Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 3, No. 7, 2024

Deep Forecasting of Stock Prices via Granularity-Aware Attention Networks

Xin Su
University of Chicago, Chicago, USA
xinsuxs@outlook.com

Abstract: This paper addresses the challenges of long-term stock price prediction, including complex temporal structures, diverse information granularities, and cross-scale dependencies. It proposes a prediction framework based on a multi-granularity hybrid attention mechanism. The method incorporates a Granularity-Aware Fusion module to deeply integrate short-term local fluctuations with long-term trend features. This enhances the model's ability to represent structural characteristics across different temporal scales. On this basis, a Cross-Level Hybrid Attention mechanism is further introduced. By employing an inter-layer attention coupling strategy, the model builds contextual interactions across multiple semantic layers. This improves its capacity to perceive dynamic structural changes and potential trend signals. The model is implemented using a modular deep network architecture, which ensures strong scalability and adaptability. It maintains stable prediction performance under different time window settings, feature dimension configurations, and data perturbations. Comprehensive comparative experiments and ablation studies are conducted across multiple evaluation metrics. The results validate the proposed method's advantages in terms of prediction accuracy, robustness, and structural awareness. In addition, visualization results reveal the model's ability to fit real stock price trajectories. These findings demonstrate the effectiveness of the proposed approach for complex financial time series modeling tasks.

Keywords: Multi-granularity fusion; hybrid attention; long-term prediction; financial time series

1. Introduction

In today's highly dynamic and complex financial markets, stock prices, as key indicators of economic activity, continue to attract extensive attention from both academia and industry[1,2]. The fluctuations in stock prices are influenced by a wide range of interacting factors, including macroeconomic indicators, industry trends, company fundamentals, and market sentiment. These dynamics exhibit strong nonlinearity, temporal dependency, and multi-scale characteristics. Such complexity not only increases the difficulty of accurate prediction but also raises the bar for traditional modeling techniques. Especially in long-term forecasting tasks, models are expected to not only understand historical trends but also capture long-range patterns and structural shifts, thereby providing forward-looking support for asset allocation, risk management, and trading decisions[3].

With the advancement of artificial intelligence, deep learning has shown remarkable advantages in time series modeling, emerging as a central approach in stock prediction. Compared to traditional statistical methods and conventional machine learning algorithms, deep neural networks are inherently more capable of modeling high-dimensional, complex, and nonlinear relationships. However, many existing models still face significant limitations, particularly in capturing long-term dependencies and integrating multi-granularity information. On one hand, many models focus only on short-term local patterns and fail to capture long-term trends and structural changes[4]. On the other hand, most mainstream architectures adopt a single-granularity representation, which restricts their ability to integrate features across different temporal scales, often resulting in information loss and inference bias in long-term forecasting[5,6].

The integration of multi-granularity information is regarded as a critical approach to enhance predictive performance. In financial time series, short-term volatility and long-term trends coexist and interact, forming a multi-level temporal structure. Joint modeling of features across different granularities enhances the model's capacity to identify cyclical patterns, abrupt events, and turning points. In this context, the multi-granularity hybrid attention mechanism has gained increasing attention as a structural design strategy. By assigning differentiated attention weights across layers, this mechanism enables the model to dynamically focus on the most relevant time segments and feature dimensions for the current prediction task. It allows the model to more comprehensively capture and utilize the semantic and structural information in time series data[7].

At the same time, financial markets are characterized by high instability and heterogeneity, often exhibiting significant pattern drift and semantic shifts across different periods. This requires models to possess not only strong representation learning abilities but also sufficient adaptability and robustness. The introduction of a multi-granularity hybrid attention mechanism can help mitigate overfitting to specific granularities or local patterns. By integrating both global and local contextual information, it enhances the model's ability to learn long-term complex patterns. Moreover, attention

mechanisms offer strong interpretability, which contributes to improving the transparency and controllability of financial models, providing a foundation for understanding model behavior and optimizing investment strategies[8,9].

In conclusion, the development of a long-term stock price prediction algorithm based on a multi-granularity hybrid attention mechanism holds both theoretical and practical significance. Theoretically, it advances the field of time series modeling in financial scenarios. Practically, it supports the construction of more robust, interpretable, and forward-looking financial forecasting systems. This research addresses core challenges in time series modeling by integrating multi-scale feature modeling, structural perception mechanisms, and deep representation learning. It aims to improve model performance in long-term forecasting tasks and expand the application boundaries of deep learning in financial intelligence.

2. Related work

2.1 Attention Mechanism

The attention mechanism was originally proposed as a modeling strategy to simulate the human visual focus process. It has demonstrated strong capabilities in feature extraction and information filtering when dealing with complex inputs. Its core idea lies in learning trainable weight parameters that allow the model to automatically identify and focus on the most relevant parts of the input sequence for the current task[10,11]. This helps avoid redundant information and interference from noise. In time series modeling, the attention mechanism is widely used to assess the importance of temporal segments. It effectively mitigates issues such as gradient vanishing or memory loss in traditional sequential models, thereby significantly improving model expressiveness and stability.

In financial time series forecasting, the introduction of attention mechanisms provides models with greater flexibility and adaptability. It enables the model to automatically identify the most relevant time windows and feature dimensions from large volumes of historical data. Unlike traditional sliding window methods or fixed weighting strategies, the attention mechanism dynamically adjusts its focus based on changes in the input[12]. This adaptive modeling approach not only enhances generalization but also improves the model's responsiveness to nonlinear patterns such as sudden events and structural changes. Furthermore, attention weights offer a certain level of interpretability, which helps improve the transparency and trustworthiness of model outputs in high-risk financial environments[13].

