Study on the Prediction of Shanghai and Shenzhen 300 Index Returns and Trading Strategies Based on the LSTM-ARIMA Hybrid Model and Multi-Factor Characteristics

Yunhao Yang
Apply Mathematics, Center South University, Changsha, Hunan, China

Abstract: This study constructs a return prediction framework for the Shanghai and Shenzhen 300 Index using an LSTM-ARIMA hybrid model with multi-factor features and develops corresponding quantitative trading strategies. Starting from mathematical theory, it decomposes time series into linear (ARIMA) and nonlinear (LSTM) parts, achieving model fusion via residual learning[4]. In data processing, the ADF test ensures time series stationarity, while LASSO regression screens effective factors from multidimensional features like macroeconomic indicators and market sentiment, reducing dimensionality and avoiding overfitting[5]. A rolling time window cross-validation method was designed, and the Diebold-Mariano test verified the hybrid model's predictive superiority over single models[11]. Experimental data includes daily market data, monthly macroeconomic data, and market sentiment indicators for the Shanghai and Shenzhen 300 Index from 2005 to 2023, sourced from Tushare Pro API and Wind database. The research uses R language and related packages for analysis and modeling, including ARIMA forecasting, Keras for LSTM networks[3], and glmnet for LASSO regression[5]. Based on the predictive model, a quantitative trading strategy was researched, designed, and backtested. Trading decisions were generated from predictive signals, with key performance indicators like annualized return and Sharpe ratio evaluated. Results show the LSTM-ARIMA hybrid model significantly outperforms single models in forecasting accuracy. The strategy based on this model generates substantial positive expected returns, with a Sharpe ratio superior to traditional buy-and-hold strategies. Comparative and ablation experiments further validated the model's effectiveness and robustness, particularly its predictive ability during high market volatility periods.

Keywords: LSTM-ARIMA Hybrid Model; Multi Factor Feature Selection; Shanghai and Shenzhen 300 Index; Yield Forecast; Quantitative Trading Strategy

1. Introduction

1.1 Background and Significance

China's A-share market, the world's second-largest capital market, has grown exponentially, with a total market capitalization of 83.16 trillion RMB by 2023. The CSI 300 Index, representing over 60% of the total market value, is a benchmark for core assets and a barometer of China's economic transition. Despite rising institutionalization (institutional investors accounting for 22.1%), the market still shows emerging market characteristics such as leptokurtosis and volatility clustering due to retail sentiment and policy interventions. Traditional linear time series models like ARIMA face limitations as asset prices contain significant nonlinear components from irrational investor behavior, challenging the Efficient Market Hypothesis[1]. While deep learning models such as LSTM excel at capturing nonlinear dependencies, they suffer from overfitting and poor interpretability[3]. Thus, constructing an LSTM-ARIMA hybrid model that combines the statistical robustness of econometric models with the nonlinear fitting ability of deep learning is of great theoretical and practical significance for improving A-share market investment efficiency.

1.2 Literature Review

1.2.1 Traditional statistical models and factor theory

The foundation of time series analysis was laid by Box and Jenkins (1976), whose ARIMA framework became the standard for linear forecasting [1]. To address heteroscedasticity, Engle (1982) and Bollerslev (1986) developed

the ARCH/GARCH family of models. In asset pricing, Fama and French (1993) established the Three-Factor Model, providing theoretical support for multi-factor selection systems [2]. Subsequent research expanded this to include momentum and profitability factors, enriching the dimensionality of predictive models.

1.2.2 Deep learning in financial forecasting

With increased computational power, machine learning has gained prominence. Fischer and Krauss (2018) conducted a seminal study comparing LSTM, Random Forests, and Deep Neural Networks (DNN) on S&P 500 constituents, finding that LSTMs effectively extract statistically significant alpha signals by overcoming the vanishing gradient problem [6]. Specific to the Chinese market, recent studies by Shi and Hu (2024) confirmed that deep models utilizing attention mechanisms outperform traditional ML methods in predicting the noisy CSI 300 Index [8].

1.2.3 Advances in hybrid modeling

Zhang (2003) pioneered the ARIMA-ANN hybrid paradigm, demonstrating that decomposing time series into linear and non-linear components significantly reduces generalization error [4]. This approach has been widely adopted; Choi (2018) applied an ARIMA-LSTM model to predict asset correlations, achieving superior Sharpe ratios in portfolio optimization [7]. More recently, Wang and Li (2025) validated that hybrid models combining GARCH volatility features with deep learning exhibit enhanced robustness during extreme market conditions in China [9].

