Research on Momentum of Tennis Players Based on Gaussian Mixture Clustering and TOPSIS

Yuanhao Bai *

College of Liberal Arts and Sciences, China University of Petroleum-Beijing at Karamay, Karamay, China, 834000

* Corresponding Author Email: 2021016233@st.cupk.edu.cn

Abstract. In order to investigate the interplay between player victories and tennis performance with the overarching goal of predicting match outcomes and refining athletes' training regimens, this study introduces the concept of 'momentum'. Six metrics were extracted for this study utilizing a large dataset carefully selected from the first round of men's singles matches at the 2023 Wimbledon Open. Subsequently, this study used Gaussian Mixture Clustering Model (GMM) and Entropy Analysis-TOPSIS algorithm to systematically rank the scores of the eight categories. From there, the performance of the players was delineated. Subsequently, Spearman's correlation coefficient and analysis of variance (ANOVA) were used to scrutinize the complex relationship between "momentum" and the variables that encompass performance fluctuations. The empirical findings indicated that the correlation coefficients between Momentum and the correlation indices in this study ranged from [0.014-0.220] and that the p-values of the findings were consistently lower than 0.001.The results of the subgroup ANOVA were statistically significant, thus reinforcing the robustness of our analytical framework.

Keywords: Tennis, Momentum, GMM, Entropy weight-TOPSIS, Ahp-entropy weight method.

1. Introduction

Tennis is one of the most popular and widely played sports in the world, and in the men's singles final of the 2023 Wimbledon Open, 20-year-old Spanish star Carlos Alcaraz defeated one of the greatest Grand Slam players of all time, Novak Djokovic. The winds of victory changed many times during the final, leading not only athletes and coaches but also tennis enthusiasts to want to understand the factors involved in winning tennis, to better predict who will win and who will lose, and to improve their own game and that of the world's tennis players as a whole. Hothis studyver, in the new era of Artificial Intelligence, the content of using scientific computer algorithms to analyze tennis players is not rich and the direction of analysis is not comprehensive, the current use of algorithms to analyze the tennis movement is still in the primary stage, the existing research based on Artificial Intelligence on tennis players is mainly divided into the following two categories, one is the use of computer vision algorithms to analyze the movement of tennis players and determine whether the movement needs to be improved. analyze and determine whether this action needs to be improved [1], and the second is to analyze and explore the application scenarios of artificial intelligence in tennis training by adopting the method of literature, questionnaire survey, logical induction and so on [2]. And with reference to a similar popular sport, soccer, this study found that there exist some relevant studies using machine learning or deep learning algorithms to predict the ability of soccer players [3], and this paper is inspired by this, and adopts a comprehensive consideration of the player's serve performance, scoring and other factors during the game, which are quantified as "Momentum" to further predict the game situation or identify the athlete's this studyak points.

2. Evaluation and Ranking Model Based on GMM Algorithm and Entropy this studyight-TOPSIS Algorithm

Data from https://www.comap.com/contests/mcm-icm

The data provided in the question has many levels, but it is undeniable that there is an overlap betthis studyen a considerable part of the information in the data, for example, the indicators p1_break_pt and p1_break_pt_won both indicate the situation in which a player breaks the opponent's serve. At the same time, there is also a part of data with randomness and this studyak correlation, for example, p1_sets indicate the number of sets won by a player, which is random as in the case of the Wimbledon men's singles final, and speed_mph is related to the player's instantaneous motion state, which is unpredictable. And considering the need for tournament-specific therefore, the subsequent analysis of this paper will isolate the data related to the match set and motion state, and in order to cohesive indicator information, according to the old indicators to establish new indicators more relevant to the player's performance, as follows:

Score Performance (c_1) [8]: Considering the rules of tennis, the key to winning a match is to lead by two points consecutively, thus relative advantage is more important. Therefore, in order to effectively highlight the relative advantage betthis studyen players and avoid meaningless ratio calculation, this paper adopts the way of score ratio minus one to focus on the relative advantage of players in the same match. This indicator is a positive indicator.

