Data Science and Engineering (DSE) Record, Volume 4, issue 1.

Chinese Stock Forecasting Based on Machine Learning

Yang Zhang[1] and Thaned Rojsiraphisal[2]

[1] Data Science Program, Chiang Mai University, Chiang Mai, Thailand, 50200
[2] Department of Mathematics, Faculty of Science, Chiang Mai University, Thailand, 50200
yang_zhang@cmu.ac.th

Abstract. In the financial market analysis field, machine learning techniques for stock price prediction have garnered considerable interest. This study investigates the effectiveness of Long Short-Term Memory (LSTM) models in predicting stock prices for growth stocks and the CSI 300 index in the Chinese A-share market. The study also explores the influence of various technical indicators of the LSTM model and models on forecasting accuracy. The experimental results demonstrate that the LSTM model is the most effective in predicting stock prices in the A-share market, while other algorithms such as WMA and ARIMA are not as successful in forecasting long-term stock market data. This study proposes some modifications further to enhance the accuracy and dependability of the prediction model.

Keywords: LSTM models, Stock Price Prediction, Chinese A-share market.

1 Introduction

The stock market plays an indispensable role in the growth and development of businesses, providing a platform for companies to raise capital and for investors to obtain returns on investment [1]. However, accurate stock price prediction is a challenging task due to the multitude of factors that influence it, making it difficult to predict future trends using traditional methods. In recent years, machine learning techniques, with their ability to handle large amounts of data and model complex non-linear relationships, have emerged as a promising approach for stock forecasting.

This research focuses on exploring the potential of machine learning techniques, specifically the Long Short-Term Memory (LSTM) model, for predicting two growth stock prices in the Chinese market, with a focus on the CSI 300 Index. The CSI 300 Index, comprising the largest 300 stocks listed on the Shanghai and Shenzhen stock exchanges, is a key indicator of the Chinese market's health and stability. The research aims to investigate the LSTM model's ability to accurately predict the opening and closing prices of the CSI 300 Index and compare its performance with other commonly used time series models.

The study's significance lies in its potential to provide valuable insights for individuals and businesses looking to invest in the Chinese stock market, contributing to the development of effective investment strategies. Furthermore, this research will enrich the existing literature on the potential of machine learning techniques for stock forecasting, addressing the limitations of traditional methods and highlighting promising avenues for future research. In summary, this study contributes to the ongoing efforts to improve the accuracy of stock price predictions and better understand the potential of machine learning techniques for stock forecasting, with significant implications for the global economy.

2 Literature Review

2.1 LSTM model to predict the stock market

Recent advances in deep learning have enabled the development of sophisticated models for financial time series prediction, with promising results reported by several research groups. Chen et al. (2015) [2] employed the Long Short-Term Memory (LSTM) model to forecast China stock returns. The authors preprocessed the historical data into 30-day sequences with ten learning features and 3-day learning rate labeling, demonstrating that the LSTM model outperformed the random prediction method, achieving an impressive accuracy improvement from 14.3% to 27.2%. Additionally, Roondiwala et al. (2017) [3] employed the LSTM model to predict the closing price of NIFTY 50 data ranging from 2011 to 2016. Their approach incorporated fundamental data such as open, close, low, and high, without incorporating macroeconomic and technical indicators. These studies provided exciting insights into the potential of deep learning for financial time series analysis and demonstrate its effectiveness in improving prediction accuracy. These articles highlight the potential of deep learning to improve prediction accuracy and demonstrate its effectiveness in financial time series analysis.

2.2 Comparison of LSTM with Other Time Series Models

Several studies have compared the performance of LSTM models with other time series models in predicting stock prices. Zhang et al. (2018) [4] compared LSTM with ARIMA and GARCH models in predicting the stock prices of six Chinese companies and found that LSTM outperformed both models. Similarly, Chen et al. (2018) [5] compared LSTM with ARIMA and random walk models in predicting stock prices of companies listed on the Shenzhen Stock Exchange and found that LSTM outperformed other models.