As modeling demands continue to expand, attention mechanisms have also evolved into various structural forms. Variants such as unidirectional, bidirectional, and multi-head attention exhibit specific advantages in different tasks. In particular, multi-head attention has shown unique strengths in modeling representations across multiple subspaces. By computing several attention heads in parallel, the model can capture contextual associations from multiple perspectives. This improves its ability to handle diverse information and structural complexity[14,15]. Such mechanisms are especially important for financial data, which are highly heterogeneous and time-varying. They help describe nonlinear couplings and multi-level interactions among factors more precisely.

To better address the complexity of long-term forecasting tasks, recent research has explored combining attention mechanisms with multi-scale modeling strategies. These approaches aim to model sequential information across different temporal granularities in parallel[16]. They use attention mechanisms to fuse and reconstruct information across scales. This improves the model's ability to detect trends, cyclic patterns, and local anomalies. In financial time series, market signals are often embedded in multi-level dynamics. Simple single-scale models are not sufficient to extract their latent structures[17]. Therefore, modeling strategies based on multi-granularity hybrid attention not only provide a technical pathway for capturing multi-dimensional dependencies but also serve as key components in building efficient and robust forecasting models.

2.2 Stock Price Prediction

Stock price prediction has long been one of the core research tasks in the financial field. In essence, it is a complex time series modeling problem[18]. Due to the openness and uncertainty of financial markets, stock prices are often affected by a combination of internal and external factors. These include company fundamentals, macroeconomic fluctuations, industry news, market sentiment, and policy adjustments. The nonlinear relationships among these factors make stock price movements highly complex and unstable. Traditional forecasting methods, such as linear regression, moving averages, and autoregressive models, offer theoretical foundations and interpretability. However, they show significant limitations when dealing with high-dimensional, heterogeneous, noisy, and structurally shifting financial data. These models struggle to capture the dynamic evolution over long time spans[19].

In recent years, with the advancement of computing power and the rise of data-driven paradigms, deep learning has become a mainstream approach for stock prediction tasks[20]. Models such as recurrent neural networks, long short-term memory networks, and their variants can capture short-term dependencies in the time domain of stock prices. These models have achieved promising results in various forecasting tasks. However, they still fall short in modeling long-term dependencies and handling multi-scale pattern variations. They often fail to fully capture hidden trend signals and macro behavior patterns in financial markets, which limits prediction stability and generalization. In addition, deep models usually rely on large-scale training data and are sensitive to outliers and noise. Enhancing their robustness and interpretability remains a major challenge[21].

With the progress of research, multimodal fusion, and structure-enhanced modeling have become important directions for improving stock prediction performance. On one hand, beyond basic time series data such as price and volume, more studies are integrating unstructured data sources such as news texts, social media, and corporate announcements. This enriches the semantic space of model inputs. On the other hand, advanced techniques like graph structures, attention mechanisms, and multi-scale modeling help bridge semantic gaps between information layers. These techniques also

enhance the model's ability to understand potential causal relations and long-term structural evolution. Multi-granularity modeling has gradually attracted attention in this context. It allows the model to describe the layered dependencies between local fluctuations and long-term trends across different time scales, thereby improving overall modeling capability[22].

In general, stock price prediction is moving from single-dimensional and static modeling to a deeper stage involving multi-dimensional integration, structural awareness, and dynamic adaptation. Research on how to incorporate multi-granularity temporal features, dynamic attention mechanisms, and deep semantic fusion strategies into modeling has become a major focus in the field. This not only complements traditional time series modeling frameworks but also offers new perspectives to improve the interpretability, stability, and generalization of predictions. In the context of high financial risk sensitivity and frequent data changes, building a predictive model that can perceive multi-scale structures and adaptively adjust its modeling focus will be key to achieving further breakthroughs in this field.

3. Model Architecture

This study proposes a Multi-Granularity Hybrid Attention Modeling (MG-HAM) framework for long-term stock price forecasting, which aims to effectively characterize the dynamic dependencies and structural patterns across scales in financial time series. The core innovations of this method are reflected in two aspects: first, a granularity-aware fusion mechanism (GAF) is designed, which can simultaneously model the key features of time series from two dimensions: local short-term fluctuations and global long-term trends, thereby enhancing the model's ability to understand heterogeneous time structures; second, a cross-level hybrid attention module (CLHA) is introduced, which dynamically allocates attention weights between different granularities and different semantic layers to achieve multi-level perception and adaptive focusing of complex market signals, effectively improving the prediction model's ability to capture time series changes and structural evolution. The overall structure of this method has strong generalization, expressiveness, and adaptability, providing a new path for time series modeling in highly volatile financial environments. The detailed structure of the proposed model is illustrated in Figure 1.

