influenced by feature selection and optimisation. An et al. (2024) employed hierarchical selection, J. Wang et al. (2024) used GBDT and an autoencoder, and Swetha & Arya (2025) utilised extended technical features together with Bayesian optimisation. An alternative model was proposed by Singh et al. (2026) combining Transformer, GRU, and ARO optimisation.

One hybridisation approach combines feature extraction with sequence modelling. Y. Chen et al. (2021) proposed CNN-BiLSTM-ECA, Vuong et al. (2024) introduced a convolutional recurrent network with attention, and Vuong et al. (2026) additionally incorporated graph convolutions to capture inter-index relationships. Safari & Ghaemi (2025) combined BiLSTM with fuzzy logic.

Another approach combines representations at different scales. S. Wang (2023) employed BiLSTM, Transformer, and TCN, Gong & Xing (2024) used decomposition, BiLSTM, self-attention, and TCN, and Li et al. (2026) utilised BiLSTM alongside a Transformer branch designed to process structural groups. Wu & Hong (2026) combined LSTM, CBAM, Time2Vec, and KAN. Zhao & Noah (2026) assigned different networks to the distinct level, trend, and seasonality components.

The accuracy of more complex models must be interpreted in the context of the forecasting task. In the research by Dipura et al. (2024), lower BiLSTM price errors were accompanied by a directional accuracy of 51.2%. T. Wang et al. (2026) also reported modest directional results, even though errors in price levels were reduced. Therefore, numerical accuracy and directional accuracy should be evaluated separately.

Previous studies have examined FTSE 100 return forecasting using ICA–LSTM (Mubarak, 2026) and DAX forecasting using CEEMDAN–LSTM–BPNN (Mutinda & Geletu, 2025) . However, these different model designs do not directly establish how a common standalone BiLSTM specification performs across the three markets considered here. This study addresses that comparison by evaluating DAX, CAC 40, and FTSE 100 returns under a common forecasting protocol and assessing whether low absolute errors are accompanied by improvements over a simple benchmark.

METHODOLOGY

This study used a quantitative empirical design to evaluate BiLSTM forecasts of next trading day logarithmic returns for DAX, CAC 40, and FTSE 100. A separate model was trained for each index using the same architecture and procedures. Daily closing prices covering October 1, 2016, to September 30, 2026, were obtained from Yahoo Finance using Python. Data preparation, model training, and evaluation were also performed in Python. Daily observations matched the one trading day forecast horizon.

The data were sorted by date and checked for duplicate entries and missing trading-day prices. Prices recorded on non-trading days according to each exchange calendar were excluded. Missing trading-day prices were not replaced with estimated values. Returns requiring a missing price and input windows containing these invalid returns were excluded. Daily logarithmic returns were calculated as (Ouyang et al., 2023):

$$r_t = ln(\frac{P_t}{P_{t-1}})$$

where: r_t – daily logarithmic return; P_t – index closing price for the current trading day; P_{t-1} – index closing price of the previous trading day. ln – natural logarithm.

Training covered October 1, 2016, to September 30, 2023; validation covered October 1, 2023, to March 31, 2025; and testing covered April 1, 2025, to September 30, 2026. These periods represent approximately 70%, 15%, and 15% of the

calendar duration. Observation counts differed across indices because of trading calendars and excluded missing-data windows. Returns were standardised using only the training-sample mean and standard deviation, with the same parameters applied to validation and test data (T. Wang et al., 2026).

Each input contained the previous 20 trading days of standardised returns, and the target was the following trading day's return. Both BiLSTM branches processed only observations available before the forecast date. Their final hidden states were combined and passed to a linear output layer, following the LSTM computational principles described by Vuong et al. (2026).

The architecture comprised one BiLSTM layer with 32 units per direction, followed by dropout of 0.2 and one linear output neuron. Training used the Adam optimiser, a learning rate of 0.001, batches of 32 observations, and mean squared error loss. Training lasted at most 100 epochs and stopped after 10 consecutive epochs without improvement in validation loss. The weights from the epoch with the lowest validation loss were restored. These settings were specified in advance; no hyperparameter search was conducted.

Each model was trained five times using seeds 11, 22, 33, 44, and 55. The same chronological partitions were used in every run. During testing, model weights remained fixed, while each input window was updated with the latest observed returns. Predictions were converted back to the original return scale. Reported BiLSTM metrics are averages of the five individual runs rather than metrics calculated from an average forecast.

