Research on Olympic Medal Table Prediction Based on Graph Convolutional Neural Network

Wenhao Lu *, Quan Zhang, Jinjiang Wang

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China, 454003

* Corresponding Author Email: xin2ele@163.com

Abstract. As a global top sports event, Olympic medal table prediction is of great significance to the development of national sports strategies. In this study, we break through the limitations of traditional regression analysis by integrating historical medal data, host country effects, athletes' characteristics and other sources of data, and innovatively constructing a hybrid feature system that includes time series features and dynamic graph structure. Through the graph convolutional neural network modeling the time evolution characteristics of inter-country competition, combined with Bootstrap resampling technology to construct a probabilistic prediction model, to achieve the dynamic prediction of the medal list of the 2028 Los Angeles Olympic Games, and concluded that the United States with a probability of 85.7% to become the most progressive country, and Russia with a probability of 70.5% to become the most regressive country. This study establishes a multi-dimensional spatio-temporal correlation feature system and develops a hybrid GCN-Bootstrap architecture to realize the probabilistic inference of the Olympic medal distribution for the first time, and reveals the deep influence mechanism of geopolitical factors on competitive sports.

Keywords: 2028 Olympic Games, Medal Prediction, GCN, Time Series.

1. Introduction

As a globally recognized sports event, the Olympic Games, the medal table has always been the focus of attention of all countries. The medal table of each Olympic Games not only reflects the athletic level of each country, but also becomes an important yardstick to measure the strength of national sports. With the continuous changes of the event programs, the prediction of the medal table becomes more and more complicated, involving various factors such as athletes' performance, program settings, and national sports advantages.

Establishing a prediction model based on historical data can help predict the distribution of medals in the future Olympic Games and provide valuable references. Bernard et al [1] used Logit model to analyze the medal table of the Olympic Games, and found that the more the population of the country or region where the Olympic team is located, the higher the per capita GDP, and the higher the host country of the Olympic Games, the more Olympic medals the team will win. Schlembach et al [2] used a random forest model to predict the performance of each representative team at the Olympic Games and assessed the contribution of different characteristic variables to the prediction. However, machine learning-based research methods are relatively backward and less effective when dealing with complex historical data. With the continuous development of deep learning, deep learning models have been widely used in various researches.

In this study, we innovatively propose a deep learning model based on graphical convolutional neural networks, which mines the core factors affecting the distribution of medals by integrating the multi-dimensional features (host country effect, athletes' performance, program setting and inter-country correlation) in the historical data and quantifies the country's Olympic athletics network in terms of its interactions (historical interactions, geographic proximity, program overlap) in the Olympic athletic network, and ultimately provide prediction results with confidence intervals, identifying the countries with the greatest progress (the U.S. tops the list with a probability of 85.7%) and the highest risk of regression (Russia with a probability of 70.5%).

The significance lies in the fact that graph neural network is introduced into the field of Olympic prediction for the first time, which breaks through the limitations of traditional machine learning models in processing complex relational data, and provides a scientific basis for the formulation of sports strategy and optimization of resource allocation; at the same time, through the multitasking learning framework and Bootstrap confidence interval technology, the dynamic mechanisms such as the influence of collective projects and the effect of the host country are revealed, and the credibility of the prediction is enhanced. This study innovatively proposes to construct a graph structure model with GCN as the core, abstracting countries as nodes and association strengths as edges to capture the interaction relationship; designing a comprehensive feature engineering covering time-series features, project importance scores and first-time award spans; optimizing the prediction of multi-medal category distribution by combining with the KL dispersion loss and quantifying the country network status through the degree matrix, which provides a reproducible and reliable solution to the complex time-series prediction problem at both the methodological and technological levels. A reusable solution is provided for the complex time series prediction problem at both methodological and technical levels.

2. Prediction study of medal table based on GCN neural network

2.1. Data preprocessing and feature engineering establishment

This study unfolds based on COMAP's relevant historical data. Missing and invalid values exist in the project data, which can have an impact on the analysis results, so they are eliminated. Because of the replacement of the name of the same country in different historical periods, this naming difference will affect the data merging and analysis, so these names are unified to ensure the consistency of the data, such as the unification of the Soviet Union into Russia. Because of the special nature of the 1906 Olympic Games, the time span does not match with other Olympic Games, so it is excluded to improve the consistency of the data and the accuracy of the analysis of the results.

