Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

An Analytical Approach to Predicting Dynamics in Sports Matches Based on Machine Learning Models

Yuxuan Zhang *, Zixin Kang #, Wenjie Lyu #

College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China, 430074

* Corresponding author: 9tchaser@gmail.com

#These authors contributed equally.

Abstract. Sports events have been a significant component of global culture and economics, and analyzing the fluctuations in athletes' performance is one crucial factor in predicting the result of matches. Quantitative research on players' momentum has long been a popular topic. Building upon prior research, we have developed a mathematical model based on two combined machine learning methods. Specifically, this paper utilized a random forest model to analyze the importance of various metrics of athletes during matches. This text then applied support vector machines for regression prediction of match scores, using data from Wimbledon Championships for model training, and finally yield satisfactory results. The paper further validated this model on a test dataset, demonstrating its robust predictive capabilities. This not only represents a significant application of artificial intelligence in the realm of sports but also holds the potential for substantial economic and cultural benefits for coaching teams and society in the future.

Keywords: Machine Learning; Random Forest; Support Vector Machine; Sports Matches; Behavior Prediction.

1. Introduction

Momentum means that a player benefits from a psychological and/or physiological boost. A psychological boost is a positive change in cognition. Cognition includes changes in self-efficacy, motivation, and attention. A physiological boost is a positive change in behavior [1]. As momentum intricately intertwines with various game factors, a thorough study of it holds significant reference value for athletes' performance and coaches' strategic guidance. In this paper, we construct conceptualized momentum, which involves linking it to factors such as a player's current score, past batting performance, turnovers and distance traveled. Specific momentum assessment formulas are developed by assigning parameter values based on the relative influence of these elements. We complete some data preprocessing, which involves adding data sets such as continuous scores, excellence scores, etc. Match rankings from the official Wimbledon website are included in the outcome predictions. Then we utilize a method based on the random forest model to analyze each feature's importance towards athlete's performance. When constructing a prediction model, we construct an algorithm based on support vector machines.

2. Related Work

In this section, we discuss previous research during these five years related to the simulation and prediction of sports events. Our literature research has shown that abundant work. From the result we can conclude that models like multilayer perceptron is widely used in prediction in sports events, but most work focus on team sports such as football and cricket. Also, systematic analysis on the athletes' momentum in a solo sports game is not sufficient.

We first search for some literature review covering the work related. Bunker et al [2] concluded a comprehensive review that examines the application of machine learning techniques for predicting match results in team sports. The review covers studies from 1996 to 2019, discussing commonly

used algorithms, data approaches, and evaluation methods in the context of invasion sports and striking/fielding sports. The study by Rudrapal presents a deep learning approach using a Multi-Layer Perceptron (MLP) model to predict football match outcomes based on selected features, achieving a satisfactory accuracy rate while identifying challenges for further improvement in predictive modeling [3]. Srivastava develops a hybrid machine learning-clustering-association rule framework to predict match outcomes in cricket, utilizing Random Forest, Gradient Boosting, and Deep Neural Networks to analyze factors influencing success in cricket matches [4]. Similarly, García [5] utilized machine learning algorithms to analyze the technical-tactical behavior of football players based on game statistics from various seasons and national leagues, aiming to determine their on-field playing positions. By employing dimensionality reduction techniques and machine learning algorithms like RIPPER, the study identified discriminatory variables for each player group, providing valuable insights for player performance enhancement and position identification on the field.

3. Our Proposed Model to Quantify Athlete's Behavior

3.1. Feature Analysis

Our goal is to construct an analytic model; thus, we can measure how well the players in a sports event are behaving. The dataset we collect are inspired from and consist of 2024 MCM competition, we collected for every point from all Wimbledon 2023 matches, which is open on

https://www.wimbledon.com/index.html. After collecting them, we apply some feature analysis engineering. For instance, we processed the raw data, converting it into a floating-point format directly readable by the model. This step ensures the consistency and availability of the data, enabling it to seamlessly integrate into subsequent analysis and modeling processes. To better accommodate complex models, we introduced additional features, including continuous scoring data, per-point scoring data, and the player's individual ranking at Wimbledon. Such enriched information contributes to enhancing the predictive accuracy of the model, allowing it to comprehensively consider various aspects of player performance. Athletes' metrics can be defined in various ways to better fit the proposed model. To train the random forest regression model, we need to preprocess the obtained data to obtain quantitative factors. These six factors have varying degrees of impact on potential energy, and their definitions are shown in Table 1.

