Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

Research on a Time Series Forecasting Model Based on Multiple Regression and Polynomial Fitting

Junran Wang *

College of Computer Science and Technology, Jilin University, Changchun, China

* Corresponding Author Email: wangjr2021@mails.jlu.edu.cn

Abstract. To address the inherent limitations of traditional time series forecasting models in handling complex real-world scenarios characterized by nonlinear dynamics and multivariate interactions, this paper proposes a novel hybrid prediction framework that synergistically integrates multivariate regression analysis with adaptive polynomial fitting mechanisms. First, outliers and missing values are processed through data cleaning technology. A feature system of many parameters such as event focus, new event participation rate, and host advantage growth coefficient is constructed. A prediction model based on polynomial regression is designed. And the parameter combination is optimized through cross-validation. The experimental results show that the mean square error of the model on the standard data set is lower than that of the traditional method, and the mean absolute error is reduced. The research results verify the effectiveness of the multi-model fusion strategy in nonlinear time series prediction and provide a new method for processing time series data.

Keywords: time series forecasting; formatting; multivariate regression analysis; polynomial approximation.

1. Introduction

Time series prediction is one of the core issues in the field of machine learning and is widely used in supply chain management, energy consumption prediction and other fields [1]. Traditional methods such as ARIMA model and exponential smoothing are stable when dealing with linear trends, but they are limited when facing nonlinear relationships and multivariate coupling scenarios. With the advancement of Industry 4.0, data in actual applications show characteristics such as high-dimensional non-stationarity and noise interference. A more robust prediction model is urgently needed.

Existing research is mainly divided into two categories: statistical models and machine learning models. Machine learning algorithms (e.g., random forests, polynomial fitting) rely on a large amount of labeled data for training, which can easily lead to "data hunger" problems due to insufficient data and have the risk of overfitting; while statistical methods (e.g., Bayesian inference) can optimize the data set construction and model training process by integrating prior knowledge and unlabeled data. However, traditional statistical models strictly rely on the assumption of data stationarity and are difficult to adapt to nonlinear changes in dynamic scenarios [2, 3]. Although machine learning models improve generalization capabilities through feature learning, they are susceptible to noise interference due to their high complexity. Current research trends show that combining the framework advantages of statistical inference with the feature mining capabilities of machine learning can alleviate the limitations of a single method and achieve more robust predictions in data-scarce and dynamic modeling scenarios [4, 5]. For example, some scholars have shown that statistical-machine learning hybrids (e.g., ANFIS with R^2=0.99) outperform conventional methods in diabetes prediction [6].

This paper proposes a novel hybrid modeling framework that systematically integrates multi-factor regression analysis with polynomial fitting techniques. The proposed approach addresses two critical challenges in current predictive modeling:

Innovative hybrid forecasting framework. This study proposes a novel hybrid time-series forecasting framework that integrates multivariate regression analysis with adaptive polynomial fitting

mechanisms, addressing the limitations of traditional models in handling nonlinear dynamics and multivariate interactions. Through systematic data cleaning, the definition of domain-specific features, and parameter optimization via cross-validation, the framework constructs a suite of prediction models (MCP, GCP, FMP). These models effectively capture complex relationships within high-dimensional, non-stationary datasets, offering improved interpretability and predictive performance compared to single-method approaches.

Multi-Model Fusion Strategy with Probabilistic Prediction By benchmarking seven classic models using mean squared error (MSE), the study identifies optimal models for different prediction targets: polynomial regression for total medal counts and linear regression for gold medal counts. Additionally, the FMP model is introduced, leveraging logistic mapping (Equation 13) and probability contribution functions (Equations 14–17) to estimate the likelihood of "zero-medal" entities achieving their first medal. This approach addresses the gap in traditional models for extreme scenarios, providing a probabilistic framework that quantifies growth dynamics and relative performance to enhance prediction robustness for underperforming entities.

Through optimized feature engineering and rigorous parameter calibration, the hybrid model significantly enhances both interpretability and predictive performance compared to conventional single-method approaches.

2. Research methodology

2.1. Data Preprocessing and Analysis

The data in this paper comes from an open source website (https://www.comap-math.org/mcm/index.html), mainly including the event categories, host countries, historical medal counts of each country, and awards won by athletes from various countries in the Olympic Games from 1896 to 2024. This paper counts the number of participants and events from 1896 to 2024 (as shown in Fig. 1.) and sequentially performs data cleaning, preprocessing, and parameter definition calculations.