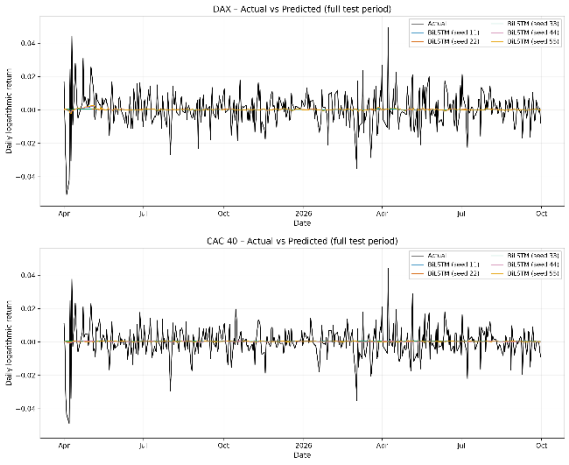
For each index, the benchmark predicted the mean daily logarithmic return calculated only from the training period, October 1, 2016, to September 30, 2023. Validation and test observations were excluded from this calculation. The resulting value remained fixed throughout testing and was therefore available before every forecast was made. Both BiLSTM and benchmark forecasts targeted the next trading day's return and were evaluated on the same test observations.

Forecasting accuracy was assessed using MAE and RMSE (Putra & Nurmawati, 2024). Directional accuracy measured the percentage of days when the predicted return direction matched the actual direction, with zero returns treated as a separate category (T. Wang et al., 2026).

Predictive R^2 was calculated as one minus the ratio of the BiLSTM sum of squared errors to the benchmark sum of squared errors. Positive values indicate lower squared errors than the benchmark, negative values indicate higher squared errors. The benchmark therefore has a predictive R^2 of zero. This measure was the primary criterion for relative performance across indices, supported by MAE, RMSE, and directional accuracy. The standard deviation of actual test returns was reported to provide context for differences in absolute errors. Statistical significance was not assessed.

RESULTS AND DISCUSSION

BiLSTM forecasts were evaluated from April 1, 2025, to September 30, 2026. Figure 1 presents actual returns and predictions from all five training runs for each index. Predictions remained close to zero and showed substantially less variation than actual returns, failing to reproduce the magnitude of large positive and negative movements. Table 1 compares average BiLSTM performance with the training-mean benchmark.



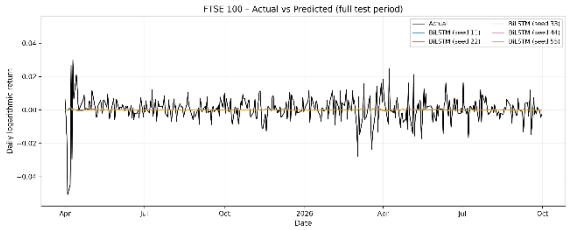


Fig. 1. *Actual and BiLSTM predicted daily logarithmic returns: (a) DAX, (b) CAC 40, (c) FTSE 100*
Source: *compiled by the authors*

Table 1. *BiLSTM and benchmark forecasting performance*

Index	MAE	RMSE	Predictive R^2	DA, %	Actual test returns SD
DAX BiLSTM	0.00776	0.01075	0.00352	50.50	0.01079
DAX Benchmark	0.00778	0.01077	0.00000	52.23	0.01079
CAC 40 BiLSTM	0.00682	0.00973	−0.00012	48.30	0.00974
CAC 40 Benchmark	0.00681	0.00973	0.00000	48.83	0.00974
FTSE 100 BiLSTM	0.00529	0.00789	−0.00134	50.29	0.00788
FTSE 100 Benchmark	0.00528	0.00789	0.00000	55.94	0.00788

Source: *compiled by the authors*

FTSE 100 the lowest BiLSTM MAE (0.00529) and RMSE (0.00789), alongside the lowest standard deviation of actual test returns (0.00788). Lower absolute errors therefore do not necessarily indicate greater predictive ability. DAX was the only index with a positive predictive R^2 (0.00352), corresponding to an approximately 0.35% reduction in mean squared error relative to the training-mean benchmark. Negative values for CAC 40 and FTSE 100 indicate slightly higher squared errors than the benchmark. BiLSTM directional accuracy was 50.50%, 48.30%, and 50.29%, respectively, below the corresponding benchmark values of 52.23%, 48.83%, and 55.94%.

Dipura et al. (2024) reported lower BiLSTM price-forecasting errors but directional accuracy of only 51.2%. Similarly, the present study found that low absolute errors were not accompanied by improvements in directional accuracy over the benchmark. Mubarak (2026) reported FTSE 100 directional accuracy of 57.4% for ICA–LSTM, compared with 50.9% for standalone LSTM. Although the standalone LSTM result is close to the 50.29% obtained here, differences in samples, inputs, and model designs prevent a direct performance ranking. The higher accuracy reported for ICA–LSTM suggests that alternative ways of preparing input data should be explored further, but their benefits would need to be tested under the same conditions.

In conclusion, the selected BiLSTM specification offered little improvement over the training-mean benchmark. The results show why absolute errors should be assessed alongside benchmark-relative performance and directional accuracy. These findings apply to the selected architecture and testing period; statistical significance was not assessed.

