Directional Predictability of the CSI 300 Index: A Walk-forward Evaluation of a BILSTM-attention Model

Yiming Hu *

School of business, Macquarie University, Sydney, Australia

* Corresponding Author Email: hym99850928@gmail.com

Abstract. This study looks at short-term predictability in China's stock market, using the CSI 300 index as the main example. A Bidirectional Long Short-Term Memory (BiLSTM) model with an Attention mechanism is built that uses recent price- and volatility-based technical indicators and is tested with an expanding-window, walk-forward method for both direction classification and return-size regression. The model choice follows earlier work showing that deep sequence models and volatility-aware hybrids can capture nonlinear patterns and changing volatility, even though the economic value of daily predictions is usually small. In the results, the BiLSTM-Attention model gives a stable but modest edge in one-day-ahead direction prediction—AUC, PR-AUC, F1, and hit ratio are all above random—while predicted return sizes are still very noisy, which is consistent with weak-form market efficiency. The paper also discusses a tradability-oriented framework that links calibrated probabilities to position sizing under transaction costs and regime shifts, and outlines possible extensions using volatility-aware hybrids and multi-scale or global attention to improve robustness across different regimes and assets. Overall, the findings show that short-horizon predictability is fragile, and that careful implementation, probability calibration, and risk control are crucial if weak statistical signals are to be turned into investable strategies.

Keywords: CSI 300 index; BiLSTM Attention; Stock index prediction; Deep sequence models; Volatility aware hybrids.

1. Introduction

Short-horizon forecasting of broad equity indices is difficult. Daily prices often show volatility clustering, regime shifts, and nonlinear relationships. These features have been studied since the early ARCH/GARCH literature in econometrics [1,2]. In China's A-share market, daily CSI 300 returns and volatility are further affected by policy changes, a large share of retail investors, and various liquidity frictions. These factors make simple linear models less effective. At the same time, deep sequence models such as LSTM variants can learn time dependence directly from sequences and have been used to predict both returns and volatility in financial markets [3,4].

Two key strands of literature are especially relevant. The first is about volatility-aware hybrids, which bring conditional volatility into deep sequence models. For example, LSTM–GARCH hybrids feed GARCH-based volatility estimates into the sequence network and often improve error measures, while also capturing leverage and asymmetry in volatility [5]. A more advanced line of work combines several GARCH-family models (e.g., EGARCH, GJR-GARCH, APARCH) in a "MULTIGARCHLSTM" structure and reports stable gains in daily prediction tasks [6].

The second strand is about attention-enhanced recurrent models and multi-scale processing. Attention helps the model focus more on informative time steps in noisy time series. Multiscale local attention, for example, can better fuse fine-grained patterns over different horizons [7].

A different family of models treats the problem as a tabular learning task instead of a sequence-learning one. Tree-based ensembles such as Random Forest, Gradient Boosting, XGBoost, and AdaBoost often work very well when they are given well-designed lagged features and technical indicators. This has been shown in many applications in operations research and forecasting [8,9]. These models can capture nonlinear relationships in tabular data, but they only encode time order in an indirect way, which may limit their use of sequential structure at daily horizons. More recently, LSTM–Transformer hybrids with global self-attention have shown strong and robust results across

different financial time series, because they combine local recurrent dynamics with long-range global dependencies in one architecture [10].

In this paper, a pragmatic approach is taken. BiLSTMAttention is used as a lightweight and relatively interpretable baseline. It is more sequence-aware than tree-based ensembles, but simpler than large hybrid stacks or Transformer models. The analysis focuses on realistic evaluation using expanding-window, walk-forward validation. Both discrimination metrics (AUC, PR-AUC, F1, hit ratio) and error metrics (MAE, RMSE) are reported. The discussion also covers how small statistical edges might be turned into tradable signals under transaction costs and changing regimes. The main finding is that there is a modest but repeatable edge in predicting direction but predicted return magnitudes are almost random. This is consistent with weak-form efficiency and the broader literature on daily index predictability [8–10].

Contributions can be summarized in a single paragraph as follows. This study builds a transparent and reproducible CSI 300 pipeline based on a BiLSTM-Attention architecture, with careful preprocessing, feature design, and walk-forward evaluation. The empirical results are placed in the context of three major model families—volatility-aware hybrids, attention-enhanced recurrent networks, and tree-based ensembles—to clarify when sequence awareness and attention add value. On top of this setup, a tradability-oriented framework is introduced that links calibrated probability thresholds to position sizing and turnover limits and clearly separates statistical significance from economic significance at the daily frequency.

