IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 26 April 2024, accepted 28 April 2024, date of publication 3 May 2024, date of current version 10 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396619

APPLIED RESEARCH

HPMG-Transformer: HP Filter Multi-Scale Gaussian Transformer for Liquor Stock Movement Prediction

LILI HUANG
School of International Economics, Anhui International Studies University, Hefei 231201, China
e-mail: 949786734@qq.com

This work was supported by the Key Research Project at Anhui International Studies University titled ''The Impact of Green Finance on the Development of Regional Low-carbon Economy'' under Project Awky2023002.

ABSTRACT Predicting financial stock prices, which are complex, volatile, and nonlinear, poses a significant challenge due to the multitude of influencing factors and inherent uncertainty in the financial market. This paper introduces a novel approach that utilizes a neural network model combining the Hodrick-Prescott (HP) filter and a Multi-Scale Gaussian transformer to tackle these challenges. The proposed method enhances the model's local features by incorporating a Multi-Scale Gaussian transformer. Initially, the stock's time series is decomposed into long-term and short-term fluctuations using the HP filter. Subsequently, the encoded long-term and short-term series are fed into a Multi-Scale Gaussian transformer. Additionally, a Multi-Scale Gaussian prior is introduced to further boost the local features of the transformer and enhance the relative positional information features of the time series. In comparison to popular recurrent neural networks like RNN, LSTM, GRU, and state-of-the-art baseline models, our model (HPMG-Transformer) offers a unique advantage in capturing both extremely long-term and short-term dependencies in stock time series. Experimental results illustrate the significant benefits of our proposed model in predicting stock trends in the China A-shares market, New York Stock Exchange (NYSE), and NASDAQ market.

INDEX TERMS Stock price, artificial neural network, time series, HP filter, transformer.

I. INTRODUCTION

The prediction of financial stock price has always been a focus research direction in academia. However, there are often many factors that affect the movement of stocks price, including government policies, politics, investment psychology and public opinion. The stock price change of the financial market is a nonlinear, volatile, high-noise time series data. Therefore, the price movement prediction for the stock market is a challenging task. At present, there are two types of financial analysis for time series data, on the one hand, traditional statistical quantification methods such as ARMA, ARIMA and GARCH methods [1], [2], [3]. on the other hand, artificial intelligence models by artificial neural network build. Recent years, more and more financial researchers get involved in the movement of stock price predict by neural network methods. At the same time, it is used state-of-the-art trading strategies to help investors make profitable decisions.

Recently, due to the rapid development of artificial neural network technology, it has been widely used in the field of finance. The artificial neural network algorithm model can be applied to the modeling of nonlinear, high-noise, long-term and short-term time series data. For recurrent neural network (RNN) [4], it can be used for long-term series data prediction. However, there is a phenomenon of gradient explosion or disappearance with long-term series data during training Hochreiter et al [5].

A long-short term artificial neural network model (Long-Short Term Memory, LSTM) for addressing gradient disappearance, and it can be applied to long-term series date. Cho et al. [6] proposed Gated Recurrent Unit (GRU) based on LSTM neural network. Compared with LSTM model GRU neural network only contains ''update gate'' and ''Reset gate'' two gated structures and fuse the ''cell state'' and

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

"hidden state", so the GRU neural network has fewer parameters, is less prone to overfitting. Vaswani et al. 2017 [7] proposed a sequence-to-sequence model Transformer, which it is a multi-head self-attention mechanism for long-term dependencies is enhanced.

And related studies have been conducted in this field. For example, Yoo et al. [8] applied DTML to learn temporal correlations, multi-level contextual information, and correlations between stocks. Muhammad et al. [9] introduced time2vec encoding to represent time series features and used a Transformer model for stock price prediction. Hu et al. [10] employed state-of-the-art Temporal Fusion Transformer (TFT) for stock price prediction. Lai et al. [11] proposed a differential transformer neural network model that utilizes time attention-enhanced bilinear layers and temporal convolutional networks (TCN) for denoising data and capturing dependencies among time series. Ramos-Pérez et al. [12] presented a Multi-Transformer model, a novel machine learning and deep learning technique for more accurate stock volatility modeling.

Chen et al. [13] developed the TPM model, a stock price trend prediction model based on the encoder-decoder framework, which can adaptively predict the volatility and duration of stock prices. Kim et al. [14] extracted temporal features of input data using CNN and explained the correlations between variables using attention mechanisms. Mishev et al. [15] explored the effectiveness and performance of various sentiment analysis methods combining text representation techniques and machine learning classifiers. Wen et al. [16] reconstructed noisy financial time series using frequent patterns and extracted the spatial structure of time series using convolutional neural networks. Li et al. [17] addressed the media-aware stock volatility problem as a multimodal problem and implemented an event-driven LSTM with tensors. Kwon et al. [18] proposed a hybrid neurogenetic system for stock trading. Li et al. [19] transferred sentimental information learned from news-rich stocks to news-poor stocks to enhance their prediction performance.

In the meantime, Ali et al. [20], [21] have proposed the Akina-EMD model, which identifies non-informative fluctuations in signals, such as noise, outliers, and ultra-high-frequency components. It decomposes clean and chaotic data into various components to avoid distortion. Additionally, they have introduced a novel algorithm, the EMD-LSTM model, for predicting financial time series in stock time series data. Furthermore, Tan et al. [22], [23], [24] have presented an analytical study of the mathematical theory addressing the stability issues of recurrent neural networks with time-varying delays, providing an effective theoretical basis for research on recurrent neural networks. Therefore, the transformer model is widely used in the application of long-term time series tasks, such as financial market stock price prediction. Since the stock price movement will be affected by different factors, some studies decompose the stock price time series into subsequences of different frequencies, and prediction them separately to improve the accuracy. HP filter (Hodrick-Prescott Filter) [25] is a filtering method proposed by Hodrick and Prescott to study the economy, which is able to effectively separate long-term trends and short-term fluctuations from economic time series, which helps in analyzing economic cycles and structural changes, and it is widely used in economic research.

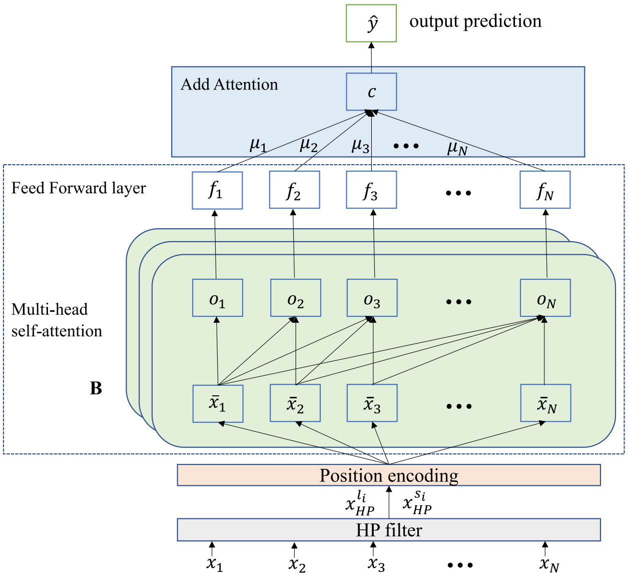

FIGURE 1. HPMG-Transformer network structure overview.