CONCLUSIONS AND RECOMMENDATIONS

A common BiLSTM architecture was applied to daily logarithmic returns of DAX, CAC 40, and FTSE 100 using data from October 2016 to September 2026. DAX achieved the best relative performance, with an approximately 0.35% reduction in mean squared error compared with the training-mean benchmark. FTSE 100 recorded the lowest absolute errors but also the lowest variability of actual test returns. BiLSTM did not outperform the benchmark in squared-error terms for CAC 40 or FTSE 100, and its average directional accuracy was below the benchmark for all three indices. These results do not establish a clear forecasting advantage for the selected specification.

The findings concern one architecture, three indices, and a single test period. Comparability is also constrained by dividend treatment: DAX incorporates dividend reinvestment, whereas the selected CAC 40 and FTSE 100 series are price

indices. Future studies should use comparable index definitions, evaluate performance over different periods, and compare BiLSTM with LSTM and simple benchmarks. Model settings should be selected using validation data, with final evaluation conducted on previously unused test data.

REFERENCE

An, Z., Wu, Y., Hao, F., Chen, Y., & He, X. (2024). A novel hierarchical feature selection with local shuffling and models reweighting for stock price forecasting. *Expert Systems with Applications*, *249*. https://doi.org/10.1016/j.eswa.2024.123482

Beniwal, M., Singh, A., & Kumar, N. (2024). Forecasting multistep daily stock prices for long-term investment decisions: A study of deep learning models on global indices. *Engineering Applications of Artificial Intelligence*, *129*. https://doi.org/10.1016/j.engappai.2023.107617

Chen, J., & Li, C. (2025). Application analysis and research of text model based on improved CNN-LSTM in the financial field. *IJACSA) International Journal of Advanced Computer Science and Applications*, *16*(6), 755–767. https://doi.org/10.14569/IJACSA.2025.0160675

Chen, Y., Fang, R., Liang, T., Sha, Z., Li, S., Yi, Y., Zhou, W., & Song, H. (2021). Stock price forecast based on CNN-BiLSTM-ECA model. *Scientific Programming*, *2021*, 1–20. https://doi.org/10.1155/2021/2446543

Dipura, R. W. A., Maulana, F. I., & Yulianto. (2024). Enhancing stock market forecasting with LSTM and BiLSTM models: A deep learning approach. *IEEE*, 50–55. https://doi.org/10.1109/ISRITI64779.2024.10963428

Dong, Z., Xu, Z., Pan, Q., Deng, Z., & Liu, J. (2026). Deep learning for multi-step financial time series forecasting with spectral energy-guided variational mode decomposition. *Neurocomputing*, *697*. https://doi.org/10.1016/j.neucom.2026.134127

Giantsidi, S., & Tarantola, C. (2025). Deep learning for financial forecasting: A review of recent trends. *International Review of Economics and Finance*, *104*. https://doi.org/10.1016/j.iref.2025.104719

Gong, H., & Xing, H. (2024). Predicting the highest and lowest stock price indices: A combined BiLSTM-SAM-TCN deep learning model based on re-decomposition. *Applied Soft Computing*, *167*. https://doi.org/10.1016/j.asoc.2024.112393

Kehinde, T. O., Khan, W. A., Chung, S. H., Wang, Y., & Piya, S. (2026). A real-time dynamic cluster-driven financial forecasting with attention-based deep learning and DEA. *Expert Systems with Applications*, *331*. https://doi.org/10.1016/j.eswa.2026.133353

Li, C., Li, Y., Hu, S., Xian, S., & Zhou, L. (2026). GBT-BiLSTM: A novel fusion network with granular-ball computing for financial time series forecasting. *Expert Systems with Applications*, *331*. https://doi.org/10.1016/j.eswa.2026.133114

Mubarak, A. (2026). Independent component analysis preprocessing for LSTM and SVR in financial engineering: Evidence from S&P 500 and FTSE 100 return forecasting. *International Journal of Intelligent Engineering and Systems*, *19*(6), 358–378. https://doi.org/10.22266/ijies2026.0630.20

Mutinda, J. K., & Geletu, A. (2025). Stock market index prediction using CEEMDAN-LSTM-BPNN-decomposition ensemble model. *Journal of Applied Mathematics*, *2025*(1). https://doi.org/10.1155/jama/7706431

Ouyang, Z., Lu, M., & Lai, Y. (2023). Forecasting stock index return and volatility based on GAVMD-Carbon-BiLSTM: How important is carbon emission trading? *Energy Economics*, *128*. https://doi.org/10.1016/j.eneco.2023.107134

