Empirical Analysis of Broad-Based Index Stock Price Forecasting Using LSTM Models: A Comparative Study with ARIMA and Ensemble Learning Models

Xinran Jia

China university of mining & technology-beijing, Beijing, China

18811529197@163.com

Abstract. Financial time series data are typically nonlinear and non-stationary, which restricts the prediction accuracy of traditional models and hardly meets practical financial decision-making needs. Long Short-Term Memory (LSTM) networks, with their special gating mechanism, provide a new technical way for accurate forecasting of stock market indices.This study uses daily trading data of the CSI 300, ChiNext, and CSI 500 indices from January 2015 to September 2025. We construct an explanatory variable system including raw price features and common technical indicators, and establish LSTM, ARIMA, Random Forest, and XGBoost models. Using regression and classification evaluation with multi-period rolling tests, we compare model performance in price direction judgment and closing price fitting.Empirical results show that the LSTM model performs best: direction prediction accuracy reaches 51.74%, AUC is 0.52, MAPE is 0.87%, and RMSE is 50.50 yuan, significantly outperforming the other models. Its gating mechanism and nonlinear structure better match the characteristics of stock price data. This paper supports the application of LSTM in broad-based index forecasting and provides a reference for financial time series model selection.

Keywords: LSTM model; Broad-based index; Stock price forecasting; Financial time series data; Multi-model comparison

1 Introduction

Financial markets' high volatility, nonlinearity, and non-stationarity make stock index forecasting a persistent challenge in quantitative investment and risk management. Traditional econometric models, based on linear assumptions, often fail to capture complex market dynamics, limiting predictive accuracy. Machine learning and deep learning offer new solutions, with Long Short-Term Memory (LSTM) networks emerging as a prominent tool for financial time series forecasting due to their ability to capture long-term dependencies via gating mechanisms. Accurate forecasts are vital for investor decisions, portfolio optimization, and risk control. However, existing studies typically focus on single markets, lack cross-index validation, and provide shallow analysis of LSTM's advantages. This paper addresses these gaps by selecting three broad-based

J. Rak et al. (eds.), *Proceedings of the 2026 11th International Conference on Social Sciences and Economic Development (ICSSED 2026)*, Advances in Economics, Business and Management Research 391,
https://doi.org/10.2991/978-94-6239-701-9_107

indices and conducting multi-model comparisons to empirically evaluate LSTM's forecasting performance and explore the sources of its advantages, offering insights for model selection in financial time series forecasting.

2 Literature Review

The high volatility and nonlinearity of financial markets make index forecasting central to quantitative investment. Traditional statistical models, limited by linear assumptions, struggle with complex dynamics; machine and deep learning have thus become the dominant paradigm due to their pattern recognition capabilities.

Research on feature engineering has established a framework combining price data with technical indicators. Modi and Yamaganti[1] demonstrate that integrating fundamental metrics (e.g., EPS) with technical indicators (e.g., RSI) enhances accuracy, noting a strong positive correlation between these variable types. Singh[3] and Weng[5] further explore refined features and binary labels to better align predictions with trading decisions.

Model comparisons reveal limitations in both traditional and ensemble approaches. For traditional models, Varadharajan[4] et al. find that ARIMA produces significantly higher errors on non-stationary stock sequences and loses market information through differencing. Ensemble models like Random Forest and XGBoost, valued for nonlinear fitting, also exhibit drawbacks: while Random Forest excels in multi-indicator fusion, XGBoost flattens sequential data into flat features, discarding inter-day correlations and underperforming when temporal dependencies are strong (Modi and Yamaganti[1], Singh and Praveen[3]).

LSTM models overcome these issues through structural advantages. Varadharajan[4] et al. show LSTMs capture long-term dependencies via gating mechanisms, achieving low training error. Weng[5] attains directional accuracy above 50% on the DAX30 with LSTM ensembles, and Piao[2] indicates that LSTM with attention mechanisms can model complex interactions among multiple variables.

Nevertheless, existing research has room for improvement. First, model comparisons often focus on single markets or specific periods, lacking robust cross-validation across multiple broad-based indices. Second, the intrinsic logic behind LSTM’s advantages remains underexplored, with insufficient analysis linking model structure to performance.

3 Samples and Data

3.1 Data Sources and Original Structure

This study selects the CSI 300, ChiNext, and CSI 500 indices as subjects. Daily data from January 2015 to September 2025 (2,656 valid observations after cleaning) are sourced from Wind. The decade-long sample, spanning multiple market phases, rigorously tests model robustness and generalization.

3.2 Variable Definition and Selection

The dependent variable is raw closing price (yuan). Predictors include 11 features: five raw features (open, high, low, close, volume) and six technical indicators (MA20, RSI14, Bollinger Bands, VWAP), all min-max normalized to [0,1].