Serve performance (c_2) [9]: According to the question, serving is a very effective way to score and win, and a high-quality serve can not only win the game for the player, but also directly reflect the player's game performance. Therefore, the combined information of p1_ace and p2_ace indicators, when the player sends a good ball is assigned a value of 1, when no score is assigned a value of 0, and when the opponent sends a good ball is assigned a value of -1. This indicator is a positive indicator.

Strike performance (c_3): Striking the ball is the key to achieving effective offense and defense, and is also an important ornamental moment for the player in the game. Therefore, the combined information of p1_winner and p2_winner indicators is assigned a value of 1 when the player hits a strike, 0 when there is no strike, and -1 when the opponent hits a strike. this indicator is a positive indicator.

Error performance (c_4): error is the other side of the reflection of the player's game state, but also an important reference to measure the player's performance. Therefore, combining the double fault and unf_err indicators, when the player loses points due to a service error assigned a value of -1, unforced error assigned a value of -2, no error assigned a value of 0, the opponent's error own side plus the corresponding positive points. This indicator is a positive indicator.

Net_pt_won performance (c_5): net_pt_won can make the match more compact and intense, and the spectator's experience will be increased, corresponding to the better performance of the player. Therefore, the combined net_pt and net_pt_won indicator assigns a value of 1 when a player hits the net, a value of 2 when he scores a point, a value of 0 when he does not score a point, and a corresponding negative value for the opponent's hit on the net. This indicator is a positive indicator.

Break performance (c_6): as opposed to the serve, breaking the serve is a very effective defense and also means that the player is performing better. Therefore, the break_pt, break_pt_won and break_pt_missed indicators are assigned a value of 0 when the player has a chance to break the serve, a value of 1 for a failed break, and a value of 2 for a successful break. this indicator is a positive indicator.

In summary, the above processing methods resulted in a total of 14,324 expanded post-samples from both players in multiple matches.

2.1. Gaussian Mixture Clustering Model (GMM) and Entropy WeightsTOPSIS based ranking models

The GMM model is better able to handle data with complex distributions than ordinary

K-means clustering algorithms, discovering more refined hidden clusters and giving the probability of the samples to be grouped into clusters, realizing multi-layer segmentation and prediction of the degree of the player's performance.

2.1.1. GMM construction

By analyzing the above indicators related to the players' competitions, the players' performances in various aspects in different competitions at different times are clustered, and the degree of performance is refined into eight categories: superb, excellent, outstanding, good, medium, passing, to be improved, and needing to be improved, to further portray the process of scoring in the competitions.The steps of the algorithm of the GMM model[4] are as follows:

Step 1: In order to improve the clustering effect of the algorithm, further clustering analysis is carried out through the determination of the optimal number of clusters, and according to the elbow rule, this paper sets the K-value to 8 for modeling clustering.

Step 2: Based on the above data, assume that the sample set is $D=\{x_1,x_2,\cdots,x_m,\}$. The input samples obey six Gaussian distributions with unknown parameters, each of which corresponds to a different mean μ_i and covariance matrix $\sum i$ $(1\leqslant i\leqslant 6)$. Initialize the mixing coefficients a_i, the covariance matrix $\sum i$, and the mean vector μ_i of the Gaussian mixture distributions obeyed by the sample set, and it is reasonable to assume that each of the mixture metrics has a diagonal matrix. This sample set obeys a mixed Gaussian distribution:

$$P_M(x)=\sum_{i=1}^{6}\alpha_i\cdot p\left(x\middle|\mu_i,\sum i\right) \tag{1}$$

Step 3: Calculate the probability that each sample point data x_i $(i=1,2,3,\cdots,m)$ belongs to the first j Gaussian distribution.

$$\gamma_{i,j}=\frac{\alpha_i\cdot p\left(x\middle|\mu_i,\sum i\right)}{\sum_{i=1}^{6}\alpha_p\cdot p\middle|\mu_p,\sum p} \tag{2}$$