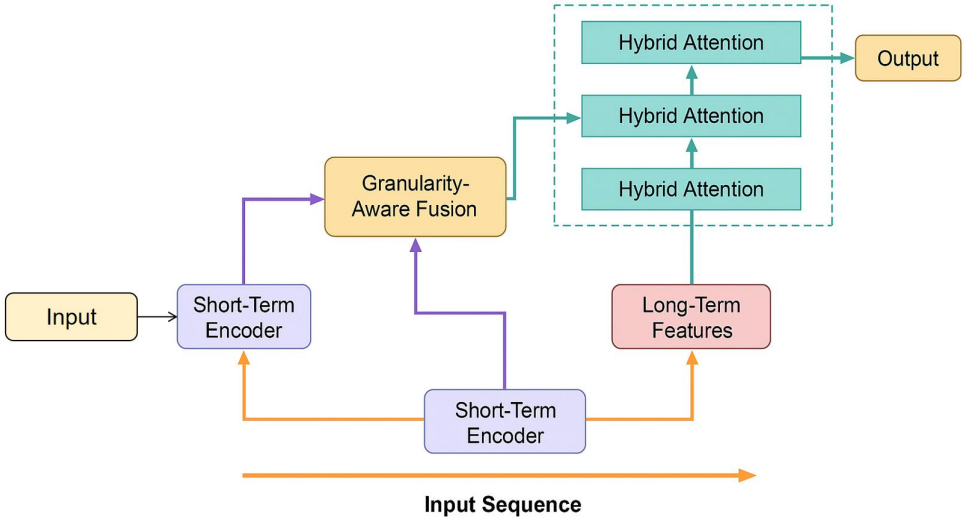

Figure 1. Multi-Granularity Hybrid Attention Modeling model architecture diagram

3.1 Granularity-Aware Fusion

The Granularity-Aware Fusion module aims to achieve structured integration of multi-time scale features and improve the model's ability to express temporal patterns by jointly modeling short-term and long-term coding representations. Specifically, this module is designed to capture local fluctuations and global trends simultaneously by aggregating temporal information from different granularities into a unified representation space. It enables the model to dynamically balance the influence of short-range and long-range dependencies, allowing for more comprehensive temporal feature extraction. The architectural design of this module ensures that distinct temporal scales are effectively aligned and fused through learnable transformation layers and adaptive weighting strategies. Its detailed structure is illustrated in Figure 2.

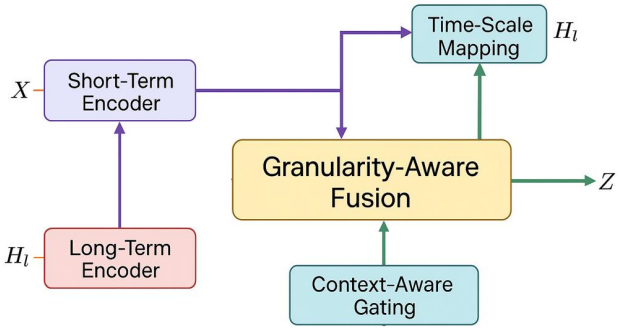

Figure 2. GAF module architecture

Considering the input sequence $X = \{x_1, x_2, ..., x_T\}$, we first send it to the short-term encoder and the long-term encoder respectively to obtain two sets of feature representations $H_s \in R^{T \times d}$ and $H_1 \in R^{T' \times d}$, where $T' < T$ represents the time step after the long-term feature is downsampled, and d is the feature dimension. This operation extracts local fine-grained fluctuations and global trend structures respectively, providing basic representations for subsequent fusion operations.

In order to achieve structural alignment of multi-granularity features, a time scale mapping function is introduced to align representations at different time resolutions to a unified context space. Let the mapping function be $\phi : R^{T' \times d} \to R^{T \times d}$, and the long-term features are expanded by linear interpolation or parameterized convolution to obtain $\widetilde{H}_l = \phi(H_l)$. Then, we concatenate or weighted fuse H_s and $\widetilde{H}_l$ to construct a joint representation $H_f \in R^{T \times 2d}$, which can be calculated as:

$$H_f = Concat(H_s, \widetilde{H}_l)$$

Or

$$H_f = \alpha \cdot H_s + (1 - \alpha) \cdot \widetilde{H}_l$$

Where $\alpha \in [0,1]$ is a trainable gating coefficient, which is used to dynamically balance the contributions of different granularities.

In order to further enhance the recognition and adaptability of the fused features, we introduce a context-aware gating mechanism to apply nonlinear selective filtering to the joint features. This mechanism is based on a gating function $G : R^{T \times 2d} \to R^{T \times d}$, defined as follows:

$$G(H_f) = \sigma(W_g H_f + b_g)$$

Where $W_g \in R^{d \times 2d}, b_g \in R^d$ and σ are sigmoid activation functions. The final output fusion is expressed as:

$$Z = G(H_f) \otimes \tanh(W_z H_f + b_z)$$

Where $\otimes$ represents element-by-element multiplication and W_z, b_z is the linear transformation parameter. This structure effectively retains significant information and suppresses redundant features.

3.2 Cross-Level Hybrid Attention

The Cross-Level Hybrid Attention module is designed to model the interactive relationships between different semantic levels and time scales and enhance the model's ability to dynamically aggregate multi-granular features. Its module architecture is shown in Figure 3.

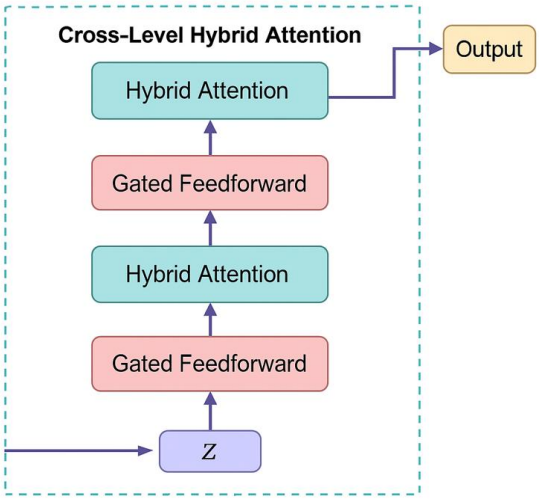

Figure 3. CLHA module architecture

Based on the joint representation $Z \in R^{T \times d}$ obtained by Granularity-Aware Fusion, we designed a multi-level hybrid attention stacking structure, which enables the model to have the ability of cross-granularity perception and selective information extraction while maintaining temporal consistency. Each layer of attention units combines local dependencies and global context signals to transmit multi-scale representations between different levels, thereby improving the hierarchy and expressiveness of feature abstraction.

