Academic Education Publishing House

Stock Index Forecasting of CSI 300 Based on a Hybrid ARIMA-LSTM Model

Yunhao Lin
Department of Applied Mathematics with Economics, Jinan University, Guangzhou, Guangdong, China
**Corresponding Author*

Abstract: As a key barometer of China's capital markets, the CSI 300 Index holds significant value for investment decision-making and risk management. In addressing the limitations of conventional forecasting models in processing nonlinear and non-stationary financial time series data, this study proposes a hybrid prediction framework that integrates Autoregressive Integrated Moving Average (ARIMA), Variational Mode Decomposition (VMD), and Long Short-Term Memory (LSTM) networks. The proposed model employs a multi-stage strategy involving signal decomposition, feature modelling, and result integration to comprehensively capture both the linear trends and nonlinear fluctuations in time series data, thereby enhancing forecasting accuracy and generalization capability. Based on historical CSI 300 data, the study compares forecasting performance across different time horizons. Results indicate that, compared with standalone ARIMA and LSTM models, the hybrid model demonstrates superior stability and adaptability in short-, medium-, and long-term forecasting. Notably, in long-term scenarios, the integration of multi-scale features effectively mitigates error accumulation, confirming the robustness of the hybrid approach under complex market conditions.

Keywords: CSI 300 Index; Time Series Forecasting; ARIMA Model; Variational Mode Decomposition (VMD); Long Short-Term Memory (LSTM); Hybrid Prediction Model

1. Introduction

In recent years, hybrid models that combine statistical time series modeling with deep learning techniques-such as ARIMA-LSTM-have attracted increasing attention from researchers and have demonstrated superior performance in financial market forecasting.

Research on time series forecasting generally falls into three categories: traditional statistical approaches, machine learning methods, and hybrid models that integrate both. Among these, the Autoregressive Integrated Moving Average (ARIMA) model is the most widely used in financial forecasting due to its interpretability and effectiveness in short-term predictions. Wu Yuxia and Wen Xin (2016) applied the ARIMA model to forecast stock prices in the ChiNext market and found it to be effective in static prediction tasks[4]. Similarly, Adebiyi et al. (2014) investigated stock prices on the New York Stock Exchange and the Nigerian Stock Exchange using ARIMA and confirmed its applicability for short-term forecasting[7]. However, as a linear model, ARIMA struggles to capture the nonlinear characteristics inherent in financial markets. To address this limitation, researchers have attempted to combine ARIMA with nonlinear models to improve forecasting accuracy. For example, Lei Kewei and Chen Ying (2007) employed an ARIMA-BP neural network model to predict inbound tourist arrivals in China, showing that the hybrid model effectively integrated linear and nonlinear information[3].

Deep learning methods have gained momentum in the field of time series prediction over the past few years, with long short-term memory (LSTM) networks receiving particular attention for their ability to capture both long-term and short-term dependencies. Junran Wu et al. (2021) transformed financial time series into graph structures and integrated them with deep learning techniques to forecast the CSI 300 Index, achieving promising results[11]. Additionally, Zhang Lei et al. (2021) demonstrated that incorporating market sentiment indicators into LSTM models could significantly enhance prediction accuracy in exchange rate forecasting[5].

Nevertheless, standalone LSTM models also have limitations, such as susceptibility to local optima and sensitivity to parameter settings. To combine the linear modeling strengths of ARIMA with the nonlinear fitting capabilities of LSTM, researchers have proposed hybrid ARIMA-LSTM models. Building upon these developments, this study introduces Variational Mode Decomposition (VMD) into the existing ARIMA-LSTM framework to construct an ARIMA-VMD-LSTM composite model, aiming to further enhance the accuracy and stability of financial time series forecasting[1].

2.Model Construction

2.1ARIMA Model

The ARIMA (Auto Regressive Integrated Moving Average) model is a classical time series forecasting method proposed by Box and Jenkins. The model combines three components-Autoregression (AR), Integration (I) through differencing, and Moving Average (MA)-to model and predict time series data. The ARIMA model is conventionally denoted as ARIMA(p, d, q), where:

p: the order of the autoregressive part (AR), representing the linear relationship between the current value and its p lagged values

d: The number of difference steps (I) employed in the process of transforming a non-stationary series into a stationary one.

q: The order of the moving average part (MA) is indicative of the linear relationship between the current error and past forecast errors.