Purnomo, K., Wijaya, R. A., Fajar, M., & Suri, P. A. (2024). Comparative analysis of BiLSTM deep learning model and random forest regressor performance on Indonesian nickel mining company stock prices. *Procedia Computer Science*, *245*,

778–786. https://doi.org/10.1016/j.procs.2024.10.304

Putra, R. A., & Nurmawati, E. (2024). Prediction-based stock portfolio optimization using bidirectional long short-term memory (BiLSTM) and LSTM. *Scientific Journal of Informatics*, *11*(3), 609–620. https://doi.org/10.15294/sji.v11i3.5941

Rani, I., Porwal, S., Verma, C. K., Srivastava, N., & Jha, M. (2026). Deep residual bidirectional learning for multi-horizon financial forecasting. *Evolutionary Intelligence*, *19*(5). https://doi.org/10.1007/s12065-026-01273-5

Safari, A., & Ghaemi, S. (2025). NeuroFuzzyMan: A hybrid neuro-fuzzy BiLSTM stacked ensemble model for financial forecasting and analysis: Dataset case studies on JPMorgan, AMZN and TSLA. *Expert Systems with Applications*, *266*. https://doi.org/10.1016/j.eswa.2024.126037

Sathya, N., Indradevi, R., & Gayathiri, R. (2025). Enhancing stock market predictions using hybrid deep learning models with sentiment analysis and feature engineering. *Informatica (Slovenia)*, *49*(33), 69–78. https://doi.org/10.31449/inf.v49i33.8079

Singh, H., Madaan, S., Kumar, J., Attri, V., Singh, S., Mittal, N., & Gared, F. (2026). A hybrid transformer GRU architecture for stock price prediction: Optimization and performance evaluation using the artificial rabbits algorithm. *International Journal of Computational Intelligence Systems*, *19*(1). https://doi.org/10.1007/s44196-026-01446-1

Swetha, B., & Arya, K. (2025). Prediction of the stock market using a hybrid model based on feature expansion and LSTM-based algorithms. *IEEE Access*, *13*, 196050–196080. https://doi.org/10.1109/ACCESS.2025.3631096

Vuong, P. H., Phu, L. H., Duy, L. N., Bao, P. T., Tran, U. T. M., & Trinh, T. D. (2026). A spatiotemporal deep learning framework based on GCNN and BiLSTM for stock price forecasting. *Journal of Engineering (United Kingdom)*, *2026*(1). https://doi.org/10.1155/je/9973211

Vuong, P. H., Phu, L. H., Duy, L. N., Bao, P. T., & Trinh, T. D. (2024). An improved convolutional recurrent neural network for stock price forecasting. *IAES International Journal of Artificial Intelligence*, *13*(3), 3381–3394. https://doi.org/10.11591/ijai.v13.i3.pp3381-3394

Wang, D., Zhao, H., Wu, L., & Li, X. (2026). Stock time series prediction correction expert: A knowledge-guided stable multiple large language models enhancement method. *Decision Support Systems*, *209*. https://doi.org/10.1016/j.dss.2026.114723

Wang, J., Cheng, Q., & Sun, X. (2024). A novel Gaussian process regression-based stock index interval forecasting model integrating optimal variables screening with bidirectional long short-term memory. *Soft Computing*, *28*(6), 5541–5556. https://doi.org/10.1007/s00500-023-09362-8

Wang, S. (2023). A stock price prediction method based on BiLSTM and improved transformer. *IEEE Access*, *11*, 104211–104223. https://doi.org/10.1109/ACCESS.2023.3296308

Wang, T., Cheng, Y., Tang, Z., Shan, Y., & Wang, H. (2026). MSATE-Net: A multi-scale attention-enhanced bidirectional temporal network for stock index forecasting. *Symmetry*, *18*(8). https://doi.org/10.3390/sym18081398

Wu, H., & Hong, X. (2026). Symmetry-aware stock prediction based on optimized multi-module collaborative features with LSTM-CBAM-Time2Vec-KAN. *Symmetry*, *18*(7). https://doi.org/10.3390/sym18071198

Zhang, T., Wang, Z., Zheng, T., & Guo, Y. (2026). A multi-stock market trend prediction model based on exponential decomposition and dual-stream networks. *Journal of King Saud University - Computer and Information Sciences*, *38*(6). https://doi.org/10.1007/s44443-026-00551-7

Zhao, B., & Mohd Noah, S. A. (2026). SeasonFuser: A parallel hybrid network with differentiable seasonal decomposition for financial time series

forecasting. *IEEE Access*. https://doi.org/10.1109/ACCESS.2026.3725334

Zheng, H., Wang, H., & Chen, J. (2021). Evolutionary framework with bidirectional long short-term memory network for stock price prediction. *Mathematical Problems in Engineering*, *2021*(1). https://doi.org/10.1155/2021/8850600