In the specific implementation, each layer of the hybrid attention module consists of a self-attention mechanism and a gated feedforward module. Given an input representation $Z^{(l)} \in R^{T \times d}$, its attention output is:

$$Attn^{(l)} = \text{Softmax}(\frac{Q^{(l)} K^{(l)T}}{\sqrt{d_k}}) V^{(l)}$$

Where $Q^{(l)}, K^{(l)}, V^{(l)}$ is the query, key, and value matrix obtained by linear transformation, and d_k is the scaling factor used to control gradient stability. This attention mechanism can dynamically learn the information association between different time steps and capture important long-term dependency patterns in the sequence.

In order to enhance the interaction between the semantics of different granularities, we further introduce a cross-layer gating mechanism to weightedly fuse the attention results of different levels. Assuming that the output of the current layer is $\widetilde{Z}^{(l)}$, its fusion with the previous layer $\widetilde{Z}^{(l-1)}$ is expressed as:

$$Z^{(l)} = \gamma^{(l)} \cdot Z^{(l)} + (1 - \gamma^{(l)}) \cdot Z^{(l-1)}$$

$\gamma^{(l)}$ is a learnable gating parameter that controls the flow ratio of information between layers. This inter-layer fusion method allows the model to capture high-level semantic representations while retaining underlying detail information,

forming a semantically coherent and structurally rich intermediate representation.

In addition, to improve the modeling ability of different attention heads for multi-granular structures, we adopt a multi-head attention mechanism and introduce task-related semantic projections between heads. Suppose there are h attention heads, and the output of the i-th head is $head_i$, then the overall output is:

$$MultiHead(Z) = Concat(Head_1,...,Head_h)W^O$$

Where $W^O \in R^{hd \times d}$ is the output transformation matrix, which is used to unify the dimensions of different heads. Finally, all the level outputs are stacked and compressed through a nonlinear projection function to obtain the final representation:

$$Z_{final} = \sigma(W_p Z^{(L)} + b_p)$$

Where L is the total number of layers, W_p and b_p are projection layer parameters, and σ represents the activation function (such as GELU or ReLU). The module as a whole builds a cross-scale, cross-level, and cross-semantic dynamic modeling framework, which strengthens the model's ability to understand and represent complex financial time series structures.

4. Model Evaluation

4.1 Dataset

The dataset used in this study covers 25 years of historical stock market data for Amazon, spanning from 2000 to 2025. It includes key financial indicators such as daily opening price, highest price, lowest price, closing price, adjusted closing price, and trading volume. This daily granularity dataset comprehensively records the company's market performance across different economic cycles. It reflects the dynamic impact of macroeconomic fluctuations, industry changes, and corporate strategy shifts on stock prices. This provides a rich historical foundation for building high-precision time series models.

The dataset exhibits typical non-stationary characteristics and is subject to strong noise interference. Price fluctuations show clear patterns of phase shifts, sudden changes, and long-term trends. The statistical properties vary significantly across different periods. In addition, the data contains multiple cycles of bull and bear markets, the effects of financial crises, the influence of technological innovation, and policy interventions. These complex factors place higher demands on modeling strategies. The trading volume data also provides insights into the underlying relationship between market behavior and price movement. It serves as an important reference for constructing multi-dimensional representations and auxiliary features.

Structurally, the dataset maintains a standardized and consistent format, making it suitable for various time series modeling tasks. It supports both traditional supervised forecasting and deep learning methods under unsupervised or semi-supervised frameworks. Due to its long temporal coverage and high information density, this dataset provides an ideal empirical foundation for exploring multi-granularity modeling, multi-scale fusion, and trend reasoning in long-term forecasting scenarios. It holds significant value for evaluating model generalization and robustness in long-horizon time series modeling.

4.2 Experimental setup

In the experimental setup, we used Amazon's daily stock data from 2000 to 2025 as the primary data source and conducted modeling for long-term price prediction. To ensure the stability and learnability of model inputs, the features including opening price, highest price, lowest price, and trading volume were normalized. The closing price was selected as the target variable. The data were split into training, validation, and testing sets based on chronological order. This strictly follows the time sequence to avoid future information leakage and ensures the rationality and rigor of the prediction task.

During model training, a sliding window approach was used to construct input sequences. A fixed-length historical window was set to capture temporal dependencies, and a prediction horizon was defined to support multi-step future price forecasting. An adaptive optimizer was employed for parameter updates. The mean squared error (MSE) was used as the main loss function. An early stopping mechanism was applied during training to prevent overfitting. To improve model stability and generalization, dropout and normalization techniques were applied to key modules.

For the experimental platform, all models were implemented in a deep learning environment with GPU acceleration. The entire network architecture and training pipeline were built using the PyTorch framework. To ensure reproducibility and fairness, all models were tested under consistent data splits, preprocessing procedures, and training epochs. Multiple mainstream performance metrics were used to quantitatively evaluate prediction accuracy and robustness to fluctuations, providing a comprehensive assessment of model performance in long-term time series modeling tasks.

