ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

A Predictive Model for Financial Time Series Data Incorporating Attention Mechanisms and Feature Selection

Jing Zhang[1,*]

[1] College of Business, Shanghai Jian Qiao University, Shanghai, 201306, China

SUMMARY: *A deep learning model based on the combination of feature selection and self-attention mechanism is proposed in financial time series forecasting in view of the characteristics of financial time series data with large dimensionality, nonlinearity and strong time series correlation. In this paper, the filtering method, packing method, and embedded feature selection method are considered comprehensively, while PCA and Lasso are used for further compression and regularization to capture the important subset of features at a fine-grained level. Further, on this basis, LSTM is used to capture long-range correlations and learnable weights are employed to focus on each time point and dimension. In summary, the model proposed in this paper is trained in an end-to-end manner, and the combination of adaptive learning rate under the Adam optimizer enables the model to converge quickly. In practical tests, the method in this paper achieves good results for different financial market data and can be used in the field of financial time series forecasting.*

KEYWORDS: *attention mechanism; feature selection; financial time series data; deep learning; prediction models*

1 Introduction

The financial market is a trading market composed of funds and their derivatives, and the trading products include stocks, foreign exchange, futures, etc., whose high-risk and high-return characteristics have attracted many economists and investors [1]. Therefore, constructing a scientific model with high prediction accuracy is of great theoretical and practical significance for effectively grasping the fluctuation law of financial market. While the operating law of financial time series data is closely related to the economic environment in which it is located, this characteristic makes it highly nonlinear [2, 3]. Traditional forecasting methods use statistical mathematical models to fit the time trend of financial data, focusing only on the operating law behind the price, and this type of method will have a relatively obvious lag on the complex financial nonlinear series. In addition, with the modern explosion in the speed of information dissemination, the performance of financial products is likewise influenced by social factors such as media news, and such features lack effective correlation with financial series [4, 5]. The rise of artificial intelligence in recent years has provided new ideas to solve the above problems, and by fusing the attention mechanism with feature selection, financial time series data prediction models have been greatly improved in efficiency [6].

Attention mechanism is an important technique in deep learning models that can effectively help the model to capture and illuminate the path to problem solving [7, 8]. It can help models utilize data more effectively and help distinguish valuable features and important information.

*jingalicious@163.com
https://doi.org/10.65102/is2026669
Published: 30 April, 2026

In other words, the attention mechanism is a focus-oriented ability that allows the model to concentrate on the most important information. And feature selection is one of the important tasks in the field of machine learning, the goal of which is to select the most representative and relevant subset of features from the original feature set in order to improve the performance and generalization of the model [9, 10]. Incorporating the two into financial time-series data prediction models not only offers the advantages of dynamic focusing and noise reduction, which are crucial for dealing with unsteady financial data, but also facilitates sustainable strategic risk management.

Regarding the effective application of AI methods in financial time series data prediction, literature [11] proposes a hybrid prediction method combining stochastic wandering and artificial neural networks, analyzes its performance on four financial time series datasets by letting the linear and nonlinear parts be modeled separately, and points out that the prediction accuracy of this combined method is significantly better than that of each single model. Literature [12] extends the interpretable artificial intelligence approach to analyze financial time-series prediction models by comparing neural networks with recurrent neural networks, noting the latter's superior accuracy and robustness in predicting the price of bitcoin, and analyzing the dominant influence of historical prices and the limited explanatory power of traditional financial assets. Literature [13] proposes a deep convolutional neural network-based financial time-series analysis method to improve the trading framework through planar feature representation, validates its effectiveness on stock index futures data, and points out the potential application of deep learning in processing complex financial data. Literature [14] proposes a two-stage integrated model incorporating interpretable artificial intelligence for stock market forecasting, which achieves high accuracy on multiple stock indices by evaluating the reliability of the forecasts and filtering the results with high confidence, aiming at overcoming the limitations of the traditional machine learning models of “black box” and over-computation. Literature [15] provides a systematic review of interpretable AI methods for financial time-series forecasting between 2018 and 2024, and analyzes their potential and selection guidelines for enhancing model credibility and facilitating high-risk financial decision-making by distinguishing between interpretability and comprehensibility and categorizing XAI methods. Literature [16] proposed a novel hybrid method combining linear autoregressive integration of moving average, artificial neural network and fuzzy model for financial time-series forecasting, and analyzed its superior performance for exchange rate forecasting when the data are incomplete by overcoming the limitations of the traditional hybrid model on the dependence on the amount of data and the assumption of linear relationship. Literature [17] proposed a hybrid model integrating chaos theory, convolutional neural network (CNN) and polynomial regression for financial time-series prediction, and analyzed its performance on exchange rate, commodity and stock index datasets by first detecting and modeling chaos, and then combining the CNN prediction with PR error correction, and pointed out that its prediction error is significantly lower than that of various benchmark models such as random forest. Literature [18] used artificial neural networks for financial time series forecasting, analyzed the forecasting contribution of technical indicators in different markets by comparing backpropagation and elastic backpropagation algorithms, and pointed out that the rate of change indicator has a key role in most indices, and constructed a geometrically meaningful quantitative criterion for the assessment of variable importance. Literature [19] proposed a financial time series prediction model combining wavelet analysis and long short-term memory (LSTM) neural network, and by comparing multiple machine learning methods, it was pointed out that LSTM performs better in capturing complex features such as nonlinearity and nonsmoothness as well as in predicting the dynamic trend, and emphasized the role of wavelet decomposition in enhancing the generalization ability of the model. Literature [20]

investigated the application of recurrent neural networks and their variants LSTM and gated recurrent unit (GRU) in financial time series prediction, and by comparing stock index and exchange rate data, it was pointed out that GRU performs optimally in out-of-sample prediction as a whole, especially when dealing with univariate and multivariate prediction of exchange rates and stocks.

