AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3039

CASCADED GLOBAL–LOCAL REPRESENTATION LEARNING FOR FINANCIAL TIME-SERIES FORECASTING

Hao Wu
School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, Anhui, China.

Abstract: Forecasting financial indices remains difficult because market observations combine persistent movements, short-lived disturbances, nonlinear interactions, and substantial noise. A single recurrent architecture may retain nearby temporal information yet fail to represent widely separated dependencies, whereas an attention-based encoder is effective at relating distant positions but does not by itself guarantee detailed sensitivity to local sequence dynamics. This paper reformulates the forecasting pipeline as a cascaded global–local learning problem. A Transformer encoder first converts normalized price windows into contextual representations through positional encoding, multi-head self-attention, residual normalization, and a feedforward sublayer. Those representations are then processed by a bidirectional long short-term memory network, so forward and reverse recurrent states refine the global context before a dense prediction head produces the output. The design was assessed on daily closing-price series for the S&P 500, Dow Jones Industrial Average, and Nasdaq Composite over 2 September 2003–13 July 2023. Preprocessing included interpolation of missing observations, interquartile-range screening of outliers, first differencing, min–max scaling, and windowed sample construction. Tests against recurrent, bidirectional recurrent, feedforward, Informer, and temporal-convolution baselines show that the hybrid system delivers the strongest overall error and goodness-of-fit profile across the three markets. The findings indicate that passing attention-derived context into a bidirectional memory module offers a practical means of combining long-horizon structure with local temporal variation, although computational cost remains relevant for latency-sensitive trading applications.
Keywords: BiLSTM; Transformer; Financial time series; Index forecasting; Feature fusion

1 INTRODUCTION

The expansion of financial data has changed both the scale and the character of index forecasting. Price histories are now observed together with trading activity and other indicators, but greater data availability does not remove the fundamental modelling problem. Market trajectories are shaped simultaneously by macroeconomic conditions, policy interventions, firm-level events, and investor responses. Their combined effect is a signal that is noisy, nonlinear, high dimensional, and often nonstationary [1]. Models designed around stable linear relations therefore struggle to describe how financial systems evolve, particularly when the mapping between recent observations and future values changes across market regimes.

Deep learning provides an alternative because hierarchical representations can be estimated directly from raw or lightly engineered sequences. Historical closing prices, transaction volume [2], and financial indicators [3] may be combined to expose patterns that fixed-form forecasting equations tend to miss. Even within deep learning, however, the relevant information is distributed over more than one temporal scale. Some variations are local and depend on the immediately preceding observations; others unfold over long intervals and reflect broad trends or economic cycles. An effective forecasting system must recover both without becoming overly sensitive to market noise.

Recurrent neural networks address sequence order directly, while long short-term memory (LSTM) units use gated state updates to reduce the vanishing-gradient difficulty associated with conventional recurrent networks [4]. Their sequential computation is nevertheless costly to parallelize, and a single forward pass may not fully exploit context on both sides of a position. Bidirectional LSTM (BiLSTM) networks mitigate the latter issue by learning forward and reverse representations. Transformer encoders take a different route: self-attention relates any sequence position to any other position and permits parallel computation, giving the architecture a natural advantage in global dependency modelling [5]. Neither family is universally sufficient. Recurrent models are strong at preserving local continuity; attention models are better suited to broad contextual association.

This complementarity has motivated hybrid forecasting systems [6-7]. Recent work has also shown that attention and recurrent components can cooperate in financial prediction pipelines [8], while combinations of convolutional and recurrent processing have proved useful for other physiological time-series tasks [9]. Many existing fusions, though, couple components in a fixed manner and provide limited flexibility in how global and local features are handed from one stage to the next. The present study adopts a cascaded dual-channel feature-fusion strategy: the Transformer encoder first establishes sequence-wide context, and the BiLSTM then reinterprets that contextual signal in both temporal directions. The resulting representation is mapped to the target through a fully connected output layer.

The empirical evaluation uses three influential United States equity indices—the S&P 500, the Dow Jones Industrial Average (DJIA), and the Nasdaq Composite (IXIC)—and compares the proposed architecture with nine alternatives. Mean absolute error (MAE), mean squared error (MSE), and the coefficient of determination (R^2) are calculated

exclusively on the test partitions. The reported aggregate comparison states that, relative to a standard LSTM, the hybrid model lowers MSE by 38% and MAE by 27% on average and produces marked gains in R^2. Beyond aggregate scores, the evaluation considers turning-point sensitivity, behaviour under heterogeneous volatility, inference latency, and memory demand. The rest of the paper develops the architecture, documents the empirical protocol, examines the numerical results, and closes with limitations and directions for further research.