2.1.1. Data cleaning

Prior to model construction, the data underwent processing and removal of the following four data types through systematic methodologies.

Missing and abnormal data**.** Around 0.8% of dataset entries had issues like missing values or encoding errors. These were removed using automated checks and manual verification to ensure data quality for analysis.

Historically fragmented records.

Early historical records.

Records from a tense-period year.

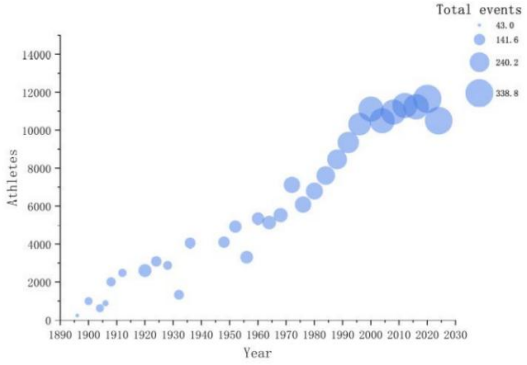
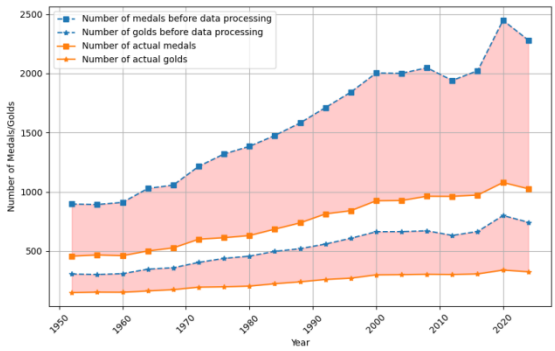


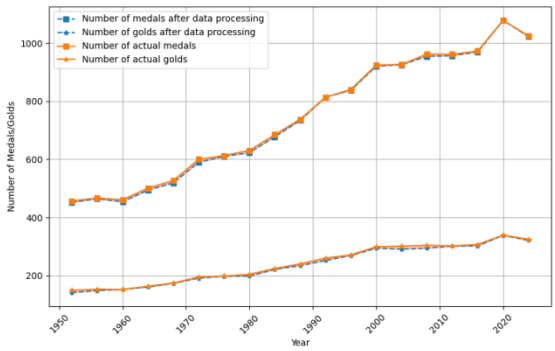
Figure 1. Number of participants and events from 1896 to 2024

2.1.2. Data preprocessing

The dataset shows that entities can take part in both single and group - based activities. In group - based activities, each member of a victorious group gets an individual award. This award - distribution system means group - based activities yield a relatively larger number of awards compared to single - entity competitions (as shown in Fig. 2(a)). To make the data accurately represent the situation, the following pre - processing steps were taken:



a. Athlete dataset before preprocessing



b. Athlete dataset after preprocessing

Figure 2. Athlete dataset preprocessing before and after

Therefore, the following preprocessing was performed on the athlete dataset:

(1) Grouping the dataset according to specified parameters. The dataset contains obvious grouping parameters including Year, NOC, and Sport. Since team events may exist within the same event, Team was also included as a grouping parameter. Additionally, as team names might repeat across different events, Event was necessary as another grouping parameter. Considering that multiple athletes in the same team could win medals, Medal was also added as a grouping parameter. The grouping parameters were prioritized as:

Year > NOC > Sport > Event > Team > Medal

with the original dataset grouped accordingly.

(2) Redundant record elimination. To mitigate the effect of group-based activity data on overall counts, only one representative record was retained per grouped dataset. This method ensures that individual entries reflect aggregated group achievements, avoiding overestimation within categorical data. As illustrated in Fig. 2, this process aligns processed counts closer to actual values, though minor discrepancies persist due to residual disqualified entries in the dataset.

2.1.3. Definition and calculation of parameters

This section defines several parameters and provides their calculation formulas through decomposition and recombination of the original datasets. These derived parameters serve as essential components for model construction.

Definition 1: Medal-winning rate PM_i^j and gold-winning rate PG_i^j

PM_i^j means the probability of entity i winning a medal in activity j. PG_i^j means the probability of entity i winning a medal in activity j. Their calculation formulas are as follows:

$$PM_i^j = \frac{M_i^j}{TM_j}, PG_i^j = \frac{G_i^j}{M_i^j} \tag{1}$$

where M_i^j denotes the medal counts of entity i in activity j, TM_j denotes the total medal counts in activity j, and G_i^j denotes the gold medal counts of entity i in activity j.

