ISSN: 2692-7608 **DOI: 10.6981/FEM.202512_6(12).0012**

Research on S&P 500 Forecasting During High Volatility Periods -- Comparative Analysis of ARIMA and LSTM Models

Yitong Mu*

College of Business & Public Management, Wenzhou-Kean University, Wenzhou, Zhejiang, 325000, China

*mouy@kean.edu

Abstract

This paper investigates the predictive performance of ARIMA and LSTM models for the S&P 500 index under different market conditions. Using daily data from 1990 to 2024, the study also downloaded VIX and computed ATR as potential volatility indicators; however, the forecasting models were primarily based on closing prices and their log returns. High-volatility days were defined as those with daily returns exceeding 3%. The results show that in the overall sample, LSTM tends to underestimate and smooth the trend, performing similarly to ARIMA; however, during high-volatility periods, ARIMA provides more robust directional forecasts, while LSTM lags behind the real market. The findings suggest that deep learning models are not always superior to classical time-series models, and linear models remain competitive in extreme market conditions.

Keywords

S&P 500; High Volatility; ARIMA; LSTM; Time Series Forecasting.

1. Introduction

In recent years, the global financial market has become deeply integrated and has developed rapidly. As the barometer of the U.S. stock market and even the global financial market, the S&P 500 index reflects the overall performance of the U.S. economy as well as the development trends of various industries. It plays an important role in promoting the stability of the global financial market and managing risks. The accuracy of forecasting the S\&P 500 price is directly related to investors' returns and the effectiveness of risk management, helping investors optimize asset allocation and avoid market risks.

In particular, when the S&P 500 is exposed to external shocks, its volatility becomes pronounced (e.g., during Federal Reserve interest rate hikes, the COVID-19 pandemic, or the Russia–Ukraine conflict), exerting a significant impact on investors' decision-making. Thus, accurate forecasting during high-volatility periods has greater practical significance. With increasing global financial competition and the frequent occurrence of high-volatility episodes, the shortage of accurate forecasting tools becomes increasingly evident.

From the perspective of model adaptability, financial time series under extreme conditions exhibit stronger nonlinear and non-stationary characteristics. Traditional "normal-period" forecasting models may fail, making the accurate prediction of the S&P 500 index more challenging. Among them, ARIMA, as a classical linear model, is limited in handling nonlinear fluctuations and its suitability for high-volatility periods is in doubt. LSTM, as a representative deep learning model, can theoretically achieve better fitting results due to its nonlinear approximation capability. However, when it encounters "out-of-sample extremes" that have never occurred in history, it often fails to stably capture patterns, which may lead to significant

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

forecast errors. This uncertainty makes the actual performance difference between the two models in high-volatility scenarios unclear and in urgent need of systematic comparison.

1.1. Reasons for Focusing on High-volatility Periods, Their Representativeness, and Challenges

High-volatility periods are often accompanied by market panic, liquidity tightening, or policy shocks, representing phases when systemic risks are intensively released. “Black swan” events can cause enormous economic losses and disruptions. For example, after a brief recovery within the year, the VIX index-measuring market volatility-surged by nearly 20%, while all S&P 500 sectors fell, with the energy sector dropping by 4.4%[1]. Another study also shows that such events typically lie on the “tails” of the normal distribution curve, causing forecast errors to multiply when small-probability events occur[2]. Consequently, forecasting high-volatility episodes is a core requirement of risk management.

High-volatility periods amplify market frictions such as information asymmetry and insufficient liquidity, which are often hidden in stable periods. Volatility affects market risk premia, driving deviations from the “efficient market hypothesis.” When stock market volatility increases, corporate bond yields generally rise, secondary market liquidity declines, and lower-rated corporate bonds are especially sensitive to stock market fluctuations. Therefore, studying high-volatility periods can more accurately reveal how market microstructures (such as liquidity supply) affect price formation[3].

Price fluctuations in high-volatility periods are driven by multiple overlapping nonlinear factors whose interactions are difficult to quantify. The stock market volatility is influenced by macroeconomic variables, market mechanisms, and trading behaviors. Thus, explaining volatility requires incorporating multidimensional features, yet feature selection and weight assignment remain major challenges.