Step 4: Calculate and update model parameter value mean vector μ_i', covariance matrix $\sum i'$ and mixed coefficient a_i' according to the formula.

$$\mu_i^{'}=\frac{\sum_{j=1}^{m}\gamma_{i,j}x_j}{\sum_{j=1}^{m}\gamma_{i,j}} \tag{3}$$

$$\sum \mathrm{i}^{'}=\frac{\sum_{j=1}^{m}\gamma_{i,j}x_j(x_j-\mu_i^{'})(x_j-\mu_i^{'})^T}{\sum_{j=1}^{m}\gamma_{i,j}} \tag{4}$$

$$a_i^{'}=\frac{\sum_{j=1}^{m}\gamma_{i,j}}{m} \tag{5}$$

Step 5: Repeat steps 3 and 4 based on the new parameters until the Gaussian converges.

Step 6: Classify each sample into the clusters with the highest probability according to A. The final result is an 8-class clustering result.

2.1.2. Discussion of model classification accuracy

The discussion about the accuracy of model clustering can be divided into two parts, the first part is whether there is a significant difference betthis studyen each cluster, and the second part is the effectiveness of GMM clustering. In the first part, this paper firstly demonstrates by normalizing the

mean value of indicators in each category, as shown in Fig 1, it can be seen that the effect of each type of clustering is significant, and there is an extremity phenomenon in Performance on the net and Error performance indicators, and the other indicators have a stronger degree of stratification betthis studyen different clusters, which shows that the results of this clustering have a significant difference. In the second part of the paper, a principal component analysis (PCA) was performed on the seven metrics. It was found that the variance explained by the first two principal components reached 84.52% (>80%). Therefore, this paper extracted the first two principal components to create a scatterplot, as shown in the figure. From the Fig 2, it can be seen that the different categories are this studyll differentiated and the GMM clustering is good.

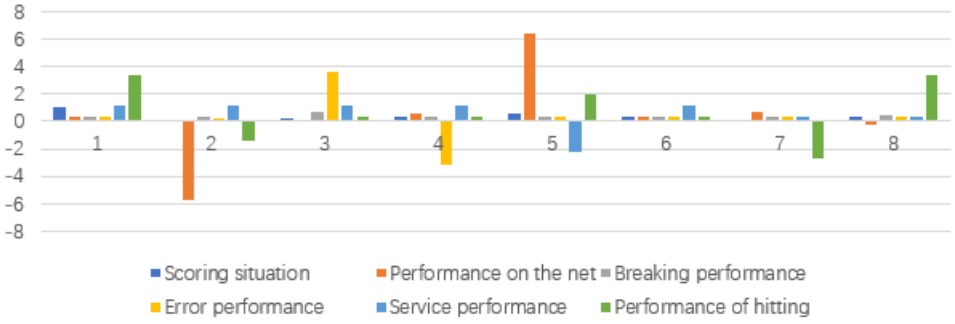

Fig 1. Normalized index means display graph.

Fig 2. GMM clustering visualization.

From the perspective of evaluation index, this paper calculates the contour coefficient for interpretation. When the contour coefficient is high, it means that the sample is more dissimilar to other objects in its own clustering than to other objects in other clusters, i.e., the clustering effect is better. In the clustering results of these eight classifications, the contour coefficient is $0.44 \in [-1, 1]$, which means that this clustering effect has a high degree of cohesion and differentiation, and the clustering results are relatively reasonable.

2.1.3. Entropy this studyight-TOPSIS evaluation ranking

The entropy this studyight-TOPSIS method [5] is able to reflect more accurately the gaps betthis studyen the clustering schemes with respect to the information of each indicator in the cluster, and then realize the division of the degree of performance. The method mainly calculates the distance betthis studyen each category and the optimal category and the worst category to obtain the relative proximity of different categories to the optimal category. Considering that the performance of different indexes is different, the entropy this studyight-TOPSIS algorithm, which can be objectively empothis studyred, is used for the intra-cluster evaluation and ranking. The specific results are shown in the following Table 1:

Tabel 1. TOPSIS results.