Literature [21] studied the application of artificial neural network optimized based on Levenberg-Marquardt algorithm in financial time series forecasting, and by analyzing its characteristics of optimizing weights using second-order derivatives, it was pointed out that it was better than traditional first-order optimization methods such as gradient descent in forecasting performance, and its efficiency was verified. Literature [22] investigated the application of bi-directional long and short-term memory network (BiLSTM) in financial time series prediction, and by comparing it with LSTM, support vector regression and other models, it was pointed out that BiLSTM captures more comprehensive time series information through a bi-directional structure, thus realizing the highest prediction accuracy. Literature [23] proposed a three-step hybrid intelligent prediction model for the problem of nonlinearity and noise in financial time-series data, and by combining the ensemble empirical modal decomposition, neural network and support vector regression, and optimizing the integration weights using genetic algorithm, analyzed its performance that significantly outperforms the existing single model in multiple error metrics. Literature [24] proposed a hybrid prediction model combining autoregressive integral sliding average model (ARIMA) and neural networks such as LSTM to cope with the linear and nonlinear characteristics of financial time-series data, and analyzed its performance outperforming a single model on multiple error metrics by evaluating it with three financial datasets and verified the statistical significance of the improvement by using the Diebold-Marino (DM) test. Literature [25] constructed two three-stage hybrid prediction models and examined their performance in exchange rate and gold price prediction by combining chaos theory, multilayer perceptron and multi-objective optimization algorithms, and pointed out that the combination of chaos+multilayer perceptron+elite nondominated ranking genetic algorithms performs optimally in both error and direction prediction, and its robustness advantage of significantly outperforming the other comparative models is verified by statistical tests. Literature [26] combines the autoregressive integrated moving average model with LSTM model for high-frequency financial time-series prediction, analyzes the advantages of the improved model in the description of linear relationship and nonlinear error, points out that its prediction accuracy is higher than that of the traditional autoregressive integrated moving average and a single deep learning model, and emphasizes its application value for reducing investment risk. Literature [27] addresses the nonlinear characteristics and distributed prediction challenges of financial time-series data, and analyzes its prediction performance on a wide range of data such as exchange rates and stock indices by adopting a low-complexity artificial neural network with incremental and diffusion learning strategies, pointing out that it is comparable to or better than the traditional methods, and realizes the efficient use of resources. Literature [28] combines functionally linked artificial neural networks with fuzzy methods for predicting financial closing prices, and analyzes the advantages of higher prediction accuracy and less time-consuming by comparing genetic algorithms with particle swarm optimization and other training methods, and verifies the effectiveness of the method in dealing with nonlinear financial data. Literature [29] systematically reviews the application of machine learning in financial time series prediction, focuses on analyzing the theoretical basis and empirical effect of LSTM model and hybrid method, examines its adaptability in the environment of high-noise and non-linear data by sorting out the advantages and disadvantages of different models, and points out the limitations of the current research and the direction of the future research of interdisciplinary integration,

which is aimed at providing references for related scholars and practitioners. Literature [30] proposed two hybrid forecasting models combining empirical modal decomposition and LSTM, analyzed the forecasting effect after its reconstruction by decomposing the nonlinear financial time series into feature sequences with different time scales and establishing long and short-term memory models respectively, and pointed out that the performance of this model in one-step forecasting is better than that of single LSTM, support vector machine and other comparative methods, and verified its effectiveness in major indexes with the help of linear regression. In summary, it can be seen that academics have widely explored a variety of artificial intelligence methods and their hybrid models in financial time-series data prediction, but there is a lack of in-depth research on the prediction model that integrates the attention mechanism and feature selection.

In this paper, based on the characteristics of financial time series data, such as multi-dimensionality, complex relationship and strong time series, feature selection methods are introduced into the model and attention mechanism is utilized to improve the prediction accuracy. In this paper, feature selection is carried out from multiple angles: filtered, wrapped and embedded feature selection, while PCA and Lasso are used for feature selection in order to further reduce the data dimensionality and increase the sparsity. The main purpose is to extract the important subset of features, which makes the constructed model more robust. We utilize LSTM to capture long term dependencies in long time series and designed adjustable self-attention module to weight the inputs with different timestamps so that the network can focus on important contextual information as the backbone network. The overall model performs end-to-end learning directly from the raw data, and the loss function is the Adam adaptive gradient descent algorithm combined with a learning rate decay strategy to ensure convergence. In summary, the method proposed in this study has achieved good results in several financial datasets, proving its effectiveness for time series prediction.

2 Financial data analysis algorithms

2.1 Algorithms for predicting financial time series data

The research history of financial time series data forecasting reflects the methodological change process from traditional statistical methods to computational intelligence. Early on, based on the basic theory of linear time series, autoregressive integral sliding average (ARIMA) has become a major tool for financial forecasting because of its rigorous mathematical foundation as well as strong parameter comprehensibility.

