Stock Price Prediction and Multi Factor Risk Quantification Evaluation based on Hybrid LSTM-GAN Model

Jiale Dong

Peking University, Beijing, 100080, China

djl2401212380@stu.pku.edu.cn

Keywords: Stock Price Prediction; LSTM-GAN Model; Multifactorial Risk; Quantitative Evaluation

Abstract: This study proposes a stock price prediction model based on the hybrid LSTM-GAN (Long Short Term Memory Generative Adversarial Network) algorithm, combined with a multi factor risk quantification evaluation method, aiming to improve the accuracy and effectiveness of stock price prediction. By combining time series prediction with deep learning techniques, this study not only captures the nonlinear characteristics of stock prices, but also quantifies risk through multi factor models, providing strong support for investment decisions. In the experiment, the mean squared error (MSE) of the LSTM-GAN model was 0.2315, and the mean absolute error (MAE) was 0.3847. In the quantitative risk assessment experiment, the model predicted a portfolio Sharpe ratio of 0.2174 and a maximum retracement of 0.2351.In the above data conclusions, the superior performance of the hybrid LSTM-GAN model in stock price prediction and risk management is demonstrated.

1. Introduction

Traditional linear models frequently struggle to capture the nonlinear aspects and complicated market dynamics in stock price data when it comes to stock market prediction and risk assessment. With the development of machine learning technology, more and more research is exploring the application of deep learning models in financial market analysis. Especially the LSTM network and GAN algorithm have shown great potential in predicting sequence data and generating realistic data. However, using these models alone still has certain limitations, such as less than ideal performance when dealing with extreme market conditions or long-term series data. Because of this, this research suggests a hybrid model that combines GAN and LSTM in an effort to increase the efficacy of multifactor risk quantification evaluation and to improve stock price forecast accuracy.

The main contribution of this paper is to propose an innovative hybrid LSTM-GAN model, which can not only effectively deal with the nonlinear characteristics of stock price data, but also improve the accuracy of prediction by generating more abundant and real training data. In addition, we have also incorporated a multi factor risk assessment method, by introducing various risk factors such as market sentiment index, to enhance the application ability and risk management effectiveness of the model in the real financial environment. This study not only provides new technological means for stock price prediction, but also offers a more comprehensive and dynamic perspective for financial risk assessment.

The article is organized as follows in terms of structure: First, the context of the research and the limitations of the previous investigations are presented in the introductory section. Next, the methodology section elaborates in detail on the combination of LSTM and GAN, the construction of the model, and the method of multi factor risk quantification. Next is the experimental results section, which showcases the performance of the model in stock price prediction and risk assessment. Finally, a summary of the study's findings is provided in the discussion and conclusion section, along with recommendations for further research. In the discussion and conclusion section, the study's findings are finally summarized and potential directions for further research are recommended. Each section aims to clearly demonstrate the logic of the research and provide

in-depth analysis of its significance and potential applications in the financial field.

2. Related Works

Scholars domestically and internationally have studied stock price prediction and risk assessment in great detail in the last few years. For example, Lu W et al. used CNN algorithm to predict the next day's stock price [1]. In light of this, Wu et al. integrated CNN and LSTM networks to increase stock price prediction accuracy [2]. The modeling and prediction of stock prices by Lin et al. is an important and challenging task in financial research, which is of great significance for investors to reduce decision risks and improve investment returns [3]. Ge et al. proposed a stock price prediction method based on sentiment analysis, which can effectively shorten training time and improve prediction accuracy [4]. Zakhidov G explored the key role of economic indicators in understanding market trends and predicting future performance, indicating that these indicators are a barometer of economic stability and crucial for risk assessment, trend identification, and developing proactive strategies[5]. Yuan et al. proposed a deep learning based method that integrates multi-source data and investor sentiment to construct a hybrid model for stock price prediction[6]. Zhan X et al. explored the application of AI and robotic process automation in financial accounting and management, highlighting their key role in promoting the digital transformation of enterprise finance [7]. In order to improve the prediction accuracy of time series models, it is necessary to comprehensively understand the linear and nonlinear characteristics of their data and model them using ARIMA and RNN models respectively [8]. However, these methods still have certain shortcomings when dealing with extreme market conditions and long time series data.