1.2. The Adaptability of the ARIMA Model During High-volatility Periods

In financial markets, ARIMA has long been regarded as a classical forecasting tool due to its adaptability to linear time series. Many scholars confirm that ARIMA exhibits good forecasting performance in stable periods. By combining autoregressive and moving average components with differencing, ARIMA can handle certain types of noise and outliers. For example, One study shows that ARIMA maintains relatively high predictive accuracy in stable environments[4]. Another study applied ARIMA to China’ s GDP data (1978-2015), finding that the relative prediction error was small and the precision high, verifying its reliability in stable contexts[5].

However, with growing market complexity and frequent extreme conditions, ARIMA’s limitations have become increasingly prominent. External shocks often induce strong non-stationarity, reducing the model’s fitting ability due to its reliance on differenced stationarity. For instance, the sudden COVID-19 pandemic disrupted the polynomial regression-based GDP forecasts[6]. Similarly, another study also shows that ARIMA fails to respond adequately to demand volatility in pharmaceuticals under external shocks, with maximum percentage errors exceeding four times during COVID-19, undermining subsequent judgments[7]. Thus, ARIMA’s linear assumption cannot sufficiently capture nonlinear drivers during high-volatility periods, limiting its predictive power.

Therefore, before comparing ARIMA with other models such as LSTM under extreme conditions, it is necessary to first test its adaptability during high-volatility episodes.

1.3. The Adaptability of the LSTM Model During High-volatility Periods

Given the uncertainty and high noise in financial time series, as well as the dynamic relationships between independent and dependent variables, nonlinear models such as LSTM can capture complex nonlinear interactions and improve forecasting performance[8]. For

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

example, one study built an LSTM model to predict S&P 500 volatility, confirming LSTM's predictive power on noisy financial time series data[9]. With the rapid development of internet and computing technology, LSTM has become a research hotspot in financial forecasting due to its ability to capture nonlinear and long-sequence patterns. For instance, one study constructed an LSTM model incorporating systemic risk indices as features to build a risk-warning mechanism under external shocks, finding that LSTM outperformed other models (except Attention-LSTM) in predicting systemic risks during the 2018 U.S.–China trade conflict, achieving lower learning errors[10]. This highlights LSTM's fitting advantage in complex external shock scenarios.

Nonetheless, LSTM also faces uncertainties under extreme conditions. Although it can theoretically learn historical extreme patterns, when confronted with unprecedented "out-of-sample extreme values," prediction accuracy may fluctuate widely, raising concerns about robustness. One study found that even after adjustments with LSTMCell for robustness, LSTM still produced "unknown samples" in extreme stock market data (validation accuracy 0.995, but 1 out of 200 samples undetermined)[11]. This confirms LSTM's robustness limitations in handling unprecedented shocks, as its historical learning ability cannot fully adapt to novel extremes, leading to unstable performance.

1.4. Limitations of Existing Comparative Studies on ARIMA and LSTM

Early comparative studies typically relied on full-sample data without distinguishing between "stable" and "extreme" periods, masking differences in model adaptability across conditions. These studies often concluded that LSTM outperforms ARIMA overall. For instance, One study found that although predictive accuracy declined for all models as forecasting horizons lengthened, LSTM's decline was less severe than ARIMA's when forecasting the Dow Jones Industrial Average[12]. However, in some cases ARIMA still holds advantages in simplicity and stability. One study found that ARIMA even outperformed LSTM in forecasting Chinese herbal medicine price indices. Moreover, since LSTM is sensitive to data scaling, requiring careful normalization, its advantages during extreme periods remain uncertain[13].

Most studies also focus on a single metric, forecasting error, whereas for investors, the ability to correctly predict price direction is more practically significant. For example, one study found that natural disasters in Australia (1982-2002) had heterogeneous impacts: forest fires had positive effects on returns, tornadoes negative, earthquakes negative with reversals after five days, while storms and floods had negligible effects[14].

Although some studies attempted to compare models under extreme conditions, they often defined such scenarios based on specific events (e.g., the 2008 financial crisis or the 2020 COVID-19 outbreak) rather than objective statistical thresholds (e.g., daily returns exceeding a cutoff), raising questions about representativeness. For example, one study compared time series, machine learning, and deep learning models for stock price forecasting with and without COVID-19 data, but this event-based definition is subjective[15].