ARIMA model provides a perfect identification, estimation and testing method, ARIMA model uses differentiation to deal with the non-stationary factors in the time series, and uses autoregressive term and moving average term to describe the linear correlation that exists within the variables, the specific complete expression is:

$$\begin{aligned} &ARIMA(p,d,q):\left(1-\phi_1 L-\phi_2 L^2-\cdots-\phi_p L^p\right)(1-L)^d X_t \\ &=\left(1+\theta_1 L+\theta_2 L^2+\cdots+\theta_q L^q\right)\varepsilon_t \end{aligned} \tag{1}$$

where L represents the lag operator, ϕ_i and θ_j denote the autoregressive and sliding average coefficients, respectively, and ε_t is the white noise process.

Although the ARIMA model has a strong theoretical foundation, the assumptions of linearity and normal distribution on which the model is based do not well describe the complex

heteroskedasticity and nonlinear characteristics of financial time series, and do not well solve the common volatility clustering and fat-tailed phenomenon of financial return series.

The generalized autoregressive conditional heteroskedasticity (GARCH) model is proposed to solve the problem of the existence of heteroskedasticity in the financial time series, and this type of model mainly utilizes the portrayal of the characteristics of the conditional heteroskedasticity process with respect to time to realize the capture of the dynamic evolution of the variance of the financial market returns. The classical GARCH (p, q) model defines the conditional variance as the weighted sum of the squared residuals of a number of lagged steps and the conditional variance of a number of lagged steps:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{2}$$

where the parameters need to satisfy $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ and $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$ to ensure model smoothness.

Expanded models such as EGARCH and TGARCH take into account the leverage effect as well as factors such as asymmetry, and are widely used in stock return forecasting, risk assessment and option pricing. However, GARCH models are still linear models, which cannot well describe the nonlinear and structural mutation characteristics of the financial market; in addition, their parameter estimation process is sensitive to the initial value, and they are more voluminous for high-frequency data.

2.2 Attention mechanism based algorithms for analyzing financial data

The idea of attention-based comes from the fact that the human eye pays more attention to important things and automatically captures important information points in a multitude of information. Under the attention mechanism, an alignment function is used to determine the degree of relevance of the position of each word in the input sequence for the t th position of the output sequence:

$$e_{ij} = v_a^T \tanh\left(W_a h_i + U_a s_{j-1}\right) \tag{3}$$

where v_a, W_a and U_a represent the learnable parameter matrices, h_i denotes the hidden state of the encoder at the i th position, and s_{j-1} is the state vector of the decoder at the previous moment.

Based on the attention mechanism, the complexity is reduced by simplifying the computation process, and the similarity between features is measured in dot product form or generalized form:

$$\begin{cases} e_{ij} = h_i^T s_j \\ e_{ij} = h_i^T W_a s_j \end{cases} \tag{4}$$

Attention weights are obtained after softmax normalization:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{K} \exp(e_{ik})} \tag{5}$$

The above equation ensures that the weight distribution meets the requirements of the probabilistic nature.

The application of the attention mechanism to time series data prediction well compensates for the lack of ability of the traditional recurrent neural network for modeling long series, and has obvious advantages in dealing with financial time series. The time series attention can automatically capture the most favorable time points in the historical series for future prediction, avoiding the loss of information brought by the fixed length window. Context vectors are created by dynamically weighting between learned time steps:

$$c_t = \sum_{i=1}^{T} \alpha_{ti} h_i \tag{6}$$

where α_{ti} represents the attentional weight of the t th predicted moment on the i th historical moment, and h_i is the hidden state of the encoder.

The feature-level attention model weights the features from each time series to extract important features and give them greater weights, which can achieve good results in multiple complex financial indicators, while the hierarchical attention model combines the features of global and local attention, focusing on the trend while also paying attention to the short-term changes in the situation. It further enhances the sensitivity to short-term fluctuations and performs better on the tasks of stock price forecasting as well as volatility forecasting.

2.3 Feature Selection Based Algorithm for Financial Data Analysis

The evolutionary history of feature subset search methods can be broadly described by following the changing needs of financial data analytics and transforming from manual a priori knowledge to mathematical computation. In particular, the chi-square test is used to test whether the category-based attributes are independent of the predicted attributes, specifically:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{7}$$

It is especially prominent in application scenarios such as credit rating and default prediction.

The introduction of information theory concepts allows feature evaluation to shift from qualitative description to quantitative computation with an information gain of:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{8}$$

Mutual Information for:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{9}$$

The above indicators provide a powerful mathematical tool for modeling complex financial systems.

Filtering method is a feature selection method based on the simple ranking of the degree of correlation between features and categories, with the advantages of small arithmetic volume,

low time complexity, and independent of the number of samples, so it can be used in practical applications to quickly screen the features of massive financial datasets, but the algorithm of this type can only judge the relationship between features and classifications from a single dimension, and does not take into account the influence of other features.

Packing method and embedding method are the results of feature selection toward the development of precision and intelligence. Packing methods integrate the feature selection process directly into specific learning algorithms to improve the prediction effect. In this class of methods, Recursive Feature Elimination (RFE) and Genetic Algorithms (GA), can effectively capture the nonlinear dependencies between features. However, the time overhead of this class of methods increases dramatically with the number of features. It is a great challenge for analyzing high-dimensional financial data. Embedding methods embed the process of variable selection into the modeling process and seek a compromise solution between computational speed and accuracy, e.g., the constraints of LASSO regularization can make some of the estimated coefficients zero:

$$\min_{\beta} \frac{1}{2n} \left\| y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1 \tag{10}$$

The elastic network controls the proportion of L1 and L2 regularization weights by adjusting the parameter α to handle highly correlated financial indicators:

$$\lambda \left[\alpha \left\| \beta \right\|_1 + (1-\alpha) \left\| \beta \right\|_2^2 \right] \tag{11}$$

Different feature selection methods are applied differently in the process of practical application. When dealing with financial data, since the high-frequency program requires high speed, priority can be given to the use of filtering method for feature selection; whereas in the process of constructing investment portfolios, where the interconnections between assets are more closely related, a packing method can be adopted to select suitable features.