2 MODEL FORMULATION AND ARCHITECTURE

2.1 Bidirectional Recurrent Representation

Financial sequences rarely move with uniform stability. A conventional LSTM reads observations in one direction and can therefore underuse information that becomes clearer when a window is viewed from both ends. BiLSTM processing places two recurrent chains over the same input: one advances chronologically, whereas the other traverses the window in reverse. At time index a, the forward state summarizes the preceding context and the backward state contributes information from the opposite direction. Concatenating the two states yields a representation that is more complete than either path alone. Figure 1 illustrates this paired flow.

$$h_a^f = o_a^f \times tanh(C_a^f) \quad (1)$$

$$h_a^b = o_a^b \times tanh(C_a^b) \quad (2)$$

$$H_a = [h_a^f; h_a^b] \quad (3)$$

Here, C_a^f and C_a^b are the cell states of the forward and reverse LSTM chains, while o_a^f and o_a^b denote their output-gate responses. H_a is the fused hidden representation. The formulation retains the gated memory behaviour of an LSTM, but it supplies the downstream predictor with two complementary readings of each input window.

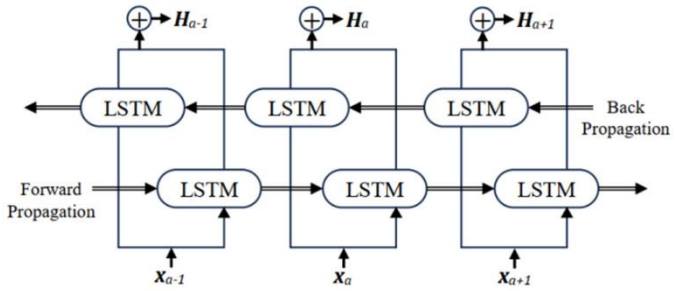

Figure 1 Bidirectional LSTM Architecture and Information Flow

2.2 Cascaded Transformer–BiLSTM Fusion

The standard Transformer is built as an encoder–decoder system for sequence transformation. Financial point forecasting does not require a full generative decoder, so only the encoder is retained here. Raw input variables are first projected into the encoder feature space and combined with positional information. Inside every encoder block, multi-head self-attention identifies relations among positions, residual connections and normalization stabilize the representation, and a position-wise feedforward network increases nonlinear capacity. Instead of sending the encoder output to a Transformer decoder, the model routes it into the BiLSTM stage. A dense layer, a linear projection, and the stated output activation then convert the recurrent features into predictions. This ordered handoff is shown in Figure 2.

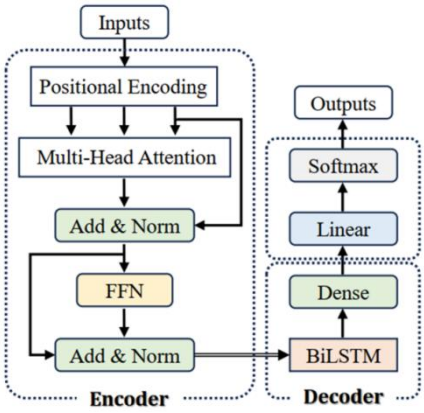

Figure 2 Structure of the Proposed BiLSTM–Transformer Cascade

Let Q, K, and V be the query, key, and value matrices for an encoder layer. Attention head p computes a weighted combination of values from the scaled query–key similarity matrix. The head outputs are concatenated and transformed by a learned matrix W^g. With d_k denoting key dimensionality, the operations are written as follows.

$$MultiHead(Q,K,V)=Con[h_1,h_2,\ldots,h_p]W^g \quad (4)$$

$$h_p=A(Q_p,K_p,V_p)=softmax(Q_pK_p^T/\sqrt{d_k})V_p \quad (5)$$

The attention output is passed through the feedforward sublayer. For input φ, weight matrices W_1 and W_2, and bias vectors a_1 and a_2, the transformation is:

$$FFN(\varphi)=ReLU(\varphi W_1+a_1)W_2+a_2 \quad (6)$$

After residual processing and normalization, the encoder representation W becomes the sequence supplied to the bidirectional recurrent block. The recurrent state at step t depends on W, the BiLSTM parameter matrix U_p, the preceding hidden output Y_p^{t-1}, and the BiLSTM hyperparameter set E_p:

$$Y_p^t=BiLSTM(W,U_p,Y_p^{t-1},E_p) \quad (7)$$

The last stage reduces dimensionality and applies the required activation to obtain the forecast. Architecturally, the sequence of operations assigns different responsibilities to the two main components. Self-attention establishes a global view of the input window, while the BiLSTM examines how that contextualized representation develops in both directions. The fusion is therefore cascaded rather than a simple late averaging of independent predictions.