Lin et al. (2021) [6] adopted a different approach by combining the phase-space reconstruction method and the LSTM model for stock price prediction. Their study analyzed various market environments, such as S&P 500, DJIA, Nikkei 225, Hang Seng Index, China Securities Index 300, and ChiNext index. By taking the historical price only, the LSTM model was compared with Multilayer Perception, Support Vector Regressor, and ARIMA models. The results indicated that the LSTM model outperformed other models in predicting the S&P 500 data.

In a comparative study by Gao et al. (2020) [7], four machine learning algorithms, namely Multilayer Perceptron, LSTM, Convolutional Neural Network, and Uncertainty-Aware Attention, were assessed for predicting the next day's stock price The S&P

500 index, CSI 300 index, and Nikkei 225 index were selected to represent the most developed market, the less developed market, and the developing market, respectively. Predictors such as open price, close price, trading volume, moving average convergence divergence, average true range, exchange rate, and interest rate were considered. The study revealed that the uncertainty-aware attention model demonstrated slightly better performance compared to other models. Furthermore, incorporating additional predictors such as the volatility index and the unemployment rate can potentially enhance the prediction model's performance.

3 Data and Methodology

We aim to develop a machine-learning model that leverages the power of LSTM techniques for forecasting the growth of Chinese companies' stock prices. The study commences with the collection of raw data from the stock market, which is then preprocessed to detect correlations between individual stocks and indexes, decompose the data, identify anomalies, and split training and test datasets.

The LSTM model's training and evaluation are carried out in three stages, starting with evaluating its performance in forecasting two growth stocks. The second stage involves creating an LSTM model suitable for the CSI 300 index, while the third stage entails creating additional time series forecasting models and comparing their strengths and weaknesses with the LSTM model. The study culminates in the evaluation and summary of all the models. The study's phases are visually represented in **Fig. 1**.

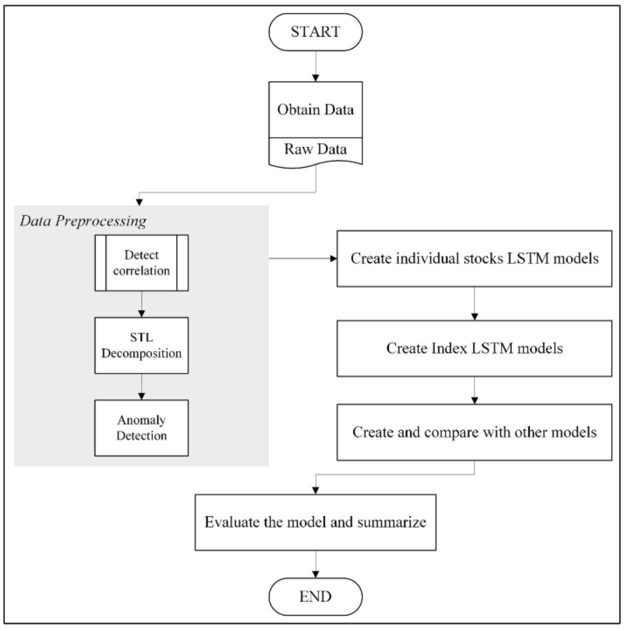

Fig. 1. Procedure

3.1 Data Capture and Pre-Processing

Before building the model, we need to preprocess the data to ensure that it meets the input requirements of the model and is in a format that the model can understand. In this section, we will go through the data preprocessing steps for the CSI 300 and two growth stock data.

The CSI 300 index was obtained from the Python package baostock, while the growth stock data was obtained using yfinance. We finalized the LSTM modeling with two growth stocks, Haier and Yonyou.

We then use the STL decomposition method to check the trend and seasonality of the data. The Seasonal and Trend Decomposition using Loess (STL) method has emerged as a versatile and robust technique for decomposing time series data into its constituent components. The method was developed by Cleveland et al. (1990) [8] and has since gained widespread adoption in numerous fields due to its effectiveness in addressing the challenge of inaccurate model predictions caused by trends and seasonality in long-time series data. The decomposition results of the CSI 300 weekly data are shown in **Fig. 2**, and the trend and seasonal intensity of each data are shown in **Table 1**.