The HP filter is grounded in spectral analysis methods of financial time series, offering a solution to short-term fluctuations and outliers in complex, nonlinear, and non-stationary time series. Given the global dependencies at play, the positional information within financial time series holds significant importance. However, the self-attention mechanism of the Transformer model exhibits a relatively weak incorporation of positional information. To address this limitation, we propose integrating a multi-scale Gaussian prior [26] into the multi-head self-attention mechanism, establishing a correlation between data that is directly linked to temporal distance. This integration enhances the extraction of local contextual information features by neural networks.

In summary, our contributions are as follows:

- We demonstrate the effectiveness of the HP filter in separating long-term trends and short-term fluctuations in economic time series, validated through experiments that aid in the analysis of economic cycles and structural changes.
- We suggest incorporating a local feature enhancement using the Multi-Scale Gaussian prior in the model.
- We introduce the HPMG-Transformer model, which significantly outperforms various state-of-the-art baseline models like LSTM, GRU, and RNN in stock prediction. This superiority is empirically proven through its application in the China A-shares liquor market.

II. PROPOSED METHOD

In this section, first describes the Transformer model by we design, then introduce the local enhancements of transformer method for liquor stocks prediction [27], final we introduce proposed the HP filer process stock series data and filer method.

A. TRANSFORMER MODEL DESIGN FOR STOCKS MOVEMENT PREDICTION

The function of HPMG-Transformer model can be seen as $f(\theta)$ for the movement of stock prediction. In our work, we design only encoder structures for transformer, which it is consist of $\mathcal{B}$ blocks of multi-head self-attention module, As shown in Figure 1, a variant of network structure in the basic transformer. For a given time series $\mathbf{X} = \{x_1, x_2, \cdots, x_N\} \in \mathbb{R}^{N\times M}$, it is required to perform HP filter processing on it to obtain long-term and short-term sequences $\mathbf{L} = \{l_1, l_2, \cdots, l_N\} \in \mathbb{R}^{N\times M}$, $\mathbf{S} = \{s_1, s_2, \cdots, s_N\} \in \mathbb{R}^{N\times M}$ and perform position encoding processing respectively, and adopt liner layer input model with *tanh* activation function as follows 1:

$$\overline{\mathbf{X}} = \sigma_{tanh}\mathbf{W}^{(input)}[\mathbf{positionencode}(X_{HP}^{L} + X_{HP}^{S})] \tag{1}$$

where $\overline{\mathbf{X}}$ represents the input of multi-head self-attention, X_{HP}^{L} and X_{HP}^{S} is long-term and short-term sequences by HP filer processing. The position encoding is adopted trigonometric function.

The input of multi-head self-attention layer is $\overline{\mathbf{X}}$, which it is input the long-term and short-term sequence obtained by HP filter of the original time series data x_i. At the same time, long-term and short-term series data is adopted position encoding function. The $\overline{\mathbf{X}}$ is computed by 2:

$$\mathbb{Q}_j = \mathbf{W}_j^{(\mathbf{Q})}\overline{\mathbf{X}}, \mathbb{K}_j = \mathbf{W}_j^{(\mathbf{K})}\overline{\mathbf{X}}, \mathbb{V}_j = \mathbf{W}_j^{(\mathbf{V})}\overline{\mathbf{X}} \tag{2}$$

where $j \in \{1, 2, 3, \cdots, J\}$ and $\mathbf{W}_j^{(\mathbf{Q})}$, $\mathbf{W}_j^{(\mathbf{K})}$,$\mathbf{W}_j^{(\mathbf{V})}$ is learn-able weights matrices for Query, Key and Value in the network model. The attention module output takes $s_j \in \mathbb{R}^{(N\times N)}$ for scores matrix, and the j^{th} of output scores matrix is computed by 3:

$$s_j = softmax(\frac{\mathbb{Q}_j\mathbb{K}_j^T}{\sqrt{d_j}} \cdot \mathcal{M}) \tag{3}$$

where $\mathcal{M}$ is position mask matrix for filter future information. The multi-head self-attention layer take $\mathbf{O}_j$ for the j^{th} of output by computed as follows 4:

$$\mathbf{O}_j = \sum_{i=0}^{N}(s_j)_i \cdot [\mathbf{V}_j]_i \tag{4}$$

The network model output of multi-head attention module is adopted concatenation operate all results by $\hat{\mathbb{O}} = concat(\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_J)$. Then the feed forward layer takes $\hat{\mathbb{O}}$ for input and convert to $\mathbf{F}$, which it is consist of two fully connected layers and *RELU* activation function layer. Finally, the output f_i is aggregated to add all the results by computed $\mathbf{c} = \sum_{i=0}^{N}\mu_i \cdot f_i$, where μ_i is a coefficient and the network model of output prediction scores $\hat{y}$ by fully connected layer and *sigmoid* function.

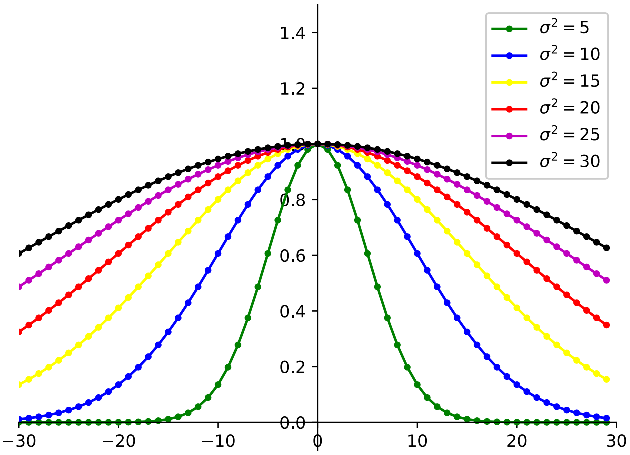

FIGURE 2. The sampling distribution of Gaussian functions, where bandwidth is $\sigma_j^2 \in \{5, 10, 15, 20, 25, 30\}$.