3.3 Data Preprocessing

A standardized workflow includes: (1) cleaning (remove suspensions, correct outliers); (2) calculating technical indicators; (3) filling missing values via bfill; (4) normalizing features only.

3.4 Time-Series Sample Construction and Dataset Partitioning

Data are transformed into supervised format using a 60-day sliding window (features from days i-60 to i-1 predict day i's close). The dataset is chronologically split into three test periods (20, 60, 250 days) for rolling evaluation.

4 Models and Performance Evaluation Metrics

This paper constructs a multi-model comparative framework centered on LSTM, with ARIMA, Random Forest, and XGBoost as benchmarks. All models are trained and tested on the same preprocessed dataset.

4.1 Model Introduction

Third Level LSTM Model.
LSTM captures long-term dependencies in financial time series via input, forget, and output gates. This study uses seven features: five foundational (ma20, rsi14, boll_mid, vwap, close) and two trend features (close_trend, rsi_trend). Training follows a 70/10/20 split, 30-day sequence length, Adam optimizer, MSE loss, with early stopping.

Third Level Benchmark Models.
1.**ARIMA**: Classic linear time series model with optimal order (1,1,1).
2.**Random Forest**: Ensemble of decision trees. Time-series windows are flattened for input; hyperparameters optimized via grid search.
3.**XGBoost**: Gradient boosting framework with hyperparameters tuned via grid search and cross-validation.

Third Level Comparative Framework.

All models predict raw closing prices. Continuous predictions are converted to directional signals (up/down) based on comparison with previous day's price, enabling evaluation of both price fitting (regression) and directional judgment (classification).

4.2 Performance Evaluation Metrics

This paper employs a dual evaluation system encompassing regression and classification. All metrics are strictly calculated using the test set. The generation rules for direction classification labels are as described in the previous section.

Third Level Classification Performance Metrics.

1. $$Accuracy = (TP + TN)/(TP + FP + FN + TN) \times 10 \quad (1)$$

Reflects the overall correct judgment rate;

2. $$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \times 100\% \quad (2)$$

Measures the ability to capture upward trends, influencing the strategy's profit-capturing potential;

3. $$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \times 100\% \quad (3)$$

Reflects the reliability of upward prediction signals, impacting transaction costs and risk exposure;

4.AUC: Area Under the ROC Curve. This metric comprehensively evaluates model performance across varying classification thresholds and is insensitive to positive/negative sample ratios.

Third Level Regression Fitting Metrics.

For predicting closing price levels, the following three common error metrics are used:

1. $$\text{MSE} = (1/\text{n}) \times \Sigma(\text{i} = 1 \text{ to n})\ (\text{y\_i} - \hat{\text{y}}\text{\_i})^2 \quad (4)$$

The mean of squared differences between predicted and actual values. This metric imposes heavier penalties on larger prediction errors.

2. $$\text{RMSE} = \sqrt{[(1/\text{n}) \times \sum(\text{i} = 1 \text{ to n})\ (\text{y\_i} - \hat{\text{y}}\text{\_i})^2]} \quad (5)$$

Retains dimensionality, making error magnitude more intuitively understandable;

3. $$\text{MAPE} = (1/\text{n}) \times \Sigma(\text{i} = 1 \text{ to n})\ |(\text{y\_i} - \hat{\text{y}}\text{\_i})/\text{y\_i}| \times 100\% \quad (6)$$

Eliminates the influence of price scales to facilitate cross-sectional comparisons.

All indicator results are summarized in the table, with values rounded to two decimal places, providing clear, quantifiable evidence for subsequent empirical analysis.

5 Empirical Results and Analysis

This section reports empirical results for four forecasting models. Analysis follows the established framework, evaluating overall performance, directional accuracy, and closing price regression.

5.1 Overview of Experimental Setup

This study uses daily data from CSI 300, ChiNext, and CSI 500 indices (2015–2025), yielding 2,656 valid samples. The target is raw closing price; explanatory variables include price, volume, and technical indicators. A 30-day sliding window constructs supervised samples. Data is split 7:1:2 into training, validation, and test sets, with a 60-day test window. LSTM uses 7 input features with early stopping; ARIMA order is (1,1,1); Random Forest and XGBoost hyperparameters are optimized via grid search.

5.2 Comprehensive Model Performance Comparison

Table 1 summarizes each model's core performance metrics over the 60-day test period. Results show a clear performance hierarchy: LSTM leads in both classification and regression tasks, Random Forest and XGBoost perform similarly in the middle tier; ARIMA lags across all measures.

Table 1. Comparison of Core Performance Metrics.