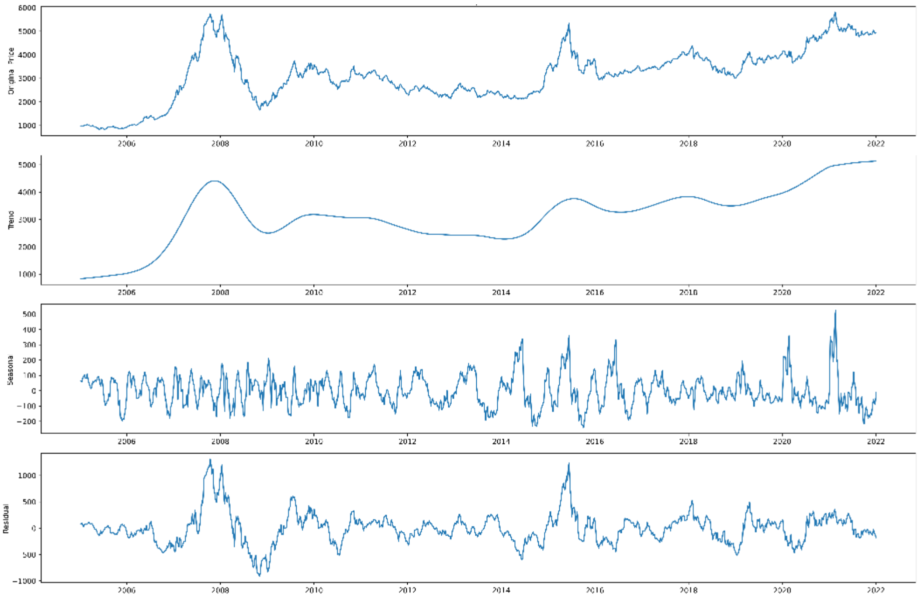

Fig. 2. CSI 300 Weekly Data STL Decomposition Result

Table 1. Trend and seasonal strength

Ticker	Freq	Trend Strength	Seasonal Strength
Haier	Daily	0.2870	0.9996
Yonyou	Daily	0.3346	0.9994
CSI 300	Daily	0.2853	0.9992
Haier	Weekly	0.3675	0.9798
Yonyou	Weekly	0.2967	0.9762
CSI 300	Weekly	0.1138	0.9294

From the above table, we can conclude that all the data trends are not particularly obvious, but the seasonal strength is close to 1, and we need to pay attention to the influence of seasonality on the model when considering modeling and verification in the later stage.

Then we performed anomaly detection for each data using the standard deviation of 3 times the residuals as a threshold. **Fig. 3** shows the anomaly detection results for the CSI 300 weekly data. For the CSI 300 index, the anomalies are mainly in 2008 and 2015. Combined with historical information, these two anomalies occurred during the global financial crisis in 2008 and the deleveraging of China's stock market in 2015, there were large fluctuations in these two years.

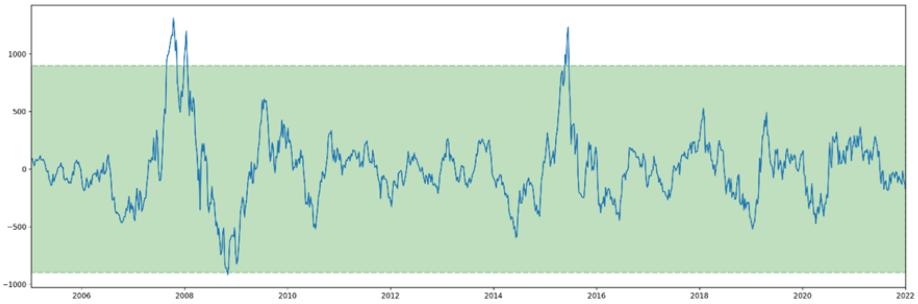

Fig. 3. Abnormal distribution of CSI 300 weekly closing prices

We then used the ADF test to assess the stationarity of each data and finally found that only the weekly data of CSI 300 was stationary.