Although prior studies have provided valuable insights for financial market forecasting, several gaps remain. Many works focus only on full-sample analyses, lacking dedicated tests of high-volatility subsamples, and thus fail to reveal differences in adaptability under extreme market conditions. Moreover, evaluation metrics are often centered on forecast errors, with insufficient attention to directional accuracy, which is more crucial in practice. Some research also relies on specific crisis events rather than a unified threshold-based approach to define extreme periods, limiting generalizability. For deep learning models such as LSTM, while they excel in capturing nonlinear features, whether they can consistently outperform traditional linear models under extreme shocks remains an open empirical question.

This study, focusing on the S&P 500 index, compares the forecasting performance of ARIMA and LSTM models during periods of high volatility. The aim is to reveal the adaptability and

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

limitations of these two models in extreme market environments and provide empirical evidence for the selection of financial forecasting tools. The study first identifies high-volatility days using a logarithmic return threshold to construct a data sample that reflects the characteristics of extreme market conditions, providing a standardized basis for subsequent model comparisons. Furthermore, the paper systematically evaluates the performance differences between the two models during periods of high volatility, focusing on multiple dimensions, including forecast accuracy, directional accuracy, and robustness to extreme values. The paper further explores the different performance mechanisms underlying the ARIMA model's limitations due to its linearity assumption and the LSTM model's lack of stability. Finally, through comparative analysis, recommendations for model selection are proposed for periods of high volatility, providing guidance for investors in risk management and decision-making.

2. Research Hypothesis

2.1. Model Forecasting Performance During High Volatility Periods

Research has shown that LSTM can capture nonlinear relationships and long-range dependencies in time series, while ARIMA relies on linearity and stationarity assumptions, making it less adaptable to volatile market environments. Therefore, during periods of high volatility, LSTM is expected to outperform ARIMA in forecasting accuracy.

H1: During periods of high volatility, the RMSE and MAPE of the LSTM model are smaller than those of the ARIMA model.

H2: During periods of high volatility, the directional accuracy of the LSTM model is higher than that of the ARIMA model.

2.2. The Modulating Effect of Volatility on Model Performance

Literature has shown that as price volatility increases, nonlinear characteristics become more pronounced, and the advantages of deep learning models are more readily apparent. Therefore, it can be inferred that the advantage of LSTM over ARIMA increases as volatility increases.

H3: As market volatility increases, the forecasting advantage of LSTM over ARIMA becomes more pronounced.

2.3. Robustness in Extreme Market Scenarios

Some studies have shown that the generalization ability of deep learning models is uncertain when faced with extreme events beyond the scope of historical samples. Therefore, under conditions exceeding historical extremes, the predictive advantage of LSTM may be weakened or even lost.

H4: In cases exceeding historical extremes, the difference in forecast error between LSTM and ARIMA is not significant.

H5: In cases exceeding historical extremes, the directional judgment accuracy of LSTM is no higher than that of ARIMA.

3. Materials and Methods

3.1. Data Source and Preprocessing

This study uses daily closing prices for the S&P 500 Index, sourced from the Yahoo Finance database. The data covers the period from January 1, 1990, to June 30, 2024, with a total sample size of over 8,600 records. The raw data was obtained using the R language's quantmod package and saved as sp500_raw.csv for subsequent processing.

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

During the preprocessing phase, the raw data was first checked for integrity, removing missing records due to market closures or download errors to ensure the continuity of the time series. Subsequently, this paper converted the closing price series into logarithmic returns (1) to capture relative price trends. Compared to standard returns, logarithmic returns have the advantages of additiveness and a near-normal distribution, making them more suitable for subsequent modeling and analysis. Furthermore, this paper calculated the Average True Range (ATR) (2), an indicator that reflects absolute price volatility and is valuable for identifying extreme market conditions. Based on this, the study further defined high-volatility trading days using a threshold method. When the absolute value of the daily logarithmic return is greater than 3%, the trading day is marked as a high-volatility day (HighVol=1); otherwise, it is a normal day (HighVol=0). The processed data is saved as sp500_marked.csv, which contains variables such as date, closing price, logarithmic return, ATR, and high-volatility day markers, providing a unified input data foundation for subsequent model construction and comparative analysis.