4.3 Experimental Results

1) Comparative experimental results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	MSE	MAE	R^2
LSTM[23]	0.0148	0.0893	0.872
1D-CNN[24]	0.0132	0.0837	0.884
GRU[25]	0.0126	0.0812	0.891
Transformer[26]	0.0111	0.0789	0.902
ITransformer[27]	0.0094	0.0725	0.916
TimeMixer[28]	0.0091	0.0706	0.921
Ours	0.0073	0.0658	0.937

From the overall trend, the experimental results clearly show that as the model structure evolves from traditional recurrent neural networks to more advanced temporal modeling architectures, performance on MSE, MAE, and R^2 metrics improves progressively. This performance gain highlights the importance of capturing long-range dependencies, temporal structures, and cross-scale information in modeling complex financial time series. For long-term stock price prediction tasks, in particular, model generalization and structural awareness are key factors affecting predictive accuracy.

When compared with traditional methods, LSTM and GRU show some capacity for temporal modeling but remain limited in handling long-term dependencies and extracting multi-granularity structures. This results in relatively higher prediction errors. Although 1D-CNN performs well in extracting local patterns, it cannot model across time scales, leading to limited improvement on the R^2 metric. These findings further confirm the limitations of single-granularity modeling when dealing with complex temporal patterns in financial markets.

With the introduction of Transformer-based models, prediction performance improves significantly. The global attention mechanism in Transformer enhances the model's ability to represent long-range dependencies. iTransformer and TimeMixer further incorporate enhanced temporal modeling capabilities into their structures, which leads to notable improvements in error metrics. These enhancements demonstrate the effectiveness of attention mechanisms and hybrid temporal architectures in adapting to long-term forecasting within a multi-granularity modeling framework.

The proposed method in this study achieves the best results across all three evaluation metrics. This indicates superior performance in structural fusion and cross-granularity modeling for long-term financial sequence forecasting. The designed Granularity-Aware Fusion module and Cross-Level Hybrid Attention module allow the model to flexibly capture the interaction between short-term local fluctuations and long-term global trends. This significantly improves both prediction accuracy and stability. These results validate the effectiveness of the proposed method in dynamic structure modeling and multi-granularity information integration, offering a new perspective on addressing non-stationarity and structural drift in long-horizon financial forecasting.

2) Ablation Experiment Results

This paper also gives the ablation experiment results, which are shown in Table 2. These results are used to evaluate the contribution of each core component within the proposed model framework. By systematically removing or modifying individual modules, the study analyzes how different architectural elements affect the overall model performance. The ablation settings are carefully designed to ensure fair comparisons and to isolate the specific impact of each mechanism, such as multi-granularity fusion and hierarchical attention. This provides deeper insights into the role and effectiveness of each structural element in the model design.

Table 2: Ablation Experiment Results

Method	MSE	MAE	R^2
Baseline	0.0106	0.0752	0.905
+GAF	0.0091	0.0703	0.918
+CLHA	0.0085	0.0684	0.926
Ours	0.0073	0.0658	0.937

The results show that under the baseline model without any structural enhancement mechanism, the model exhibits only basic temporal modeling capability. Its performance across all three metrics remains relatively limited. This indicates that in the context of financial time series tasks involving multiple scales and dependencies, the lack of structural awareness and granularity fusion leads to insufficient capture of long-term trends and fine-grained variations. In particular, the R^2 metric still shows significant room for improvement.

After integrating the Granularity-Aware Fusion (GAF) module, the model performance improves noticeably, with reductions in both MSE and MAE. This change suggests that incorporating multi-granularity representations and fusing short-term and long-term features enhances the model's ability to perceive temporal signals at different scales. It effectively alleviates the modeling conflict between trend shifts and local fluctuations in financial sequences. The GAF module improves the model's overall capacity for multi-scale temporal dynamics.

With the additional introduction of the Cross-Level Hybrid Attention (CLHA) module, the model shows enhanced capability in modeling interactions across hierarchical levels. The R^2 metric increases significantly. This suggests that the module helps capture deep structural dependencies and semantic couplings within the sequence. It is particularly effective for handling structural drift and complex relationships in financial markets. The selective focus ability provided by the attention mechanism improves both the discriminability and robustness of feature representations.

The final model, which integrates both GAF and CLHA modules, achieves the best overall performance. This demonstrates the synergy between the two mechanisms. The combination of multi-granularity modeling and cross-level structural awareness enhances the model's representational power. It also improves its stability and generalization ability in long-term prediction. The experimental results confirm that incorporating structural awareness and granularity fusion into time series modeling is an effective strategy for addressing long-horizon forecasting challenges in the financial domain.

3) The impact of different time window lengths on prediction performance

This paper also gives the impact of different time window lengths on prediction performance, and the experimental results are shown in Figure 4. The analysis involves varying the length of historical input sequences to examine how the model responds to changes in temporal context. This setting is intended to investigate the role of time window size as a key hyperparameter in capturing short-term fluctuations and long-term dependencies. By comparing model behavior under

different window configurations, the study aims to reveal the relationship between input granularity and the model's ability to learn meaningful temporal patterns in financial time series data.