Table 2. ADF test results

Ticker	Ferq	ADF Statistic	p-value	Stationary or not
Haier	Daily	0.3488	0.9795	Non-stationary
Yonyou	Daily	-0.5941	0.8723	Non-stationary
CSI 300	Daily	-2.0359	0.2711	Non-stationary
Haier	Weekly	0.3821	0.9808	Non-stationary
Yonyou	Weekly	-0.4558	0.9004	Non-stationary
CSI 300	Weekly	-3.6017	0.0057	Stationary

3.2 Modeling and evaluation

In the model creation section, we utilized a deep learning model known as LSTM to create predictive models for stock market data. Specifically, we created forward propagation and backward propagation models using daily and weekly data for growth stocks (Haier & Yonyou) and the CSI 300 index. The aim was to develop models that could accurately forecast future trends in the stock market.

To ensure the accuracy of the models, we employed various techniques such as varying the number of EPOCHS and learning rates and evaluating the performance of different models. By doing this, we were able to identify the optimal model that provided the highest level of accuracy in predicting future trends.

In addition, we also created two other models, namely the Weighted Moving Average (WMA) and Auto ARIMA models to compare their performance with LSTM.

4 Results

In the model creation process, we experimented with different approaches to improve the model's performance. We found that using the first-order difference data did not result in a good fit for the model. Therefore, we chose to use the original data for training the model.

To optimize the model's performance, we tested different EPOCHS and learning rates. We found that an EPOCHS value of 20 was more appropriate as increasing the EPOCHS beyond this point did not improve the model's performance, but instead caused overfitting. Furthermore, we selected learning rates of 0.01 and 0.001 for training the model. We found that too high a learning rate precision caused the loss to change slowly, while too low a learning rate precision caused the loss to change drastically. By selecting an optimal learning rate, we were able to train a model with better performance.

4.1 LSTM models

Table 3. Test Loss for different LSTM models

Ticker	Frep	Model	LR	EPOCHS	Test Losses
Haier	Daily	Forward	0.01	7	2.3978
Haier	Daily	Backward	0.001	19	4.1068
Yonyou	Daily	Forward	0.001	19	0.6294
Yonyou	Daily	Backward	0.001	19	1.1457
Haier	Weekly	Forward	0.01	3	13.7853
Haier	Weekly	Forward	0.001	7	9.5219
Haier	Weekly	Backward	0.01	1	12.8730
Haier	Weekly	Backward	0.001	13	14.7881
CSI 300	Daily-Open	Forward	0.001	19	0.0299
CSI 300	Daily-Open	Forward	0.01	5	0.0278
CSI 300	Daily-Close	Forward	0.001	16	0.0405
CSI 300	Daily-Close	Forward	0.01	19	0.0314
CSI 300	Weekly-Open	Forward	0.01	18	0.1085
CSI 300	Weekly-Open	Forward	0.001	19	0.1580
CSI 300	Weekly-Close	Forward	0.01	18	0.0897
CSI 300	Weekly-Close	Forward	0.001	18	0.1418

From **Table 3**, we can see that for individual stocks, the forward propagation LSTM models generally perform better than the backward propagation models, in addition, the weekly data has less predictive power than the daily data. The optimal LSTM model for the growth stock prediction model is the forward propagation LSTM model for yonyou daily data with a learning rate of 0.001 and an EPOCH of 19, and the test loss is 0.6294. The training processes of the model and the predicted shape of the optimal model for Yonyou are depicted in **Fig. 4-6**.