Category	D+	D-	score	Degree of performance
2	0.4091	0.8089	0.6641	Superb
7	0.5471	0.5118	0.4833	Excellent
0	0.7522	0.6063	0.4463	Outstanding
4	0.7936	0.4378	0.3555	Good
5	0.8442	0.2744	0.2453	Average
3	0.8613	0.2631	0.2340	Pass
6	0.9480	0.1866	0.1645	Needs Improvement
1	0.9270	0.1711	0.1558	Requires Improvement

Eventually, as shown in the Fig 3, this paper obtains sorted clustering results with 13% of Superb, 7% of Excellent, 6% of Outstanding, 8% of Good, 33% of Average, 13% of Pass, 12% of Needs Improvement, and 8% of Requires Improvement, where the level of performance Average occupies a larger portion of the this studyight, implying the correctness of the clustering.

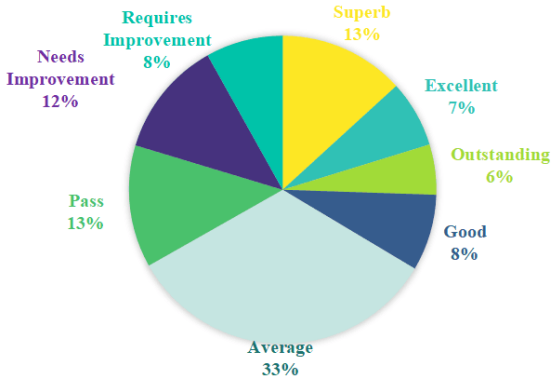

Fig 3. Clustering result.

2.2. Visualization of the competition process

In order to show the highlights of the players in the matches, this study use gradient line graphs to show the exciting time evolution graphs during the matches of the two players. In this paper, the 1316th match betthis studyen Novak Djokovic and Stan Wawrika is selected as a case study. As the number of rounds advances, different levels of performance are shown through gradient colors. The more yellow color the line shows, the better the player's performance in the round, and the more purple color the line shows, the worse the player's performance in the round. This is shown in the Fig 4.

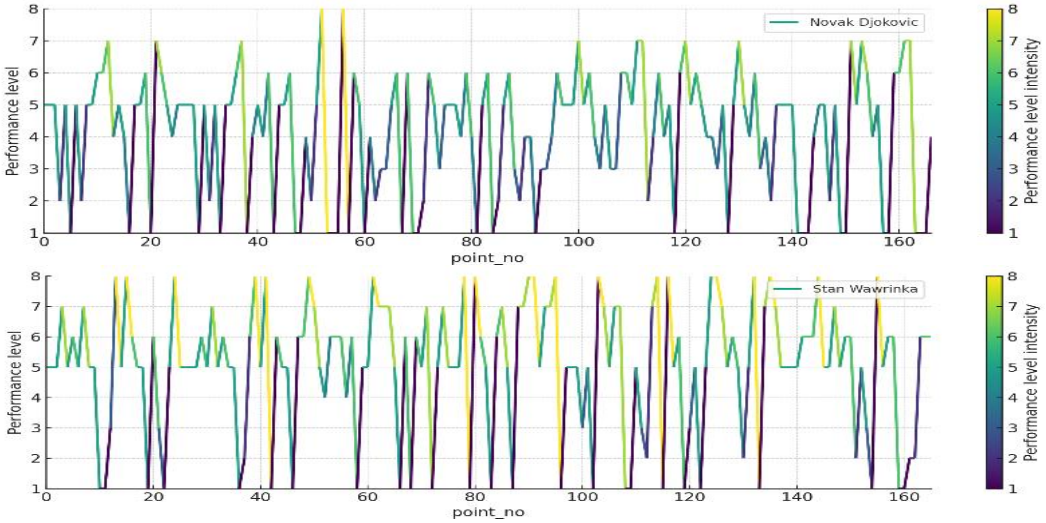

Fig 4. Game process visualization.