To solve the above problems, some researchers have introduced techniques such as GAN algorithms. For example, Wang Lei used Bayesian search method to optimize hyperparameters of four machine learning algorithms and selected 34 input factors suitable for the Chinese market [9]. Wang provides a simple and practical method for investors to make investment decisions [10]. However, their approach still has some limitations when dealing with multi-factor risk quantification. For this reason the hybrid LSTM-GAN model proposed in this study, combined with a multifactor risk assessment approach, aims to address these issues.

3. Methods

3.1 Data Preprocessing

3.1.1 Data Collection

Firstly, we collected historical data of the stock market from several reliable data sources, including daily opening price, closing price, high price, low price and volume. In addition, we also collected multi factor data related to stocks, such as market sentiment indices and other data. The time span of these data covers the last ten years of trading records to ensure the adequacy and representativeness of the data [11].

3.1.2 Data cleaning

After data collection was completed, we cleaned and processed the data. Data cleaning includes the following steps:

Missing value processing: for missing values, this paper uses forward-filling and backward-filling methods to fill in the missing values. If there are more missing data in a certain time period, they are deleted to avoid affecting the LSTM-GAN model training.

Outlier detection: records with obvious errors will be selected for deletion; for non-obvious outliers, this paper adopts the median replacement method for processing.

3.1.3 Data normalization

Due to the large difference in the data magnitude of different indicators, we normalized the data in order to eliminate the impact of magnitude differences on model training.

3.1.4 Data display

In order to demonstrate the effect of data preprocessing, we selected a part of the processed data for display, as shown in Table 1:

Table 1: Processed data

Date	Open	High	Low	Close	Volume	P/E Ratio	P/B Ratio	Market Sentiment Index
2023-01-01	10.5	10.8	10.2	10.6	1200000	15.2	1.5	0.65
2023-01-02	10.6	10.9	10.3	10.7	1150000	15.3	1.52	0.66
2023-01-03	10.7	11	10.5	10.8	1300000	15.4	1.55	0.67
2023-01-04	10.8	11.1	10.6	10.9	1400000	15.5	1.57	0.68
2023-01-05	10.9	11.2	10.7	11	1250000	15.6	1.6	0.69

3.2 LSTM Model Construction

3.2.1 Model architecture design

This study developed a multi-layer LSTM network to represent the temporal dependence of prices. The basic architecture includes an input layer, multiple LSTM layers and an output layer. Each layer of the LSTM is primarily responsible for extracting features from the time series data and layer by layer capturing higher level feature information [12].

Input layer: the input layer accepts preprocessed time series data. The shape of the input data is (time step, number of features), where the time step represents the length of the time series and the number of features represents the data dimension of each time step.

LSTM layer: We stack two LSTM layers to improve the model's capacity for learning. After receiving data from the input layer, the first LSTM layer outputs the hidden state of the data. After receiving the first layer's output, the second LSTM layer extracts additional features.

Fully connected layer: after the LSTM layer, this paper adds a fully connected layer that maps the output of the LSTM layer to the predicted stock prices.

Output layer: the output layer is used to generate the final stock price prediction.

3.2.2 Model hyperparameters setting

This work carefully sets and optimizes the model's hyperparameters to guarantee the LSTM-GAN model's training effect. These hyperparameters include:

Number of LSTM units: the number of units contained in each LSTM layer, we set it to 50 to ensure that the model has enough capacity to learn complex time series features.

Batch size: the amount of data in each batch during training, we choose 32 as the batch size to balance the training time and model performance.

Number of training rounds: This article will train the model 100 times to fully understand the data types.

3.2.3 Model training

In order to optimize the model parameters during the training phase of the LSTM-GAN model, the back-propagation technique is applied to minimize the loss function. This study used MSE as the loss function in the training process, which calculates the discrepancy between the expected and actual values. The specific steps are as follows:

Forward propagation: To produce the anticipated values, the input data is routed via the fully connected layer and the LSTM layer.

Calculate loss: the value of the loss function is calculated based on the difference between the predicted and actual values.

Backpropagation: the gradient of the loss function with respect to the model parameters is calculated and the parameters are updated to minimize the loss function.

3.3 GAN Model Integration

Until recent years, the rise of GAN algorithmic modeling has led to a solution to the problem of generating financial time series. GAN models can learn the distribution of data to generate new financial time series data. GAN models are extensively utilized in the financial industry for a variety of purposes, including data improvement, anomaly detection, and the creation of financial time series. In this paper, we will research on the GAN algorithm and BOOTSTRAP and other models to generate financial time series to simulate domestic and international stock trend situation.