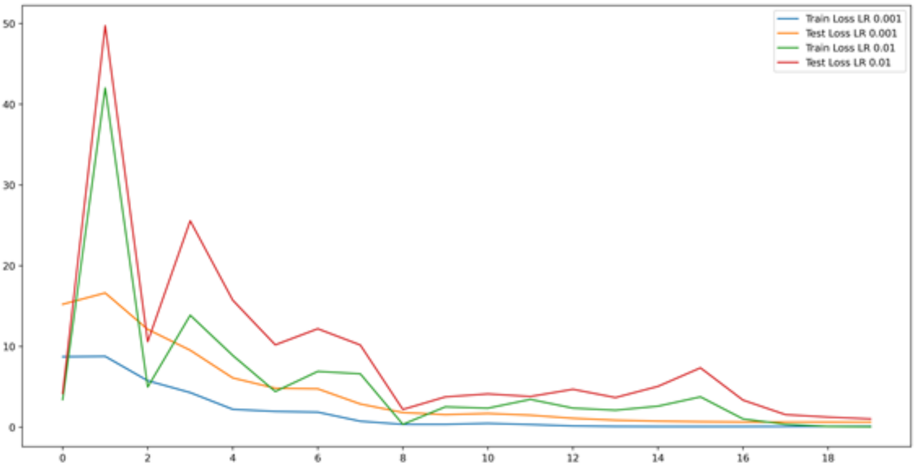

Fig. 4. Daily Yonyou Forward Propagation LSTM Loss with Different LR at EPOCH=20

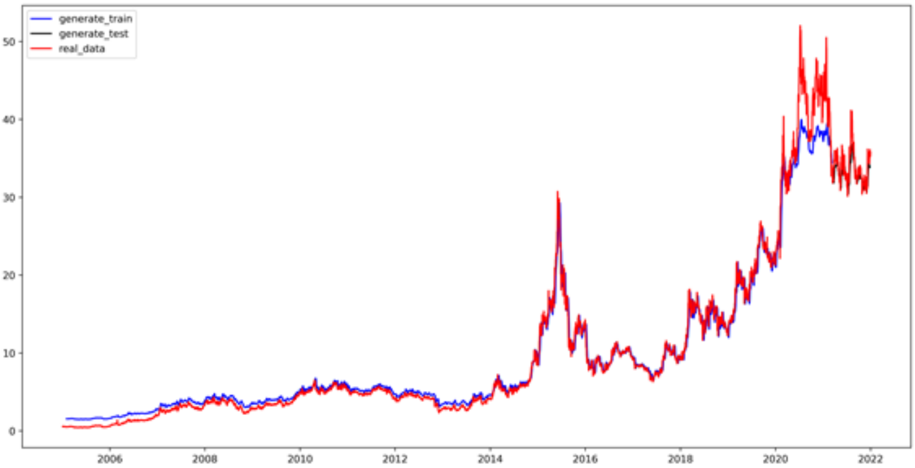

Fig. 5. The shape of daily Yonyou

The LSTM models yield better results for the CSI 300 index, with lower daily data Losses overall (see **Fig. 6**). However, to prevent overfitting problems, we show the prediction results of the weekly optimal model in **Fig. 7-8**.

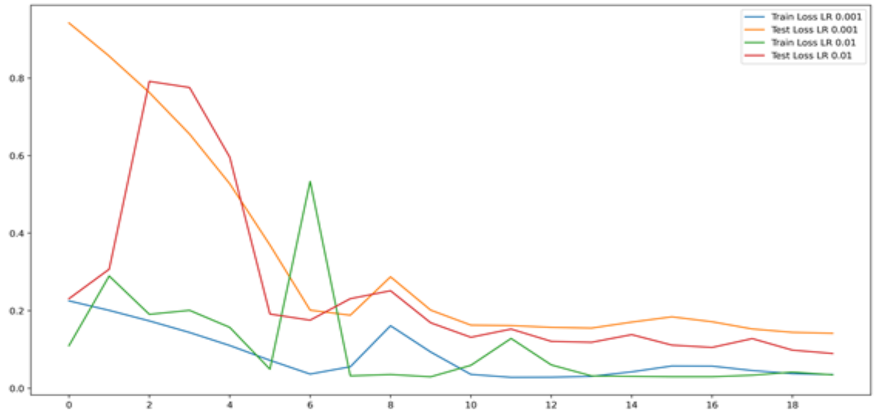

Fig. 6. CSI 300 weekly close price forward propagation LSTM loss for different LRs at EPOCH=20