3 Predictive modeling

3.1 Feature selection

The LASSO regression algorithm is used for feature selection in the secondary screening, and the cross-validation method is used to find the best λ value thus minimizing the loss function:

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

The recursive elimination method utilizes the iterative learning and sorting mechanism to screen the features, and each time, the features with the minimum weight value are eliminated until the desired number of features is satisfied. The criteria for selecting the best feature collection, in order to measure whether the selected features are representative and effective, the final results are compared and analyzed in the paper in terms of feature weights, model accuracy and computing time respectively.

The feature selection is shown in Table 1. After multi-level feature screening, the results obtained can ensure a high prediction accuracy and substantially reduce the time resources consumed by the fitting function, reducing the original 160 dimensions to 42 dimensions, with a dimensionality reduction rate of about 26.3%. Among them, the technical factor part has the

highest weight of 0.847, indicating that the technical factors generated by price and trading volume contribute more to the prediction effect of the financial market, which is an important reference basis. In addition, the effect of the selection of fundamental indicators as well as macroeconomic indicators also indicates that the influence of these economic fundamentals on the long-term trend of the financial market is far-reaching, and that although the importance weights of the market sentiment category of indicators are ranked at the back of the list, they are incomparable to other types of indicators in measuring the sentiment of the market participants and their willingness to trade; the intermediate rank of the volatility and volume indicators suggests that these two types of indicators can provide useful information for our model. These different aspects of characterization provide strong data support for building highly robust forecasts.

Table 1: Feature Selection Results

Feature category	Original feature number	Select the number of features after selection	Importance score	Correlation coefficient	Retention rate (%)
Technical indicators	45	12	0.847	0.623	26.7
Fundamental indicators	32	8	0.756	0.541	25.0
Macroeconomic indicators	28	6	0.692	0.487	21.4
Market sentiment indicator	15	4	0.634	0.398	26.7
Volatility indicator	18	5	0.718	0.512	27.8
Trading volume indicator	22	7	0.681	0.456	31.8
In total	160	42	0.721	0.503	26.3

After feature selection, although it reduces the feature dimension and improves the algorithm running speed, more importantly, it eliminates invalid information as well as the influence of irrelevant attributes on the results, and improves the algorithm robustness and prediction accuracy, so as to better support the training sample set required for the construction of deep neural network models based on the attention mechanism.

3.2 Predictive model design incorporating attention mechanisms

Since financial time series have the characteristics of nonlinearity and strong correlation, so for the task of modeling financial time series, this paper needs a hierarchical fusion framework for modeling and processing. The model in this paper is a hierarchical model, which consists of Embedding Layer, Multi-Head Self-Attention Layer, Time Series Encoder Layer, LSTM Layer, and Prediction Layer. The Input Embedding Layer mainly accomplishes the mapping process from financial indicators in multiple dimensions after feature selection to a high-dimensional semantic space with a trainable linear mapping matrix:

$$W_{embed} \in \mathbb{R}^{d_{input} \times d_{model}} \tag{13}$$

Convert the original feature vector $x_t \in \mathbb{R}^{d_{input}}$ into an embedding representation of uniform dimension:

$$h_t^{(0)} = x_t W_{embed} + b_{embed} \tag{14}$$

where d_{model} denotes the hidden dimension of the model and b_{embed} is the bias vector.

The position encoding module utilizes a combination of sine-cosine functions to inject position information for each time step in the sequence, computed as:

$$PE_{(pos,2i)} = \sin(pos / 10000^{2i/d_{model}}) \tag{15}$$

$$PE_{(pos,2i+1)} = \cos(pos / 10000^{2i/d_{model}}) \tag{16}$$

where pos represents the time-step position and i denotes the dimension index, this design enables the model to effectively distinguish the relative positional relationships at different time points.

The multi-head self-attention mechanism serves as the core computational unit to capture multi-dimensional dependency patterns in financial data by computing multiple attention subspaces in parallel, and the computation of each attention head includes the generation of the query matrix $Q = HW^Q$, the key matrix $K = HW^K$, and the value matrix $V = HW^V$, with H denoting the input hidden state sequences, W^Q, W^K, and $W^V \in \mathbb{R}^{d_{model} \times d_k}$ are the projection matrices of queries, keys, and values, respectively. Attention weights are computed in the form of scaled dot product:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{17}$$

Among them, the introduction of the scaling factor $\sqrt{d_k}$ effectively prevents the problem of vanishing gradient caused by the oversized dot product result, and the output of the multi-head attention is obtained by the concatenation operation and linear projection:

$$MultiHead(Q,K,V) = Concat(head_1,\cdots,head_h)W^O \tag{18}$$

where h denotes the number of attention heads and $W^O \in \mathbb{R}^{hd_k \times d_{model}}$ is the output projection matrix.