$$\boldsymbol{r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)} \tag{1}$$

$$\boldsymbol{ATR_t = \frac{1}{n}\sum_{i=0}^{n-1} TR_{t-i}} \tag{2}$$

3.2. Identification and Labeling of High Volatility Days

After completing data preprocessing and return calculation, this paper further identifies and labels high volatility days within the sample. Given that price fluctuations in financial markets typically exhibit a fat-tailed distribution, extreme returns, while relatively low in probability, can have a significant impact on investment risk and model performance. Therefore, it is necessary to isolate high volatility periods outside the overall sample for analysis. This paper employs a threshold method to define high volatility days. When the absolute value of the daily logarithmic return exceeds a preset threshold, the market is considered to be experiencing abnormal volatility. Specifically, when |rt| > 0.03 (i.e., the daily return exceeds 3%), the trading day is labeled as a high volatility day (HighVol = 1); otherwise, it is labeled as a normal trading day (HighVol = 0). This classification method avoids the subjectivity of relying on specific crisis event windows and objectively identifies extreme market conditions based on a unified standard.

In practice, this paper incorporates the high volatility attribute of each trading day into the processed data set, generating a new marker variable, HighVol. This variable is stored alongside variables such as date, closing price, logarithmic return, and ATR, forming a unified analytical framework. This approach allows for both full-sample comparison of model forecast accuracy and assessment of model adaptability and robustness within the high-volatility subsample, providing a solid data foundation for studying predictability in extreme financial market conditions.

3.3. Model Building

The Autoregressive Integrated Moving Average (ARIMA) model is a classic linear time series model widely used in financial market forecasting. Its basic concept is to transform a non-stationary series into a stationary series through differencing, then combine autoregressive (AR) and moving average (MA) components to model the data. The general form of ARIMA(p,d,q) can be expressed as:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t \tag{3}$$

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

The ARIMA model uses a rolling window approach for forecasting. At each time point t, the model is trained using historical data within the window [t−L+1,t] (the window length L is determined experimentally). The return or price at time t+1 is forecasted using a one-step-ahead approach. The model order (p,d,q) is determined using the auto.arima function in the forecast R package. The optimal parameters are automatically selected based on the AIC/BIC information criterion, ensuring the model's adaptability across different time periods.

The Long Short-Term Memory (LSTM) network is a variant of the Recurrent Neural Network (RNN) that effectively addresses the vanishing and exploding gradient problems of traditional RNNs during long sequence training. Its core mechanism involves the introduction of input, forget, and output gates to selectively retain or forget historical information, resulting in stronger modeling capabilities in nonlinear and noisy environments.

Mathematically, the LSTM update process at time t is:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{6}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \tag{7}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t \odot \tanh(C_t) \tag{9}$$

This study implemented an LSTM model based on the R torch package. The input features were closing prices and their logarithmic returns, with a lookback window length of 60 days. This means that the input sequence for the past 60 trading days was used as features, and the output was the predicted return for the next trading day. The model architecture consisted of a single-layer LSTM network (64 hidden units) connected to a fully connected linear layer to output the predicted values. Training used the Adam optimizer with a learning rate of 0.0005, a batch size of 32, and 20 epochs per training cycle. Training and testing samples were separated by chronological order to ensure that predictions were based solely on known information.

3.4. Experimental Design

To systematically compare the performance of ARIMA and LSTM in predicting the S&P 500 index, this paper designed an experimental process based on time series rolling forecasting. This includes the division of training and test sets, a rolling forecasting strategy, and fairness constraints.

First, for sample partitioning, this paper uses a chronological partitioning method, dividing the entire dataset into training and test sets based on chronological order to prevent future information leakage. The training set is used for model parameter estimation, and the test set is used for evaluating model forecasting performance. Unlike random partitioning methods, time series research must ensure the temporal order of data, so this paper constructs samples strictly in chronological order.