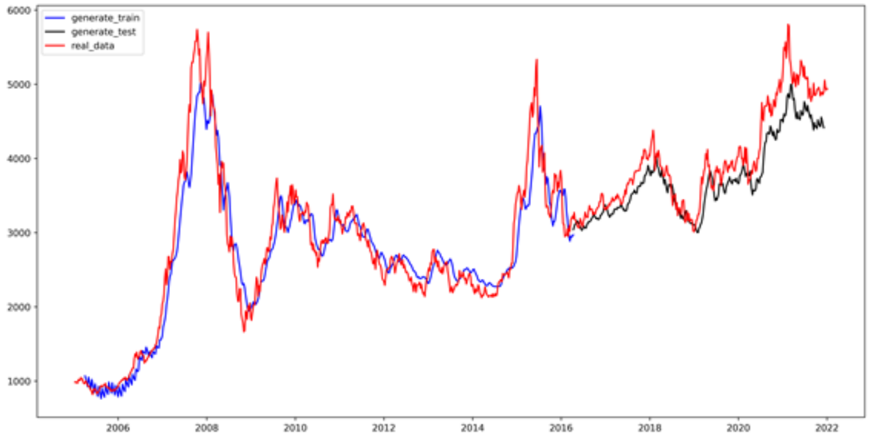

Fig. 7. The shape of weekly CSI 300 close price training, testing, and real data

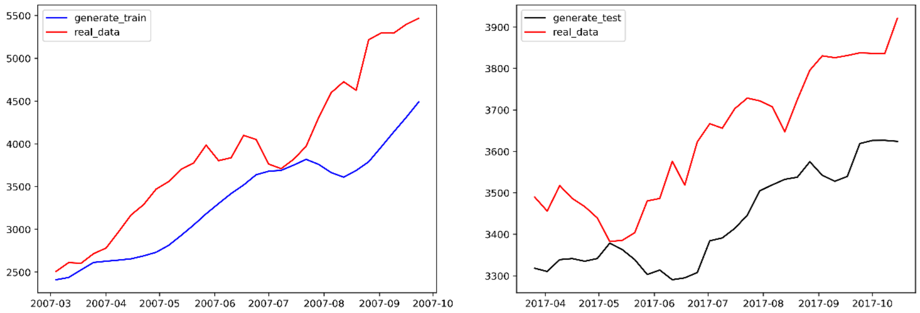

Fig. 8. Weekly CSI 300 close price of real data vs training data (left) and vs test data (right)

From **Fig. 7**, we can see that the model as a whole fits the real data trend better, but there is a certain lag, and later studies can consider how to eliminate such lags by adjusting the data or model parameters.

4.2 Other time series analysis models

For the training of other time series models, we use the CSI 300 index weekly data, using the first 80% of the normalized data as training data and the last 20% as test data, and the final results are shown in **Fig. 9**.