Model	Accuracy (%)	Recall (%)	Precision (%)	AUC	MSE	RMSE	MAPE (%)
LSTM	51.74	54.15	53.88	0.52	2550.72	50.5	0.87
ARIMA	47.03	3.95	38.46	0.49	162213.06	402.76	9.24
Random Forest	51.33	53.36	52.94	0.51	2680.85	51.78	0.94
XGBoost	51.12	51.78	52.82	0.51	2825.44	53.15	0.93

LSTM achieves the highest accuracy and AUC in classification, while in regression, its MSE is over 98% lower than ARIMA with MAPE below 1%. This hierarchy reflects methodological characteristics: ARIMA struggles with nonlinear, non-stationary data; ensemble models lose temporal structure through flattening; LSTM's architecture captures long-term dependencies, providing structural advantages in forecasting.

5.3 Directional Trend Prediction Analysis

LSTM leads with 51.74% accuracy and 0.52 AUC, edging Random Forest (51.33%, 0.51). ARIMA falls below 50% random baseline. Examining recall and precision

(**Figure 1**), LSTM achieves highest recall (54.15%) while maintaining reasonable precision (53.88%), indicating reliable bullish signals. ARIMA's recall drops to 3.95%, missing uptrends. Ensemble models balance both metrics but slightly trail LSTM.

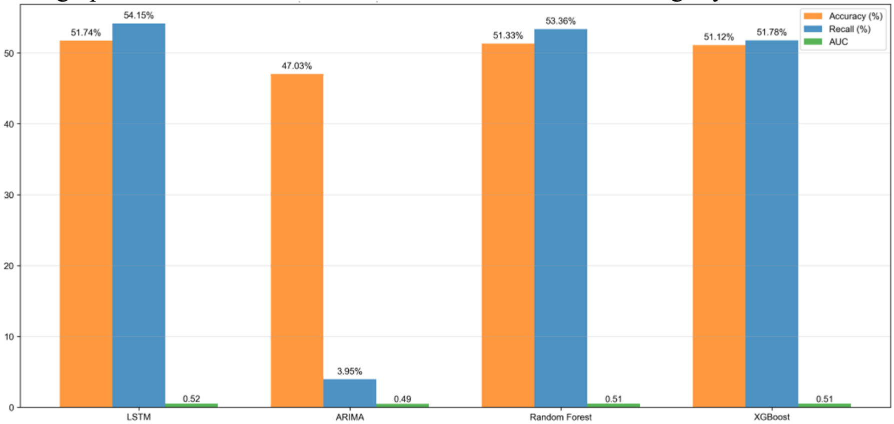

Fig. 1. Comparison of Bullish/Bearish Prediction Metrics Across Models.

The ROC curve offers an additional evaluation dimension (**Figure 2**). The LSTM curve lies closest to the upper-left corner, achieving the highest AUC value (0.52). This indicates its ability to maintain a relatively optimal trade-off between true positives and false positives across different decision thresholds. This further validates its comprehensive advantage in directional prediction.

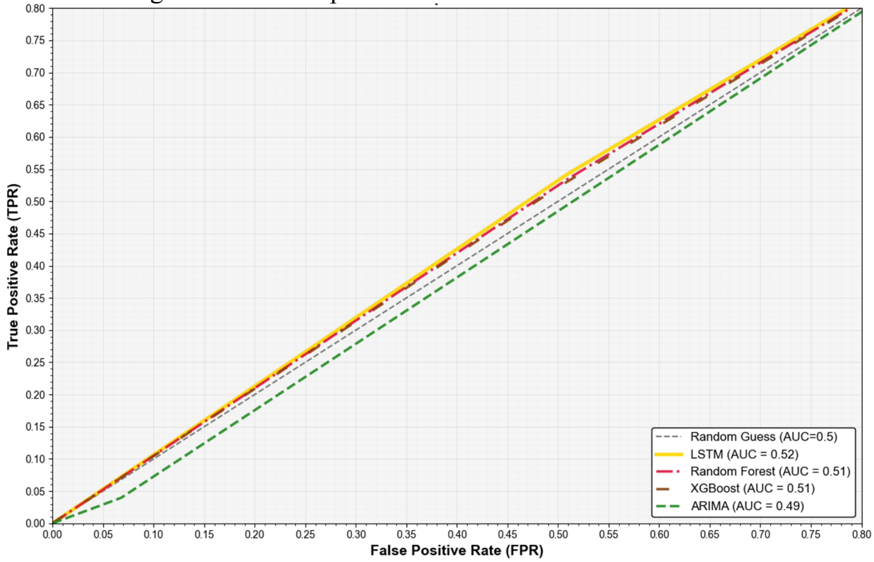

Fig. 2. Comparison of ROC Curves Across Models.