3.3.1 GAN model architecture design

The GAN model consists of two main components:

Generator: the output of a generator is a high dimensional vector, such as a generated image etc. Since the input vectors of the generator are randomly sampled through a distribution, the input vectors are different each time and hence the output of the generator is different each time and will form a complex distribution. Where the generator can be represented by equation (1):

$$G(z) = \sigma(W_g \cdot z + b_g) \tag{1}$$

Where in equation (1), $G(z)$ denotes the output of the generator, z denotes the input random noise vector, W_g is the weight matrix of the generator, b_g denotes the bias of the generator and σ is the activation function.

Discriminator: the discriminator receives real and generated data as input and outputs a probability value indicating the likelihood that the input data is real. By opposing the generator, the discriminator, which is made up of a multi-layer completely connected network, increases the created data's legitimacy.

3.3.2 Generator and discriminator training

The training process of the GAN consists of alternating training of the generator and the discriminator:

Discriminator training: first, the discriminator is trained using real stock price data and generated fake data. By maximizing the output probability of the discriminator on real data and minimizing the output probability on generated data, the discriminator's discriminative ability is improved.

Generator training: the generator is then fed with random noise and the gradient is back-propagated through the discriminator's discrimination of the generated data. By reducing the discriminator's output probability on the produced data, the generator progressively produces more realistic data.

3.3.3 Integration of GAN and LSTM models

The generator is linked with the LSTM model to improve stock price prediction once the GAN model has finished training. The specific integration steps are as follows:

Data generation: using the trained generator to generate a large number of stock price series similar to the real data distribution to expand the training dataset.

LSTM training: To improve the LSTM model's ability to generalize and make accurate predictions, it is trained using an expanded dataset.

Joint optimization: further improving the prediction performance by jointly optimizing the GAN and LSTM models. The specific method is to optimize the generator and the LSTM model alternately to make the generated data more realistic, and at the same time to improve the prediction ability of the LSTM model.

3.4 Multifactor Risk Quantification

3.4.1 Selection of risk factors

The first step in multi-factor risk quantification is to select appropriate risk factors. Based on financial theory and actual market experience, we have selected the following main factors:

Price/Earnings Ratio (P/E Ratio): it reflects the valuation level of a stock, and an overly high P/E

ratio may imply that there is a bubble in the market.

P/B Ratio: It calculates the assets of a company's market worth. A low P/B ratio could indicate that the company's assets are being undervalued by the market.

Market Sentiment Index (Market Sentiment Index): it reflects investor sentiment and market expectations, and high market sentiment may indicate an overheated market.

3.4.2 Construction of multifactor models

Based on the selected risk factors, a multi-factor risk model is constructed. To calculate the effect of each component on stock prices, we employ multiple linear regression. The form of the multiple regression model is shown in equation (2):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon (2)$$

In equation (2), Y is the stock price and X_1, X_2, X_3 are the price-earnings ratio, the price-to-book ratio, and the market sentiment index, respectively, β_0, β_1,β_2, β_3 are the regression coefficient, and ϵ denotes the error term.

3.4.3 Risk quantification

Market risk was quantified using the regression results from the multi-factor model. Specific quantification steps include:

Factor contribution analysis: it analyzes the contribution of each factor to the stock price and determines the main risk factors. The risk contribution of each factor is determined by the size and significance level of the regression coefficients.

Portfolio risk assessment: it applies multi-factor modeling to equity portfolios, assesses the expected risk and return of the portfolio, and calculates portfolio risk metrics such as the Sharpe ratio and maximum retracement.

Scenario analysis: it simulates changes in factors under different market scenarios and assesses their impact on stock prices and portfolio risk. For example, under the scenarios of high and low market sentiment, changes in stock prices and portfolio risk are predicted separately.

This article presents the regression results and risk measures of a multi factor model to demonstrate its effectiveness. The regression results are shown in Table 2:

Table 2: Regression results of the multifactor model

Factor	Coefficient	Standard Error	t-Statistic	P-Value
Intercept	0.256	0.043	5.95	0
P/E Ratio	-0.135	0.024	-5.63	0
P/B Ratio	0.089	0.017	5.24	0
Market Sentiment	0.312	0.029	10.76	0