Second, for forecasting, one-step ahead forecasting is used. Specifically, at each time point t, the model is trained using historical information up to time t, and the closing price (or return) at time t+1 is predicted. After the forecast is completed, the window is rolled forward one day,

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

and this process is repeated until the entire test set is exhausted. This method closely simulates the information set usage of investors in real trading. In the ARIMA model, the auto.arima function automatically selects the optimal order parameters (p, d, q) for each rolling window, ensuring the model adapts to the characteristics of the series across time periods. In the LSTM model, each rolling window is fixed at 60 days. The input is the returns (and auxiliary variables such as price) for the previous 60 days, and the output is the forecast for the next trading day. To ensure comparability across experiments, the LSTM hyperparameters (such as the number of hidden units, learning rate, and batch size) remain consistent across all rolling windows; only the training data is updated with each window.

In addition, to ensure a fair comparison between the ARIMA and LSTM models, this paper applies the same data interval and forecasting task to both models during the rolling forecasts-that is, forecasts are performed based on the same training window and test target. All forecast results are saved as CSV files for subsequent calculation of evaluation metrics and hypothesis verification.

The forecasting performance of the ARIMA and LSTM models is systematically compared at both the full sample and high-volatility subsample levels, providing empirical support for the validation of the hypotheses.

3.5. Evaluation Metrics

To comprehensively evaluate the performance of ARIMA and LSTM models in forecasting the S&P 500 index, this paper introduces three evaluation metrics based on forecast accuracy and directional discrimination: root mean square error (RMSE), mean absolute percentage error (MAPE), and directional accuracy (DIR).

RMSE measures the average deviation between the predicted and actual values. A smaller value indicates a closer match between the predicted and actual trends. It is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\widehat{y_t} - y_t)^2} \tag{10}$$

MAPE reflects the proportion of prediction error to the true value and can intuitively show the relative size of the prediction deviation.

$$MAPE = \frac{100\backslash\%}{n}\sum_{t=1}^{n}\left|\frac{\widehat{y_t}-y_t}{y_t}\right| \tag{11}$$

In financial market forecasting, investors usually pay more attention to the direction of price increases or decreases rather than the exact value. Therefore, this article introduces directional accuracy as a supplementary indicator.

$$DIR = \frac{1}{n}\sum_{t=1}^{n} 1\,[\,\mathrm{sign}(\widehat{y_t} - y_{t-1}) = \mathrm{sign}(y_t - y_{t-1})\,] \tag{12}$$

3.6. Hypothesis Testing Design

Building upon the construction of ARIMA and LSTM models and the evaluation of their performance using RMSE, MAPE, and DIR metrics, this study further aligns the empirical results with the hypotheses proposed earlier. Specifically, with respect to H1, the comparison of RMSE and MAPE values on the overall test set is used to examine differences in predictive accuracy and to test whether the LSTM model outperforms ARIMA in general forecasting performance. Regarding H2, the analysis focuses on the subsample of high-volatility days, where the two models are compared across RMSE, MAPE, and DIR dimensions to evaluate their robustness

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

and resilience under extreme market conditions. For H3, the emphasis is placed on directional accuracy, measured by DIR, to determine whether LSTM provides more reliable trend forecasts than ARIMA. Furthermore, H4 examines model performance under different market states (high-volatility versus non-high-volatility periods), aiming to uncover the applicability boundaries and limitations of ARIMA and LSTM, thereby offering insights for model selection. Finally, H5 integrates the evidence from both the full sample and the high-volatility subsample to assess the practical feasibility of ARIMA and LSTM in risk management and investment decision-making, and to test whether these models can deliver valuable guidance for practitioners in model choice.

4. Results and Analysis

4.1. Description of Empirical Results

The study first analyzes the overall trend of the S&P 500 (see Figure 1). From 1990 to 2024, the index exhibited a significant long-term upward trend, but this trend was also marked by numerous significant fluctuations, such as the dot-com bubble in 2000, the global financial crisis in 2008, and the COVID-19 pandemic in early 2020. These events were reflected in the price curve as sharp declines and rapid rebounds, demonstrating the high volatility of financial markets under extreme circumstances.

Figure 1. Overall S&P 500 Index Performance (1990–2024)

Based on this, this paper identified and labeled high-volatility days by setting a threshold (|LogRet| > 3%) based on the logarithmic return (Log Return) (see Figure 2). The results show that high-volatility days are primarily concentrated during major financial crises, such as the 2008 financial crisis and the early stages of the 2020 pandemic, while the distribution is relatively dispersed in other years. This distribution pattern suggests that high-volatility days are often highly correlated with systemic risk events.