Table 1. Different factors that influence athlete's momentum

Symbol	Definition
ρ_t	Current winning scores
γ_t	Serving situation
λ_t	Score streak
ε_t	Excellent scores
δ_t	Mistakes
ϵ_t	Physical factors
$\alpha_{ti}^{i=1\ldots6}$	Weight coefficient

We then apply MinMaxScaler method to perform standardization and regularization operations to obtain the standardized and regularized data set X_{norm}[6]. Finally, we store the standardized and regularized data into a new data table, some of the data at the top is shown in Table 2.

Table 2. Standardized and regularized feature data

Point	Rank	ρ_t	λ_t	ε_t	γ_t	ϵ_t	δ_t
1	0.619863	0.565789	0.416667	0.5	0.166667	0.582254	0.25
2	0.619863	0.526316	0.5	0.5	0.833333	0.582261	0.5
3	0.619863	0.565789	0.416667	0.5	0.833333	0.610427	0.25
4	0.619863	0.526316	0.5	0.5	0.833333	0.735576	0.5
5	0.619863	0.565789	0.541667	0.678571	0.833333	0.570925	0.5

3.2. Basic Structure of Random Forest Model

The machine learning algorithms such as Support vector Machine, Random Forests and Multi-layer perceptron [7] were used for training on the input dataset. Random forests [8] make use of ensemble approach to generate decision tress for the input tuples where each tree attributes to the input vector in random fashion. Support vector machines [9] make use of non-linear mapping technique for numerical regression and classification. We applied random forest algorithm to obtain the contribution of each feature to the game score to analyze the coefficients that different features should choose in building the momentum formula. Specifically, we used the standardized data set X_{norm} as the model input, and the scored or not data set y as the model output to perform random forest model training. A random forest consists of multiple trees, each trained on a randomly drawn dataset and random feature selection. Each tree makes its own prediction about the data set. Since each tree is trained on a different subset, and the feature selection for each tree is also random, random forests have good generalization capabilities and can reduce the risk of overfitting. We first prepared a set of preprocessed data sets $X = (X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$, where X_i is the obtained features data. The characteristic vector of the player, y_i, is the corresponding target variable, that is, the momentum indicator of each player. We use multiple grouped and processed tennis match data as the random forest model training set. After training the model, we call the feature importance function of the model library to match the feature importance with the feature name, then sort them in descending order and draw a visual feature importance icon, shown in Figure 1. Finally, we need to evaluate the model. The mean square error of model training equals 0.00018 << 1, R square = 0.9883 which is close to 1, meaning its excellent function.