B. MULTI-SCALE GAUSSIAN PRIOR FOR ENHANCEMENT LOCALITY

Transformers are widely used in time series modeling scenarios. The positional encoding in the self-attention mechanism has a weaker dependence on global information. To address the above problems, it is necessary to add a multi-scale Gaussian prior strategy to the temporal position information, which it improves enhancements the feature of locality information. A Gaussian bias term is added to the calculation of the self-attention score matrix, so the score of the self-attention matrix obeys the assumption of a Gaussian normal distribution. The self-attention score matrix by *softmax* activation function as follows 5:

$$s_j = softmax[(\frac{\mathbb{Q}_j\mathbb{K}_j^T}{\sqrt{d_j}} + \mathbf{B}_j^{(G)}) \cdot \mathcal{M}] \tag{5}$$

In other words, adding a bias $\mathbf{B}_j^{(G)}$ to Equation 3,where $\mathbf{B}_j^{(G)} \in \mathbb{R}^{N\times N}$ is a vector matrix as follows 6:

$$[\mathbf{B}_j^{(G)}]_{n,m} = \begin{cases} 0, & n < m \\ exp(-\frac{(n-m)^2}{2\sigma_j^2}), & n \geq m \end{cases} \tag{6}$$

where σ_j is the bandwidth of the Gaussian distribution, it's set $\Delta = \{\sigma_1, \sigma_2, \cdots, \sigma_d\}$ and where is $d = 6$. In stock prices movement, it is from last 5-day, 10-day, 15-day, 20-day, 25-day and 30-day are usually considered in close trading time strategies. In other word, a 6-head self-attention layer, we can take Gaussian bandwidth scale set $\Delta = \{5, 10, 15, 20, 25, 30\}$ to σ_j with $j = \{1, 2, 3, 4, 5, 6\}$ respectively as is shown in Figure 2.

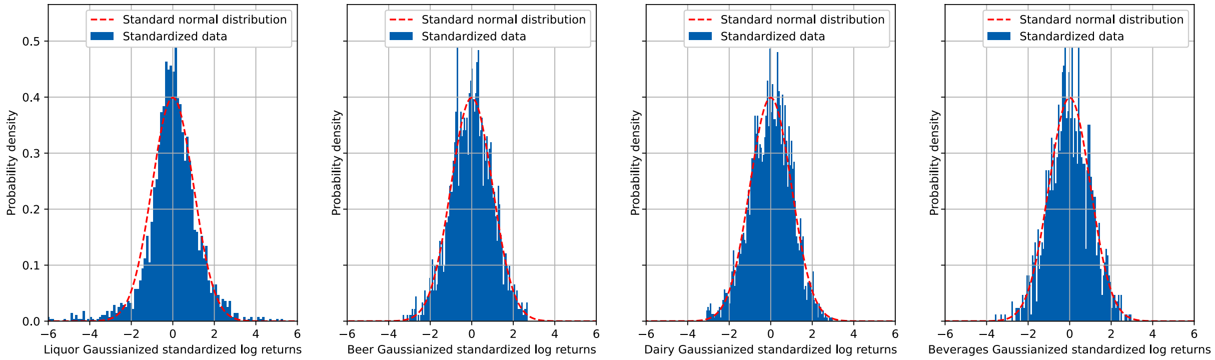

FIGURE 3. The dataset standardized distribution of the closing prices of stocks in the four industry categories, liquor, beer, dairy, and beverages.

III. MODEL TRAINING AND EVALUATION METRICS

A. DATASET AND PLATFORM

In this paper, the dataset is adopted the trading stock price of the liquor industry of China A-shares main board. It includes all trading data of 20 companies since listing in the China A-shares market. This article utilizes stock trading data from four industry sectors of publicly listed companies, including liquor, beer, beverages, and dairy products. In total dataset from 104 listed companies are included. The dataset is daily opening price, highest price, lowest price, closing price and turnover rate for network model analysis. Additionally, the data is divided into training, validation, and test sets in an 8:1:1 ratio.

To further validate the generalization and effectiveness of the model, this paper selects stock data from foreign exchanges, specifically from the New York Stock Exchange (NYSE) and NASDAQ, for verification experiments. The data includes information from the alcoholic beverages sector, the beverage industry, and the dairy products industry. Daily trading data from January 4^{st}, 2010, to October 31^{st}, 2023, have been collected. The beer industry 23 listed companies, the beverage industry includes 19 listed companies, and the dairy products industry consists of 22 listed companies.

This experiment is carried out under the Linux18.04 operating system, the Python version used is 3.6, and the model is built under the Pytorch1.8.0 framework that supports GPU. The CPU is Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz, the GPU is NVIDIA Tesla V100, and the memory is 32 GB.

B. EVALUATION METRICS

To evaluate the prediction effect of the model, we proposed root mean square error (Root Mean Square Error, *RMSE*), *SMAPE*(Symmetric Mean Absolute Percentage Error) and R^2*score* (R^2 coefficient of determination) as evaluation metrics. The formulas for computing *RMSE*,*SMAPE* and R^2*score* are as follows 7,8 and 9:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} \tag{7}$$

$$SMAPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{|y_i' - y_i|}{(|y_i'| + |y_i|)/2} \tag{8}$$

$$R^2 score = 1 - \frac{\sum_{i=1}^{N}(y_i - y_i')^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{9}$$

where the N is sample numbers, y_i and y_i' is true label and prediction respectively, $\overline{y}_i$ is the average value of the true value. The value of R^2*score* is between 0 and 1, and the closer to 1, the better the fitting effect of the model. The smaller the *RMSE* the better the prediction effect of the model. *SMAPE* can effectively avoid the issue of excessively large calculation results due to small true values of y_i.

C. EVALUATION METRICS DIEBOLD-MARIANO TEST

If we evaluate the model based on the above performance indicators, we can only obtain its accuracy, but cannot statistically test whether the results are significant. This paper uses the Diebold-Mariano test (DM test) method [28] to compare the models. The DM test is a statistical principle of asymptotic normal distribution. Assuming that the prediction result of the i-th model is $\{\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{iT}\}$, and j-th model is $\{\hat{y}_{j1}, \hat{y}_{j2}, \ldots, \hat{y}_{jT}\}$. The true value is $\{y_1, y_2, \ldots, y_T\}$, and the error of the i-th model is $E_i = \{\phi(y_1 - \hat{y}_{i1}), \phi(y_2 - \hat{y}_{i2}), \ldots, \phi(y_T - \hat{y}_{iT})\}$, the error of the j-th model is $E_j = \{\phi(y_1 - \hat{y}_{j1}), \phi(y_2 - \hat{y}_{j2}), \ldots, \phi(y_T - \hat{y}_{jT})\}$, where $\phi(*)$ is error function, T represents the length of the sequence.

The obtained error sequence is $\Delta = E_i - E_j = \{d_1, d_2, \ldots, d_T\}$, $d_t = \phi(y_t - \hat{y}_{it}) - \phi(y_t - \hat{y}_{jt})$, and obtained mean is $\overline{\Delta}$ and std is $\sigma(\Delta)$ are as follows 10 and 11.

$$\overline{\Delta} = \frac{\sum_{i=1}^{T} d_i}{T} \tag{10}$$

$$\sigma(\Delta) = \sqrt{\frac{\sum_{i=1}^{T}(d_i - \overline{\Delta})^2}{T - 1}} \tag{11}$$

the final result of the DM statistic is as follows 12:

$$DM = \frac{\overline{\Delta}}{\sigma(\Delta)} \sim \mathcal{N}(0, 1) \tag{12}$$