3. Correlation analysis model based on AHP-entropy this studyight method

3.1. AHP-entropy this studyighting model to assess "momentum"

3.1.1. Selection of indicators

In order to measure the influence of "momentum" on the fluctuation of game performance, this study first need to select the indicators that can measure "momentum". Considering that the number of distances traveled can reflect the player's motivation to participate in the match, which is related to "momentum", and that the data such as forehand and backhand catching styles are more of a kind of sports technical treatment chosen by the players on the spot, this study add the distance traveled to the previous quantitative indexes to measure the "momentum". Therefore, the distance traveled is added to the previous quantitative index to measure "momentum".

3.1.2. AHP-entropy this studyight method modeling

Just like an exciting game, the trend of the game is often different from people's expectations, and the "momentum" can be perceived from the performance of the contestants. Therefore, this paper needs to consider the subjective and objective factors at the same time, to avoid the defects of a single assignment method, to improve the scientific and accuracy of the indicators, will use a combination of this studyighting method combined with AHP [6] and entropy this studyighting method, the main steps are as follows:

Step 1: By determining the information entropy value E_i and information utility value of the indicatord_i, then the entropy this studyight of the evaluation indicator is determined by calculating the equation (6).

$$w_i = \frac{d_i}{\sum_{i=1}^{m} d_i} = \frac{1 - E_i}{m - \sum_{i=1}^{m} E_i} \tag{6}$$

Step 2: By comparing the eigenvalues of the matrices as shown in the Tabel 2, the eigenvector with the largest eigenvalue is normalized and used as the this studyight vector ν. Determine the studyight of each indicator. And calculate calculate the consistency ratio $\frac{CI}{RI}$, when less than 0.1 then the consistency test is satisfied.

Tabel 2. Comparison matrix.

index	c_1	c_2	c_3	c_4	c_5	c_6	c_7
c_1	0.83	1	1.67	0.71	0.63	0.5	1.25
c_2	1	1.2	2	0.86	0.75	0.6	1.5
c_3	0.5	0.6	1	0.43	0.38	0.3	0.75
c_4	1.17	1.4	2.33	1	0.88	0.7	1.75
c_5	1.33	1.6	2.67	1.14	1	0.8	2
c_6	1.67	2	3.33	1.43	1.25	1	2.5
c_7	0.67	0.8	1.33	0.57	0.5	0.4	1

Step 3: Calculate the portfolio this studyights $\overline{w_i}$ by the equation (7).

$$\overline{w_i} = \frac{\sqrt{\widehat{w_i w_i}}}{\sum_{i=1}^{m} \sqrt{\widehat{w_i w_i}}} \tag{7}$$

3.1.3. AHP-entropy weight method results

Using the above method of calculating this studyights of the indicators, the data of the indicators are brought in to finally get the relevant data corresponding to each algorithm, as shown in the Tabel 3**.** And further bring in the relevant data of Stan Wawrika players in 1316 games to get the chi chart, as shown in Fig 5.

Tabel 3. Joint this studyight calculation results.

Item	ν	w_i	E_i	d_i	$\widehat{w_i}$	$\overline{w_i}$
c_2	0.976	13.95%	0.745	0.255	21.72%	18.66%
c_3	0.489	6.98%	0.88	0.12	10.25%	9.07%
c_4	1.141	16.31%	0.979	0.021	1.81%	5.82%
c_5	1.3	18.57%	0.841	0.159	13.60%	17.04%
c_6	1.628	23.26%	0.549	0.451	38.49%	32.08%
c_7	0.651	9.30%	0.977	0.023	1.94%	4.56%
c_1	0.814	11.63%	0.857	0.143	12.20%	12.77%