After the data preprocessing to establish the feature engineering [3], this study analyzes the data of opening sports as the feature quantity for calculating the opening of sports in the next session; obtains the data of each country's advantageous sports as the feature quantity to be used for the reinforcement of prediction model; analyzes the effect of the host country, because the host country can get more support and resources through the home advantage, so it can be assumed that the host country usually performs better than other years in the current Olympic Games. Because the host country can get more support and resources through the home advantage, it can be assumed that the host country usually performs better than other years in this Olympics, define the historical host country and analyze the performance of the host country before and after hosting the games.

$$E=\frac{P_{t+n}-P_t}{P_t} \tag{1}$$

Where P_t indicates the performance of the host country in year t.

Analyze first-time medal wins, obtain the year in which each country first won a medal, and calculate the time span that each country has experienced from its first participation to its first medal win. Calculate each country's medal wins in different events, analyze each country's strengths in different events, combine the medal data, calculate each country's performance in different events, and calculate the event importance score.

$$IS=0.4\times R+0.3\frac{G}{G_{max}+1}+0.3\times\frac{T}{T_{max}+1} \tag{2}$$

Where R is the award rate, G is the number of gold medals, G_{max} is the maximum number of gold medals, T is the total number of medals.

The analyzed athlete characteristics include the number of athletes, the average experience of the athletes, and the average number of athlete participations, which are used as feature vectors for the construction of time series features.

To better understand the performance trends of each country in the Olympics, this study collects the total number of medals each country has won since their first participation, analyzing their long-term performance trend. It calculates the medal growth rate between consecutive Olympic Games to assess short-term performance improvements. Additionally, the moving average of the medals is calculated to smooth fluctuations and observe trends over a longer time span. The moving average formula is as follows:

$$P_i=\sigma(\omega_1 MA_i+\omega_2 \frac{d_i}{\sum_j d_j}+\omega_3 f(x_i)) \quad (3)$$

Where MA_i is the composite moving average index, $\frac{d_i}{\sum_j d_j}$ is the degree centrality, $f(x_i)$ is the other characteristic function, ω_1, ω_2, ω_3 are the weighting parameters.

In this study, time series features are created in a comprehensive manner by analyzing data on opening events, data on dominant events in different countries, host country effects, combining medal data, obtaining information on first-time winners, calculating the number of years until first-time winners, calculating cumulative statistics on medals, analyzing athlete characteristics, calculating moving averages of medals, and calculating medal growth rates.

2.2. GCN Network Modeling and Solving

In this study, a variety of features were extracted from each country's historical data, including the number of gold, silver and bronze medals, the total number of medals, the historical ranking, the number of events participated in, while more complex features, the moving average of gold and total medals and whether it is a factor of the host country or not, were also taken into account. After the feature preparation, this study then normalizes the data to ensure that the data is balanced, so as to avoid certain features dominating the model training due to excessive scaling. For each country node, a multidimensional feature vector is constructed:

$$x_i=[f_{recent},f_{hist},f_{trend},f_{dev},f_{env},\ldots] \quad (4)$$

The data was divided by year into a training set (1896-2016), a validation set (2020) and a test set (2024). Dividing the data chronologically thereby simulating medal prediction scenarios for future Olympics ensures the generalization ability of the model.

GCN is a deep learning architecture specialized in processing graph-structured data [4]. Different from traditional neural networks, graph convolutional neural networks are able to perform feature learning and information transfer directly on graph structures. In the graph $G=(V,E)$ in this problem, each node $v \in V$ not only contains its own feature information, but also exchanges information with neighboring nodes through the edge E. The core operation of graph convolutional neural network is represented as:

$$H^{l+1}=\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (5)$$

Where, $\tilde{A}=A+I_N$ is the adjacency matrix of the added self-loop, $\tilde{D}$ is the corresponding degree matrix, $H^{(l)}$ is the node feature representation at layer 1, $W^{(l)}$ is the learnable weight matrix, and σ is the nonlinear activation function.

In this study, GCN is used to process graph-structured data, and each layer of GCN is able to aggregate the neighbor information of the nodes through graph convolution operation, and then learn the feature representation of each country. The GCN model in this paper consists of a three-layer graph convolutional network, each layer contains ReLU activation function [5] and uses Dropout regularization technique to prevent overfitting. With this hierarchical structure, the model can progressively integrate the information between countries, which allows for better prediction of the number of medals.