Definition 2: Sports concentration SC_i^j

Sports concentration is used to reflect the degree of participation of a entity in a certain activity. It is composed of the Heffendaal index and the medal-winning rate. Heffendaal index HHI_j^i can be calculated as:

$$HHI_i^j = (\frac{NA_i^j}{TNA_j})^2 \tag{2}$$

where NA_i^j denotes the number of entities from entity i participating in activity j, and TNA_j denotes the total number of entities from activity j.

So sports concentration can be calculated as:

$$SC_i^j = \alpha \cdot HHI_i^j + \beta \cdot PM_i^j \tag{3}$$

where α and β are the correction parameters and α + β = 1.

Definition 3: Delta events (ΔNE), historical events and novel events.

Delta events represents activities added or removed compared to the previous set, ΔNE is calculated as ΔNE = NE(Y+1) - NE(Y). Historical events are those that existed prior to a certain point. Novel events are new activities with no historical precedent.

Definition 4: Cumulative number in novel events $NANE_i$ (Y) and total cumulative number of entities participating in novel events TANE(Y).

For a set of activities in year Y, the cumulative number of entities from entity i participating in novel events is:

$$NANE_i(Y) = \sum_{y=1960}^{Y} \sum_{j=j_1}^{j_n} NA_i^j(y) \tag{4}$$

Where $NA_i^j(y)$ is the number of entity i participating in activity j in year y.

For the set of activities in year Y, the total cumulative number of entities participating in novel events is:

$$TANE(Y) = \sum_{i=i_1}^{i_n} NANE_i(Y) \tag{5}$$

Definition 5: Participation rate in novel events $PNE_i(Y)$

The participation rate in novel events for entity i in year Y can be calculated as:

$$PNE_i(Y) = \frac{NANE_i(Y)}{TANE(Y)} \tag{6}$$

Definition 6: Change in Delta Events ΔDE

Compared to the previous set of activities, ΔDE is the number of delta events increased or decreased.

Definition 7: Medal/Gold Predictions in Delta Events (MC_{DE}/GC_{DE})

For historical events:

$$MC_{DE} = 3 \cdot PM_j^i \cdot \Delta NE, GC_{DE} = PG_i^j \cdot \Delta NE \tag{7}$$

For novel events:

$$MC_{DE} = 3 \cdot PNE_i(Y) \cdot \Delta NE, GC_{DE} = PNE_i(Y) \cdot \Delta NE \tag{8}$$

2.2. Idea of Model Construction

To predict certain metrics related to entity performance in activities, various models were tested [7-10], including Linear Regression Model, Poisson Regression Model, Support Vector Regression Model, Random Forest Regression Model, Polynomial Regression Model, Negative Binomial Regression Model, and Multilayer Perceptron Model, using Mean Squared Error (MSE) as the evaluation indicator.

2.3. Model Evaluation Metrics

To estimate the predictive uncertainty of the model, this paper employed the methodology of calculating Mean Squared Error (MSE). The higher value of MSE indicates a greater average squared deviation between model predictions and actual values, a higher uncertainty, and a lower precision of the model. MSE can be calculated as follows:

$$MSE_M = \frac{1}{n}\sum_{i}^{n}\frac{1}{m}\sum_{j}^{m}(M_i^j(Y) - \overline{M_i^j(Y)})^2 \tag{9}$$

$$MSE_G = \frac{1}{n}\sum_{i}^{n}\frac{1}{m}\sum_{j}^{m}(G_i^j(Y) - \overline{G_i^j(Y)})^2 \tag{10}$$

3. Model building and solution

To predict the medal counts and gold counts of each country in the 2028 Los Angeles Summer Olympic Games, this paper constructs the MCP (Medal Counts Prediction) model, GCP (Gold Counts Prediction) model and FMP (First Medal Prediction) model. Fig. 3 shows the composition of the MCP and GCP Model, each factor will be discussed in detail in the following sub-sections.