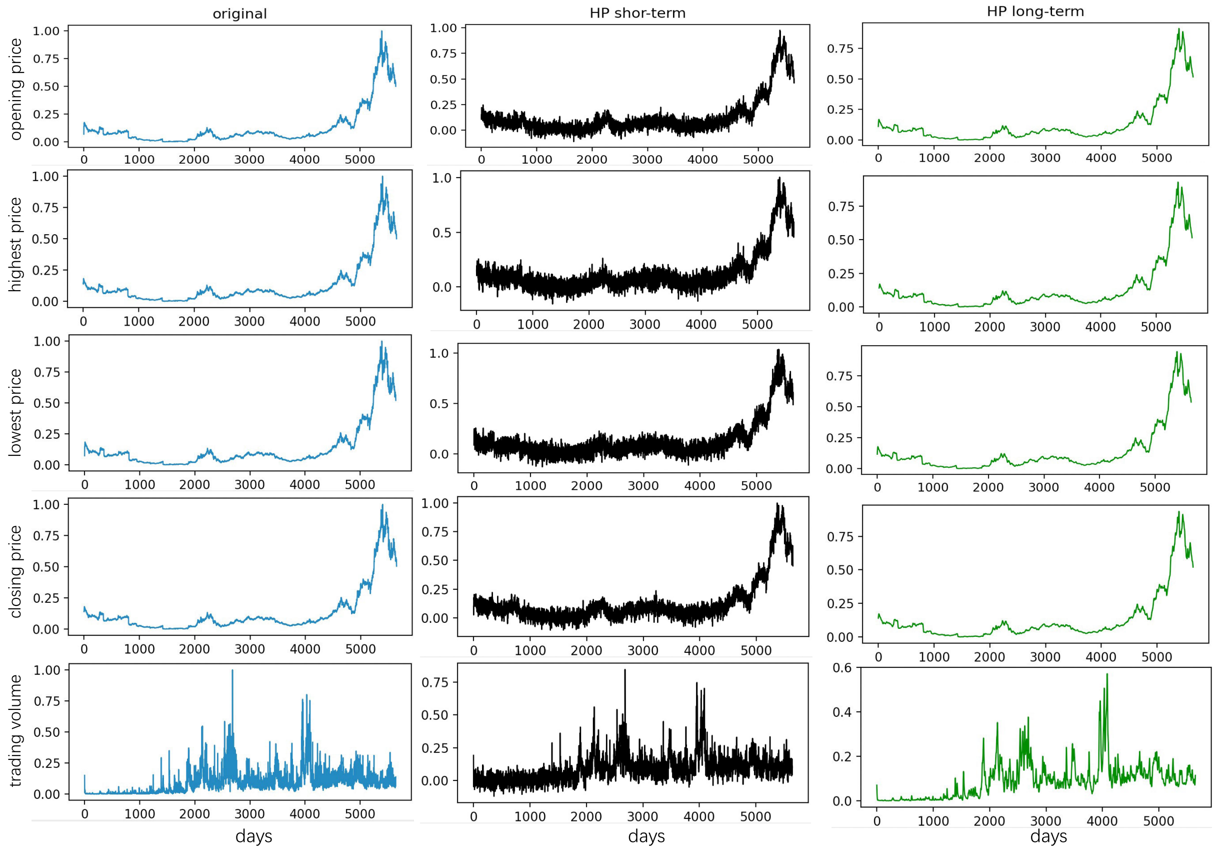

FIGURE 4. Visualizing the decomposition of standardized stock time series information into short-term and long-term components using the HP filter.

After standardization, a standard normal distribution with $\mathcal{N}(0, 1)$ can be obtained.The null hypothesis and alternative hypothesis are provided as follows:

$\mathbb{H}_0 : \overline{\Delta} = 0$ and $\mathbb{H}_1 : \overline{\Delta} \neq 0$ respectively represent that the effects of the two models' predictions are the same, which means the null hypothesis is accepted. If there is a difference in the performance of the two models' predictions and $\overline{\Delta} > 0$, it indicates that Model 1 is inferior to Model 2. Conversely, if $\overline{\Delta} < 0$, it suggests that Model 1 has better predictive ability than Model 2.

D. TRAIN MODEL

To train the model, data preprocessing is performed initially. For stock time series data, the data is first logarithmically transformed and then standardized. The standardized distribution of the closing prices of stocks in the four industry categories, namely liquor, beer, dairy, and beverages, is shown in Figure 3.

Simultaneously, utilizing the HP filter for preprocessing the raw data provides data smoothing functionality, which effectively suppresses and handles certain outliers. This reduces sensitivity to outliers and ensures the model's generalizability.

In model training, multiple hyperparameters need to be set to achieve optimal results. Firstly, the Batch size is set to 64, which refers to the number of samples sampled in each iteration. A larger batch size can improve training speed but may also increase memory requirements. Next, we choose the Adam optimizer, which is a commonly used gradient descent optimization algorithm. It can adaptive adjust the learning rate and usually exhibits good performance. We set the initial learning rate to $lr = 0.001$, which represents the learning rate at the beginning of the training process.

To further optimize the learning rate variation, we employ the cosine annealing algorithm. This algorithm periodically adjusts the learning rate, allowing the model to better adapt to complex data distributions and loss function curves. Through this approach, we can more effectively search the parameter space and improve the model's performance and convergence speed.

For the choice of the objective optimization loss function, It utilize mean squared error *MSE* is a common loss function for regression tasks, and to enhance the model's generalization capability and prevent over-fitting, a regularization term is added to the loss function, the loss function is represented as follows 13. It measures the average squared difference between predicted values and true values. By minimizing the mean squared error, we can make the model fit the training data more accurately and expect good generalization performance on test data.

$$\mathcal{L}_{loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \xi||\mathcal{W}||_2^2 \tag{13}$$

where, N represents the number of training data in the dataset, where x_i denotes each training data, represents the true output of y_i, and $\hat{y}_i$ represents the predicted value of x_i by the model, and where $\xi = 0.4$.

In addition, we apply HP filter offers a method for local trend estimation, capable of capturing structural breakpoints and changes within time series. Through this processing, we can better capture periodic and trending information in the data, thereby improving the model's ability to model time series data. At the same time, by using the original data preprocessing of the HP filter, which has the function of data smoothing, it can effectively suppress and handle certain outlier data. This ensures that the model is not sensitive to outliers data and maintains models generalization capability. HP filtering is a decomposition method that separates the signal into a low-frequency trend component l_t and a higher-frequency component representing cycles or noise h_t, as shown in equation 14.

$$y_t = h_t + l_t \tag{14}$$

The HP filter optimization function is represented as follows 15.

$$\begin{aligned}\Gamma_{\underset{l}{argmin}} &= \sum_{i=1}^{N}(y_i - l_i)^2 + \lambda\sum_{i=1}^{N-2}(l_i - 2l_{i+1} + l_{i+2})^2 \\ &\Longrightarrow \underset{l}{argmin}\,||y-l||^2 + \lambda'||\triangledown^2 l||^2\end{aligned} \tag{15}$$

On the one hand, it is fitting: reducing the filtered signal l Error from the original signal y; On the other hand, smoothing limits the second-order difference size of the smoothed signal. Where, we set the hyperparameter $\lambda = 100$ for the optimization function of the HP filter, performing decomposition and obtaining the long-term and short-term sequence information from the original data, as shown in Figure 4.