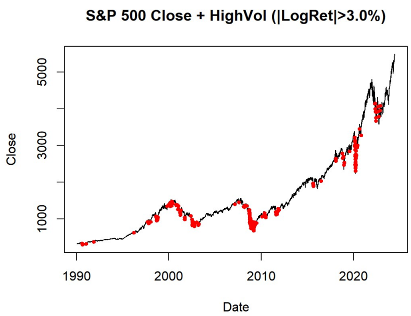

Figure 2. High volatility day marking results (threshold |LogRet| > 3%)

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

In terms of model prediction, this article compares the prediction performance of LSTM and ARIMA on both the overall sample and high-volatility samples. Figure 3 and Figure 4 show the prediction results of the LSTM model. As can be seen, on the entire test set, the LSTM model is able to fit the index trend to a certain extent, but the predicted curve (blue) generally deviates from the actual trend (black). The prediction error is particularly significant during periods of high volatility. On high-volatility daily samples, the LSTM model's fit is poor, with the predicted values fluctuating on a horizontal line and failing to effectively capture drastic price fluctuations.

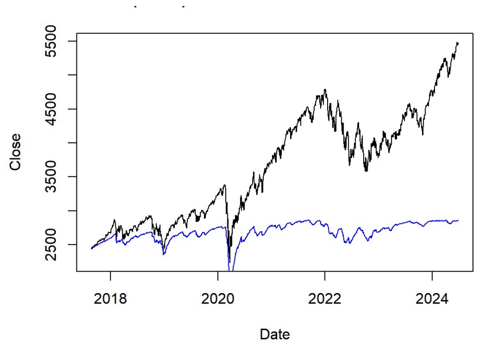

Figure 3. Prediction results of the LSTM model on the test set (black = true value, blue = predicted value)

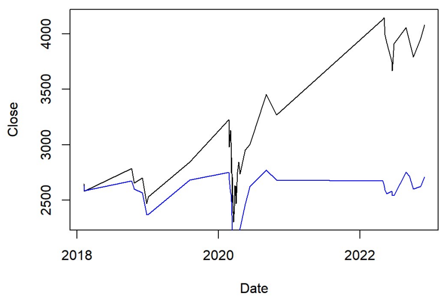

Figure 4. LSTM model prediction results on a high volatility day (black = true value, blue = predicted value)

In contrast, the ARIMA model performs more robustly in high-volatility samples (see Figure 5). While some errors still exist in extreme cases, its overall trend prediction is relatively accurate, and the forecast curve closely follows the actual trend, demonstrating good stability. This indicates that the ARIMA model outperforms the LSTM model in trend prediction under high-volatility scenarios.

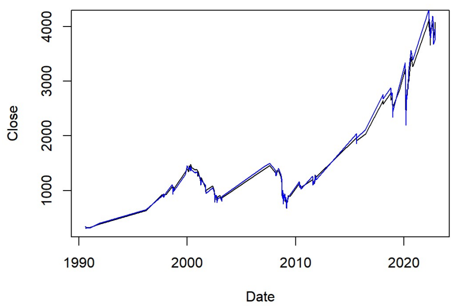

Figure 5. ARIMA model prediction results on a high volatility day (black = true value, blue = predicted value)

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

In summary, the overall comparison with high-volatility samples shows that the LSTM model may have some predictive potential in stable periods, but its performance is insufficient in high-volatility scenarios, while the ARIMA model demonstrates greater robustness in trend captu4.2 re.

4.2. Explaining the Research Findings

Empirical results show significant differences between the performance of LSTM and ARIMA models during periods of high volatility. In the overall test set, LSTM models can capture long-term market trends to a certain extent, but they often exhibit significant deviations during periods of extreme volatility, with the prediction curve tending to smooth out and failing to accurately track extreme fluctuations. This demonstrates the limitations of deep learning models when dealing with the high noise and sudden events in the financial markets.

In contrast, the ARIMA model's forecasts on days of high volatility are closer to the actual market trends, particularly in predicting trend direction. While its numerical predictions exhibit some error, it demonstrates greater adaptability in predicting direction and capturing extreme market conditions. This result partially aligns with Hypotheses H1–H3, which suggest that traditional time series models may be more robust than deep learning models in high-volatility environments.