5.4 Analysis of Closing Price Numerical Prediction Accuracy

LSTM excels in numerical fitting, with RMSE of 50.50 yuan versus ARIMA's 402.76 yuan, whose smooth predictions lag actual fluctuations. Random Forest and XGBoost achieve RMSE of 51–53 yuan, approaching but trailing LSTM. MAPE confirms LSTM's accuracy at 0.87%—below 1% of actual price and within typical 1–2% daily volatility—while ARIMA's 9.24% confirms its unsuitability, as shown in **Figure 3**).

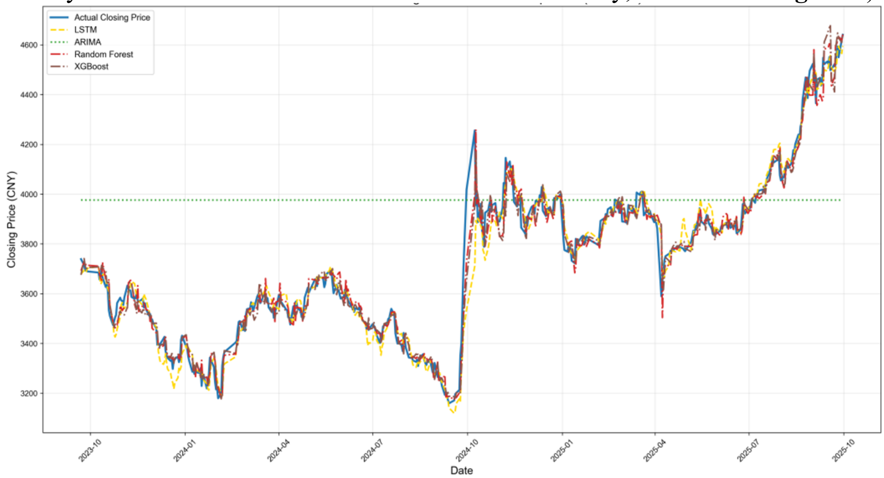

Fig. 3. Comparison of Stock Closing Price Forecasts for 2023-2025.

6 Supplementary Analysis

LSTM significantly outperforms ARIMA, Random Forest, and XGBoost in both directional accuracy and numerical forecasting. This advantage stems from three structural characteristics: its gating mechanism retains long-term dependencies while filtering noise; its data-driven nature handles nonlinearity without flattening sequences; and its sequential processing captures continuous market dynamics. Unlike ARIMA, which relies on linear assumptions, and tree-based models, which discard temporal structure, LSTM aligns with the inherent characteristics of financial time series.

7 Conclusion and Outlook

Using three indices, this multi-model study shows LSTM optimally predicts stock price direction and values, due to its gating mechanism, nonlinear adaptability, and alignment with financial data dynamics. Random Forest and XGBoost are robust nonlinear benchmarks; ARIMA suits short-term linear trends only. This study supports model selection, though macroeconomic conditions and sentiment were excluded. Future research should enrich features and expand asset coverage to assess generalization.

References

1. MODI Y K. 2025. INTEGRATING FUNDAMENTAL AND TECHNICAL ANALYSIS FOR ENHANCED STOCK MARKET PREDICTION: A MACHINE LEARNING APPROACH[J]. 25(06).Welcome To IJESAT
2. PIAO C, ZHU T, BALDEWEG S E, et al. 2024. GARNN: An Interpretable Graph Attentive Recurrent Neural Network for Predicting Blood Glucose Levels via Multivariate Time Series[EB/OL]. arXiv[2026-02-27].https://doi.org/10.1016/j.neunet.2025.107229
3. SINGH S K, PRAVEEN D S. 2025. A COMPARATIVE ANALYSIS OF TREE-BASED ENSEMBLES AND DEEP LEARNING FOR EQUITY PRICE FORECASTING USING FEATURE-ENGINEERED ADVANCE TECHNICAL INDICATORS[J]. International Journal of Applied Mathematics, 38(2) https://doi.org/10.12732/ijam.v38i2s.124
4. VARADHARAJAN V, SMITH N, KALLA D, et al. 2024. Stock Closing Price and Trend Prediction with LSTM-RNN[J]. Journal of Artificial Intelligence and Big Data, 4(1): 1-13.DOI:10.31586/jaibd.2024.877.
5. WENG Y, SU G, CHEN H, et al. 2025. Long Short-Term Memory Neural Network for Different Regional Financial Time Series[M/OL]//LEE G M, WANG H, EROKHIN V, et al. Proceedings of the 2025 4th International Conference on Bigdata Blockchain and Economy Management (ICBBEM 2025): Vol. 195. Dordrecht: Atlantis Press International BV: 166-174[2026-02-27] https://doi.org/10.2991/978-94-6463-742-7_19