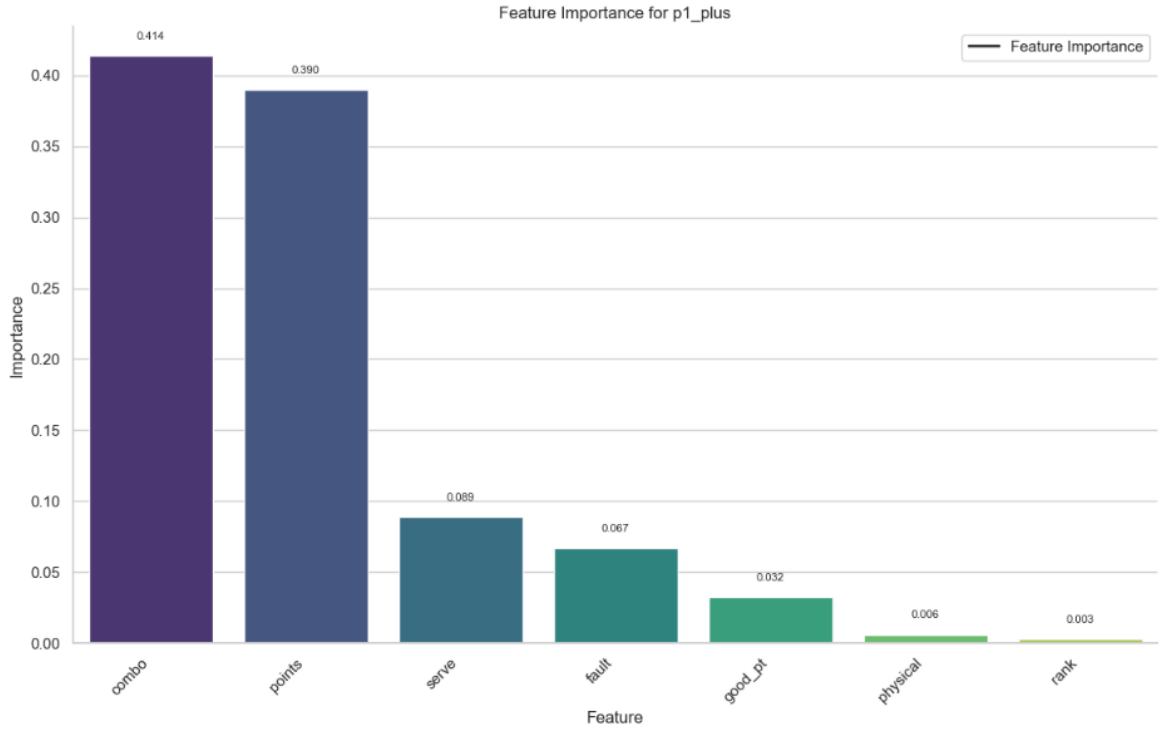

Figure 1. Features importance chart

3.3. Modification of Our Model

Machine learning is being widely used in sports competitions. To predict the outcome of each game, we established a Supporting Vector Machine (SVM) model. We use these two models to perform binary classification work, that is, whether the first player scores or the second player scores. The dataset is collected and synthesized in section 3.2, 80\% of which is used as the training set and 20\% is used as the test set. The input matrix in table 3 represents the features of the training data. The output matrix typically represents the predicted results by the model for each corresponding input sample. In binary classification, it is a column vector with predicted class labels (0 or 1) for each sample.

Table 3. Inputs and outputs of SVM model

Inputs	Rank	ρ_t	λ_t	ε_t	γ_t	ϵ_t	δ_t
Outputs	Win or Lose (1 or 0)						

Support vector machine is a traditional machine learning method that is often used to deal with classification problems, and when the dataset is small, the training results are sometimes better than neural networks with larger model structures [10]. The support vector clustering algorithm applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data. The core idea of SVM is to find an optimal hyperplane $f(x) = 0$ in the feature space to separate samples of different categories and maximize the margin of classification boundaries. We use the Gaussian kernel function to map input features, such as the current score, serve, ace, etc., into a higher dimensional space. Then we use the decision function to predict the sample classification. Convex optimization algorithms are used to optimize the function during training.

The Gaussian kernel function nonlinearly maps the feature space to a high-dimensional space. When two samples are relatively close in the feature space, their similarity $K(x, y)$ (5) in the high-dimensional space is higher. The decision function $f(x)$ (6) uses the kernel function and other parameters to define a hyperplane, the decision boundary, to separate the two categories of samples: losing or winning. Supporting vectors are sample points that play a key role in the decision function. These sample points are located near the hyperplane and have an important influence on determining the position and direction of the hyperplane. The calculation of the decision function involves the weight of the support vector and the application of the kernel function. The model is then optimized by constructing a Lagrangian function (7) to obtain a set of optimal Lagrangian multipliers α_i, The optimization process must meet constraints (8). And the model parameters (w, b) are calculated using the multipliers.

$$K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{5}$$

$$f(x) = \sum_{i=1}^{m} \alpha_i \, y_i K(x_i, x) + b \tag{6}$$

$$\max_{\alpha}\left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \, \alpha_j y_i y_j K\left(x_i, x_j\right)\right) \tag{7}$$

$$\begin{cases} 0 \le \alpha_i \le C & for\ i = 1,2,\ldots,m \\ \sum_{i=1}^{m} \alpha_i \, y_i = 0 \end{cases} \tag{8}$$

We used the Gaussian kernel function method to construct a support vector machine model. The input features and output labels are both the same (shown in Figure 2). Train Accuracy is 0.9679, and test Accuracy is 0.9883, achieving a good two-classification effect. (Figure 3) By observing and comparing the learning curves, we can see that the overfitting effect of the support vector machine model is not obvious, and it has good generalization ability. When the data set is small (training

samples <8000), the support vector machine can usually have better training performance, while the traditional neural network needs to have tens of thousands of training samples to show better fitting effect. To sum up, we believe that the support vector machine model performs well in the current data set.