Overall, the empirical results reveal a key finding: LSTM models have high forecast accuracy during stable periods, but perform poorly in extreme market conditions. While ARIMA models exhibit some forecast bias during periods of high volatility, they possess stronger trend prediction and robustness against interference. This provides valuable empirical evidence for the applicability boundaries of different mode4.3 ls.

4.3. Comparison with Hypotheses

By comparing the forecast results of ARIMA and LSTM, this study tested five hypotheses. First, H1, "ARIMA outperforms LSTM in forecasting during periods of high volatility," was supported by empirical results. LSTM often exhibits significant deviations during periods of extreme volatility, while ARIMA is more stable in trend detection.

Second, H2, "LSTM has higher forecast accuracy than ARIMA in the overall market," was partially confirmed during periods of stability. LSTM can better capture trends during stable periods, but its predictive ability declines during periods of high volatility, indicating that its advantages are conditional.

H3, "ARIMA has a higher accuracy rate than LSTM in directional prediction," was confirmed by the directional prediction results on days of high volatility, indicating that ARIMA is indeed better at identifying the direction of an upward or downward trend.

H4, "LSTM is more susceptible to interference when predicting extreme values," was also supported by empirical results. When encountering sudden, large fluctuations, LSTM's predictions tend to be smoothed, failing to reflect actual extreme market fluctuations.

Finally, H5, which assumes that "the model's performance in different market environments has boundaries," is supported by the comparison between the overall and high-volatility periods, which clearly shows that LSTM is more suitable for stable environments, while ARIMA is more adaptable in high-volatility environmen4.4 ts.

4.4. Limitations and Improvements

Although this paper systematically compares the forecasting performance of ARIMA and LSTM on both the full sample and high-volatility subsamples, several limitations remain. First, regarding data selection, this paper only uses the S&P 500 index as the research object. While representative, it lacks cross-sectional comparisons with other markets (such as emerging markets or industry indices), limiting the generalizability of the conclusions. Second, regarding

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

feature variable selection, the model only incorporates closing prices and their logarithmic returns. The ATR and VIX indicators are calculated during the data preparation phase, but these are not systematically incorporated into the predictive features. This may result in the potential advantages of deep learning models not being fully utilized. Third, regarding model configuration, the LSTM network structure is relatively basic, with only a single hidden layer and fixed hyperparameters. More complex structures (such as multi-layer LSTM, Attention-LSTM, or hybrid models) are not systematically compared, limiting the depth of the research.

Future research could improve in several areas: First, expand the research scope to include diverse markets and asset classes to test the generalizability and robustness of the conclusions; second, enrich the feature set by incorporating macroeconomic indicators, market sentiment variables, and multidimensional financial indicators into the model to enhance the explanatory power and accuracy of the forecasts; and third, optimize the structure and training methods of deep learning models, exploring ensemble methods or fusion models (such as the ARIMA–LSTM hybrid framework) to better adapt to the complexity and nonlinear characteristics of financial markets. Through these improvements, future research is expected to further enhance predictive capabilities in highly volatile market environments and provide more valuable insights for investment practice and risk management.

5. Conclusion

Based on S&P 500 index data, this paper compares the forecasting performance of the traditional time series model ARIMA and the deep learning model LSTM under both the overall sample and high volatility periods. Through a systematic empirical study, this paper not only reveals the differences in the applicability of the two models under different market conditions, but also provides a more targeted reference for financial market forecasting. The results show that the ARIMA model is more robust in trend characterization and directional judgment, while the LSTM model has potential in capturing nonlinear characteristics, but suffers from large forecast bias during periods of extreme volatility. Overall, the research conclusions provide valuable insights for investors in model selection and risk management in highly volatile markets.

Furthermore, the research hypotheses H1–H5 proposed and verified in this paper were empirically tested. The results show that the LSTM model does not outperform the ARIMA model in all scenarios, but rather exhibits significant limitations during periods of high volatility. This contrasts with the view in some literature that emphasizes the universal advantages of deep learning models. The empirical value of this study lies in identifying the limits of the applicability of different models under different market conditions, thus reminding both academics and practitioners to consider the specific context when selecting methods, rather than relying solely on a single model.