3 EMPIRICAL DESIGN

3.1 Data Construction and Preprocessing

The empirical data were obtained from the Choice Financial Database and cover three major United States stock indices: S&P 500, DJIA, and IXIC. A common observation interval—2 September 2003 through 13 July 2023—was imposed so that results across markets would be comparable. The interval spans several distinct regimes, including the global financial crisis in 2008 and the COVID-19 market shock in 2020. The analysed variable is the unadjusted daily closing price. Dividends, stock splits, and other corporate actions were not back-adjusted. Each index contributes roughly 5,000 observations, giving approximately 15,000 records in total. Chronological partitions allocate 80% of each series to training and 20% to testing.

Preprocessing was designed to address incompleteness, extreme observations, scale differences, and nonstationarity. Missing entries were filled by linear interpolation. Outliers identified with the interquartile-range rule were replaced by median values. Each price series was then first-differenced and rescaled to the [0, 1] interval. If x is an observation and x_{min} and x_{max} are the minimum and maximum of the relevant data range, min–max normalization is defined by Equation (8). A sliding window of 16 observations, advanced one step at a time, was used to create model inputs.

$$x_scaled=(x-x_{min})/(x_{max}-x_{min}) \quad (8)$$

3.2 Comparator Models and Training Settings

Nine baseline architectures were used to test whether the proposed cascade provides an advantage beyond standard sequence learners. The comparison set contains a feedforward neural network (FNN) [10], a recurrent neural network (RNN), a bidirectional RNN, LSTM [11], BiLSTM [12], a gated recurrent unit (GRU), a bidirectional GRU, Informer, and a temporal convolutional network (TCN). Settings for Informer and TCN followed their respective publications and were adapted to the present data. Other hyperparameters were chosen through grid search and cross-validation on the training partition, with prediction quality and training efficiency considered jointly (Table 1).

Table 1 Hyperparameter Configuration Used in the Comparative Experiments

Model group	Parameter	Value
Transformer	Hidden dimension	64
Transformer	Number of attention heads	4
Transformer	FFN hidden units	128
BiLSTM/LSTM	Hidden units	32
BiLSTM/LSTM	Dropout rate	0.1
Other	Gradient-clipping threshold	1.0
Other	Epochs	100
Other	Learning rate	0.001

For the classical baselines, the shared network layout contained three hidden layers with 64, 32, and 16 units. ReLU served as the activation function, and Adam optimization used $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 0.001. The proposed architecture used a Transformer encoding dimension of 64, four attention heads, a BiLSTM hidden size of 32, 128 hidden units in the feedforward sublayer, dropout of 0.1, a gradient-clipping threshold of 1.0, and no more than 100 training epochs.

3.3 Evaluation Criteria

All performance statistics were computed on the held-out test sets. MAE summarizes absolute deviation, MSE gives additional weight to large errors, and R^2 measures the proportion of variation explained by the predictions. Let x_i be the observed value, $\hat{x}_i$ the corresponding estimate, $\bar{x}$ the mean of observed test values, and n the number of test samples. The three criteria are:

$$MAE=(1/n)\Sigma_{i=1}^{n}|x_i-\hat{x_i}| \quad (9)$$

$$MSE=(1/n)\Sigma_{i=1}^{n}(x_i-\hat{x_i})^2 \quad (10)$$

$$R^2=1-[\Sigma_{i=1}^{n}(x_i-\hat{x_i})^2/\Sigma_{i=1}^{n}(x_i-\bar{x})^2] \quad (11)$$

4 RESULTS AND DISCUSSION

4.1 Cross-Market Predictive Performance

All experiments used the 16-step sampling window described above. Tables 2–4 reproduce the test-set results for the three indices. Taken as a whole, the hybrid model provides the most consistent balance of low errors and high R^2. Its advantage is visible across markets with different constituent structures and volatility profiles, suggesting that the cascaded representation is not tied to a single index. The numerical values below are retained exactly from the experimental record.