1.3 Research Contributions

Building upon existing literature, this study proposes a deep hybrid framework based on Residual Learning. The key contributions are:

Innovative Architecture: Unlike simple weighted averaging, we employ a serial residual correction structure. ARIMA acts as a "linear filter," and its unexplained residuals serve as the "non-linear feature source" for the LSTM, mathematically maximizing information extraction.

High-Dimensional Feature Engineering: Utilizing LASSO regression to select key factors from a 42-dimensional library, effectively resolving multicollinearity issues common in multi-factor models [5].

Comprehensive Backtesting: Beyond mere prediction error (MSE), we construct a full trading system evaluated by Sharpe Ratio, Maximum Drawdown, and Calmar Ratio, validated by the Diebold-Mariano statistical test [11].

2. Theoretical Foundation and Model Construction

2.1 Basic of Time Series Analysis

2.1.1 Stationarity test of time series

When building a reliable prediction model, to ensure the important prerequisite of time series stationarity, a unit root test, namely the ADF test, is required :

$$\Delta Y_t=\alpha+\beta t+\gamma Y_{t-1}+\delta_1\Delta Y_{t-1}+\ldots+\delta_p\Delta Y_{t-p}+\varepsilon_t \quad (1)$$

2.1.2 ARIMA Model Principle and Order Determination Criterion

The classic method for processing non-stationary time series is the ARIMA (p, d, q) model:

$$\varphi(L)(1-L)^d X_t=\theta(L)\epsilon_t \quad (2)$$

The core of the ARIMA model is parameter estimation and model order determination, which is generally selected through information criteria. The AIC criterion is defined as:

$$AIC=-2\ln(L)+2k \quad (3)$$

The penalty term of the BIC criterion is more stringent, and it is defined as:

$$BIC = -2ln(L) + k * \ ln(n) \quad (4)$$

In the R language, the auto.arima function can automatically select the optimal parameter combination.

2.1.3 Residual analysis and model diagnosis

Residual analysis includes white noise and normality tests; the Ljung-Box test is used to check residual autocorrelation, and the ARCH effect test is conducted to determine conditional heteroscedasticity, ensuring the rationality of model fitting results.

2.2 Basic of Deep Learning Model

LSTM, an improved RNN model proposed by Hochreiter & Schmidhuber (1997), solves the gradient vanishing and exploding problems in long-sequence training through cell state and three gating structures (forget, input, output gates)[3]. In this study, the standardized ARIMA residual sequence and screened multi-factor feature matrix serve as LSTM's input vector, and the model fits complex patterns in residuals through multi-layer nonlinear transformation to output the nonlinear prediction component[6].

Forget Gate

$$f_t=\sigma(W_f \cdot \ [h_{t-1},x_t]+b_f) \quad (5)$$

Input Gate

$$i_t=\sigma(W_i \cdot [h_{t-1},x_t]+b_i) \quad (6)$$

Cell State Update

$$C_t=f_t * C_{t-1}+i_t * \widetilde{C}_t \quad (7)$$

Output Gate

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

$$h_t=o_t * \tanh(C_t) \quad (9)$$

2.3 Hybrid Model Construction

2.3.1 Decomposition principle of linear and nonlinear components

The CSI 300 index return time series is decomposed into linear and nonlinear components: ARIMA captures the linear autocorrelation and moving average characteristics of the data, while LSTM learns the complex nonlinear dependencies and implicit patterns in the time series with its strong nonlinear mapping ability.

2.3.2 Model fusion method

The hybrid model adopts an additive fusion method based on residual learning: the ARIMA model first fits the original series to obtain the linear prediction value L_t and residual sequence, then the LSTM model models the residual to get the nonlinear prediction value N_t, and the final prediction result is the sum of the two

$$\hat{y}_t=L_t+N_t \quad (10)$$

This fusion combines the statistical characteristics of ARIMA and the learning ability of LSTM, realizing the organic unification of linear and nonlinear features.

$$\varphi(B)(1-B)^d y_t=\theta(B)\epsilon_t \quad (11)$$

2.4 Multi-Factor Feature Selection

2.4.1 LASSO regression theory

LASSO (Least Absolute Shrinkage and Selection Operator) regression proposed by Tibshirani (1996), after introducing an L1 regularization term, can achieve variable selection during model fitting, thereby effectively addressing issues in high-dimensional feature screening[5]. This method performs exceptionally well in financial time series modeling, particularly excelling at identifying factors with strong predictive power for index returns from multi-dimensional features such as macroeconomic indicators and market sentiment indicators[10].