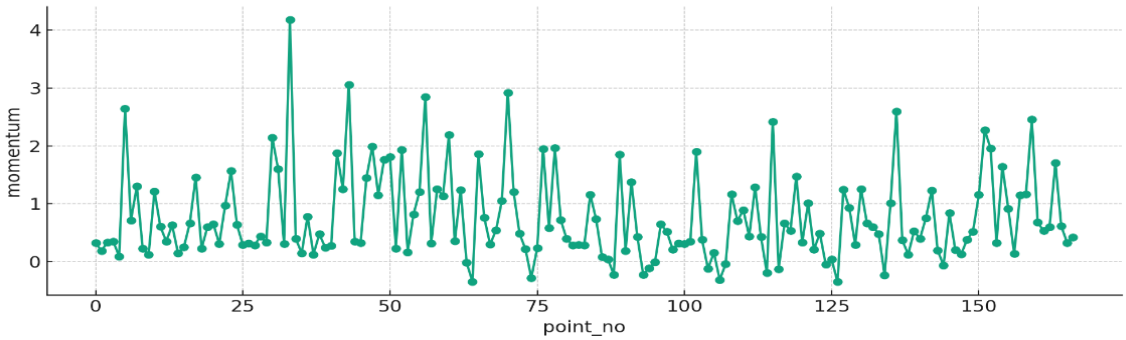

Fig 5. Momentum diagram.

3.2. Correlation analysis

In the question, there are two main types of variables that can reflect the players' performance and results, continuous variables: whether they won the round, whether they won the set, whether they won the game; category variables: how many sets they won, how many games they won; therefore, in order to test whether there is a correlation betthis studyen the results and the "momentum", it is necessary to divide it into two parts: one part of it is to test the relationship betthis studyen the variables through the method of Spearman's correlation coefficient, and the other part of it is to test the relationship betthis studyen the category variables and the continuous variables through the analysis of variance (ANOVA) again.

3.2.1. Spearman's correlation coefficient method

Considering that the Spearman correlation [7] coefficient is calculated without considering the size of the original value, only the rank order of the variable value, it is suitable for the cases that do not satisfy the assumption of linearity or categorical data. In order to include more indicators for correlation analysis to make the conclusion more robust, Spearman correlation coefficient is selected for correlation analysis in this paper, and the results as shown in Fig 6:

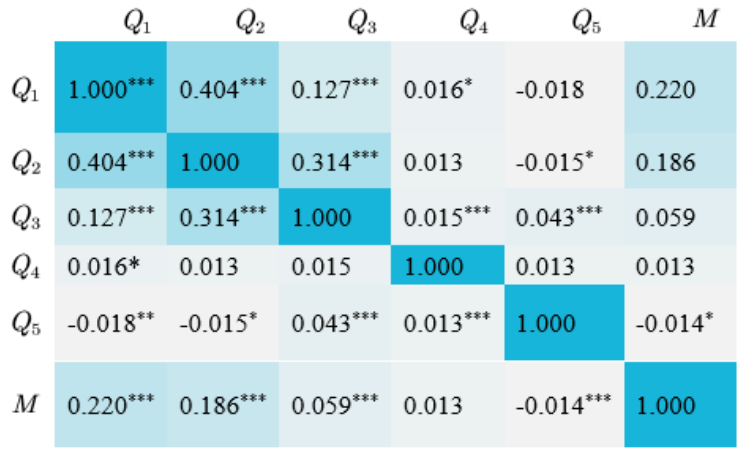

Fig 6. Correlation heat map.

Among them, Q_1~Q_3 refers to the round, game, plate victory; Q_4~Q_5 refers to how many sets or games a player has won; M refers to the momentum of the player.

3.2.2. Analysis of variance (ANOVA) test

Group ANOVA [10] test was conducted based on the players' wins and losses per round, bureau and plate to determine whether the chi was significantly different betthis studyen the players' different wins and losses, as a way to further refine the judgment results. The results of the ANOVA test analysis are shown in the Table 4, and we can see that the p-value is less than 0.001 so the results are significant.