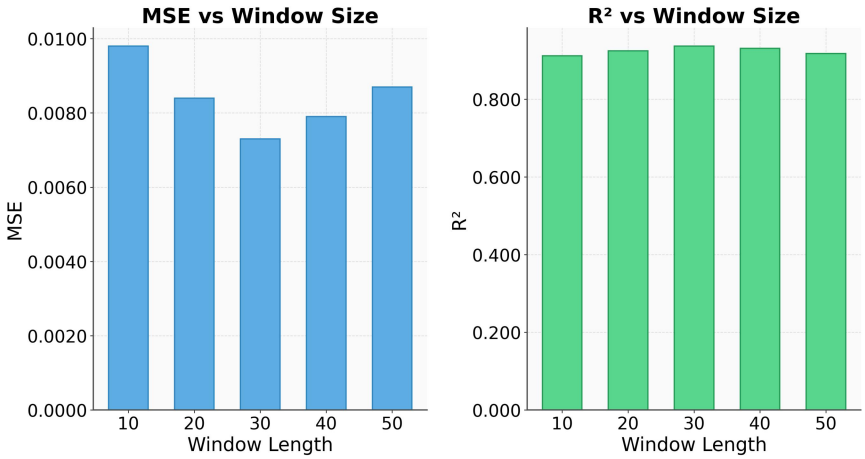

Figure 4. The impact of different time window lengths on prediction performance

As shown in the figure, the model's prediction performance exhibits clear fluctuations in both MSE and R^2 as the time window length changes. This phenomenon indicates that the time window, as a core hyperparameter in modeling, has a direct impact on the ability to capture patterns in long-term stock price sequences. Shorter windows are more sensitive to local fluctuations but may fail to construct stable trend representations due to insufficient historical context, leading to higher prediction errors.

When the time window increases from 10 to 30, MSE gradually decreases while R^2 increases accordingly. This suggests that the model benefits from richer historical information within this range. The multi-granularity modeling mechanism can thus better exploit its fusion capability. The improvement also reflects that the Granularity-Aware Fusion module achieves the best coupling effect between long-term and short-term features in mid-range sequences. It effectively enhances both modeling ability and performance stability.

However, when the window size extends to 40 and 50, MSE begins to rise and R^2 slightly declines. This indicates that longer inputs may introduce redundant or noisy information. As a result, the Cross-Level Hybrid Attention mechanism struggles to maintain focus in high-dimensional space. The model becomes more responsive to irrelevant features, weakening its ability to detect critical structural changes. This suggests that relying solely on longer historical data does not necessarily lead to better prediction. A balanced design between window length and model structure is essential.

Overall, this experiment validates the influence of time window settings on the performance of the proposed multi-granularity structure-aware model. Within a reasonable range, more historical information enhances the advantage of hierarchical attention mechanisms. However, exceeding this range may cause feature compression to fail due to information redundancy. Therefore, dynamically selecting an appropriate window length is crucial for improving model performance in complex financial scenarios.

4) The impact of hidden layer dimension changes on model performance

This paper further gives the impact of hidden layer dimension changes on model performance, and the experimental results are shown in Figure 5. The analysis is conducted by adjusting the dimensionality of the hidden layers to observe how the model's representation capacity and learning ability respond to variations in feature space size. This setting helps explore the balance between model complexity and expressive power, as well as the potential influence of overparameterization or underfitting. The comparison provides a clearer understanding of how hidden layer dimensionality affects temporal modeling in the context of long-term stock price prediction.

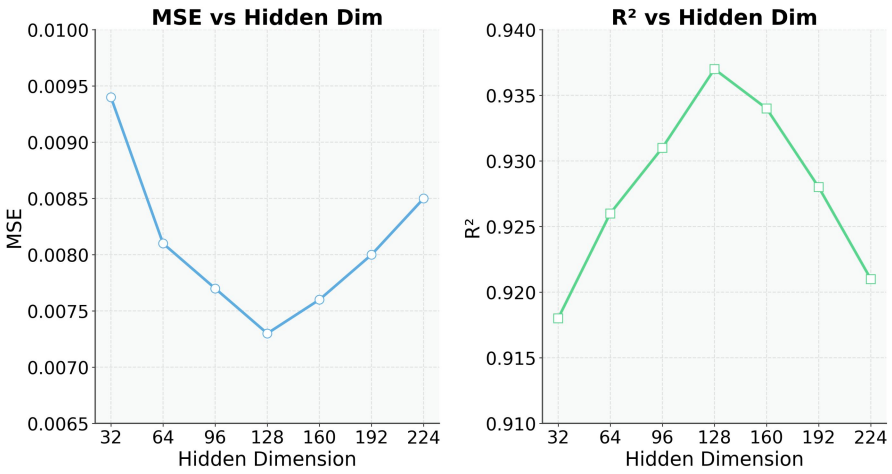

Figure 5. The impact of hidden layer dimension changes on model performance

This experiment focuses on the impact of hidden layer dimensions on model performance. The results show that different feature space sizes lead to significant differences in the effectiveness of multi-granularity structural modeling. When the dimensionality is low, the model's representation capacity is limited. It fails to extract complex cross-scale features adequately, resulting in higher MSE and lower R^2. This suggests that excessive compression harms the model's ability to capture long-term dependencies and semantic interactions, often causing feature loss.

As the hidden layer dimension increases, the model's learning capacity improves. Specifically, at 128 dimensions, MSE reaches its lowest value and R^2 its highest. This indicates that this dimensionality achieves optimal information representation and structural alignment under the current architecture. The trend confirms that the Cross-Level Hybrid Attention module performs better in high-dimensional space, where the multi-head attention mechanism can focus on finer-grained features. At the same time, it forms a positive synergy with the dynamic feature fusion in the Granularity-Aware Fusion module.

However, when the dimension is further increased to 192 and 224, model performance degrades. MSE rises and R^2 declines. This suggests that an overly large hidden space introduces redundant features or noise, disrupting the semantic structure. The overfitting trend implies that high-dimensional representations without effective feature selection can interfere with the attention module's ability to model selectively. As a result, both prediction stability and generalization performance are negatively affected.