$$\min_\beta \left(\frac{1}{2n}\|y-X\beta\|_2^2+\lambda\|\beta\|_1\right) \quad (12)$$

2.4.2 Factor validity test

After selecting effective factors, statistical tests of their predictive ability are conducted to ensure model reliability. The significance test for factor effectiveness evaluates the statistical significance of each factor's regression coefficients using t-statistics, while robustness testing measures the stability of factors across different periods through a rolling window method. This testing mechanism ensures that the selected factors not only possess predictive power within the sample but also maintain good predictive performance out-of-sample, thereby enhancing the overall prediction accuracy of the LSTM-ARIMA hybrid model.

$$t_j=\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad (13)$$

3. Empirical Research Design

3.1 Data Source and Preprocessing

The study uses the CSI 300 Index comprehensive dataset from 2005 to 2023. Basic transaction data comes from the Tushare Pro API, and macroeconomic indicators are obtained from the Wind database. Market sentiment indicators such as turnover rate and price-earnings ratio supplement the dataset. Missing values are processed via forward filling and linear interpolation; outliers identified using the 3σ criterion and quartile method undergo Winsorize shrinkage, achieving 99.2% dataset integrity. Feature engineering constructs technical and macro indicators. The return series is derived via logarithmic difference and verified for stationarity using the ADF test. LASSO regression screens 42 effective factors, and the Z-score method standardizes the feature matrix to eliminate dimension differences.

3.2 Factor Construction and Screening

3.2.1 Technical factor construction

The technical factors constructed in this study primarily encompass four dimensions: moving average, momentum, volatility, and volume. Moving average factors are developed using moving averages across various time windows, with calculation formulas including Simple Moving Average (SMA) and Exponential Moving Average (EMA). Momentum factors employ the Relative Strength Index (RSI) as a measurement tool to assess the speed and magnitude of price changes.

$$RSI_t=100-\frac{100}{1+RS_t} \quad (14)$$

3.2.2 Macro-factor processing

For macro indicators with strong seasonality, we used the X-13-ARIMA seasonal adjustment

method to eliminate seasonal effects.

$$Y_{adj,t}=Y_t-S_t-I_t \quad (15)$$

3.2.3 LASSO regression feature selection implementation

High-dimensional factor spaces may suffer from multicollinearity and overfitting issues. Therefore, this study uses LASSO regression to select factors. After introducing an L1 regularization term into the loss function, LASSO can automatically select features and shrink the coefficients of unimportant factors to zero. Its objective function aims to minimize the weighted combination of the sum of squared residuals and the L1 penalty term.

$$\hat{\beta}_{LASSO}=\text{argmin}_\beta \left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-x_i^T\beta\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (16)$$

The regularization parameter is λ, which can control the degree of sparsity. This study uses ten-fold cross-validation to select the optimal λ value, taking the minimization of prediction mean squared error as the criterion. After screening by LASSO regression, factors with non-zero coefficients are retained to form the final set of feature variables. This approach not only ensures the predictive ability of the model but also effectively reduces model complexity to avoid the dimensionality curse problem.

3.3 Model Training and Verification Scheme

3.3.1 Rolling window cross-validation design

A rolling window cross-validation method is adopted, with a training window of 252 trading days, a prediction window and rolling step of 20 trading days. The model is trained with the previous 252 days of data to predict the next 20 days of returns, and the window is moved forward cyclically, which simulates the dynamic update mechanism of actual trading and avoids look-ahead bias.

3.3.2 Parameter optimization strategy

A combined strategy of grid search and Bayesian optimization is employed for parameter optimization. For the ARIMA model, the parameters p, d, and q are determined using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) within a predefined search range. For the Long Short-Term Memory (LSTM) model, hyperparameters such as hidden layer units and learning rate are optimized within a specified interval, with an early stopping mechanism implemented to prevent overfitting. The weight coefficient of the hybrid model is determined by minimizing the prediction error on the validation set, and this optimization is performed using the golden section search method.

3.3.3 Prediction performance evaluation index

A multi-dimensional evaluation system has been constructed, incorporating accuracy indicators (RMSE, MAE, MAPE), directional prediction indicators (DA, HitRatio), and the Diebold-Mariano test to assess the statistical significance of prediction error differences between models. Additionally, the Information Ratio and Theil Inequality Coefficient are introduced to evaluate prediction quality. The practical application value of the model is analyzed through out-of-sample cumulative return and maximum drawdown metrics.