We design a three-layer GCN with the following expressions:

$$H^{(1)}=\mathrm{ReLU}(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}XW^{(0)}) \tag{6}$$

$$H^{(2)}=\mathrm{ReLU}(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}HW^{(1)}) \tag{7}$$

$$H^{(3)}=\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{2}W^{(2)} \tag{8}$$

Dropout layers were added after each of these layers to prevent overfitting:

$$H^{(l)}=\mathrm{Dropout}(H^{(l)},p=0.2) \tag{9}$$

In this model architecture, this study has separate prediction heads for each medal category including Gold, Silver, Bronze and Total Medals. Each prediction head is a multilayer neural network consisting of a fully connected layer, a ReLU activation function and a Dropout layer. The multi-head predictor allows the model to be optimized specifically for different medal categories, while also capturing the interrelationships between medal prediction tasks from multiple perspectives.

Optimize for different medal categories and capture the interrelationships between medal prediction tasks from multiple perspectives. The medal count prediction formula is as follows:

$$\begin{cases} \hat{y}_{gold}=\sigma(W_gH^{(3)}+b_g) \\ \hat{y}_{silver}=\sigma(W_sH^{(3)}+b_s) \\ \hat{y}_{bronze}=\sigma(W_bH^{(3)}+b_b) \end{cases} \tag{10}$$

For the building blocks of the graph, this study introduces edge weights to measure the correlation between countries by considering historical performance, ranking differences and program similarity (the stronger the correlation, the higher the weight). In this study, the participating countries are regarded as nodes in the graph, and the relationships between countries are used as edges. The construction process of the graph is as follows:

$$\begin{cases} V=\{v_i|i\in C\} \text{ (node definition)} \\ D_{ij}=\sum_k A_{ik} \text{ if } i=j \text{ (degree matrix definition)} \\ L_t=\lambda_1L_m+\lambda_2L_i+\lambda_3L_f+\lambda_4L_r \text{ (loss function)} \\ \omega_{ij}=\alpha\cdot sp(i,j)+\beta\cdot sg(i,j)+\gamma\cdot ss(i,j) \text{ (weight calculations)} \end{cases} \tag{11}$$

The degree matrix D not only reflects the connection strength of each country, but also reveals its central position in the Olympic athletic network. A_{ik} denotes the connection strength between countries i and k. The degree matrix is the medal prediction loss, is the improvement prediction loss, and is the regularization term. L_m denotes medal prediction loss, L_i denotes improvement prediction loss, L_f denotes first-time winner prediction loss, and L_r denotes the regularization term. $sp(i,j)$ denotes historical performance similarity, $sg(i,j)$ denotes geographic similarity, $ss(i,j)$ denotes dominant program overlap, α, β, γ are weighting coefficients, and $\alpha+\beta+\gamma=1$.

Under the multi-task learning framework, this study not only predicts the number of gold, silver and bronze medals, but also the distribution of medals. The model is allowed to learn the interactions between tasks from multiple perspectives to improve the overall prediction ability. Meanwhile, independent loss functions are designed for each medal category separately, and KL scatter loss [6] is introduced to optimize the probability distribution among medal categories. In order to ensure the reasonableness of the prediction results, the total number of each medal category was also normalized to ensure that the total number of medals was equal to the sum of the number of medals in each category.

During training, the initial learning rate is set to 0.001 and a learning rate scheduler is introduced to improve stability and efficiency. The number of training sessions was 200, and an early stopping mechanism was used to prevent overfitting. The Adam optimizer [7] and mean square error loss function are used to optimize the model, while the KL scatter loss is added to optimize the probability distribution to ensure that the number of medals and categories are predicted accurately. Combining

the count loss and distribution loss and using weighting coefficients to balance the two optimizes the performance of the model in terms of count and distribution. Finally, the model performance is evaluated using validation and test sets.