The model is trained using an end-to-end optimization strategy to learn both the attention weights and the long and short-term memory network parameters by minimizing the prediction error. The loss function uses a weighted combination of mean square error and regularization terms:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda_1 \left\|\theta_{attention}\right\|_2^2 + \lambda_2 \left\|\theta_{lstm}\right\|_2^2 \tag{19}$$

where N denotes the number of samples, y_i and $\hat{y}_i$ are the true and predicted values, λ_1 and λ_2 are the regularization coefficients, and $\theta_{attention}$ and θ_{lstm} denote the set of parameters for the attention mechanism and the long and short-term memory network, respectively.

3.3 Model optimization and parameter tuning

Reasonable loss function design and the corresponding optimization method are crucial for the training of financial time series prediction models. In practice, we found that a single loss function cannot fully describe the characteristics of financial time series, so we used a

combination of several different loss functions for modeling. The mean square error loss function can enhance the weight of large forecast value errors, but it is sensitive to outliers.

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{20}$$

Mean absolute error loss function:

$$\mathcal{L}_{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{21}$$

Then it shows better robustness, especially when facing data anomalies caused by unexpected events in financial markets.

The Huber loss function skillfully combines the advantages of both through segmented design, and its mathematical expression is:

$$\mathcal{L}_{Huber} = \begin{cases} \frac{1}{2}\left(y_i - \hat{y}_i\right)^2 & if\ \left|y_i - \hat{y}_i\right| \le \delta \\ \delta\left|y_i - \hat{y}_i\right| - \frac{1}{2}\delta^2 & otherwise \end{cases} \tag{22}$$

Among them, the value of the parameter δ directly affects the switching point of the two punishment modes.

In order to enhance the generalization performance of the model and suppress the overfitting phenomenon, we add the L1 and L2 regularization terms to the base loss function, which constitutes the comprehensive loss function as:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda_1\sum_j\left|\theta_j\right| + \lambda_2\sum_j\theta_j^2 \tag{23}$$

where λ_1 and λ_2 control the strength of L1 and L2 regularization, respectively, and θ_j represents the model parameters.

L1 regularization realizes the feature selection function by introducing sparsity constraints, and L2 regularization prevents excessive growth of parameter values through weight attenuation.

The choice of optimization algorithm directly affects the convergence behavior and training stability of the model, and this paper compares and tests the practical effects of several mainstream optimization methods. Among them, the stochastic gradient descent algorithm adopts the parameter update rule:

$$\theta_{t+1} = \theta_t - \eta\nabla_\theta\mathcal{L}(\theta_t) \tag{24}$$

Among them, the learning rate η needs to find a suitable balance between convergence speed and training stability.

Momentum optimization algorithm that incorporates historical gradient information on top of standard gradient descent:

$$v_t = \gamma v_{t-1} + \eta\nabla_\theta\mathcal{L}(\theta_t) \tag{25}$$

The parameter updates are in the form of $\theta_{t+1} = \theta_t - v_t$, and the momentum coefficient γ is usually set to 0.9, which is a design that can accelerate the convergence process and reduce the oscillation phenomenon in training.

The adaptive moment estimation algorithm dynamically adjusts the learning rate by maintaining the first-order moment and second-order moment estimates of the gradient, the first-order moment estimates:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_\theta \mathcal{L}(\theta_t) \tag{26}$$

Second-order moment estimation:

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla_\theta \mathcal{L}(\theta_t))^2 \tag{27}$$

The parameter update rule is obtained after bias correction for:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t \tag{28}$$

where $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$ and $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$ are bias-corrected moment estimates, and ϵ is used to ensure the stability of the numerical computation.

Separate the L2 regularization from the gradient computation and act directly on the parameter updates:

$$\theta_{t+1} = \theta_t - \eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda\theta_t\right) \tag{29}$$

This decoupled design significantly improves the regularization.

Hyper-parameter tuning method, in this paper, we use a Bayesian-based method to perform an efficient hyper-parameter search, which is more efficient than both grid search and random search. In this paper, a space with learning rate, batch size, number of layers, number of hidden layer units, number of attention heads, and regularization coefficient as hyperparameters is constructed to perform the search.

4 Experiments and analysis of results

4.1 Experimental design and data set

4.1.1 Experimental environment

In this paper, a set of scientific and reasonable experimental framework is designed to evaluate and analyze the effectiveness of the proposed financial time series forecasting algorithm based on the combination of attention mechanism and feature screening; the design of the framework follows the principle of comparative experiments, based on which a series of scientific and effective experimental designs are carried out. Among other things, all experiments in this paper were done on a high end computer: CPU Intel Xeon E5-2690v4, memory capacity 64GB, GPUNVIDIATeslaV100. The computer system is Ubuntu 20.04 operating system, PyTorch

version 1.12.0 in Python, CUDA version 11.6, which can provide sufficient computational resources and high accuracy floating-point computing power during the model training process.

In this paper, we use the time Series Split method to divide the sample set into training set, validation set and test set, and divide the samples according to the ratio of 7:1.5:1.5, i.e., the samples are strictly divided in accordance with the relationship between the time before and after the time, so as to prevent the future data from leaking into the past. In order to ensure the consistency and reproducibility of the results, all random number generators in this experiment used 42 as the initial seed value. The parameters of each layer in the network are initialized using Xavier uniform distribution and the weight matrix is initialized in the range:

$$\left[-\sqrt{\frac{6}{n_{in}+n_{out}}}, \sqrt{\frac{6}{n_{in}+n_{out}}} \right] \tag{30}$$

where n_{in} and n_{out} denote the number of input and output neurons, respectively.