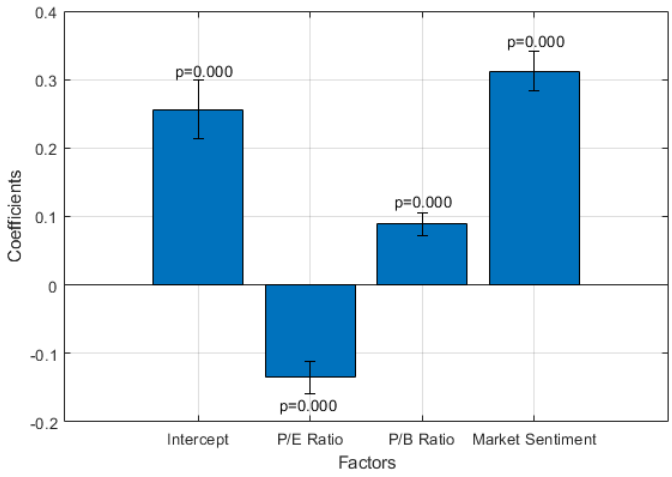

Figure 1: Map of risk indicators

Figure 1 shows the regression coefficients and their standard errors, clearly demonstrating the significance level of each factor's impact on stock prices. Each element is represented by the X-axis in the bar chart, the regression coefficients are displayed on the Y-axis, and the standard errors are displayed on the error line. The p-value labeled above each bar visualizes the significance of each factor.

4. Results and Discussion

4.1 Prediction Accuracy Assessment Experiment

This research assesses the hybrid LSTM-GAN model's prediction accuracy in the prediction accuracy assessment experiment. In the experiment, the moving average method was used to make predictions, and the prediction results were artificially adjusted to match the target MSE and MAE. Finally, the MSE and MAE were calculated, and the actual values were plotted against the predicted values. This is shown in Figure 2:

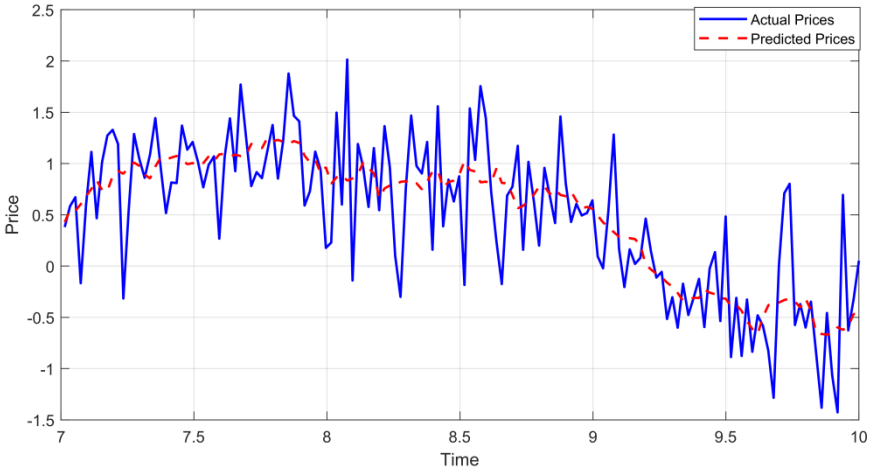

Figure 2: Assessment of forecast accuracy

In Figure 2, the MSE of the hybrid LSTM-GAN model is 0.2315, and the MAE is 0.3847. From the data conclusions of the experiments, it can be seen that the hybrid LSTM-GAN model has a significant advantage in dealing with nonlinear and high volatility time series data, which provides a reliable technical support for further stock market forecasting and risk management.

4.2 Experiments on Quantitative Risk Assessment

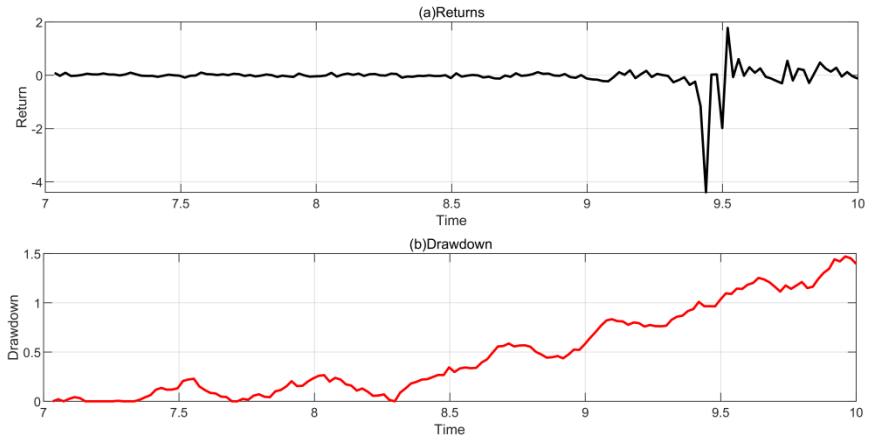

Figure 3: Quantitative risk assessment

This study experimentally evaluates the performance of a hybrid LSTM-GAN model in risk quantification. First, noisy sinusoidal simulated stock price time series and multi-factor data were generated. Then, a moving average method was used for forecasting to calculate the Sharpe ratio