Based on the above model, this study predicts the medals of the 2028 Olympic Games. The predicted values of the number of gold, silver and bronze medals for each country were obtained, and these results were sorted according to the number of gold medals, and the predicted probabilities were calculated. In order to obtain a more accurate prediction value, this study utilizes the Softmax function [8] to output the probability distribution of gold, silver, and bronze medals, and the three probability values output represent the probability of gold, silver, and bronze medals, respectively, and then modifies the feature preparation function in order to increase the computation of medal distributions, and after the output of the function, it is concluded that the probability distributions of gold, silver, and bronze medals are correlated with each other. The formula is as follows:

$$\text{Softmax}(x)=\frac{e^{x_i}}{\sum_i e^{x_i}} \tag{12}$$

In this study, Bootstrap sampling [9] is introduced to compute prediction intervals and the probability of the country's progress in future Olympics, setting the prediction results to have 95% confidence intervals. The expression is as follows:

$$\hat{\theta}_{\text{bootstrap}}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)} \tag{13}$$

After the training evaluation of the above model and calculating its prediction intervals, the results are shown in figure 1:

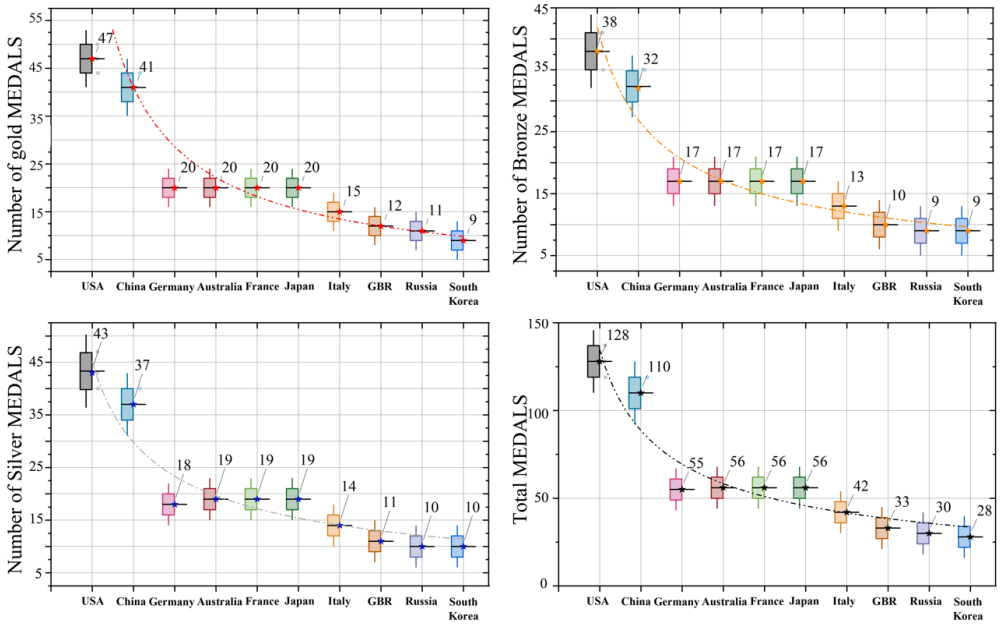

Figure 1. Medal table prediction chart with confidence intervals (top 10)

As can be seen in Figure 1, the United States and China continue to be far ahead of other countries in the total number of medals, and the top ten rankings have not changed significantly from the previous edition and most of them are developed countries.

For a certain country, the D matrix in the model reflects the degree of the country's connection with other countries, which contains multi-dimensional information such as historical interactions, sports exchanges, and geographic location. The strength of each country's connection also reveals its centrality in the Olympic athletic network. In order to determine which countries are most likely to improve and which will be worse off than in 2024, so this study utilizes the degree matrix to reflect the position and influence of countries in the Olympic system. The degree matrix characteristics are analyzed to establish the relationship between the country improvement predictions and the degree matrix, which is expressed as follows:

$$H=\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}XW) \tag{14}$$

In the formula, $\tilde{D}$ reflects the degree matrix of countries in the Olympic network, $\tilde{A}$ indicates the interactions between countries, and X contains country characteristics.

In addition, in order to anticipate how many countries will win their first medal at the next Olympic Games and the estimated magnitude of the likelihood. Using the previously established characteristic equations, this study counts the characteristics of countries in terms of first-time participants and first-time winners. Combined with the Sigmoid function [10] this enables to derive the countries and their probability of winning a first medal. For different countries, The relationship between each program and the likelihood of a country winning a medal can be obtained by considering the program's impact on the country. Formula:

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{15}$$

$$P(f)_i=\sigma(\sum * j\epsilon N(i)\frac{d_j}{\sum_k d_k}a_{ij}x_j) \tag{16}$$

$$I_{sport}=\sum_{i\epsilon V}\frac{d_i}{\sum_j d_j}\cdot f_{sport}(i) \tag{17}$$

Where d_j denotes the degree value of country j, a_{ij} denotes the strength of association between countries, and x_j is the country eigenvector.