The general mathematical expression of the ARIMA model is:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t \qquad (1)$$

Where:

B is the lag operator, defined as $By_t = y_{\{t-1\}}$

$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the autoregressive polynomial

$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ is the moving average polynomial

ε_t denotes the white noise series

The ARIMA model is constructed through the following steps for the purpose of financial time series forecasting:

Initially, the series is subjected to a stationarity test using the Augmented Dickey-Fuller (ADF) test. In the event of non-stationarity, the application of differencing is employed to ascertain the appropriate value of d. Subsequently, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are utilized to preliminarily identify suitable values for p and q. Thereafter, the Akaike Information Criterion (AIC) is employed to select the optimal order. The estimation of model parameters is then conducted through the utilization of the Maximum Likelihood Estimation (MLE) method. Finally, the residuals are subjected to the Ljung-Box test to verify that they resemble white noise. Once the model has passed these checks, it can be used for forecasting.

However, the ARIMA model, while effective for linear time series, struggles with the nonlinear and volatile nature of financial data. It cannot fully capture complex patterns such as regime shifts and long-range dependencies, which are common in financial markets. As a result, its forecasting accuracy may decline in such contexts.

To overcome these limitations, this study integrates ARIMA with the Long Short-Term Memory (LSTM) network, which is well-suited for modeling nonlinear relationships and temporal dependencies. The following section details the construction of the LSTM model[6].

2.2 LSTM Model

The Long Short-Term Memory (LSTM) network is a specialised form of Recurrent Neural Network (RNN) proposed by Hochreiter and Schmidhuber. It was designed to address the vanishing or exploding gradient problems encountered by traditional RNNs when processing long sequences. In comparison with standard RNNs, LSTM networks incorporate a sophisticated gating mechanism that enables them to effectively capture long-term dependencies in time series data, thereby demonstrating strong performance in financial forecasting tasks[8].

The fundamental unit of LSTM is the memory cell, which is comprised of three gating mechanisms: the Forget Gate, the Input Gate, and the Output Gate. The components function collectively to determine which information is retained or discarded.

The subsequent expression provides a formal representation of the computations conducted by the LSTM cell at time step t:

The forget gate decides which information to discard from the cell state:

$$f_t = \sigma\left(W_f \cdot \ \left[h_{\{t-1\}}, x_t\right] + b_f\right) \qquad (2)$$

In this equation, f_t denotes the output of the forget gate, whilst σ represents the sigmoid activation function. The weight matrix and bias term of the forget gate are denoted by W_f and b_f, respectively. The previous hidden state is represented by $h_{\{t-1\}}$, and $x_t$denotes the current input.

The function of the input gate is to determine which new information is added to the cell state:

$$i_t = \sigma\left(W_i \cdot \ \left[h_{\{t-1\}}, x_t\right] + b_i\right) \qquad (3)$$

$$\widetilde{C}_t = \tanh\left(W_C \cdot \ \left[h_{\{t-1\}}, x_t\right] + b_C\right) \qquad (4)$$

Here, i_t controls the update level of each component, and $\widetilde{C}_t$represents the candidate cell state.

The cell state is updated as follows:

$$C_t = f_t \ C_{\{t-1\}} + i_t \ \widetilde{C}_t \qquad (5)$$

Where denotes the Hadamard product (element-wise multiplication), enabling selective forgetting and addition of memory.

The output gate decides which parts of the cell state are output:

$$\mathrm{o_t} = \sigma\left(\mathrm{W_o} \cdot \ \left[h_{\{\mathrm{t-1}\}}, \mathrm{x_t}\right] + \mathrm{b_o}\right) \qquad (6)$$

$$h_t = o_t \tanh\left(C_t\right) \qquad (7)$$

The output $h_t$serves as the output of the current time step and is passed to the next time step.

2.3 VMD Signal Decomposition

Variational Mode Decomposition (VMD) was proposed by Dragomiretskiy and Zosso as a novel adaptive signal processing method. Compared with traditional Empirical Mode Decomposition (EMD), VMD decomposes signals by formulating and solving a variational problem, supported by a solid mathematical foundation. It effectively overcomes the issues of mode mixing and end effects commonly seen in EMD. In the context of financial time series analysis, VMD has been demonstrated to facilitate the decomposition of non-stationary and nonlinear price sequences into a series of relatively stationary Intrinsic Mode Functions (IMFs). This approach has been shown to provide more discriminative feature representations for the subsequent forecasting models. The following is a detailed description of the decomposition process:

Set the number of mode components K and the penalty parameter α, initialize each mode component $\left\{u_k^{1(t)}\right\}$ and its central frequency $\{\omega_k^1\}$, and initialize the Lagrange multiplier $\lambda^{1(t)}$. VMD obtains IMF components by solving the following optimization problem:

$$\min_{\{u_k\},\{\omega_k\}} \left\{\sum_{k=1}^{K} // \partial_t \left[\left(\delta(t) + \frac{j}{\pi t}\right) * \ u_k(t)\right] e^{-j\omega_k t} //_2^2\right\} \qquad (8)$$

Subject to the constraint: ∑{k=1}$^{\mathrm{K}}$$\mathrm{u}_{\mathrm{k(t)}}$= f(t)

The Alternating Direction Method of Multipliers (ADMM) has been utilised for the purpose of iterative optimization:

Update of mode components:

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha\left(\omega - \omega_k^n\right)^2} \qquad (9)$$

Update of central frequency:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \mid \hat{u}_k^{n+1}(\omega) \mid {}^2 d\omega}{\int_0^\infty \mid \hat{u}_k^{n+1}(\omega) \mid {}^2 d\omega} \qquad (10)$$

Update of Lagrange multiplier:

$$\hat{\lambda}^{\{n+1\}(\omega)} = \hat{\lambda}^{n(\omega)} + \tau \left[\hat{f}(\omega) - \sum_{\{k=1\}}^{K} \hat{\mathrm{u}}_k^{\{n+1\}(\omega)}\right] (11)$$

The iteration stops when the following condition is met:

$$\sum_{k=1}^{K} \frac{// u_k^{n+1} - u_k^n //_2^2}{// u_k^n //_2^2} < \epsilon \qquad (12)$$

where ε is a preset convergence threshold.

The values of K and α play a critical role in the quality of decomposition. The number of modes K is typically chosen based on prior knowledge of the signal's frequency content or determined empirically through spectral analysis. Alternatively, a range of K values can be evaluated using cross-validation to identify the setting that yields the best reconstruction performance or lowest forecasting error. The penalty parameter α, which controls the bandwidth of each mode, is similarly selected by balancing mode smoothness and separation. A larger α enforces narrower bandwidths, while a smaller α allows more overlap. In practice, both parameters can be tuned using grid search or heuristic optimization methods to achieve optimal decomposition results.

Through the above steps, the original signal is adaptively decomposed into K relatively stationary IMF components, which serve as effective features for subsequent time series modeling[9].

2.4 ARIMA-LSTM Model Construction

The hybrid ARIMA-VMD-LSTM prediction model proposed in this study adopts a three-stage "decomposition–prediction–integration" architecture. First, the original financial time series is decomposed by VMD into several relatively stationary IMF components and a residual term. Then, ARIMA models and LSTM networks are used respectively to forecast components with different characteristics. Finally, the prediction results are reconstructed through a nonlinear integration strategy. This framework fully exploits the strengths of each method: VMD provides high-quality signal decomposition, ARIMA captures linear trends, and LSTM models nonlinear relationships, thus achieving complementary advantages[10].

The model is mathematically formulated as follows:

$$\hat{y}_t = F_{ensemble}\left(F_{ARIMA(r_t)}, \{F_{LSTM(imf_k)}\}_{\{k=1\}}^{K}\right)(13)$$

where $\hat{y}_t$ is the predicted value at time t, $\mathscr{F}_{\mathrm{ARIMA}}(r_t)$ is the prediction result from the ARIMA model applied to the residual component, and $\mathscr{F}_{\mathrm{LSTM}}(\mathrm{imf}_k)$ denotes the LSTM-based prediction for the k-th IMF component.

K is the total number of IMF components produced by VMD.

The integration function $\mathscr{F}_{\mathrm{ensemble}}$ is implemented as a weighted average of the individual predictions, with weights empirically assigned or optimized based on validation performance. This aggregation strategy ensures that each component contributes proportionally to the final output according to its predictive accuracy or importance.