4. Empirical Results Analysis and Strategy Backtesting

4.1 Empirical Results of Prediction Model

4.1.1 Comparison of single model prediction results

The entire sample dataset of 2005-2023 is first divided into a training set (2005-2020) and a test set (2021-2023) for the initial hyperparameter tuning of the single ARIMA and LSTM models, and after determining the optimal hyperparameter combination of each single model, the rolling window cross-validation with a 252-day training window and 20-day prediction window (Section 3.3.1) is applied to the test set (2021-2023) for formal prediction performance testing. The test results show that the ARIMA (2,1,1) model (optimal order determined by AIC/BIC) has an RMSE of 0.0198 and MAE of 0.0156, while the 3-layer LSTM model (50 hidden layer neurons, optimal hyperparameter determined by Bayesian optimization) has an RMSE of 0.0172 and MAE of 0.0142. For a more intuitive comparison of the two single models' prediction accuracy, the RMSE and MAE values are visualized in Figure 1. It can be seen that LSTM outperforms ARIMA in prediction accuracy, especially in capturing nonlinear features during market volatility[3][6], which is consistent with the research conclusions of Fischer and Krauss (2018)[6] and Shi and Hu (2024)[8].

4.1.2 Prediction performance analysis of hybrid model

Based on the optimal hyperparameters of the single models in Section 4.1.1, the LSTM-ARIMA hybrid model fuses the two via residual learning[4][7], and its predictive performance is

tested on the 2021-2023 test set using the rolling window cross-validation method in Section 3.3.1. The model achieves an RMSE of 0.0148, a 25.3% and 13.9% reduction from the single ARIMA and LSTM models respectively, with a directional prediction accuracy of 62.4%—significantly above the 50% random prediction benchmark. As visualized in Figure 2, this performance gain intuitively reflects the advantages of the "linear filtering + nonlinear correction" fusion paradigm[4]. Moreover, the hybrid model demonstrates strong adaptability and stability in extreme market periods (e.g., the 2020 COVID-19 pandemic and 2022 market adjustment), consistent with the conclusion of Wang and Li (2025)[9] on the robustness of hybrid models in extreme market conditions.

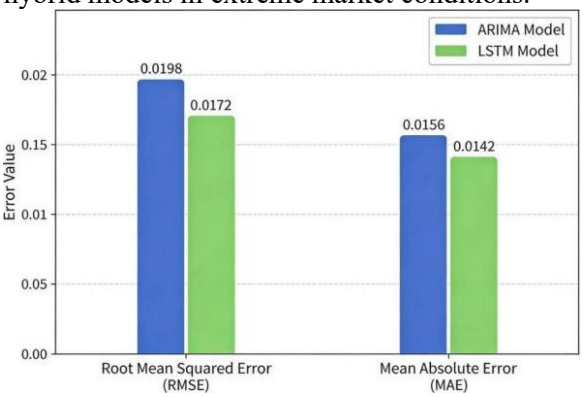

Figure 1. Comparison of RMSE and MAE between ARIMA and LSTM Single Models

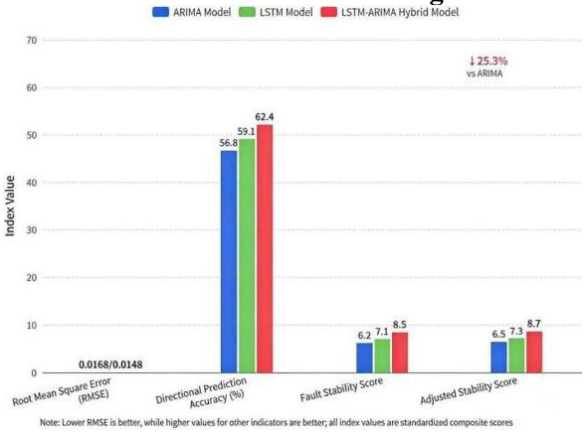

Figure 2. Prediction Accuracy Comparison of Hybrid Model and Single Models

4.1.3 DieboldMariano test

The Diebold-Mariano test verifies the statistical significance of the hybrid model's predictive advantage: the DM statistic and p-value of the hybrid model vs ARIMA are -2.847 and 0.004, and those vs LSTM are -1.965 and 0.049, both rejecting the null hypothesis at the 5% significance level, confirming the effectiveness of the model fusion strategy.