A weight of 0.3 was assigned to the frequency of historical interactions, 0.2 to geographic proximity, and 0.5 to the density of sports exchanges for the calculation.

Network centrality was calculated to be significantly associated with increased athletic strength. The higher degree centrality of the United States (0.152) reflects its centrality in the Olympic network, with a predicted dominance of 85.7% for the 2028 Los Angeles Games. The network position of some countries fluctuates significantly, with Russia's degree centrality declining (-0.025) and its risk of declining athletic strength at 70.5%. Regional network centrality is key to predicting first-time winners, with Singapore's higher regional connectivity (0.045) giving it a 75.3% probability of first-time winners. Programs with high network spread coefficients have a greater impact on the distribution of medals, and can raise the level of athleticism in each country through network effects.

In the GCN prediction model, the number of projects and project types are introduced as new characteristics, arguing that more project participation usually means that the country has more chances to win medals. Project types can be categorized into individual and team projects. Team projects usually produce more medals, while individual projects have a more balanced number of medals. In order to better reflect these differences, this study categorizes projects by type and considers their contribution to the number of medals separately, and finally derives the effect of projects and number on the total number of medals as shown in the figure 2:

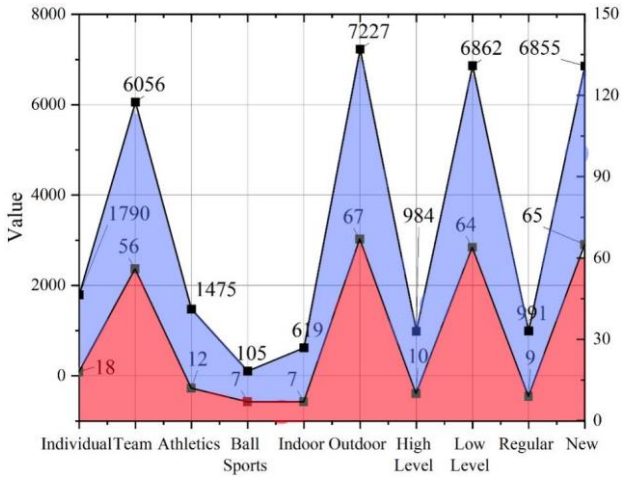

Figure 2. Impact of item count and medal count on total medals

From Figure 2, it can be concluded that the total number of medals in collective sports is much higher than in individual sports, indicating that collective sports are larger and have more events. The number of medals in outdoor sports is significantly higher than in indoor sports, mainly because many important sports such as athletics and swimming are outdoor sports. The number of medals in low-level competitive sports is much higher than in high-level sports, indicating wider participation in these sports, and the total number of medals in newcomer sports is much higher than in regular sports, reflecting the growth of emerging sports in the Olympics.

Finally based on the historical data, this paper defines the strength items for each country. Certain countries have significant strengths and therefore these strong items contribute more to the country's medal count. Comprehensively analyzing each country's historical performance in specific events, quantifying its strengths in strong events, and using this characteristic as one of the input characteristics as well, GCN derives each country's strengths by learning the relationship between countries and each country's performance in different event types, using the United States, China, and Germany as examples, the results are shown in Figure 3:

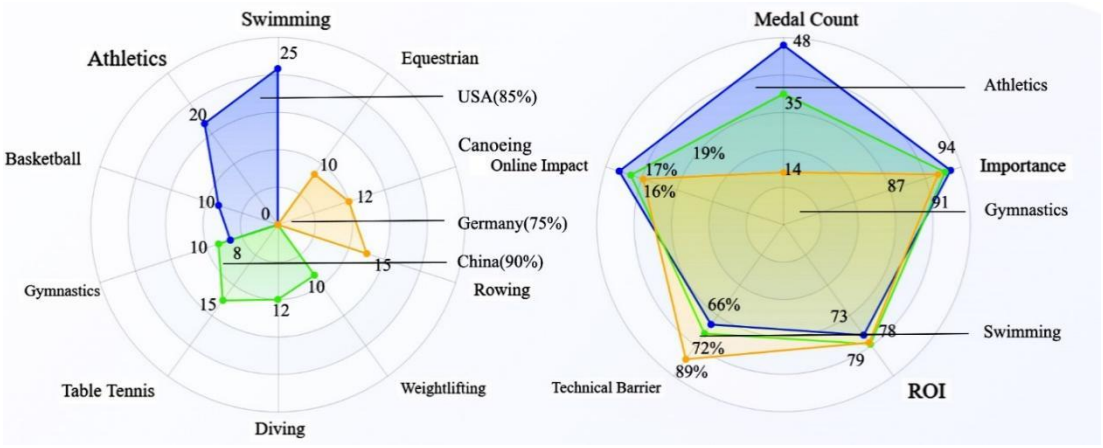

Figure 3. Radar chart of important projects

As can be seen from Figure 3, the dominant sport in the United States is swimming, China's dominant sport is table tennis, and Germany's dominant sport is rowing, while gymnastics, track and field, and swimming as high-return sports have a more significant impact on the number of medals.

3. Conclusion

In this study, Olympic medals are predicted by building a graphical convolutional neural network model, which firstly removes missing values, unifies country names and time spans from the data. The host country, athletes, events, and statistical features are analyzed by feature engineering analysis thus constructing the time series features. A graph convolutional neural network is established to process the graph structure data, and the model is combined with Dropout regularization through a three-layer convolutional network to prevent overfitting and set independent prediction heads for each medal category. During the training process, the Adam optimizer and KL scatter loss function were used, combined with an adaptive learning rate and early stopping mechanism, to derive predictions for the distribution of medals across countries. Confidence intervals are determined by Bootstrap and it is derived that the number of medals in collective events is much higher than in individual events and that the United States is the most improved country with a probability of 85.7%, with the recommended events being Swimming, Athletics and Basketball. Russia has a 70.5% risk of decline. Singapore has a 75.3% probability of winning its first bronze medal in swimming. This study not only improves the accuracy of medal prediction, but also reveals the complex impact of multidimensional factors on Olympic performance, providing strong support for the prediction of medal distribution in future Olympic Games.

References

[1] Andrew B. Bernard; Meghan R. Busse. Who Wins the Olympic Games: Economic Resources and Medal Totals [J]. The Review of Economics and Statistics, 2004, 86 (1).

[2] Schlembach Christoph; Schmidt Sascha L.; Schreyer Dominik; Wunderlich Linus. Forecasting the Olympic medal distribution – A socioeconomic machine learning model [J]. Technological Forecasting & Social Change, 2022, 175.

[3] Akter R, Islam R M, Debnath K S, et al. A hybrid CNN-LSTM model for environmental sound classification: Leveraging feature engineering and transfer learning [J]. Digital Signal Processing, 2025, 163105234-105234.

[4] Du Y, Ding N, Lv H. Spatio-temporal prediction of terrorist attacks based on GCN-LSTM [J]. Journal of Safety Science and Resilience, 2025, 6 (2): 186-195.

[5] Sooksatra K, Rivas P. Dynamic-Max-Value ReLU Functions for Adversarially Robust Machine Learning Models [J]. Mathematics, 2024, 12 (22): 3551-3551.

[6] Alp T G, Vasi İ, Alp E, et al. PRDX-4: a novel biomarker similar to KL-6 for predicting the occurrence and progression of systemic sclerosis-ILD. [J]. Biomarkers in medicine, 2025, 1-7.

[7] Wang J, Kumar D M, Mamatha S, et al. Deep learning-based Adam optimization for magnetohydrodynamics radiative thin film flow of ternary hybrid nanofluid with oscillatory boundary conditions [J]. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena, 2025, 196116448-116448.

[8] Santoso B I, Utama N S, Supriyono. Meta-learning based softmax average of convolutional neural networks using multi-layer perceptron for brain tumour classification [J]. Array, 2025, 26100398-100398.

[9] Wang H, Fan J, Wang Y, et al. Bootstrap Masked Visual Modeling Via Hard Patch Mining. [J]. IEEE transactions on pattern analysis and machine intelligence, 2025, PP.

[10] Venkatesh S, Sindhu R, Arunachalam V. Hardware efficient approximate sigmoid activation function for classifying features around zero [J]. Integration, 2025, 103102421-102421.