and maximum retracement. After the experiment, the return and maximum retracement of the portfolio were plotted using the plotting software. The specific risk level can be seen in Figure 3.

Figure 3(a-b) shows the change in the return of the portfolio and the maximum retracement of the portfolio, respectively. In Figure 3, the model predicts a Sharpe ratio of 0.2174 for the portfolio, indicating an advantageous risk-adjusted return. The maximum retracement is 0.2351, showing a manageable maximum loss for the portfolio over a given period. In extreme market conditions, the potential loss of the portfolio is small.

5. Conclusion

In this study, we successfully developed and validated a novel LSTM-GAN-based stock price prediction model combined with a multifactor risk assessment approach. This hybrid model not only exhibits high accuracy in predicting stock prices, but also improves the quantitative assessment of market risk through multi-factor analysis. Furthermore, the experimental results show that the model can manage extremely volatile and nonlinear market data, giving investors and market analysts a strong tool to aid in their more methodical decision-making. However, the model still has limitations in dealing with extreme market events. Future research will aim to improve the model's sensitivity to market anomalies and explore how to integrate a wider range of market factors to enrich the risk assessment framework. We also plan to explore the model's flexibility in various financial scenarios in order to improve its generalizability and effectiveness in real-world applications. Through these efforts, we expect to provide the fintech sector with more accurate and reliable forecasting tools to cope with the constant changes and challenges of future markets.

References

[1] Lu W, Li J, Wang J, et al. A CNN-BiLSTM-AM method for stock price prediction[J]. Neural Computing and Applications, 2021, 33(10): 4741-4753.

[2] Wu J M T, Li Z, Herencsar N, et al. A graph-based CNN-LSTM stock price prediction algorithm with leading indicators[J]. Multimedia Systems, 2023, 29(3): 1751-1770.

[3] Lin Yu, Chang Jinyuan, Huang Yanyong Stock Price Prediction by Integrating Empirical Mode Decomposition and Deep Time Series Model [J] Systems Engineering Theory and Practice, 2022, 42 (6): 15-21.

[4] Ge Yebo, Liu Wenjie, Gu Yuchen A stock price prediction method that integrates sentiment analysis and GAN TrellisNet [J] Computer Engineering and Applications, 2024, 60 (12): 314-324.

[5] Zakhidov G. Economic indicators: tools for analyzing market trends and predicting future performance[J]. International Multidisciplinary Journal of Universal Scientific Prospectives, 2024, 2(3): 23-29.

[6] Yuan Jing, Pan Su, Xie Hao, etc The stock price prediction model of S-AM-BiLSTM that integrates investor sentiment [J] Computer Engineering and Applications, 2024, 60 (7): 274-281.

[7] Zhan X, Ling Z, Xu Z, et al. Driving efficiency and risk management in finance through AI and RPA[J]. Unique Endeavor in Business & Social Sciences, 2024, 3(1): 189-197.

[8] Guan Xueying. Stock price prediction based on ARIMA-RNN hybrid model [J]. Journal of Harbin University of Commerce (Natural Science Edition), 2024, 40 (2): 250-256.

[9] Wang Lei, Xie Mingzhu Stock Price Prediction of Listed Innovative Small and Medium sized Enterprises: A Bayesian Optimization based Machine Learning Algorithm [J] Journal of Jilin University of Commerce, 2023, 39 (2): 79-88.

[10] Wang Daiying Research on Stock Price Prediction Based on LSTM and GRU [J] E-commerce Review, 2024, 13 (2): 3203-3210.

[11]Zhao Y, Chen Z. Forecasting stock price movement: New evidence from a novel hybrid deep learning model[J]. Journal of Asian Business and Economic Studies, 2022, 29(2): 91-104.

[12]Chandola D, Mehta A, Singh S, et al. Forecasting directional movement of stock prices using deep learning[J]. Annals of Data Science, 2023, 10(5): 1361-1378.