Therefore, this experiment shows that hidden layer dimensionality must be carefully balanced in multi-granularity temporal modeling. The model should have sufficient capacity to represent complex patterns, enabling it to capture both fine-grained details and high-level temporal structures across different time scales. At the same time, it is necessary to avoid introducing excessive redundancy, which can lead to noise accumulation, increased computational cost, and degraded feature selectivity. Maintaining an appropriate representation dimensionality not only supports stable learning but also preserves the model's ability to focus on relevant structural signals. This balance is essential for achieving reliable performance and strong structural awareness in dynamic financial prediction scenarios.

5) Comparison of actual value and predicted value

This paper also gives a comparison between the true value and the predicted value, and the experimental results are shown in Figure 6. The comparison is conducted by aligning the predicted stock prices with the corresponding ground truth over a continuous time horizon. This setting is designed to assess the model's ability to track the underlying trends and directional changes in the financial time series.

By visualizing the predicted and actual trajectories, the analysis provides an intuitive understanding of how well the model captures dynamic behaviors in real market conditions. This also allows for the observation of the model's responsiveness to local fluctuations and its consistency in following broader temporal patterns across different stages of the input sequence.

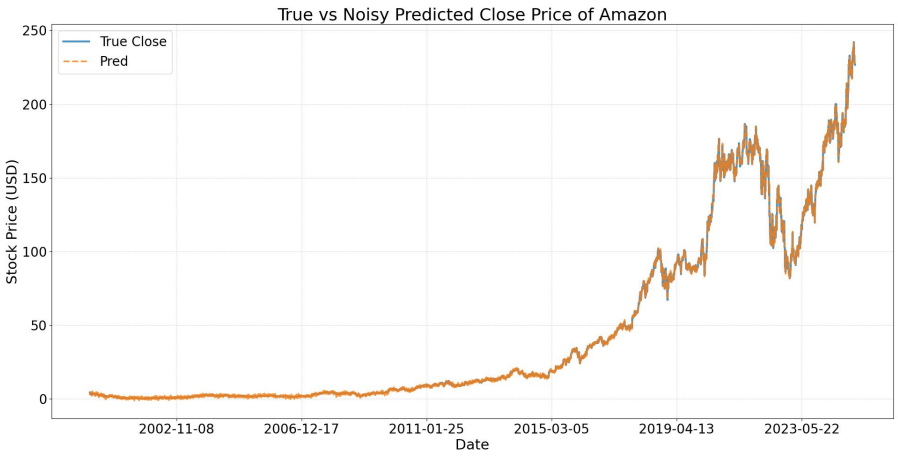

Figure 6. Comparison of actual value and predicted value

This figure presents the fitting results between the model predictions and the actual closing prices of the stock. From the overall trend, the model successfully captures both upward and downward movements across multiple periods. This indicates that the proposed multi-granularity modeling mechanism is effective in identifying key structural patterns in financial time series. Especially during phases of significant price changes, the predicted curve remains aligned with the trend of the actual values, demonstrating strong dynamic response capabilities.

It is worth noting that in certain local regions, the predicted curve appears smoother than the ground truth. This is particularly evident near points of sharp price fluctuation. Such behavior may be related to the model's conservative strategy when handling high-frequency variations. Although the Granularity-Aware Fusion module can effectively integrate long-term trends and short-term changes, it may exhibit mean-reversion tendencies under the influence of anomalies or sudden events. This often leads to delayed responses, which is a common phenomenon in financial modeling and reflects the balance between model stability and sensitivity.

In the later part of the series, the model maintains an accurate representation of the overall upward trend. It also adjusts the prediction direction promptly around multiple turning points. This reflects the significant advantage of the Cross-Level Hybrid Attention module in multi-level semantic modeling and contextual focus. By dynamically selecting features of different granularities, the model becomes more adaptable to structural shifts and trend reversals, which enhances its ability to depict the trajectory of price changes.

Overall, the experiment demonstrates that the proposed method can achieve stable prediction in long-term stock price modeling tasks. The model maintains consistent trend alignment while also showing a degree of structural sensitivity. Although there are slight deviations at points of extreme fluctuation, the overall fitting performance remains strong. This highlights the practical value of multi-granularity fusion and structure-aware attention mechanisms in financial time series forecasting.

5. Conclusion

This paper focuses on the task of long-term stock price prediction and proposes a deep modeling framework based on a multi-granularity hybrid attention mechanism. The aim is to address the limitations of traditional models in capturing long-range dependencies, cross-scale structures, and local fluctuations. By introducing the Granularity-Aware Fusion module, the model can perceive both short-term local variations and long-term global trends. This enhances its ability to represent features across multiple temporal scales. In parallel, the Cross-Level Hybrid Attention module connects semantic interactions between different granularity layers, improving the model's capacity to detect complex structures and evolving patterns in financial time series.

Experimental results show that the proposed method outperforms mainstream baseline models across multiple evaluation metrics, demonstrating strong stability and generalization. The model maintains robust performance under different input structures, hyperparameter settings, and data sampling frequencies, indicating high adaptability in practical applications. In fitting tests using real stock price trajectories, the method not only preserves trend consistency but also shows strong structural sensitivity at key inflection points. These characteristics enable the model to extract potential decision signals from complex financial time series, making it promising for use in risk warning, strategy development, and asset allocation.

This study contributes to both methodology and empirical analysis. It explores multi-granularity modeling and structural attention mechanisms in depth and systematically examines the model's behavior under varying parameter settings and environmental perturbations. The findings enhance the current understanding of structure awareness and scale coupling in financial time series modeling. The results suggest that relying solely on single-scale inputs or shallow attention mechanisms is insufficient for long-horizon forecasting in complex financial systems. Instead, a structured and hierarchical fusion approach provides a stronger foundation for expressive and decision-oriented modeling.