2. Related Work

Volatility modeling in financial time series has long used ARCH/GARCH-family models to describe clustering and asymmetry in conditional variance [1,2]. With the rise of deep learning, researchers began to use sequence models that reduce the need for heavy manual feature engineering. Early work shows that LSTM-based architectures can improve direction or level prediction compared with linear baselines in different markets and assets [3,4].

Integrative hybrids combine deep learning with classical volatility models. LSTM–GARCH models, for example, inject volatility dynamics into recurrent networks and often improve error measures and performance stability across regimes [5]. More broadly, MULTIGARCHLSTM uses several GARCH-type conditional volatilities as extra input channels to an LSTM. This design has produced consistent improvements in R^2, RMSE, MAE, and MAPE on daily futures and index data [6].

Attention mechanisms further improve recurrent models by assigning different weights to time steps within the input window. Multiscale local attention models, such as BiLSTMMLAM, have reduced prediction errors across many time-series datasets. Ablation studies show that bidirectional processing, multiscale fusion, and attention each add independent gains [7]. At the same time, LSTM–Transformer hybrids such as LSTMmTransMLP combine local LSTM processing with global self-attention and have reported strong robustness across different assets (indices, large-cap stocks, cryptocurrencies) with mostly fixed hyperparameters [10].

Parallel work in operations research and applied forecasting has repeatedly shown that tree-based ensembles (RF, Gradient Boosting, XGBoost, AdaBoost) tend to outperform classical statistical baselines when given rich tabular features [9]. As a result, they are often used as strong non-sequence baselines for financial classification and regression problems.

Given this background, BiLSTMAttention is used as a simple and interpretable sequence baseline, tree-based ensembles serve as tabular comparators, and volatility-aware hybrids and LSTM–Transformer models are treated as promising future extensions for robustness.

3. Methods

The analysis uses daily CSI 300 index data from 2013 to 2022 (about 2430 trading days). The raw variables include open, high, low, close, previous close, absolute and percentage returns, volume, and amount. From these, several groups of technical features are constructed: trend indicators (5- and 10-day simple and exponential moving averages), momentum and oscillator features (14-day RSI, MACD and its signal line, and 5- and 20-day momentum), volatility and band features (5- and 20-day rolling volatility and Bollinger band position), and short-term dependency encoders (recent return and momentum lags). Erroneous observations, such as zero volume on normal trading days, are removed, and all features are standardized using statistics from the training set to avoid look-ahead bias. Two prediction targets are considered. The first is a direction classification task where the label is the sign of the next-day return, with a ±0.1% neutrality band to reduce label noise. The second is a magnitude regression task where the target is the 3-day average future return, used to study short-horizon return predictability. The sample is split chronologically into training (2013-02-05 to 2019-12-31), validation (2020-01-02 to 2020-12-31), and test (2021-01-04 to 2022-12-29) periods. An expanding-window walk-forward protocol is adopted: an initial window is used to fit the model and select thresholds and hyperparameters on the validation fold, then the window is rolled forward and performance is evaluated on the next segment, and this process is repeated over the full sample. Inputs to the neural architecture are 30-day lookback sequences over the multivariate feature set, which are processed by a BiLSTM layer with 64 units and sequence output, a temporal attention mechanism that applies softmax-weighted aggregation over time steps, a dropout layer with rate 0.2, and a 16-unit ReLU-activated dense layer. Task-specific output heads then use a sigmoid activation for the classification task and a linear activation for the regression task. Optimization relies on the Adam algorithm, early stopping based on validation loss, and mild L2 and dropout regularization to mitigate overfitting, with a unidirectional LSTM with 50 units used as a minimal sequence-modeling baseline.

4. Results

4.1. Dataset and Evaluation Overview

Under the above protocol, the BiLSTMAttention model is evaluated across walk-forward windows for both tasks. The main interest is the ability to discriminate direction at the next-day horizon. Regression results are also reported to study the common gap between direction and magnitude predictability at the daily frequency [8–10].

4.2. Directional Classification

Across walk-forward windows, the BiLSTMAttention model achieves average AUC ≈ 0.620, PR-AUC ≈ 0.645, F1 ≈ 0.683, and hit ratio ≈ 0.590, which is clearly above the random hit rate of 0.5. The tuned decision threshold is around 0.516 and yields a maximum F1 ≈ 0.761. This suggests that a more conservative cutoff can extract higher-confidence signals, but at the cost of fewer trades.