The schematic diagram of the combined model is presented below Figure 1:

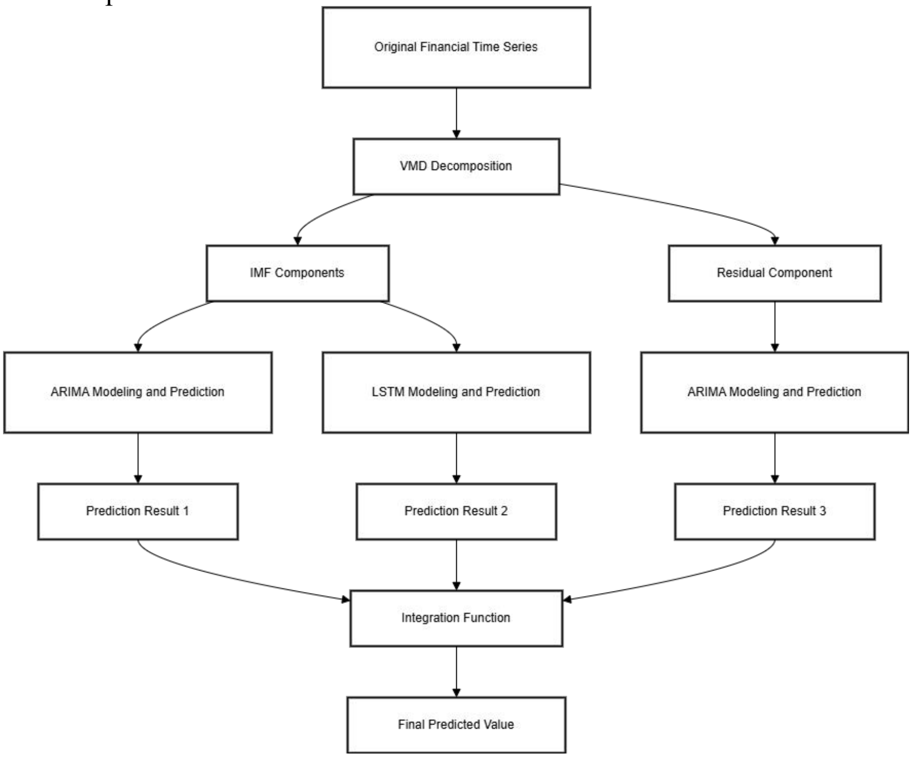

Figure 1. Schematic Diagram of the ARIMA-LSTM Hybrid Model

3. Empirical Analysis

3.1 Experimental Data and Preprocessing

The present study has focused on the daily closing prices of the CSI 300 Index as the primary object of research. The historical data is divided into three categories according to the time frame over which it is predicted: short-term, medium-term and long-term, corresponding to time spans of 28, 68, and 126 trading days, respectively. All intervals end on March 14, 2025, with the corresponding start dates being February 5, 2025; December 2, 2024; and September 2, 2024. For each interval, the preceding three trading days are utilized to assess the model's short-, medium-, and long-term predictive efficacy, with the residual portion constituting the test input.

To ensure the breadth and depth of training samples, this study uses all available daily closing prices of the CSI 300 Index from the first trading day of 2010 up to the starting date of each test period as the training dataset. All data is sourced from the Wind database.

During preprocessing, missing values are first imputed using a sliding window method to ensure the continuity of the time series. Then, Min-Max normalization is applied to rescale the data, improving the stability of model training.

3.2 Model Environment and Parameter Settings

It is imperative to note that all models utilized in this study are implemented using Python 3.13 in the PyCharm development environment. The development of deep learning components is informed by the PyTorch framework. The selection of the optimal order of the ARIMA model is made for each prediction interval based on the Akaike Information Criterion (AIC). This determines the best-fitting (p, d, q) parameters to capture linear patterns in the time series.

The LSTM model is designed following insights from existing literature and empirical tuning. A two-layer architecture is adopted, with each hidden layer consisting of 128 neurons. Dropout layers are inserted after each hidden layer to prevent overfitting, with the dropout rate set to 0.2. The model training utilizes the Adam optimizer, with a batch size of 64. The number of training epochs is adjusted between 50 and 2000, depending on the data scale, with the objective of ensuring adequate learning across different prediction intervals.

To further enhance prediction performance, Bayesian optimization is employed to fine-tune the LSTM hyperparameters. Tuned parameters include the number of layers, neurons per layer, and dropout rate. The model structure is adjusted within a reasonable range, while the dropout rate is automatically selected via the optimization algorithm to balance overfitting prevention with learning capability.

The forecasting performance of the standalone LSTM model and the integrated ARIMA-VMD-LSTM model is compared under different hyperparameter combinations. All results presented are based on the models with the best empirical fit.

In terms of input-output design, the standalone LSTM model uses only the raw closing price series as input. In contrast, the ARIMA-VMD-LSTM hybrid model constructs a multidimensional input consisting of the trend component extracted by ARIMA and the decomposed components obtained via VMD. The output of the model is the predicted closing prices for the next three consecutive trading days, enabling joint modeling of both linear and nonlinear features.