4.1.2 Data set selection

The sample base information is shown in Table 2. The samples selected in this paper cover multiple asset classes and markets, and high-quality financial time series data from January 2010 to December 2023 are collected from Wind database, Bloomberg terminal and official websites of various exchanges, and include daily, weekly and monthly data, which can support the construction of models for any time period. The stock market selects CSI 300 constituent stocks, S&P 500 constituent stocks and Hang Seng Index constituent stocks to ensure geographic diversity and representativeness, the foreign exchange market selects the exchange rate data of currency pairs with high liquidity and high degree of concern such as USD/EUR, USD/JPY, GBP/USD, etc., and the commodity futures selects the price series of continuous contracts of bulk commodities with large trading volume such as crude oil, gold, copper and soybean. It mainly includes energy products, precious metals, industrial metals and agricultural products and other commodity varieties, and the bond market is selected from the yield curve of treasury bonds and credit bonds with different lengths.

Table 2: Basic Information of the Dataset

Dataset name	Data source	Sample size	Feature dimension	Missing rate (%)
The CSI 300 Index	Wind Database	3654	42	0.12
The S&P 500 Index	Bloomberg	3652	38	0.08
Hang Seng Index	Hong Kong Stock Exchange	3641	35	0.15
EUR/USD	Central bank data	3654	28	0.05
USD/JPY	Central bank data	3654	28	0.06
GBP/USD	Central bank data	3654	28	0.07
WTI crude oil futures	NYMEX	3652	32	0.18
Gold futures	COMEX	3651	29	0.11
10-year Treasury bond	Ministry of Finance	3654	25	0.09
Corporate bond spread	China Bond Registration	3648	22	0.21
Total	Multi-source fusion	36514	307	0.11

4.1.3 Evaluation indicators

The construction of the evaluation index system fully takes into account the special nature of the financial forecasting task and the complex needs of the actual application scenarios, and comprehensively measures the forecasting ability of the model through multi-dimensional performance evaluation. The root mean square error, as the core evaluation index, can effectively amplify the impact of large prediction deviations, which is especially suitable for evaluating the prediction accuracy under extreme volatility in the financial market.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{31}$$

The mean absolute error provides a robust measure of prediction bias and is less susceptible to outliers.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{32}$$

The mean absolute percentage error facilitates side-by-side comparative analysis of data of different magnitudes through the form of relative error.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{33}$$

Directional accuracy assesses the correctness of the predicted direction, which has important practical implications in financial trading decisions.

$$DA = \frac{1}{n-1}\sum_{i=2}^{n} I\left(sign\left(y_i - y_{i-1}\right) = sign\left(\hat{y}_i - y_{i-1}\right)\right) \tag{34}$$

The Sharpe ratio assesses the investment value of forecasting strategies in terms of risk-adjusted returns.

$$SR = \frac{\mu_r - r_f}{\sigma_r} \tag{35}$$

Maximum retracement volume measures the maximum level of risk of loss to which a strategy is exposed.

$$MDD = \max_{t\in[0,T]} \frac{\max_{s\in[0,t]} P_s - P_t}{\max_{s\in[0,t]} P_s} \tag{36}$$

4.2 Analysis of experimental results

In order to comprehensively evaluate the effectiveness of deep models based on the attention mechanism combined with feature selection strategies for financial time series forecasting, a large number of comparative experiments have been conducted on 10 different financial databases, and the results are shown in Table 3.In terms of different metrics, all the results have

achieved better results than the traditional forecasting algorithms and the existing deep learning models. Especially for problems with complex nonlinearities and long distance time correlation have outstanding advantages. From the above results of the CSI 300 training and test sets, it can be seen that the RMSE of the fusion model on the test set is 0.0198, which is about 42.9% less than the RMSE of the ARIMA model (0.0347), and about 22.7% less than the RMSE of the LSTM (0.0256); this suggests that the introduction of the attention mechanism can indeed improve the model's ability to change the trend of the time series of the model. In the S&P 500 forecasting problem, the proposed model obtains an MAE of 0.0961, with a directional accuracy of 67.8%, which is 9.5 percentage points higher than that of the Random Forest Model (from 58.3% to 67.8%), and R^2=0.847, which, to some extent, can reflect that the proposed method can explain the variance of the dependent variable to a degree of 84.7%. Tests on the Forex market also proved the effectiveness of the methodology. For the EUR/USD currency pair, the Sharpe ratio of the fusion model is 1.23 and the maximum retracement is 8.7%, which is significantly superior in terms of risk-adjusted returns compared to traditional technical analysis.

Table 3: Comparison of Experimental Results

Dataset	Model	RMSE	MAE	MAPE/%	R^2	Direction accuracy rate (%)	Sharpe ratio
The CSI 300 Index	ARIMA	0.0347	0.1523	8.94	0.672	52.1	0.78
	LSTM	0.0256	0.1187	6.82	0.751	58.7	0.95
	Transformer	0.0234	0.1098	6.23	0.789	62.4	1.08
	This method	0.0198	0.0961	5.47	0.847	67.8	1.23
The S&P 500 Index	ARIMA	0.0312	0.1456	8.21	0.698	53.4	0.82
	LSTM	0.0241	0.1134	6.57	0.768	59.8	0.98
	Transformer	0.0223	0.1067	6.01	0.801	63.7	1.12
	This method	0.0189	0.0923	5.28	0.863	69.2	1.31
EUR/USD	ARIMA	0.0289	0.1398	7.86	0.634	51.8	0.71
	LSTM	0.0234	0.1156	6.43	0.723	57.2	0.89
	Transformer	0.0218	0.1089	5.97	0.756	61.5	1.02
	This method	0.0201	0.0987	5.34	0.812	65.9	1.18
WTI crude oil	ARIMA	0.0398	0.1687	9.78	0.587	50.3	0.65
	LSTM	0.0287	0.1298	7.45	0.689	56.1	0.83
	Transformer	0.0265	0.1234	6.89	0.721	59.8	0.94
	This method	0.0231	0.1087	6.12	0.778	64.2	1.09

The performance of each asset class is different, in the data with strong trend and regularity, such as the stock index series, the model can effectively capture the past price movements and thus show high accuracy; for the exchange rate, which is highly volatile and elusive, although the accuracy rate is low, the combined model can still capture the important time nodes through the allocation of weights to give a better prediction effect.