With the detailed settings described above, we can better understand the hyperparameter choices and data processing methods used in the model training process, as well as the expected effects and performance optimization strategies.

IV. EXPERIMENTAL AND ANALYSIS

A. NUMERICAL EXPERIMENTS

We conducted a comparative analysis using two different methods: one is the HP-Transformer, which solely utilizes the HP filter technique, and it has a solid theoretical foundation in economics, especially when analyzing economic cycles and long-term growth trends. The second method is the MG-Transformer, which exclusively employs a multi-scale Gaussian prior to accentuate the features related to both position and time information. Finally, we introduced the HPMG-Transformer model, which combines the use of both the HP filter and multi-scale Gaussian prior modules in the Transformer framework.

TABLE 1. Ablation experiment results, use to different methods for comparative experimental validation in the opening price prediction of China's A-shares Liquor market dataset.

Models	Methods		Metrics(@**Acc.**)		
	HP	MG	*RMSE*	*SMAPE*	R^2score
Basic-Transformer	×	×	104.76	0.6311	0.9133
HP-Transformer	✓	×	99.65	0.5928	0.9345
MG-Transformer	×	✓	97.35	0.5027	0.9471
HPMG-Transformer	✓	✓	**95.23**	**0.3982**	**0.9639**

TABLE 2. Ablation experiment results, use to different methods for comparative experimental validation in the opening price prediction of China's A-shares Liquor market dataset.

Models	Metrics(@**Acc.**)		
	RMSE	*SMAPE*	R^2score
Transformer	**104.76** ↓	**0.6311** ↓	**0.9133** ↑
LSTM	112.71	0.6684	0.9012
GRU	109.92	0.6871	0.9055
RNN	120.33	0.7508	0.8798
HP-Transformer	**99.65** ↓	**0.5928** ↓	**0.9345** ↑
HP-LSTM	105.70	0.6521	0.9275
HP-GRU	102.45	0.6689	0.9271
HP-RNN	114.81	0.7022	0.8923
MG-Transformer	**97.35** ↓	**0.5027** ↓	**0.9471** ↑
MG-LSTM	100.28	0.5439	0.9332
MG-GRU	99.46	0.5531	0.9360
MG-RNN	109.15	0.5721	0.9017
HPMG-Transformer	**95.23** ↓	**0.3982** ↓	**0.9639** ↑
HPMG-LSTM	99.29	0.4328	0.9524
HPMG-GRU	98.93	0.4246	0.9557
HPMG-RNN	105.82	0.4725	0.9298

Through comparative experiments, we observed that the performance improvements primarily stem from the integration of the HP filter and multi-scale Gaussian prior into the Transformer model. The HP filter effectively reduces short-term fluctuations, enabling better forecasting of long-term trends. On the other hand, the multi-scale Gaussian prior enhances the manifestation of features associated with position and time information. By leveraging both techniques, the HPMG-Transformer achieves further performance enhancements.

Overall, our results from the ablation experiments, as shown in Table 1, demonstrate that the integration of HP filtering and multi-scale Gaussian prior in Transformer-based methods has a significant impact, leading to improved predictive capabilities and feature representation.

B. ABLATION EXPERIMENTS

This paper compares the advantages of the proposed Transformer architecture by contrasting it with RNN, LSTM, GRU architectures. The model accuracy on the test set is compared through different evaluation metrics including *RMSE*, *SMAPE* and R^2score. To further verify the effectiveness of

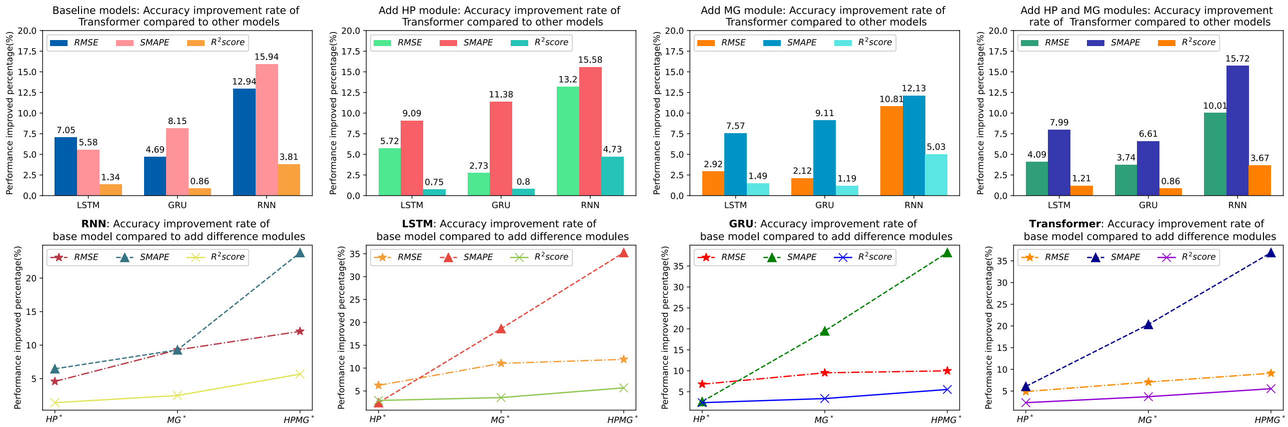

FIGURE 5. Ablation experiment visualization of baseline models Transformer, RNN, LSTM, and GRU, and comparative experiment on the HP, MG, and HPMG modules.

TABLE 3. DM test of HPMG-Transformer with the other 4 models(DM test statistics).

	Transformer	LSTM	GRU	RNN
MSE	18.34***	26.41**	25.38***	34.66*
MAPE	2.71***	7.73***	5.95***	12.83*

Notes :*** indicates significance at the 1% level. ** indicates significance at the 5% level. * indicates significance at the 10% level.

the HP filtering module and multi-scale Gaussian module proposed in this work, ablation experiments are conducted on different baseline models. For convenience, the experiments only use opening prices for models prediction, and the results are shown in Table 2.

To facilitate understanding of the data in Table 2, it is visualized to compare the percentage improvements of various evaluation metrics by the baseline models, as shown in Figure 5. An analysis of the data obtained from the ablation experiments in Figure 5 suggests that both the HP filtering module and the multi-scale Gaussian model proposed in this work are effective.

To further validate the effectiveness of the model proposed in this paper, the DM test method is used for experimental result analysis, as shown in Table 3. The model's prediction error is evaluated using two indicators, *MSE* and *SMAPE*. Based on these two evaluation metrics, the DM test is conducted for pairwise comparison between HPMG-Transformer and other models Transformer, LSTM, GRU, and RNN models, as shown in Table 3. The results significantly reject the null hypothesis, indicating that HPMG-Transformer prediction ability is superior to the other four models.

C. APPROACHES IN COMPARISON

We compare our approaches HPMG-Transformer with the baselines TEANet [29], FDGRU-transformer [30], TEGRU [31], KHIT [32] and PCA-IGRU [33].