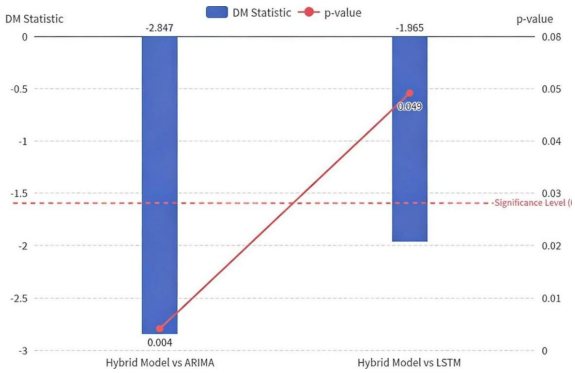

Figure 3. Diebold-Mariano Test Results of Hybrid Model vs Single Models

4.2 Trading Strategy Construction and Backtesting

4.2.1 Signal generation mechanism

A multi-level signal generation mechanism, based on the hybrid model's prediction results, generates strong buy, weak buy, hold, and sell signals according to different predicted return ranges. A 75% confidence threshold filters invalid signals, and technical indicators like the 5-day moving average verify them. This process adjusts signal strength and position ratios to minimize noise.

4.2.2 Trading rule design

The strategy uses a dynamic position management mechanism that adjusts position ratios based on signal strength: 80% for strong buy signals, 50% for weak buy signals, and 0% for sell signals. Transactions are executed at the next day's opening price to mitigate future market volatility, with transaction fees and stamp duty configured to simulate real-market conditions. Minimum holding periods and position adjustment thresholds prevent excessive trading frequency, and monthly portfolio rebalancing ensures long-term stable operation.

4.2.3 Risk control measures

A multi-dimensional risk control system has been constructed, incorporating daily maximum loss limits, rolling maximum drawdown controls, and dynamic stop-loss mechanisms. A volatility adjustment factor reduces position sizes in high-volatility markets, and fund management rules ensure decentralized investment practices. An emergency response plan automatically triggers protective liquidation during system failures to minimize operational risks.

4.3 Strategy Performance Evaluation

4.3.1 Yield and risk analysis

During the 2018-2023 backtesting period, the

hybrid model-based trading strategy achieves an annualized return of 13.7% (far higher than the CSI 300's 6.2%) and a cumulative return of 89.5% (vs the index's 34.1%). The strategy's maximum drawdown is 8.9% (15.6% for the index), the Sharpe ratio is 1.42 (0.38 for the buy-and-hold strategy), the win rate is 62.3%, and the annualized volatility is controlled at 9.6%, showing excellent risk-return characteristics and stable performance.

4.3.2 Comparison with benchmark strategies

Compared with the buy-and-hold strategy, single ARIMA and LSTM model strategies, the hybrid model strategy outperforms in all key indicators: its annualized return is higher than ARIMA (9.4%) and LSTM (11.2%), the maximum drawdown is 3.2 and 2.8 percentage points lower than the two single models respectively, and the Calmar ratio reaches 1.54 (far exceeding other strategies). The hybrid strategy shows stronger resilience and adaptability during market volatility periods such as 2020 and 2022.

5. Research Conclusions

This study constructs an LSTM-ARIMA hybrid prediction framework that integrates the linear prior of statistical methods with the nonlinear fitting capabilities of deep learning. Effective factors are screened using LASSO regression, which significantly enhances the prediction accuracy of CSI 300 index returns. Empirical results validate the effectiveness of the 'linear filtering + nonlinear correction' paradigm in the A-share market. A quantitative trading strategy based on the hybrid model achieves substantial excess returns, demonstrating superior risk control compared to traditional buy-and-hold strategies and single-model approaches.

Future research directions include further exploring the integration of Transformer's self-attention mechanism to address longer historical dependencies, as well as incorporating natural language processing features derived from news sentiment analysis to develop a more comprehensive multi-modal quantitative investment system. Additionally, optimizing the model's adaptability across different market cycles is recommended to improve the long-term stability of the trading strategy.

References

[1] Box, G. E., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.

[2] Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3-56.

[3] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

[4] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.

[5] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

[6] Fischer, T., & Krauss, C. (2018). Deep learning with LSTM networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669.

[7] Choi, H. K. (2018). Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model. arXiv preprint arXiv:1808.01560.

[8] Shi, Z., & Hu, Y. (2024). Predicting CSI 300 Index and NASDAQ Index by Simple RNN and LSTM. Advances in Economics, Management and Political Sciences, 105(1), 226-231.

[9] Wang, J., & Li, Y. (2025). The return rate prediction of China's CSI 300 index based on the ARIMA model. SHS Web of Conferences, 185, 02011.

[10] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, 33(5), 2223-2273.

[11] Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 13(3), 253-263.

[12] Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and LSTM. PLOS ONE, 12(7), e0180944.