Table 2 Predictive Results on the S&P 500 Dataset

Model	MAE	MSE	R^2
FNN	0.01425	0.00036	0.54903
RNN	0.01605	0.00004	0.49749
BiRNN	0.01517	0.00040	0.51437
LSTM	0.02073	0.00073	0.35421
BiLSTM	0.01994	0.00071	0.40221
GRU	0.01224	0.00031	0.57494
BiGRU	0.01826	0.00059	0.44807
Informer	0.01455	0.00034	0.53121
TCN	0.01531	0.00042	0.05002
Ours	0.01128	0.00025	0.60144

Table 3 Predictive Results on the DJIA Dataset

Model	MAE	MSE	R^2
FNN	1.54554	4.25311	0.47873
RNN	1.91432	5.46021	0.42019
BiRNN	1.23305	2.77725	0.54741
LSTM	1.93787	6.13278	0.39477
BiLSTM	1.46902	3.76762	0.49358
GRU	1.90172	5.65019	0.41787
BiGRU	1.49552	3.79167	0.49112
Informer	1.11699	2.61088	0.58025
TCN	1.44290	3.62073	0.48876
Ours	1.11672	2.59975	0.58418

Table 4 Predictive Results on the IXIC Dataset

Model	MAE	MSE	R^2
FNN	1.48457	3.52967	0.46709
RNN	2.07437	6.15389	0.38490
BiRNN	1.37267	2.69122	0.52130
LSTM	1.63022	4.37994	0.44942
BiLSTM	1.58221	3.81701	0.45902
GRU	2.12519	6.69163	0.36953
BiGRU	1.43098	3.14887	0.48664
Informer	1.17517	2.34984	0.56595
TCN	1.41756	3.09012	0.49257
Ours	1.12666	2.18065	0.58953

On the S&P 500, the hybrid model records an MAE of 0.01128 and an R^2 of 0.60144, both better than the corresponding values for every listed comparator. Its MSE is 0.00025. The accompanying interpretation identifies the hybrid as the leader on both error measures; Table 2 nevertheless contains an apparent inconsistency because the RNN MSE is printed as 0.00004 while its MAE and R^2 are weaker. That entry should be verified against the underlying experiment before publication. The reported MAE reduction relative to BiGRU is approximately 30%. Against Informer, the stated MSE improvement is 12.6%, supporting the view that recurrent refinement adds value after global attention encoding.

The DJIA results also favour the cascade. Its MAE of 1.11672 narrowly improves on Informer's 1.11699, while its MSE of 2.59975 and R^2 of 0.58418 are the best values in the table. Directionality matters even among conventional recurrent networks: BiRNN attains an MAE of 1.23305, substantially below the RNN result of 1.91432. Yet the bidirectional baselines do not match the combined attention–memory architecture overall. The analysis places the hybrid model's MAE advantage over BiGRU at about 25% for this index.

The IXIC series presents another demanding case because technology-heavy constituents can respond sharply to new information. The proposed model reaches an MAE of 1.12666, an MSE of 2.18065, and an R^2 of 0.58953. Compared with BiRNN, the MAE is lower by 17.9%; compared with BiGRU, the reported reduction is roughly 20%. Informer is the nearest competitor, but its three statistics remain less favourable. These outcomes suggest that the attention stage captures broad shifts while the bidirectional recurrent stage improves the local placement of the forecast.

4.2 Interpretation of the Global–Local Interaction

The advantage of the cascade is most plausibly explained by division of labour between the modules. A standard LSTM stores information through gated recurrence, yet its local memory path can lose sensitivity as dependencies stretch over long windows. Self-attention avoids that strictly sequential route and can assign weight directly to distant positions. Once those global relations have been encoded, BiLSTM processing introduces forward and backward local structure. Predictions near turning points consequently track the observed trajectory more closely than those of a single LSTM in the experiments. The reported R^2 improvement is about 15% relative to BiLSTM for the forecasting task, interpreting the gain as evidence that economic-cycle information is represented more effectively.

Performance is not uniform across every metric and dataset. The MSE advantage is described as especially clear for the S&P 500, which may indicate greater resistance to extreme deviations. On the other indices, the margin is less stable. Attention may not always concentrate sufficiently on the most consequential dates, and each index combines constituents with different volatility and event sensitivity. The daily sampling interval also hides intraday movement. That loss of high-frequency information is particularly consequential for technology-focused markets, where abrupt responses can occur between daily closes. Accordingly, better aggregate accuracy should not be interpreted as proof that the model has completely solved sudden-shift adaptation.