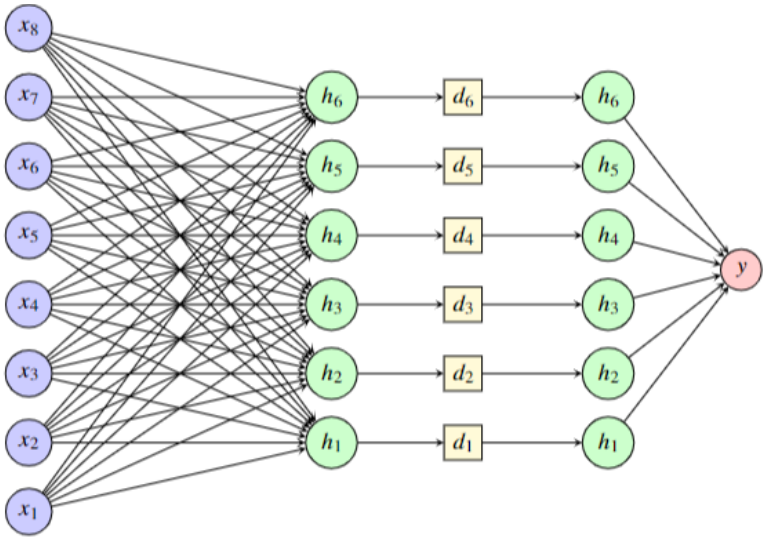

Figure 2. Schematic picture of SVM model

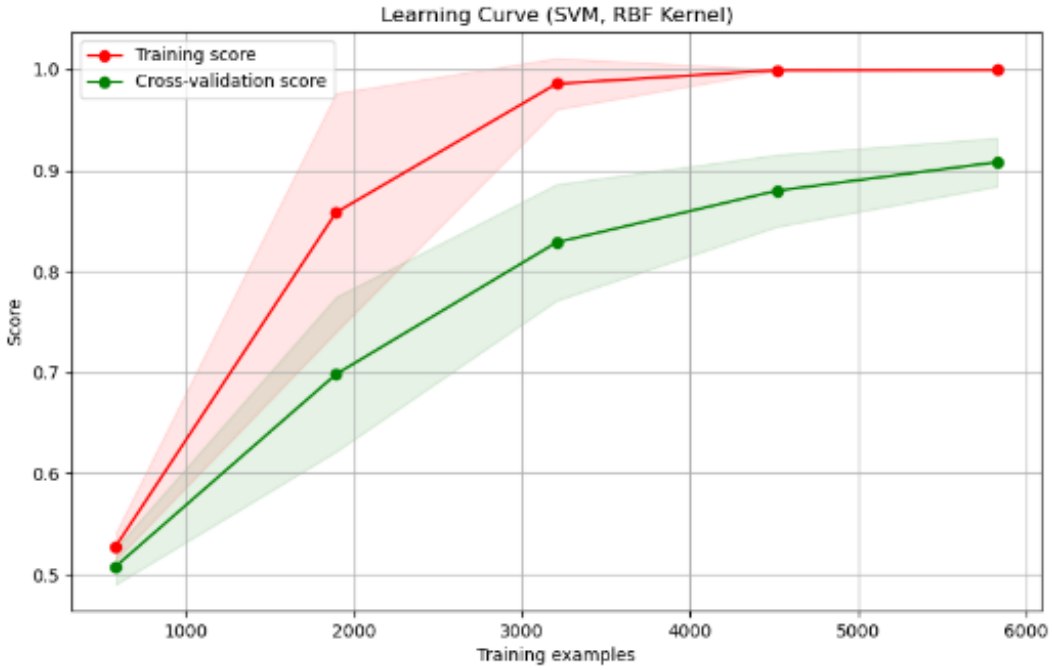

Figure 3. Learning curve of SVM model

4. Results

4.1. Prediction Model to Simulate the Match Outcome

To confirm the hypothesis that the larger the absolute value, the higher the winning percentage, we associate the momentum of all matches with their outcomes. We examine the winning percentages of players under different absolute values of momentum. The resulting data is illustrated in Figure 4, and its variation appears to be describable by a function. Therefore, we perform non-linear regression on the data. The criteria for the generated graph are such that if the x-coordinate is 5, the y-coordinate represents the winning percentage in matches where the momentum is greater than 5. It can be

observed that even when defining positive momentum as a win and negative as a loss, there is an accuracy of over seventy percent. Furthermore, it is noticeable that as the momentum increases, the winning percentage shows a significant improvement.