- TEANet *[Zhang Q,et.al, 2022]*: This algorithm achieves the temporal dependency of financial data through precise feature engineering on small samples.
- FDGRU-transformer *[Li C,et.al, 2022]*: It decomposes the complete set of chaotic data into trend components and several information-rich and mutually independent modal components.
- TEGRU *[Haryono A T,et.al, 2023]*: Frequency Decomposition Guided Recurrent Unit, utilizes sentiment indicators for stock price prediction.
- KHIT *[Lin F,et.al, 2022]*: Kernel-based Hybrid Interpretable Transformer model, combined with a novel loss function, used for handling prediction tasks in non-stationary stock markets.
- PCA-IGRU *[Wang J,et.al, 2023]* Reduce model training and prediction time through feature engineering PCA technology, and then introduce an improved gated recurrent unit model with an anti-oversaturation conversion module to prevent over-saturation and improve model learning sensitivity.

In this paper, we will test the performance of an improved prediction model compared with the baseline model on the Chinese A-shares Liquor market dataset. We take the closing price of Liquor stock as the prediction value, and comprehensively evaluate the prediction performance of the model by comparing *RMSE*, *SMAPE* and R^2*score* evaluation metrics.The final results are shown in Table 4.

By comparing the experimental results, we can draw the conclusion that our model shows significant advantages over the benchmark model in the dataset of China's A-shares Liquor market. This indicates that our model has higher accuracy in predicting stock prices. Therefore, we can use this model to guide investment decisions to achieve better investment returns. Among them, HPMG-Transformer compared with the best KHIT accuracy performance in the benchmark models, the three different evaluation metrics of *RMSE*, *SMAPE*, and R^2*score* performance improved by 0.51%, 8.19%, and 1.87%, respectively.

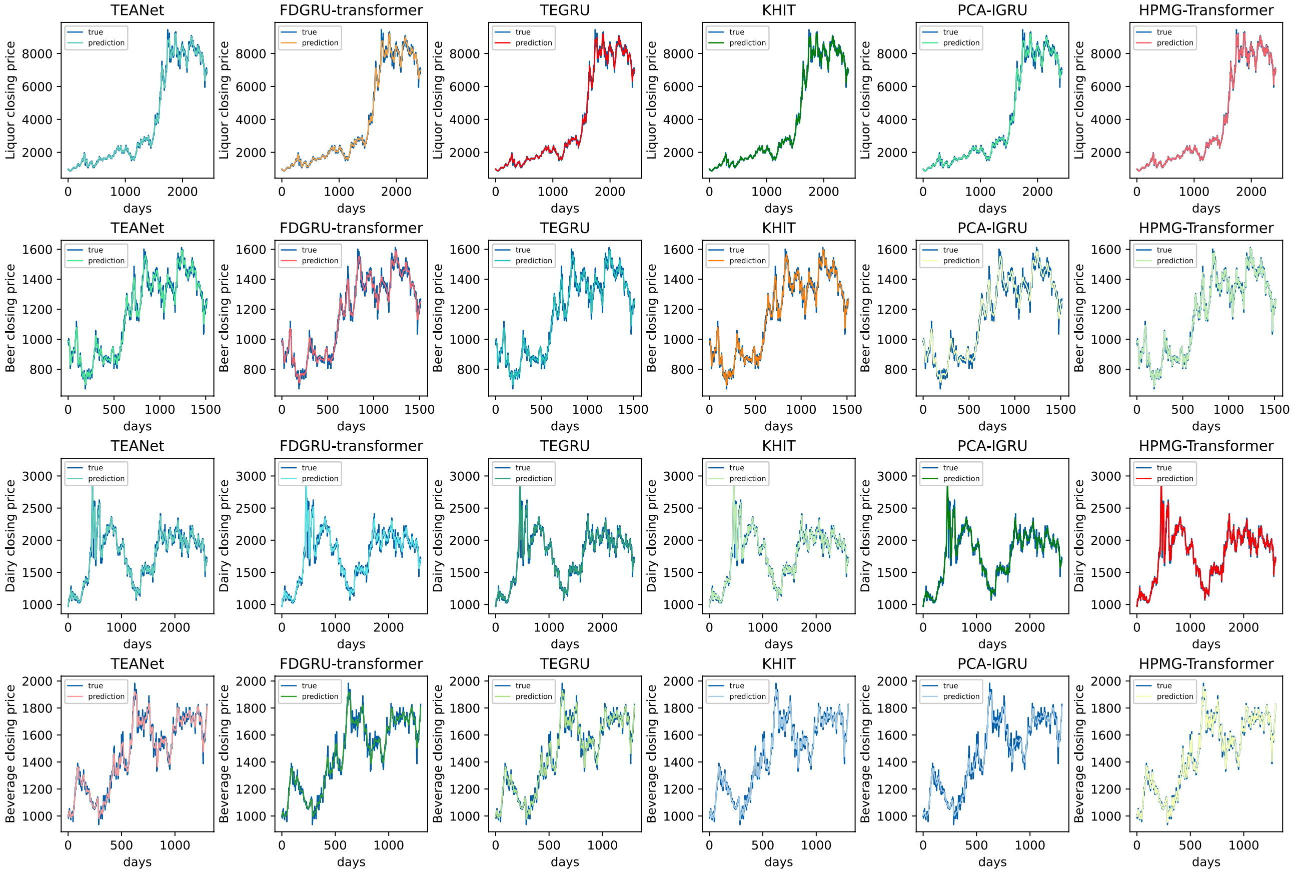

FIGURE 6. The HPMG transformer and the baseline model are prediction closing price. Visualization of the closing price of Liquor,Beer,Dairy, Beverage in China's A-share market.

TABLE 4. Comparison of the accuracy performance of HPMG-Transformer and benchmark model in the closing price prediction of China's A-shares Liquor market dataset.

Models	Metrics(@**Acc.**)		
	$RMSE$	$SMAPE$	R^2score
TEANet	96.56	0.4881	0.9378
FDGRU-transformer	97.14	0.4494	0.9411
TEGRU	94.23	0.5932	0.9229
KHIT	93.37	0.4337	0.9524
PCA-IGRU	98.85	0.6349	0.9006
HPMG-Transformer	**92.89**	**0.3982**	**0.9702**

TABLE 5. Comparison of the accuracy performance of HPMG-Transformer and benchmark model in the closing price prediction of China's A-shares Liquor, Beer, Dairy and Beverage market and in the New York Stock Exchange (NYSE) and NASDAQ market Beer, Dairy and Beverage.