On the final test window (2022), the ROC curve in Figure 1 shows an AUC of 0.604. This is clearly above the no-skill diagonal, but still far from perfect discrimination. It supports the idea that daily index dynamics contain weak but persistent directional structure. In the ROC curve, the vertical axis is the true positive rate, and the horizontal axis is the false positive rate. The dashed diagonal line represents random guessing.

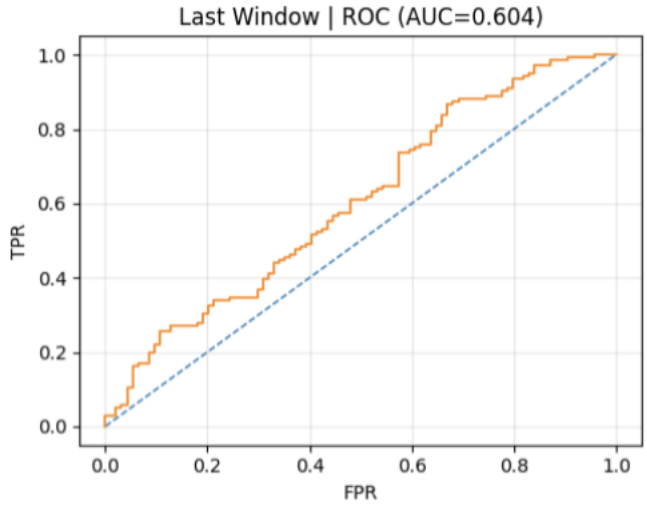

Fig. 1 ROC curve of the BiLSTM Attention model on the final test window, AUC = 0.604.

4.3. Return Magnitude Regression

For the 3-day average return, the model achieves MAE ≈ 0.01223 and RMSE ≈ 0.01507. The sign-hit ratio of the regression outputs is about 0.492 across walk-forward windows, which is very close to a nave or random benchmark. The small variation of these metrics across folds suggests that the regression performance is stable but weak. On a daily horizon, the model has almost no useful information about return size.

This contrast—directional metrics noticeably above chance but magnitude metrics close to random—is consistent with previous LSTM-based studies and volatility-aware hybrids. In many markets, it is somewhat easier to extract weak directional signals than to accurately forecast return size in the presence of high noise and frequent regime changes [3–6].

4.4. Baselines and Qualitative Comparisons

The plain unidirectional LSTM baseline has an AUC of about 0.51, which is basically noise. This highlights the benefit of using both bidirectionality and temporal attention for short-horizon direction prediction.

The literature also suggests that LSTM–GARCH and broader MULTIGARCHLSTM models can further improve regression performance by adding conditional volatility as extra inputs, and they often show better error metrics [5,6]. Tree-based ensembles (RF, XGBoost, Gradient Boosting, AdaBoost) remain strong tabular baselines when they receive rich lagged features and indicators and often beat classical statistical models in comparative studies [8,9]. Finally, LSTM–Transformer hybrids have achieved robust results across different markets by adding global self-attention [10].

5. Discussion

The discussion focuses on three themes: the gap between statistical and economic value, the difference between direction and magnitude predictability, and possible methodological extensions and evaluation choices.

First, the empirical pattern is clear. There is a small but repeatable edge in predicting direction, while return magnitudes are almost random. Turning this edge into investable alpha is not straightforward. A threshold-based trading rule could act only when the predicted probability of an upward move is higher than a tuned cutoff. This focuses on high-confidence signals and helps control turnover. Position sizing and risk control can then scale exposures using predicted probabilities, combined with volatility targeting, turnover caps, and drawdown limits, to keep realized performance

more stable across regimes. In a broader portfolio context, even though this study uses a single index, a two-stage pipeline could be used: first a predictive filter that pre-selects candidate opportunities, then an allocation layer that concentrates capital where the signals are strongest.

However, AUC gains at daily horizons are small. Transaction fees, slippage, and capacity limits can easily absorb or reverse the statistical edge. Therefore, discrimination metrics cannot be directly translated into economic significance. Instead, explicit, cost-aware backtests are needed to judge economic value.

Second, direction tends to be easier to predict than magnitude for several reasons. Daily returns are heavy-tailed and strongly driven by shocks. Even if the sign of the move is slightly predictable, the size is often dominated by noise. Many technical indicators, such as moving averages, momentum, and band positions, are also highly collinear, which limits their ability to support precise size forecasts and pushes the model to focus on coarse directional patterns instead. In addition, policy changes and macro shocks continuously change the relationship between inputs and outputs. In such a non-stationary environment, a binary direction label is more robust to misspecification than a continuous magnitude target and is therefore easier to learn.