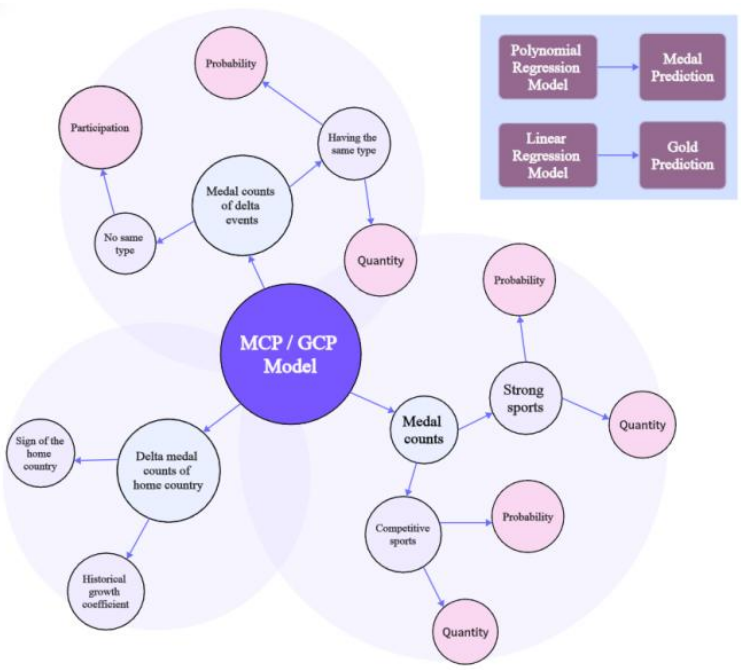


Figure 3. Composition of MCP Model

3.1. MCP (Medal-Counts-Prediction) Model

By comparing the MSE of the results of the seven models, this research find the Polynomial Regression Model performs best in medal counts prediction. The mathematical expression of the Polynomial Regression Model is as follows:

$$\begin{aligned} M_i^j(Y) = {} & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1^2 + \\ & \beta_7 X_2^2 + \beta_8 X_3^2 + \beta_9 X_4^2 + \beta_{10} X_5^2 + \beta_{11} X_1 X_2 + \beta_{12} X_1 X_3 + \\ & \beta_{13} X_1 X_4 + \beta_{14} X_1 X_5 + \beta_{15} X_2 X_3 + \beta_{16} X_2 X_4 + \beta_{17} X_2 X_5 + \\ & \beta_{18} X_3 X_4 + \beta_{19} X_3 X_5 + \beta_{20} X_4 X_5 + \varepsilon \end{aligned} \tag{11}$$

The symbol and meaning of the formula is shown in TABLE I.

Table 1. Symbols and meanings

Symbol	Meaning
X_1	M_i^j (Y-1) + M_i^j (Y-2)
X_2	G_i^j (Y-1) + G_i^j (Y-2)
X_3	MC_{DE}
X_4	GC_{DE}
X_5	$\gamma \times \varepsilon$
$\beta_0, \beta_1, ..., \beta_{20}$	Regression Coefficient

In order to reduce the error between the predicted value and the actual value, research estimates the regression coefficient through the dataset training. The regression coefficient and its value is shown in TABLE II.

Table 2. Regression Coefficient and Value in MCP Model

Coefficient	Value
0	0.272843
1	-0.067000
2	2.653846
3	0.201857
4	0.145227
5	-0.001590
6	-0.750000
7	-2.580000
8	-0.018200
9	0.284369
10	-1.900000
11	0.378582
12	-0.018300
13	-0.448000
14	-0.047700
15	0.112758
16	0.012302
17	-0.877000
18	1.920448
19	-0.043900
20	0.571484

3.2. GCP (Gold-Counts-Prediction) Model

By comparing the MSE of the results of the seven models, this research finds the Linear Regression Model performs best in gold counts prediction. The mathematical expression of the GCP Model is as follows:

$$\begin{aligned} &G_i^j(Y) = \beta_0 + \beta_1 \times (M_i^j(Y-4) + M_i^j(Y-8)) + \\ &\beta_2 \times (G_i^j(Y-4) + G_i^j(Y-8)) + \beta_3 \times MC_{DE} + \\ &\beta_4 \times GC_{DE} + \beta_5 \times \gamma \times \varepsilon \end{aligned} \tag{12}$$

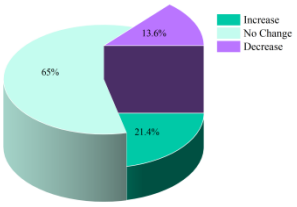
where β_0, β_1, ... , β_5 refers to the regression coefficient. The value of the regression coefficient is shown in TABLE III.