China Stock	Models	Metrics(@**Acc.**)		
		$RMSE$	$SMAPE$	R^2score
Liquor	ours	**92.89** ↓	**0.3982** ↓	**0.9702** ↑
Liquor	KHIT	93.37	0.4337	0.9524
Beer	ours	**93.32** ↓	**0.3993** ↓	**0.9659** ↑
Beer	KHIT	95.27	0.4468	0.9447
Dairy	ours	**93.74** ↓	**0.4012** ↓	**0.9641** ↑
Dairy	KHIT	94.98	0.4946	0.9437
Beverage	ours	**95.76** ↓	**0.5372** ↓	**0.9465** ↑
Beverage	KHIT	97.21	0.6196	0.9382
US Stock	Models	Metrics(@**Acc.**)		
		$RMSE$	$SMAPE$	R^2score
Beer	ours	**98.45** ↓	**0.6924** ↓	**0.9113** ↑
Beer	KHIT	102.23	0.8192	0.8794
Dairy	ours	**97.22** ↓	**0.6692** ↓	**0.9015** ↑
Dairy	KHIT	101.08	0.7995	0.8682
Beverage	ours	**99.01** ↓	**0.8203** ↓	**0.8875** ↑
Beverage	KHIT	104.73	0.9421	0.8421

D. ANALYSIS OF EXPERIMENTAL RESULTS

In order to demonstrate the generalizability of the proposed HPMG-Transformer model architecture, several relevant datasets were selected for validation. The experiment involved predicting stock indices of liquor, beer, dairy, and beverage companies, comprising a total of 104 companies' stock data. By comparing the baseline model (TEANet, FDGRU-transformer, TEGRU, KHIT, PCA-IGRU) with the HPMG-Transformer model proposed in this paper, the results shown in Figure 6 were obtained. The comparison was made with the best baseline model, KHIT, to evaluate the performance using metrics such as $RMSE$, $SMAPE$, and R^2score. The proposed HPMG-transformer model, as compared to the KHIT model, exhibits improved accuracy

TABLE 6. Comparison of the accuracy performance of HPMG-Transformer and benchmark model in the closing price prediction of China's A-shares Liquor market dataset.

Models	Prediction Values	Metrics(@**Acc.**)		
		RMSE	*SMAPE*	R^2score
KHIT	opening price	102.14	0.4082	0.9414
ours	opening price	**101.98** ↓	**0.3873** ↓	**0.9427** ↑
KHIT	highest price	112.35	0.8574	0.9123
ours	highest price	**105.17** ↓	**0.8011** ↓	**0.9216** ↑
KHIT	lowest price	98.56	0.9936	0.9305
ours	lowest price	**97.44** ↓	**0.9374** ↓	**0.9486** ↑
KHIT	trading volume	142.18	0.7151	0.8762
ours	trading volume	**131.45** ↓	**0.6745** ↓	**0.8891** ↑
KHIT	closing price	93.37	0.4337	0.9524
ours	closing price	**92.89** ↓	**0.3982** ↓	**0.9702** ↑

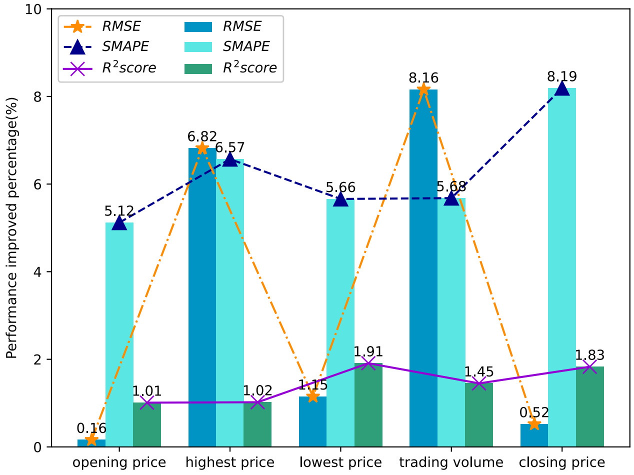

FIGURE 7. The percentage of performance metrics improvement for each prediction value visualization.

across experimental data from the China A-shares market test set and the stock data of the U.S. New York Stock Exchange (NYSE) and NASDAQ markets, as detailed in Table 5.

On the test set, the prediction results of various models exhibit similar trends. However, our HPMG-Transformer model performs the best in terms of prediction accuracy and stability, with its prediction results closest to the actual closing price. This indicates that our model has good performance in dealing with the complexity and outliers of time series data.

In order to further verify and analyze the advantages of our proposed model HPMG-transformer, we further prediction the model in the Liquor of China's A-share market, and obtain the corresponding opening price, closing price, highest price, lowest price, and trading volume for prediction. The comparison of the three accuracy performance of model output prediction, *RMSE*, *SMAPE*, and R^2*score*, between the optimal performance baseline model KHIT and the proposed model HPMG-transformer is shown in Table 6, showing the comparison of each prediction value.

Figure 7 shows the percentage of performance accuracy improvement for each prediction value visualization, with the *SMAPE* performance accuracy improvement for closing prices and lowest prices being the most significant, with a 8.19% improvement in closing prices.

V. CONCLUSION

In this research article, we present a novel approach to effectively predict stock index movements in the complex, non-linear, and non-stationary in the Chinese A-shares, the U.S. New York Stock Exchange (NYSE) and NASDAQ market. We utilize the classical HP filter, a common tool in macroeconomics, to extract valuable features from the volatile financial time series data. By implementing a long-short term decomposition, we filter out short-term fluctuations in the correlated time series data while also providing smoothing capabilities to handle outliers. The foundation of our approach is rooted in spectral analysis methods for financial time series, which enhance feature representation when fed into neural networks. Additionally, to tackle the challenge of weak positional information in the self-attention mechanism of the Transformer model, especially in capturing global dependencies and the significance of positional information in financial time series, we propose integrating a multi-scale Gaussian prior into the multi-head self-attention mechanism. This incorporation leverages the correlation between data and temporal distance, aiding in extracting local contextual information features. Through ablation experiments, detailed in Table 1, we demonstrate the superior performance of the HP and MG modules by evaluating three key metrics: Root Mean Square Error (*RMSE*), Symmetric Mean Absolute Percentage Error (*SMAPE*), and R^2*score*. Our proposed HPMG-Transformer model achieves the most favorable results.

To further demonstrate the advantages of our proposed model, we conduct comparative experiments with advanced baseline models and ablation experiments. The baseline models include TEANet, FDGRU-transformer, TEGRU, KHIT, and PCA-IGRU, and the evaluation metrics used are *RMSE*, *SMAPE*, and R^2*score*. Through experimental verification, as shown in Table 4 and 5, our HPMG-Transformer outperforms the highest-performing KHIT model by improving the corresponding *RMSE*, *SMAPE*, and R^2*score* by 0.51%, 8.19%, and 1.87% respectively. Furthermore, we compare RNN, LSTM, GRU, and Transformer through ablation experiments, evaluating them with *RMSE*, *SMAPE*, and R^2*score* metrics. Additionally, we employ the MD test to compare and evaluate the models, as depicted in Figure 5, Table 3 and 4, and find that the Transformer architecture demonstrates advantages. To validate the generalization capability of our proposed model, we select 104 listed companies in the Chinese A-share market and conduct experimental predictions for four major industry sector indices (liquor, beer, dairy, and beverage). As shown in Figure 6, through the comparison with different baseline models, we find that our HPMG-Transformer achieves the best performance in all three evaluation metrics: *RMSE*, *SMAPE*, and R^2*score*, as shown in Table 6. To validate the strong generalization capability of the HPMG-transformer model proposed in this paper, data from the Chinese A-share market and the four major industries, as well as the U.S. New York Stock Exchange (NYSE) and NASDAQ market segments, were used for verification. This demonstrates its superiority in comparison to the baseline model KHIT in terms of experimental accuracy, as shown in Table 5.