The broader comparison reinforces this interpretation. Unidirectional RNN and LSTM models remain vulnerable when long-sequence gradients weaken. FNN contains no explicit temporal-state mechanism and therefore produces relatively large errors in several tasks. TCN offers high computational efficiency, but the experiments indicate weaker long-range representation. Bidirectional recurrent baselines improve contextual access, while Informer strengthens long-sequence modelling; the proposed cascade is the only evaluated architecture that applies both operations in sequence.

4.3 Computational Requirements and Deployment Limits

Accuracy is accompanied by a measurable deployment cost. With an RTX 3080 and a batch size of 64, the proposed model uses an average of 7.2 ms for one prediction and occupies 349 MB of GPU memory. These requirements are acceptable for many analytical forecasting settings, but high-frequency trading systems operate under tighter latency constraints. The additional attention operations and bidirectional recurrence may therefore restrict real-time use even when predictive scores are favourable. Lightweight attention, parameter pruning, or related compression strategies offer reasonable paths for reducing this burden without abandoning the global–local design.

5 CONCLUSION

This study reorganizes financial index forecasting around a staged interaction between global attention and local bidirectional memory. The Transformer encoder first represents long-range relations across the input window; the BiLSTM subsequently refines that contextual signal by reading temporal structure in both directions. A dense prediction head completes the mapping to the forecast. Experiments on the S&P 500, DJIA, and IXIC show that the architecture achieves the strongest overall performance among the reported feedforward, recurrent, bidirectional, Informer, and temporal-convolution alternatives. The evidence supports the central premise that multi-head attention and gated memory are complementary rather than interchangeable mechanisms for noisy financial sequences.

The findings also define the boundary of that conclusion. Daily data cannot reveal intraday shocks, heterogeneous index composition affects the stability of metric gains, and the combined model requires more computation than lighter baselines. Future extensions should incorporate forward-looking information, including option-implied volatility surfaces, to strengthen anticipation of changing risk. Structural causal models are another promising addition because they could distinguish genuine drivers of volatility from variables that are merely correlated with market disturbances. Pursuing these directions may improve interpretability, risk identification, and deployment efficiency while preserving the extensible global–local forecasting paradigm developed here.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Cavalcante R C, Brasileiro R C, Souza V L, et al. Computational intelligence and financial markets: A survey and future directions. Expert Systems with Applications, 2016, 55: 194–211.

[2] Do H X, Brooks R, Treepongkaruna S, et al. How does trading volume affect financial return distributions? International Review of Financial Analysis, 2014, 35: 190–206.

[3] Iacuzzi S. An appraisal of financial indicators for local government: a structured literature review. Journal of Public Budgeting, Accounting & Financial Management, 2022, 34(6): 69–94.
[4] Cao J, Li Z, Li J. Financial time series forecasting model based on CEEMDAN and LSTM. Physica A: Statistical Mechanics and its Applications, 2019, 519: 127–139.
[5] Yang L, Li J, Dong R, et al. NumHTML: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 11604–11612.
[6] Kabir M R, Bhadra D, Ridoy M, et al. LSTM–Transformer-based robust hybrid deep learning model for financial time series forecasting. Sci, 2025, 7(1): 7.
[7] Mishra A K, Renganathan J, Gupta A. Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture. Engineering Applications of Artificial Intelligence, 2024, 133: 108223.
[8] Li J C, Sun L P, Wu X, et al. Enhancing financial time series forecasting with hybrid deep learning: CEEMDAN–Informer–LSTM model. Applied Soft Computing, 2025: 113241.
[9] Swapna G, Kp S, Vinayakumar R. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Computer Science, 2018, 132: 1253–1262.
[10] Ojha V K, Abraham A, Snášel V. Metaheuristic design of feedforward neural networks: A review of two decades of research. Engineering Applications of Artificial Intelligence, 2017, 60: 97–116.
[11] Waqas M, Humphries U W. A critical review of RNN and LSTM variants in hydrological time series predictions. MethodsX, 2024: 102946.
[12] Siami-Namini S, Tavakoli N, Namin A S. The performance of LSTM and BiLSTM in forecasting time series. 2019 IEEE International Conference on Big Data, 2019: 3285–3292.