Third, several methodological extensions look promising. One is volatility-aware augmentation. Here, rolling EGARCH, GJR-GARCH, or APARCH conditional volatilities could be added as extra channels to the BiLSTMAttention encoder or integrated through a MULTIGARCHLSTM block that feeds estimated volatility states into the network. Another direction is the use of global attention to better handle regime shifts. For example, inserting a small Transformer block after the recurrent encoder, in an LSTM–Transformer–MLP style model, may capture longer-range dependencies and improve probability calibration near regime changes, which matches recent work on volatility-aware and attention-based hybrids. The third extension is multi-scale temporal fusion. In this case, each input window is decomposed into several temporal scales, local attention is applied within each scale, and the resulting representations are fused. Ablation results in related work suggest that bidirectionality, multi-scale fusion, and attention each add robustness.

From an evaluation point of view, disciplined procedures are essential. Expanding-window, walk-forward splits should be kept to respect time order and changing regimes. Reporting should always include both discrimination metrics (ROC and PR curves, F1, hit ratio) and error metrics (MAE, RMSE), so that directional and magnitude performance can be assessed together. Visualizing attention weights over time can support interpretability and help identify where the model fails in certain regimes. For any trading-related claims, cost-aware backtests must include explicit assumptions about transaction costs, turnover and drawdown limits, and clear entry and exit rules that correspond to specific operating points on the ROC and PR curves. Only then is the link between statistical discrimination and realized PnL transparency.

Several limitations also need to be mentioned. This study focuses on a single index, so applying the approach to other indices or asset classes might require new features, horizons, and regularization settings. The feature set is mainly based on technical price and volume variables. Adding macroeconomic variables, options-implied information, or text and sentiment features—similar to hybrid GARCH–deep learning or sentiment-enhanced frameworks—could enrich the signal space but would also raise model complexity and data requirements. Finally, the study stops at statistical evaluation and does not include real or paper trading simulations. Future research should extend the pipeline to capacity- and cost-aware backtests with realistic execution assumptions, so as to narrow the gap between weak statistical signals and robust, implementable trading strategies.

6. Conclusion

Using a transparent BiLSTMAttention pipeline on CSI 300 daily data, this paper finds a modest but consistent directional edge (average AUC ≈ 0.62; final-window AUC $= 0.604$) and almost random magnitude predictability (MAE ≈ 0.012; RMSE ≈ 0.015). These results match the idea of weak-form

efficiency at daily horizons, and agree with broader evidence that sequence-aware deep models can extract limited directional regularities, while accurate size forecasts remain very difficult.

From a methodological perspective, BiLSTMAttention offers a reasonable trade-off between performance and simplicity. It can later be extended with volatility-aware channels, multiscale attention, or global self-attention to improve robustness. From a practical perspective, turning small statistical edges into real economic value requires threshold-based deployment, risk-aware position sizing, and tight cost controls.

Future work will first add volatility-aware channels and LSTM–Transformer hybrids as robustness baselines. It will then extend the prediction horizon to 5–10 days to study the stability of directional and magnitude signals. Next, it will explore multimodal features, such as macroeconomic variables, options-implied measures, and selected text-based information. Finally, it will link calibrated predictive probabilities to portfolio construction rules under explicit transaction cost assumptions, turnover limits, and drawdown controls.

References

[1] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 1982, 50(4), 987–1007.

[2] Bollerslev T. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 1986, 31(3), 307–327.

[3] Fischer T, & Krauss C. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 2018, 270(2), 654–669.

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

[5] Kim H Y, & Won C H. Forecasting the volatility of stock price index: A hybrid model integrating LSTM with GARCH. Expert Systems with Applications, 2018, 103, 25–37.

[6] Pan H, Tang Y, & Wang G. A stock index futures price prediction approach based on the MULTIGARCHLSTM mixed model. Mathematics, 2024, 12, 1677.

[7] Yu F, Tong Q, et al. BiLSTMMLAM: A multiscale feature fusion model with local attention for time series prediction. Sensors, 2024, 24, 3962.

[8] Mitra A, Jain A, Kishore A, & Kumar P. A comparative study of demand forecasting models for a multichannel retail company: A novel hybrid machine learning approach. Operations Research Forum, 2022, 3, 58.

[9] Ejaz A, Ahmed A D, et al. Stock price prediction using machine learning and deep learning methods during COVID-19. European Journal of Management and Business Economics, 2022, 31(3), 234–252.

[10] Khan M U G, Yu H, Tariq M, Javed M A, & Baik S W. LSTMmTransMLP: A hybrid model for robust financial time series prediction. Sci, 2025, 7(7), 1–22.