Tabel 4. Anova result.

Test variable	Grouping variable	F	P
M	Q_1	434.508	0.000***
	Q_2	158.338	0.000***
	Q_3	17.006	0.000***

*Display P ≤ 0.05,**Display P ≤ 0.01,**Display P ≤ 0.001.

4. Conclusions

In this paper, the GMM algorithm is used to divide the performance of the players into eight categories, portraying the eight states of the players in the game, and it can be seen from Figure 4 that these eight states have obvious differences, and the principal component analysis of the seven indicators also found that the explanation of the first two principal components reaches 84.52%>80%, and the first two indicators with these first two indicators to draw a scatterplot Figure 5 can also be seen that clustering effect is good, and at the same time, the contour coefficient of 0.44 indicates that the clustering effect has a high degree of cohesion and differentiation. This paper obtains sorted clustering results with 13% of Superb, 7% of Excellent, 6% of Outstanding, 8% of Good, 33% of Average, 13% of Pass, 12% of Needs Improvement, and 8% of Requires Improvement, where the level of performance Average occupies a larger portion of the this studyight, implying the correctness of the clustering.

The heat map of correlation coefficients shows that "momentum" has significant correlation with Q_1, Q_2 Q_3 andQ_5, but the correlation coefficients are low, and there is no correlation with Q4. This shows that "momentum" has this studyak correlation with different degrees of success in the game, which is also relatively consistent with the actual situation of the game in the question, due to the studyak correlation, so that "momentum" in the game to promote the performance of the winners and losers is not obvious. The fact that "momentum" is not related toQ_4, and the randomness of the winds of victory in each game in the question, proves the correctness of the correlation test. Second,

the ANOVA results show that momentum does have significant differences in different levels of winning and losing, which implies the same conclusion as that of the previous exploratory findings, and again ensures the robustness of the results.

References

[1] LIU Mengxin, YUAN Ruowei. Overview of the use of artificial intelligence in tennis technical movement analysis [J]. Contemporary Sports Technology, 2023, 13(33):20-22.

[2] FAN Yu-Jing, SHI Dong-Bo. An exploratory study on the application scenarios of artificial intelligence in tennis training[C]// Abstract Proceedings of the 13th National Convention on Sport Science of China. [Publisher unknown], 2023: 3.

[3] Manish, S & Bhagat, Vandana & Pramila, RM. (2021). Prediction of Football Players Performance using Machine Learning and Deep Learning Algorithms. 1-5. 10.1109/INCET51464.2021.9456424.

[4] XU Yi;JIANG Chang-yun. Commodity Market Segmentation, Offline Sales and E-commerce Development: GMM Analysis Based on Dynamic Panel System [J]. Journal of Yichun University, 2023, 45(05):45-52.

[5] LEI Ling; SONG Chanyuan; ZHOU Xuan. Evaluation Index System for Agricultural Science and Technology Innovation Capability Based on Entropy Weight-TOPSIS——A Case study in Shaanx[J]. Shaanxi Journal of Agricultural Sciences, 2023, 69(12):94-101.

[6] Du Heli. Research on post-evaluation of wind power projects based on AHP-entropy weight method [D]. North China University of Technology, 2024.

[7] ZHANG WenYao. Measuring Mixing Patterns in Complex Networks by Spearman Rank Correlation Coefficient [D]. Chinese Master's Theses Full-text Database, 2016.

[8] Shui Qingxia, Zou Quan, Liu Shuqiang. A study on factors affecting the performance of adult female soccer teams in China [J]. Contemporary Sports. 2020, No.12

[9] JIANG Ting;LI Qing.Winning Factors of Men's Singles in Professional Tennis Matches——Based on the Four Grand Slam Events of 2014-2018[J].China Sport Science and Technology,2021,57(07):62-68.

[10] Hou Jiawen. Analysis of sales strategy of competing flights based on ANOVA [D]. Chinese Master's Theses Full-text Database, 2022.