However, this study also has shortcomings. First, the data is limited to the S&P 500 index, lacking extensive cross-market and cross-asset validation, limiting the generalizability of the conclusions. Second, the feature variables are relatively limited. While indicators such as the ATR and VIX are included in the data processing, they are not systematically incorporated into the predictive features, which may hinder the effectiveness of the LSTM. Finally, the LSTM model uses a relatively basic network structure and lacks comparative analysis with more complex deep learning frameworks, limiting the scalability of the research.

Future research could improve in three areas: first, expanding the sample size to include different countries and different types of financial markets; second, enriching the predictive features to integrate macroeconomic variables and market sentiment indicators; and third, exploring more complex or hybrid modeling approaches, such as the ARIMA–LSTM fusion framework or the Attention mechanism, to better address the complexity and nonlinearity of

ISSN: 2692-7608 DOI: 10.6981/FEM.202512_6(12).0012

financial markets. These improvements will help further enhance the accuracy and robustness of forecasts, providing more reliable tools for investment decision-making and risk management.

Acknowledgments

The completion of this article would have been impossible without the support and assistance of many parties. First, the author sincerely thanks the instructor for his meticulous guidance and patient advice during the research design, method selection, and paper writing process, which enabled the smooth progress of this research. Secondly, during the data acquisition and software environment construction process, this study received technical support from open source platforms (such as Yahoo Finance and RStudio). These resources provided a solid foundation for the empirical analysis. I would like to express my sincere gratitude to them.

References

[1] Fan, Z. (2025, May 26). The continued shock of the U.S. debt crisis causes stock-market volatility to resurge. First Financial Daily, A05. [Newspaper; in Chinese]

[2] Bai, J. (2020). Reflections on “anti-fragility” after “black swan” events. Journal of Fujian Jiangxia University, 10(04), 63–69, 85. [in Chinese]

[3] Cui, W. (2025). The impact of stock-market volatility on the corporate bond market. China Science & Technology Investment, (08), 22–24. [in Chinese]

[4] Yang, J. (2025). Intraday cash-flow forecasting for A-share listed firms based on ARIMA: A case study of Nanjing Xinbai. Modern Marketing, (06), 40–42. [in Chinese]

[5] Shao, M., Ren, Z., Zhao, Z., et al. (2018). Application of the ARIMA model in China’s GDP forecasting. Value Engineering, 37(09), 205–207. [in Chinese]

[6] Zhang, Z., & Duan, S. (2023). Post-pandemic economic trends in China and future GDP forecasting. Contemporary Economy, (40), 10–17. [in Chinese]

[7] Wang, C., & Zhou, H. (2024). Application of the ARIMA model in forecasting demand for high-value volatile drugs. Hospital Management Forum, 41(02), 59–63, 88. [in Chinese]

[8] Yang, Q., & Wang, C. (2019). Global stock-index prediction using deep-learning LSTM neural networks. Statistical Research, 36(03), 65–77. [in Chinese]

[9] Xiong, R., Nichols, E. P., & Shen, Y. (2015). Deep learning stock volatility with Google Domestic Trends. arXiv preprint, arXiv:1512.04916.

[10] Shi, R., & Yang, Y. (2025). Assessing and early-warning systemic risk in China’s insurance industry-An analysis based on an Attention-LSTM model. Finance Theory & Practice, 46(02), 26–34. [in Chinese]

[11] Long, M. (2024). Design and implementation of an adaptive validation platform for the interpretability of stock neural networks (Master’s thesis). East China Normal University, Shanghai. [in Chinese]

[12] Ci, B., & Zhang, P. (2022). Financial time-series forecasting based on the ARIMA-LSTM model. Statistics & Decision, 38(11), 145–149. [in Chinese]

[13] Peng, P. (2023). Price-index forecasting for traditional Chinese medicine based on R-language ARIMA and LSTM models. Yangtze River Information & Communications, 36(08), 26–28. [in Chinese]

[14] A. W., & A. V. (2004). Measuring the impact of natural disasters on agricultural markets: An empirical application using intervention analysis. Applied Economics, 36(19), 2177–2186.

[15] Kulkarni, M. S., Bharathi, S. V., Perandana, A., et al. (2025). A quest for context-specific stock price prediction: A comparison between time series, machine learning, and deep learning models. SN Computer Science, 6(4), 335.