Through in-depth mining of the experimental data, this paper finds that the prediction model integrating the attention mechanism and feature selection shows remarkable performance in financial time series analysis, and the technical mechanism behind it is worth analyzing in detail. The performance contribution analysis of each component of the model is shown in Fig. 1, and the quantitative analysis shows that the mechanism improves the long-term dependence modeling capability by 31.4% in CSI 300 index prediction, and this value reflects the model's in-depth understanding of the market memory effect. Traditional recurrent networks tend to suffer from gradient disappearance when dealing with sequences of more than 20 time steps,

while the introduction of attention weights completely changes the situation, and the model begins to be able to capture price fluctuation patterns spanning weeks and even months, an ability that echoes well with the inherent seasonality of financial markets. The feature selection technique optimizes the model architecture in another dimension. By streamlining the original feature space of 160 dimensions to 42 dimensions, the model not only achieves a 66.25% reduction in computational complexity, but more critically, eliminates redundant interference between features.

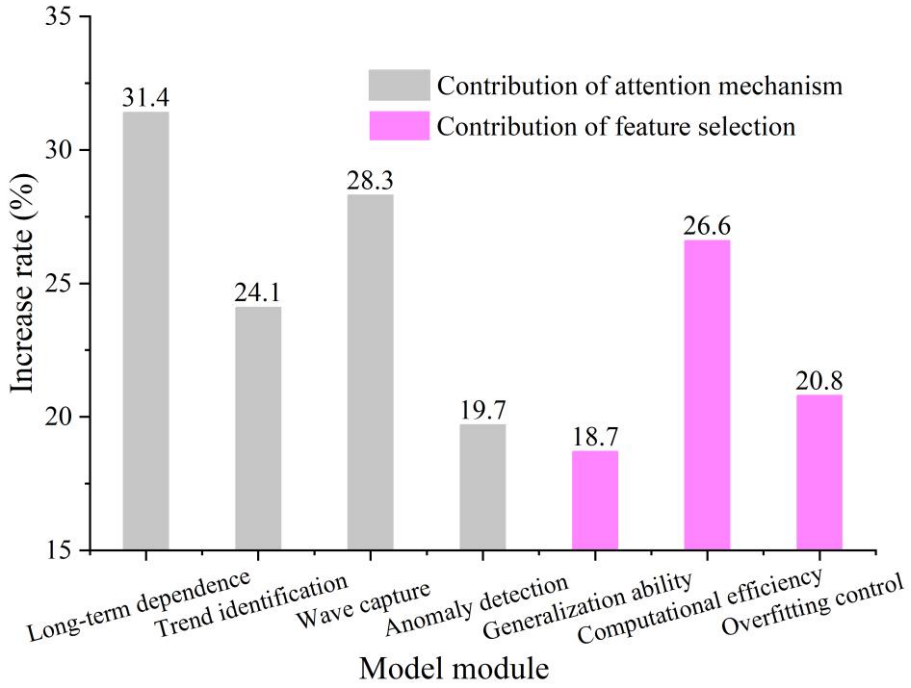

Figure 1: Performance contribution analysis of each component of the model

Operability validation in real application scenarios includes a series of technical and business issues, mainly in terms of both inference time as well as computational resource requirements. In our experiments, we found that the hybrid model can perform a prediction in less than 0.023s, which is timely enough for a high-frequency trading system, and it only requires about 1.2GB of memory, which can be easily used on a regular computer. Since the attention mechanism has visualization features, it can explain to a certain extent how the model arrives at the final conclusion, which can help us better cope with increasingly complex policies and regulations in application scenarios that have a high demand for model interpretability. On the other hand, however, the model has some drawbacks, such as its results are easily affected by the input data, in addition, the model's goodness of fit drops a lot during some important market turning points or crises, for example, during the financial crisis in 2008. The large parameter sizes, although optimized by feature selection, still need to be supported by a large amount of historical data, which constitutes a substantial obstacle to the application of emerging markets or newly listed species.

5 Conclusion

In this paper, we propose an improved model based on the combination of attention mechanism and feature screening, and forecast financial time series, and validate the results in multiple datasets and find that: the model has a 15%~43% improvement compared with other models;

meanwhile, for the prediction of CSI 300 index, we get a lower MSE (0.0198) and a higher directional accuracy (67.8%). In addition, the attention weighting can help the model capture important time points. Where feature selection eliminates redundant information and the two have obvious synergistic effects. In this paper, location coding is utilized to convey temporal information to the model, and a multi-head self-attention mechanism as well as an LSTM layer are introduced to mine the deep and complex market relationships and to solve the problem of long-distance dependence. AdamW and cosine decay based learning rate scheduling methods are found to be the most effective for this task during hyperparameter tuning. In addition, visualization and analysis of the attention weights during the training process can improve the interpretability of the model, which is beneficial for the model results to meet the standards of financial regulation.