Future research can be extended in two directions. First, multi-source heterogeneous information such as graph structures and event evolution can be introduced to build higher-level frameworks for modeling financial causality. This may improve the model's ability to recognize nonlinear feedback mechanisms in complex economic systems. Second, integrating explainable learning with decision support systems could enable risk assessment and behavior tracing of model outputs, further enhancing the practical value and trustworthiness of the model in financial regulation, investment assistance, and intelligent trading.

References

[1] Soni P, Tewari Y, Krishnan D. Machine Learning approaches in stock price prediction: A systematic review[C]//Journal of Physics: Conference Series. IOP Publishing, 2022, 2161(1): 012065.

[2] Hu Z, Zhao Y, Khushi M. A survey of forex and stock price prediction using deep learning[J]. Applied System Innovation, 2021, 4(1): 9.

[3] Lu W, Li J, Wang J, et al. A CNN-BiLSTM-AM method for stock price prediction[J]. Neural Computing and Applications, 2021, 33(10): 4741-4753.

[4] Ji X, Wang J, Yan Z. A stock price prediction method based on deep learning technology[J]. International Journal of Crowd Science, 2021, 5(1): 55-72.

[5] Lu M, Xu X. TRNN: An efficient time-series recurrent neural network for stock price prediction[J]. Information Sciences, 2024, 657: 119951.

[6] Wu S, Liu Y, Zou Z, et al. S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis[J]. Connection Science, 2022, 34(1): 44-62.

[7] Kanwal A, Lau M F, Ng S P H, et al. BiCuDNNLSTM-1dCNN — A hybrid deep learning-based predictive model for stock price prediction[J]. Expert Systems with Applications, 2022, 202: 117123.

[8] Gülmez B. Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm[J]. Expert Systems with Applications, 2023, 227: 120346.

[9] Ouyang Z, Yang X, Lai Y. Systemic financial risk early warning of financial market in China using Attention-LSTM model[J]. The North American Journal of Economics and Finance, 2021, 56: 101383.

[10] Kong X, Du X, Xue G, et al. Multi-step short-term solar radiation prediction based on empirical mode decomposition and gated recurrent unit optimized via an attention mechanism[J]. Energy, 2023, 282: 128825.

[11] Wan A, Chang Q, Khalil A L B, et al. Short-term power load forecasting for combined heat and power using CNN-LSTM enhanced by attention mechanism[J]. Energy, 2023, 282: 128274.

[12] Zhang J, Jiang Y, Wu S, et al. Prediction of remaining useful life based on bidirectional gated recurrent unit with temporal self-attention mechanism[J]. Reliability Engineering & System Safety, 2022, 221: 108297.

[13] Choi J, Yoo S, Zhou X, et al. Hybrid information mixing module for stock movement prediction[J]. Ieee Access, 2023, 11: 28781-28790.

[14] Lin J, Ma J, Zhu J, et al. Short-term load forecasting based on LSTM networks considering attention mechanism[J]. International Journal of Electrical Power & Energy Systems, 2022, 137: 107818.

[15] Abbasimehr H, Paki R. Improving time series forecasting using LSTM and attention models[J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 13(1): 673-691.

[16] Chhajer P, Shah M, Kshirsagar A. The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction[J]. Decision Analytics Journal, 2022, 2: 100015.

[17] Yifan, Y., Ju'e, G., Shaolong, S., & Yixin, L. (2020). A new hybrid approach for crude oil price forecasting: Evidence from multi-scale data. arXiv preprint arXiv:2002.09656.

[18] Wang, Xing, et al. "Stock2Vec: a hybrid deep learning framework for stock market prediction with representation learning and temporal convolutional network." arXiv preprint arXiv:2010.01197 (2020).

[19] Shi, Z., Hu, Y., Mo, G., & Wu, J. (2022). Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction. arXiv preprint arXiv:2204.02623.

[20] Zou, C. (2023). The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China. Highlights Sci. Eng. Technol, 39, 327-333.

[21] Xiao, R., Feng, Y., Yan, L., & Ma, Y. (2022). Predict stock prices with ARIMA and LSTM. arXiv preprint arXiv:2209.02407.

[22] Ekambaram, V., Jati, A., Nguyen, N., Sinthong, P., & Kalagnanam, J. (2023, August). Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting. In Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining (pp. 459-469).

[23] Selvin S, Vinayakumar R, Gopalakrishnan E A, et al. Stock price prediction using LSTM, RNN and CNN-sliding window model[C]//2017 international conference on advances in computing, communications and informatics (icacci). IEEE, 2017: 1643-1647.

[24] Yi J, Chen J, Zhou M, et al. Analysis of stock market public opinion based on web crawler and deep learning technologies including 1DCNN and LSTM[J]. Arabian Journal for Science and Engineering, 2023, 48(8): 9941-9962.

[25] Qi C, Ren J, Su J. GRU neural network based on CEEMDAN−wavelet for stock price prediction[J]. Applied Sciences, 2023, 13(12): 7104.

[26] Muhammad T, Aftab A B, Ibrahim M, et al. Transformer-based deep learning model for stock price prediction: A case study on Bangladesh stock market[J]. International Journal of Computational Intelligence and Applications, 2023, 22(03): 2350013.

[27] Liu Y, Hu T, Zhang H, et al. itransformer: Inverted transformers are effective for time series forecasting[J]. arXiv preprint arXiv:2310.06625, 2023.

[28] Wang S, Wu H, Shi X, et al. Timemixer: Decomposable multiscale mixing for time series forecasting[J]. arXiv preprint arXiv:2405.14616, 2024.