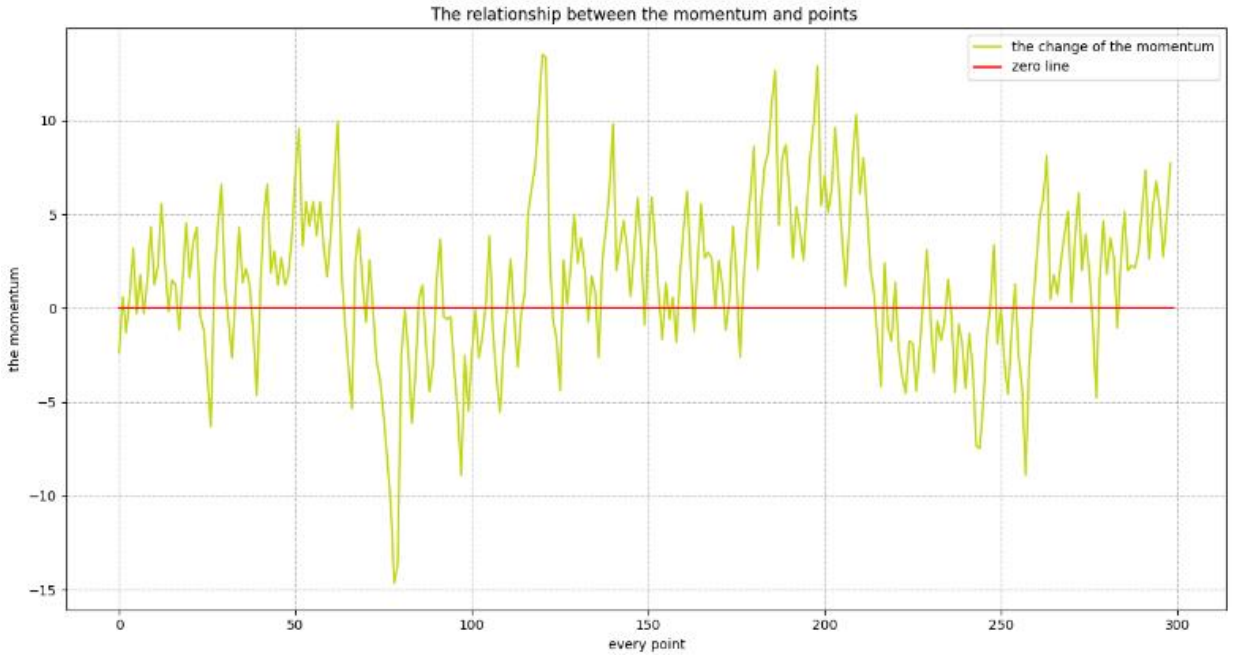

Figure 4. Athlete's momentum flow during a game

4.2. Analysis of Experimental Results

To validate the hypothesis that the greater the absolute value of momentum, the higher the winning probability, we correlate the momentum of all games with their outcomes. The results are presented in the following graph. Due to its nearly linear trend, we apply linear regression. The relationship between momentum and victor is shown in Figure 5. If the x-axis is 5, the y-axis represents the winning percentage in games where the momentum is greater than 5. It can be observed that even if we consider momentum greater than 0 as a win and less than 0 as a loss, we achieve an accuracy exceeding seventy percent. It is evident that as momentum continues to increase, the winning percentage shows a significant improvement. Therefore, our momentum evaluation model can effectively quantify the changes in players' status during the game will further help predict the trend of the game and help coaches and players make decisions.