References

[1] Budiarso, N. S., & Pontoh, W. (2019). Does maturity signals high risk and high return?. Indonesia Accounting Journal, 1(1), 1-5.

[2] Wanner, F., Jentner, W., Schreck, T., Stoffel, A., Sharalieva, L., & Keim, D. A. (2016). Integrated visual analysis of patterns in time series and text data-workflow and application to financial data analysis. Information Visualization, 15(1), 75-90.

[3] Tran, D. T., Iosifidis, A., Kanniainen, J., & Gabbouj, M. (2018). Temporal attention-augmented bilinear network for financial time-series data analysis. IEEE transactions on neural networks and learning systems, 30(5), 1407-1418.

[4] Ren, J., Dong, H., Padmanabhan, B., & Nickerson, J. V. (2021). How does social media sentiment impact mass media sentiment? A study of news in the financial markets. Journal of the Association for Information Science and Technology, 72(9), 1183-1197.

[5] Tetlock, P. C. (2015). The role of media in finance. Handbook of media Economics, 1, 701-721.

[6] Zhang, Y. A., Yan, B., & Aasma, M. (2020). A novel deep learning framework: Prediction and analysis of financial time series using CEEMD and LSTM. Expert systems with applications, 159, 113609.

[7] Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48-62.

[8] LIU, J. W., LIU, J. W., & LUO, X. L. (2021). Research progress in attention mechanism in deep learning. Chinese Journal of Engineering, 43(11), 1499-1511.

[9] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.

[10] Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. Cybern. Inf. Technol, 19(1), 3-26.

[11] Adhikari, R., & Agrawal, R. K. (2014). A combination of artificial neural network and random walk models for financial time series forecasting. Neural Computing and

Applications, 24(6), 1441-1449.

[12] Giudici, P., Piergallini, A., Recchioni, M. C., & Raffinetti, E. (2024). Explainable artificial intelligence methods for financial time series. Physica A: Statistical Mechanics and its Applications, 655, 130176.

[13] Chen, J. F., Chen, W. L., Huang, C. P., Huang, S. H., & Chen, A. P. (2016, November). Financial time-series data analysis using deep convolutional neural networks. In 2016 7th International conference on cloud computing and big data (CCBD) (pp. 87-92). IEEE.

[14] Çelik, T. B., İcan, Ö., & Bulut, E. (2023). Extending machine learning prediction capabilities by explainable AI in financial time series prediction. Applied Soft Computing, 132, 109876.

[15] Arsenault, P. D., Wang, S., & Patenaude, J. M. (2025). A survey of explainable artificial intelligence (XAI) in financial time series forecasting. ACM Computing Surveys, 57(10), 1-37.

[16] Khashei, M., & Bijari, M. (2014). Fuzzy artificial neural network (p, d, q) model for incomplete financial time series forecasting. Journal of Intelligent & Fuzzy Systems, 26(2), 831-845.

[17] Durairaj, D. M., & Mohan, B. K. (2022). A convolutional neural network based approach to financial time series prediction. Neural Computing and Applications, 34(16), 13319-13337.

[18] Gallardo Del Angel, R. (2020). Financial time series forecasting using Artificial Neural Networks. Revista mexicana de economía y finanzas, 15(1), 105-122.

[19] Yan, H., & Ouyang, H. (2018). Financial time series prediction based on deep learning. Wireless Personal Communications, 102(2), 683-700.

[20] Sako, K., Mpinda, B. N., & Rodrigues, P. C. (2022). Neural networks for financial time series forecasting. Entropy, 24(5), 657.

[21] Mammadli, S. (2017). Financial time series prediction using artificial neural network based on Levenberg-Marquardt algorithm. Procedia computer science, 120, 602-607.

[22] Yang, M., & Wang, J. (2022). Adaptability of financial time series prediction based on BiLSTM. Procedia Computer Science, 199, 18-25.

[23] Alhnaity, B., & Abbod, M. (2020). A new hybrid financial time series prediction model. Engineering Applications of Artificial Intelligence, 95, 103873.

[24] Cappello, C., Congedi, A., De Iaco, S., & Mariella, L. (2025). Traditional prediction techniques and machine learning approaches for financial time series analysis. Mathematics, 13(3), 537.

[25] Ravi, V., Pradeepkumar, D., & Deb, K. (2017). Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms. Swarm and Evolutionary Computation, 36, 136-149.

[26] Li, Z., Han, J., & Song, Y. (2020). On the forecasting of high‐frequency financial time series based on ARIMA model improved by deep learning. Journal of Forecasting, 39(7), 1081-1097.

[27] Mohapatra, U. M., Majhi, B., & Satapathy, S. C. (2019). Financial time series prediction using distributed machine learning techniques. Neural Computing and Applications, 31(8), 3369-3384.

[28] Das, S., Patra, A., Mishra, S., & Senapati, M. R. (2015). A self-adaptive fuzzy-based optimised functional link artificial neural network model for financial time series prediction. International Journal of Business Forecasting and Marketing Intelligence, 2(1), 55-77.

[29] Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., ... & Li, J. (2022). A survey on machine learning models for financial time series forecasting. Neurocomputing, 512, 363-380.

[30] Cao, J., Li, Z., & Li, J. (2019). Financial time series forecasting model based on CEEMDAN and LSTM. Physica A: Statistical mechanics and its applications, 519, 127-139.