3.3 Evaluation Metrics

In order to provide an objective and reliable assessment of the model's performance, this study adopts several evaluation metrics from different perspectives. These include Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The following mathematical definitions have been established for these metrics:

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \quad (14)$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \quad (15)$$

$$MAPE=\frac{100\backslash\%}{N}\sum_{i=1}^{N}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \quad (16)$$

Where the symbol N is used to denote the number of samples, y_i represents the actual observed value at time i, and $\hat{y}_i$ is the corresponding predicted value.

3.4 Experimental Results

Table 1. Error Metrics for Short-Term Prediction Models on the CSI 300 Index (Unit: %)

Forecast Horizon	Model	RMSE	MAE	MAPE
Short-term (28 days)	ARIMA	3.07	2.98	2.21
	LSTM	2.54	2.36	1.91
	ARIMA-VMD-LSTM	1.11	0.88	0.89

As shown in Table 1, the ARIMA-VMD-LSTM hybrid model demonstrates superior performance across all error metrics in the short-term prediction task, significantly outperforming the standalone ARIMA and LSTM models in terms of forecasting accuracy. Specifically, over the 28-day forecasting horizon, the ARIMA model yields an RMSE of 3.07%, an MAE of 2.98%, and a MAPE of 2.21%. The LSTM model shows improved performance, with all error metrics reduced and MAPE falling to 1.91%, indicating its better capacity for modeling nonlinear structures compared to traditional statistical methods.

The ARIMA-VMD-LSTM model, which incorporates VMD-based decomposition, achieves the most remarkable results: the RMSE drops to 1.11%, the MAE decreases to just 0.88%, and the MAPE is controlled within 0.89%. These results indicate that the integration of linear and nonlinear components significantly enhances the model’s generalization ability and short-term forecasting accuracy.

Table 2. Error Metrics for Medium-Term Prediction Models on the CSI 300 Index (Unit: %)

Forecast Horizon	Model	RMSE	MAE	MAPE
Medium-term (68 days)	ARIMA	8.27	7.51	4.94
	LSTM	3.01	2.57	2.84
	ARIMA-VMD-LSTM	3.59	3.06	3.20

As shown in Table 2, the performance differences among models become more pronounced in the medium-term forecast. The standalone ARIMA model’s errors increase significantly, with an RMSE of 8.27%, MAE of 7.51%, and MAPE rising to 4.94%, indicating poor adaptability to price fluctuations over longer time spans. In contrast, the LSTM model demonstrates more stable performance, with RMSE, MAE, and MAPE of 3.01%, 2.57%, and 2.84%, respectively, reflecting its advantage in modeling medium-term trends.

Although the ARIMA-VMD-LSTM hybrid model does not achieve the best results in the medium-term forecast, it still delivers competitive performance, with an RMSE of 3.59%, MAE of 3.06%, and MAPE of 3.20%. While these metrics are slightly higher than those of the LSTM model, they are significantly better than those of the ARIMA model, suggesting that the hybrid model benefits from the complementary strengths of its components and offers a balanced approach in handling moderately complex market dynamics.

Overall, while the LSTM model shows the best performance in the medium-term prediction, the ARIMA-VMD-LSTM model's error metrics are close, demonstrating its strong comprehensive fitting ability and robustness, providing a potential foundation for further improving medium- and long-term forecasting accuracy.

Table 3. Error Metrics for Long-Term Prediction Models on the CSI 300 Index (Unit: %)

Forecast Horizon	Model	RMSE	MAE	MAPE
Long-term (126 days)	ARIMA	13.85	11.30	9.41
	LSTM	3.79	3.67	4.15
	ARIMA-VMD-LSTM	3.10	1.95	2.40

According to Table 3, the performance gap among models widens further in the long-term forecast. The ARIMA model’s prediction errors increase substantially, with RMSE reaching 13.85%, MAE 11.30%, and MAPE 9.41%, indicating its weak ability to model long-term trends and difficulty in accurately capturing the nonlinear volatility patterns of financial markets over extended periods. By comparison, the LSTM model maintains relatively stable performance with RMSE of 3.79%, MAE of 3.67%, and MAPE controlled at 4.15%, showing its advantage in handling long-term temporal dependencies.