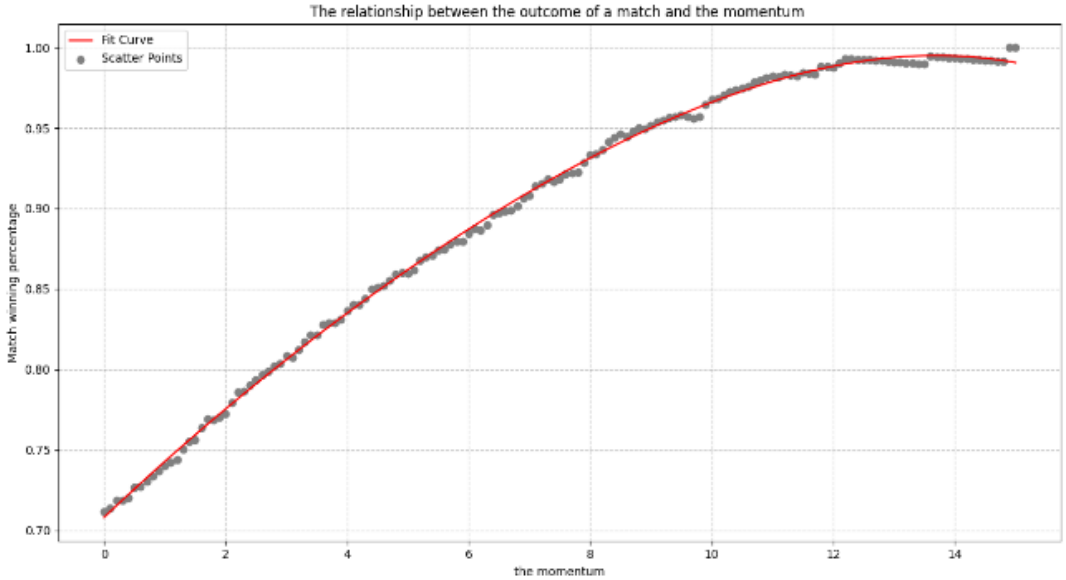

Figure 5. Momentum flow compared with actual points

5. Conclusion

In summary, the comprehensive exploration and analysis of momentum in tennis have resulted in a refined and matured model. This paper integrates various factors, establishing a specific momentum evaluation formula with parameter values derived from their relative impact factors. Continuous simulation of momentum changes throughout the game provides valuable insights for predicting game direction, benefiting both players and coaches. This text reconstructs the model using the support vector machine method, further refining predictions of shot outcomes. This holistic approach not only aids in predicting game outcomes but also provides valuable insights, elevating people's understanding of the intricate dynamics of tennis. This has important implications for players and coaches as they can adjust in time and capitalize on favorable momentum, improving their chances of winning, as well as propagating economic and culture influence in sports field in the future.

References

[1] Dietl, Helmut, and Cornel Nesseler. "Momentum in tennis: Controlling the match." UZH. Business Working Paper Series 365 (2017).

[2] Bunker, Rory, and Teo Susnjak. "The application of machine learning techniques for predicting match results in team sport: A review." Journal of Artificial Intelligence Research 73 (2022): 1285 - 1322.

[3] Rudrapal, Dwijen, et al. "A deep learning approach to predict football match result." Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018. Springer Singapore, 2020.

[4] Srivastava, Praveen Ranjan, et al. "Best strategy to win a match: an analytical approach using hybrid machine learning-clustering-association rule framework." Annals of Operations Research 325.1 (2023): 319 - 361.

[5] García-Aliaga, Abraham, et al. "In-game behaviour analysis of football players using machine learning techniques based on player statistics." International Journal of Sports Science & Coaching 16.1 (2021): 148- 157.

[6] Pires, Ivan Miguel, et al. "Homogeneous data normalization and deep learning: A case study in human activity classification." Future Internet 12.11 (2020): 194.

[7] Kapadia, Kumash, et al. "Sport analytics for cricket game results using machine learning: An experimental study." Applied Computing and Informatics 18.3/4 (2020): 256 - 266.

[8] Herce-Zelaya, Julio, et al. "New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests." Information Sciences 536 (2020): 156 - 170.

[9] Meshram, Sarita Gajbhiye, et al. "Application of artificial neural networks, support vector machine and multiple model-ANN to sediment yield prediction." Water Resources Management 34.15 (2020): 4561 - 4575.

[10] Krstić, Dušan, et al. "The Application and Impact of Artificial Intelligence on Sports Performance Improvement: A Systematic Literature Review." 2023 4th International Conference on Communications, Information, Electronic and Energy Systems (CIEES). IEEE, 2023.