Table 3. Regression Coefficient and Value in GCP Model

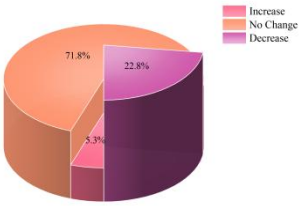
Regression Coefficient Value	Value
β_0	0.030909
β_1	0.483643
β_2	-0.066432
β_3	0.498357
β_4	-0.030147
β_5	0.009573

3.3. FMP (First-Medal-Prediction) Model

This paper makes statistics on medal changes in all countries (as shown in Fig. 4.). It finds that 65% of countries have never won a medal and 71.8% of countries have never won a gold medal. Considering such special cases, this paper proposes an optimization model based on the MCP/GCP model: the FMP model.



a. Change of total medal



b. Change of gold medal

Figure 4. Changes in total and gold medal

3.3.1. Probability mapping based on $M_j^i(2028)$

The predicted medal count is converted into a probability value using a Logistic Function, which effectively maps real-valued inputs to the [0,1], thus reflecting the intrinsic relationship between predicted medal counts and the probability of winning the first medal. Generally, a higher predicted medal count corresponds to a relatively higher probability of securing the first medal. The formula of mapping probability P_{2028} is as follows:

$$\pi_{2028} = \frac{1}{1+e^{-k(M_i^j(2028)-M_0)}} \tag{13}$$

In Equation 13, k is an adjustment parameter that controls the slope of the function curve, M_0 denotes the threshold parameter indicating that the probability of winning the first medal. It is 0.5 when the expected number of medals reaches M_0. This parameter can be adjusted according to actual data and experience and is taken as the median in the model $M_j^i(2028)$.

3.3.2. Probability contribution based on r_{arg} and r_{max}

Let two sets of data be N1 and N2. For entities that haven't met a certain target (as of a specific point in time, here 2024), the average growth rate r_{avg} between these two data - points can be calculated as:

$$r_{\arg} = \frac{N_2 - N_1}{N_1} \tag{14}$$

A linear mapping function is employed to map r_{avg} to the interval [0,0.5]. Assume that the mapping value is 0.5 when $r_{avg} = r_{max}$ and 0 when $r_{avg} = 0$. The formula for π_{pr} is as follows:

$$\pi_{pr} = \min(0.5, \frac{r_{\arg}}{r_{\max}} \times 0.5) \tag{15}$$

where r_{max} is the maximum growth rate threshold, which can be determined based on the maximum value of the actual growth rate in the dataset or through empirical methods. π_{pr} reflects the relative situation of non-target-achieving entities in terms of growth between the two data-points compared to the entities that met the target at the first data-point. A larger π_{pr} indicates that non-target-achieving

entities perform better in terms of growth, and the probability of them achieving the target may increase.

3.3.3. Radio mapping of $r_{country}$ and r_{median}

First, filter out the units that have grown in a certain metric between two relevant periods. Then calculate the growth rate of this metric for each unit, denoted as $r_{country}$. Calculate the median of the average growth rates of units that reached a specific milestone in the previous two periods as r_{median}.

Then, the ratio R is calculated as follows:

$$R = \frac{r_{country}}{r_{median}} \tag{16}$$

The same linear mapping is used to map R to [0,0.5]. Assuming that the mapping value is 0.5 when $R=R_{max}$ and 0 when R=0, the expression for π_{rpg} is as follows:

$$\pi_{rpg} = \min(0.5, \frac{R}{R_{\max}} \times 0.5) \tag{17}$$

where R_{max} is the maximum growth rate threshold, which can be determined either by the maximum value of the actual calculated ratio or empirically. π_{rpg} reflects the relative situation of non-milestone-reaching units in terms of growth of the metric compared to the new units that reached the milestone. If π_{rpg} is larger, it means that these non-milestone- reaching units are performing better in terms of growth of the metric and may have a correspondingly higher probability of reaching the milestone.

3.3.4. FMP model establishment

Suppose that the probability of the first medal in 2028 for countries that do not win a medal in 2024 is P, and it can be calculated as follows:

$$P = \alpha \times \pi_{2028} + \beta \times (\pi_{pr} + \pi_{rpg}) \tag{18}$$

where a and b are control parameters, a + b = 1, a >= 0, b >= 0.

3.4. Model Application

3.4.1. For MCP/GCP model

Based on the MCP Model, GCP Model, and relevant data collected, the predicted gold medal and total medal standings for the 2028 Los Angeles Summer Olympics are presented in TABLE IV, which shows the top ten countries.

Table 4. Predicted Medal Table in 2028 Summer Olympics

Rank	NOC	Gold	Total
1	USA	66	147
2	CHN	29	82
3	GBR	18	60
4	JPN	16	48
5	AUS	15	45
6	FRA	11	43
7	NED	11	35
8	ITA	9	36
9	GER	8	35
10	CAN	5	27

3.4.2. For FMP model

Based on the FMP Model, countries that will win their first medal is shown in TABLE V.