Notably, the ARIMA-VMD-LSTM hybrid model continues to demonstrate strong adaptability in long-term prediction, with RMSE reduced to 3.10%, MAE lowered to 1.95%, and MAPE only 2.40%, significantly outperforming the individual models. This result indicates that by incorporating VMD decomposition to effectively extract signal components at different frequencies, combined with ARIMA capturing linear trends and LSTM modeling nonlinear structures, the model can comprehensively capture the long-term evolution characteristics of the index, thereby substantially improving forecasting accuracy.

Considering the results across the short-, medium-, and long-term forecast horizons, the ARIMA-VMD-LSTM hybrid model exhibits strong robustness and generalization ability across different time scales, with particularly pronounced advantages in long-term forecasting. These findings validate the practical value and significance of multi-model fusion in financial time series prediction.

4. Conclusion

This paper systematically investigates the forecasting capability of the ARIMA-VMD-LSTM nonlinear hybrid model on the closing prices of the CSI 300 Index across different time scales. By dividing historical data into three representative forecasting intervals and introducing Variational Mode Decomposition (VMD) to enhance feature extraction, the model combines the linear modeling advantages of ARIMA with the nonlinear fitting ability of LSTM to construct a multi-level forecasting framework. Comparative analyses with baseline models such as ARIMA and LSTM were conducted.

The results show that the ARIMA-VMD-LSTM hybrid model is highly robust and capable of generalization in the short-, medium- and long-term forecasting of the CSI 300 Index. The MAPE errors for the three forecast intervals are 0.89%, 3.20%, and 2.40%, respectively, all significantly outperforming the single models, reflecting the adaptive advantage of nonlinear hybrid strategies in modeling complex time series. Compared with the ARIMA model, the hybrid model reduces errors substantially across all forecast horizons, especially in long-term prediction where the ARIMA model's errors increase sharply while the hybrid model maintains high accuracy. Compared with the LSTM model, the hybrid model achieves further improved forecasting accuracy in both short- and long-term predictions, indicating that VMD decomposition enhances the deep model's capability to capture multi-scale information.

Moreover, as the forecast horizon lengthens, the accuracy of single models, particularly ARIMA, declines noticeably, whereas the hybrid model demonstrates greater stability, highlighting its advantage in handling financial market non-stationarity and structural changes.

In summary, the ARIMA-VMD-LSTM hybrid model shows promising generalization and practical applicability in trend forecasting of the CSI 300 Index. Beyond the CSI 300, this model can be extended to predict other financial products such as government bond futures and cryptocurrencies, where high volatility and complexity provide rich scenarios for model validation and optimization, further expanding its practical value across diverse financial markets.

References

[1] Ci, B., Zhang, P. (2022) Financial time series prediction based on ARIMA-LSTM model. Stat. Decis., 38(11): 145–149.

[2] Guo, J. (2020) Research on CSI 300 index prediction based on VMD-EEMD-LSTM model. Mod. Finance (Tianjin Univ. Finance Econ. J.), 40(08): 31–44.

[3] Lei, K., Chen, Y. (2007) Prediction of China's inbound tourist volume based on BP neural network and ARIMA hybrid model. J. Tourism, (04): 20–25.

[4] Wu, Y., Wen, X. (2016) Short-term stock price prediction based on ARIMA model. Stat. Decis., (23): 83–86.

[5] Zhang, L., Sun, S., Wang, Y. (2021) Exchange rate prediction based on deep learning LSTM model. Stat. Decis., 37(13): 158–162.

[6] Zhou, Z., He, X. (2023) Stock price prediction method based on optimized LSTM model. Stat. Decis., 39(06): 143–148.

[7] Ariyo, A.A., Adewumi, A.O., Ayo, C.K. (2014) Stock price prediction using the ARIMA model. Procedia Comput. Sci.

[8] Dave, E., Leonardo, A., Jeanice, M., et al. (2021) Forecasting Indonesia exports using a hybrid model ARIMA-LSTM. Procedia Comput. Sci., 179: 480–487.

[9] Pierre, A.A., Akim, S.A., Semenyo, A.K., et al. (2023) Peak electrical energy consumption prediction by ARIMA, LSTM, GRU, ARIMA-LSTM and ARIMA-GRU approaches.

[10] Wen, T., Liu, Y., Bai, Y.H., et al. (2023) Modeling and forecasting CO_2 emissions in China and its regions using a novel ARIMA-LSTM model. Heliyon, 9(11): e2023.

[11] Wu, J., Xu, K., Chen, X., et al. (2021) Price graphs: Utilizing the structural information of financial time series for stock prediction. Papers.