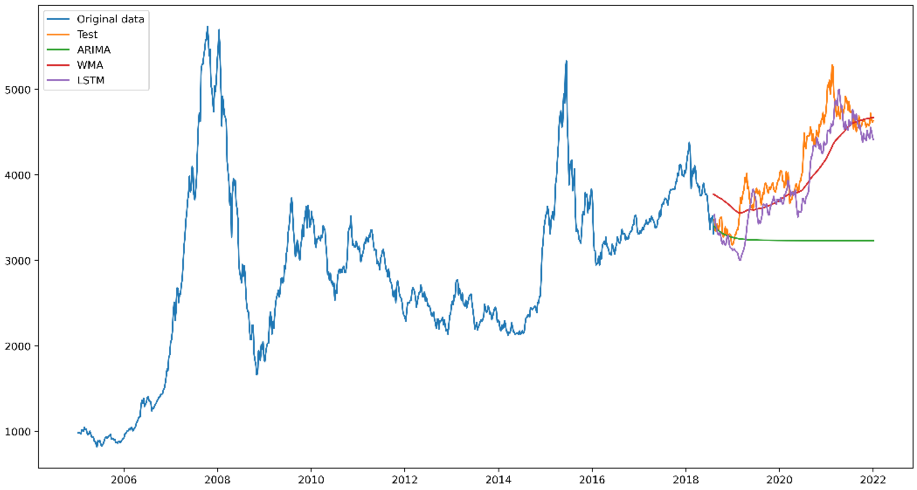

Fig. 9. Weekly CSI 300 Closing Price Models

From the comparison in **Fig. 9**, we clearly see that the LSTM model has the best prediction performance for the data, followed by the WMA model, while the ARIMA model can predict only the last few cycles of the data, and the late prediction results show a straight line. We can also get from **Table 4** that the MSE of LSTM is 0.0586, which is the smallest among the three models, while ARIMA is 1.0177, which is the maximum.

Table 4. MSE results for each model

Ticker	Ferq	Model	MSE
CSI 300	Weekly	LSTM	0.0586
CSI 300	Weekly	WMA	0.1178
CSI 300	Weekly	ARIMA	1.0177

5 Conclusion

The application of machine learning techniques to the prediction of the Chinese stock market is a challenging yet important area of research.

For growth stocks, several LSTM models were developed, using daily and weekly data, as well as forward and backward propagation, and various parameter tuning methods. Through careful analysis, an optimal Loss of 0.6294 was achieved for the forward propagation prediction of Yonyou, indicating the effectiveness of the LSTM model in this area. However, the forward propagation prediction for Haier achieved a loss of 2.3978, which is significantly higher than that of Yonyou.

In the prediction of the CSI 300 index, the study found that weekly data performed better in fitting the LSTM model, with the smoothed-out view of the stock prices allowing for a more accurate capture of underlying patterns and trends. However, the study also highlighted the importance of taking measures to avoid overfitting when working with weekly data.

In the comparison with other models, the study showed the WMA and ARIMA models did not perform as well as the LSTM model. Nevertheless, the study also emphasized that each model has its unique advantages and disadvantages and that the appropriate model should be chosen based on specific conditions and requirements.

References

1. Pettinger, T. (2015). How does the stock market affect the econmy. *Stock market*.
2. Chen, K., Zhou, Y., & Dai, F. (2015, October). A LSTM-based method for stock returns prediction: A case study of China stock market. In *2015 IEEE international conference on big data (big data)* (pp. 2823-2824). IEEE.
3. Roondiwala, M., Patel, H., & Varma, S. (2017). Predicting stock prices using LSTM. *International Journal of Science and Research (IJSR)*, *6*(4), 1754-1756.
4. Zhang, C., Ji, Z., Zhang, J., Wang, Y., Zhao, X., & Yang, Y. (2018, October). Predicting Chinese stock market price trend using machine learning approach. In *Proceedings of the 2nd International Conference on Computer Science and Application Engineering* (pp. 1-5).
5. Chen, L., Qiao, Z., Wang, M., Wang, C., Du, R., & Stanley, H. E. (2018). Which artificial intelligence algorithm better predicts the Chinese stock market?. *IEEE Access*, *6*, 48625-48633.
6. Lin, Y., Yan, Y., Xu, J., Liao, Y., & Ma, F. (2021). Forecasting stock index price using the CEEMDAN-LSTM model. *The North American Journal of Economics and Finance*, *57*, 101421.
7. Gao, P., Zhang, R., & Yang, X. (2020). The application of stock index price prediction with neural network. *Mathematical and Computational Applications*, *25*(3), 53.
8. Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition. *J. Off. Stat*, *6*(1), 3-73.