In conclusion, the proposed novel HPMG-Transformer model offers an effective method for predicting complex, non-linear, and non-stationary financial time series data. This model presents a fresh approach to addressing these challenges and demonstrates promising outcomes.

REFERENCES

[1] A. I. McLeod and W. K. Li, "Diagnostic checking ARMA time series models using squared-residual autocorrelations," *J. Time Ser. Anal.*, vol. 4, no. 4, pp. 269–273, Jul. 1983.

[2] K. Gilbert, "An ARIMA supply chain model," *Manage. Sci.*, vol. 51, no. 2, pp. 305–310, Feb. 2005.

[3] L. Bauwens, S. Laurent, and J. V. K. Rombouts, "Multivariate GARCH models: A survey," *J. Appl. Econometrics*, vol. 21, no. 1, pp. 79–109, Jan. 2006.

[4] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[6] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[7] A. Vaswani et al., "Attention is all you need," *ArXiv*, 2017, doi: 10.48550/arXiv.1706.03762.

[8] J. Yoo, Y. Soun, Y.-C. Park, and U. Kang, "Accurate multivariate stock movement prediction via data-axis transformer with multi-level contexts," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 2037–2045.

[9] T. Muhammad, A. B. Aftab, M. Ibrahim, M. M. Ahsan, M. M. Muhu, S. I. Khan, and M. S. Alam, "Transformer-based deep learning model for stock price prediction: A case study on Bangladesh stock market," *Int. J. Comput. Intell. Appl.*, vol. 22, no. 3, Sep. 2023, Art. no. 2350013.

[10] X. Hu, "Stock price prediction based on temporal fusion transformer," in *Proc. 3rd Int. Conf. Mach. Learn., Big Data Bus. Intell. (MLBDBI)*, Dec. 2021, pp. 60–66.

[11] S. Lai, M. Wang, S. Zhao, and G. R. Arce, "Predicting high-frequency stock movement with differential transformer neural network," *Electronics*, vol. 12, no. 13, p. 2943, Jul. 2023.

[12] E. Ramos-Pérez, P. J. Alonso-González, and J. J. Núñez-Velázquez, "Multi-transformer: A new neural network-based architecture for forecasting S&P volatility," *Mathematics*, vol. 9, no. 15, p. 1794, Jul. 2021.

[13] Y. Chen, W. Lin, and J. Z. Wang, "A dual-attention-based stock price trend prediction model with dual features," *IEEE Access*, vol. 7, pp. 148047–148058, 2019.

[14] D.-K. Kim and K. Kim, "A convolutional transformer model for multivariate time series prediction," *IEEE Access*, vol. 10, pp. 101319–101329, 2022.

[15] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662–131682, 2020.

[16] M. Wen, P. Li, L. Zhang, and Y. Chen, "Stock market trend prediction using high-order information of time series," *IEEE Access*, vol. 7, pp. 28299–28308, 2019.

[17] Q. Li, J. Tan, J. Wang, and H. Chen, "A multimodal event-driven LSTM model for stock prediction using online news," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 10, pp. 3323–3337, Oct. 2021.

[18] Y.-K. Kwon and B.-R. Moon, "A hybrid neurogenetic approach for stock forecasting," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 851–864, May 2007.

[19] X. Li, H. Xie, R. Y. K. Lau, T.-L. Wong, and F.-L. Wang, "Stock prediction via sentimental transfer learning," *IEEE Access*, vol. 6, pp. 73110–73118, 2018.

[20] M. Ali, D. M. Khan, I. Saeed, and H. M. Alshanbari, "A new approach to empirical mode decomposition based on Akima spline interpolation technique," *IEEE Access*, vol. 11, pp. 67370–67384, 2023.

[21] M. Ali, D. M. Khan, H. M. Alshanbari, and A. A.-A.-H. El-Bagoury, "Prediction of complex stock market data using an improved hybrid EMD-LSTM model," *Appl. Sci.*, vol. 13, no. 3, p. 1429, Jan. 2023.

[22] G. Tan and Z. Wang, "Stability analysis of recurrent neural networks with time-varying delay based on a flexible negative-determination quadratic function method," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–6, 2023, doi: 10.1109/TNNLS.2023.3327318.

[23] J. Hu, G. Tan, and L. Liu, "A new result on H_∞ state estimation for delayed neural networks based on an extended reciprocally convex inequality," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 71, no. 3, pp. 1181–1185, Mar. 2024.

[24] W.-J. Lin, G. Tan, Q.-G. Wang, and J. Yu, "Fault-tolerant state estimation for Markov jump neural networks with time-varying delays," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 71, no. 4, pp. 2114–2118, Apr. 2024.

[25] M. O. Ravn and H. Uhlig, "On adjusting the HP-filter for the frequency of observations," *CEPR Discussion Papers*, vol. 84, no. 5, pp. 371–376, 2001, doi: 10.2139/ssrn.289197.

[26] Q. Ding, S. Wu, H. Sun, J. Guo, and J. Guo, "Hierarchical multi-scale Gaussian transformer for stock movement prediction," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 4640–4646.

[27] M. G. Sousa, K. Sakiyama, L. d. S. Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, "BERT for stock market sentiment analysis," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2019, pp. 1597–1601.

[28] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *J. Bus. Econ. Statist.*, vol. 13, no. 3, p. 253, Jul. 1995.

[29] Q. Zhang, C. Qin, Y. Zhang, F. Bao, C. Zhang, and P. Liu, "Transformer-based attention network for stock movement prediction," *Expert Syst. Appl.*, vol. 202, Sep. 2022, Art. no. 117239.

[30] C. Li and G. Qian, "Stock price prediction using a frequency decomposition based GRU transformer neural network," *Appl. Sci.*, vol. 13, no. 1, p. 222, Dec. 2022.

[31] A. T. Haryono, R. Sarno, and K. R. Sungkono, "Transformer-gated recurrent unit method for predicting stock price based on news sentiments and technical indicators," *IEEE Access*, vol. 11, pp. 77132–77146, 2023.

[32] F. Lin, P. Li, Y. Lin, Z. Chen, H. You, and S. Feng, "Kernel-based hybrid interpretable transformer for high-frequency stock movement prediction," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2022, pp. 241–250.

[33] J. W. J. Wang, D. L. J. Wang, L. J. D. Liu, Q. S. L. Jin, and Z. X. Q. Sun, "A PCA-IGRU model for stock price prediction," *J. Internet Technol.*, vol. 24, no. 3, pp. 621–629, May 2023.

LILI HUANG received the master's degree in economics with a specialization in finance from Nanjing University of Science and Technology. She is currently employed with Anhui International Studies University. She has published two papers. Her research interests include AI in digital finance and securities investment. She has led three educational research projects and two scientific research projects. She received the Third Prize of Anhui Provincial Teaching Achievement Award.