Table 5. Countries get their first medal in 2028 Summer Olympics

NOC	Medal Counts	Probabiity
ESA	1	0.515093
HON	1	0.527990
LBR	1	0.524226
LIE	1	0.565879
SAM	1	0.670038

3.4.3. Strong Sports

For a sport in a country, the MCP Model determines the importance of the sport through the SC index. Definition 2 states that the SC index is determined by the Hefendal index HHI and the PM of recent sports. Therefore, the MCP Model divides all sports into three categories by setting an upper threshold (thresholdα) and a lower threshold (thresholdβ). If SC >= thresholdα, the sport is classified as a strong sport, if SC <= thresholdβ, then the sport is classified as a weak sport. Otherwise, the sport is a competitive sport.

To verify the MCP Model classification, the study examined the strong sports of China and the United States in 2024 along with their actual total and gold medal counts. The correction coefficient α of SC was set at 0.8 for this analysis, as shown in TABLE VI.

Table 6. Medal counts of China and the United States in strong sports

CHN Sports	Gold Medal	Medal	USA Sports	Gold Medal	Medal
Aquatics	12	25	Aquatics	8	30
Gymnastics	2	9	Athletics	14	34
Shooting	5	10	Basketball	2	3
Table Tennis	5	6	Gymnastics	3	9
Weightlifting	5	5	Wrestling	2	7

4. Conclusion

This paper proposes a hybrid predictive model that combines multifactor regression with polynomial fitting, significantly improving time-series forecasting accuracy. The dynamic fusion strategy of the feature engineering system effectively addresses nonlinear relationship modeling and multivariable coupling issues. The model's superiority is validated through experiments using a carefully designed data division method (training-test split by year), grouping calculations (aggregation by NOC and Sport), and rich feature selection. These methodological choices ensure robustness in handling sparse medal data while capturing historical trends and patterns.

However, limitations such as fixed window size, lack of cross-validation, and comparisons with benchmark models indicate opportunities for refinement. Future work will focus on (1) deeper integration of deep learning with traditional methods to enhance feature extraction, (2) development of adaptive learning frameworks for dynamic window sizing and improved flexibility, and (3) expansion into multimodal time-series prediction to further advance forecasting capabilities in complex scenarios. These directions aim to address current weaknesses while extending the model's applicability to broader domains.

References

[1] KURISU D, FUKAMI R, KOIKE Y. Adaptive deep learning for nonlinear time series models[J]. Bernoulli, 2025, 31(1): 240-270.

[2] Dunwang Qin, Zhen Peng, Lifeng Wu. Deep attention fuzzy cognitive maps for interpretable multivariate time series prediction[J]. Knowledge-Based Systems, 2023, 275: 110700.

[3] Bonas M, Datta A, Wikle C K, et al. Assessing predictability of environmental time series with statistical and machine learning models[J]. Environmetrics, 2025, 36(1): e2864.

[4] Z. Han, J. Zhao, H. Leung, K. F. Ma and W. Wang, "A Review of Deep Learning Models for Time Series Prediction," in IEEE Sensors Journal, vol. 21, no. 6, pp. 7833-7848, 15 March15, 2021.

[5] M. Sui, C. Zhang, L. Zhou, S. Liao and C. Wei, "An Ensemble Approach to Stock Price Prediction Using Deep Learning and Time Series Models," 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 2024, pp. 793-797.

[6] Almutairi E, Abbod M, Hunaiti Z. Prediction of diabetes using statistical and machine learning modelling techniques[J]. Algorithms, 2025, 18: 145.

[7] Khedr A, Arif I, P V PR, et al. Cryptocurrency price prediction using traditional statistical and machine learning techniques: A survey[J]. Intelligent Systems in Accounting, Finance and Management, 2021, 28(1): 3-34.

[8] Yao Z, Su J N, Fan G, et al. GACA: a gradient-aware and contrastive-adaptive learning framework for low-light image enhancement[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-14.

[9] LI Yuyao, ZHANG Xiangwen, LI Ziyang, et al. Accurate and adaptive state of health estimation for lithium-ion battery based on patch learning framework[J]. Measurement, 2025, 250: 117083.

[10] LIU Haoxin, XU Shangqing, ZHAO Zhiyuan, et al. Time-MMD: multi-domain multimodal dataset for time series analysis[C]//GLOBERSON A, MACKEY L, BELGRAVE D, et al. Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2024: 77888-77933.