Football Match Analysis and Prediction Based on LightGBM Decision Algorithm

Chen Chen
Physical Education Institute, Yantai Institute of Science and Technology, Yantai 265600, China
E-mail: chenchen_yt@163.com

Keywords: LightGBM decision algorithm; Football matches; Grey relational analysis; Event prediction; Fuzzy theory

Received: May 21, 2024

The advent of the digital age has created new opportunities for the development of the sports industry, especially with data mining technology promoting the informatization process of the sports industry. However, there are many factors that influence football matches, and predicting the result is extremely difficult. Therefore, firstly, a dataset is constructed using a crawler algorithm and processed through various data processing techniques. Then, an improved algorithm combining the random forest algorithm and the light gradient boosting machine decision tree algorithm is proposed. Finally, a fuzzy grey relational method is designed by combining fuzzy theory and grey relational model. From the research results, in the two groups before and after performing feature engineering operations, although the feature count decreased by 48.8% after the operation, the accuracy and area under the curve of the improved algorithm were the highest, with 95.31% and 86.74%, 0.9124 and 0.9767, respectively. In comparison with other mainstream algorithms, the fusion improvement algorithm and fuzzy grey relational method had the highest accuracy, F1 value, and area under the curve, corresponding to 97.26%, 93.71%, and 0.9885, which were 0.12% and 0.06% higher than the accuracy of all features and area under the curve results, respectively. The above results indicate that the proposed method has superior analysis and prediction performance, which can further explore effective information, providing an effective analysis and prediction method for football related personnel and enterprises.

Povzetek: Narejena je analiza in napoved rezultatov nogometnih tekem, ki temelji na izboljšanem algoritmu LightGBM, optimiziranem z naključnim gozdnim algoritmom in relacijsko analizo. Uporaba metode omogoča izboljšano taktično analizo, podporo za načrtovanje treningov, oblikovanje strateških odločitev ter prinaša podlago za odločanje trenerjem, športnim strokovnjakom in navijačem.

1 Introduction

At present, China's sports industry is in the primary stage. The greatest driving force for industrial development comes from the strong support of national top-level policies [1]. In recent years, the country has clearly proposed the strategic goal of making the sports industry a pillar industry of the national economy. The General Administration of Sport of China and relevant departments has jointly issued 12 industry plans for 10 sports projects [2-3]. In competitive sports, both professional athletes and sports enthusiasts face the same problem, where the competition process is difficult to quantify and valuable information cannot be efficiently obtained from massive data to improve the training effectiveness of athletes [4]. Artificial intelligence technology, big data technology, and other technologies provide new methods for sports training, event analysis and prediction, and also add new impetus to the sustainable development of China's sports industry [5-6]. However, Football Match (FM) is extremely difficult to analyze and predict due to the ambiguity, complexity, and uncertainty of related data [7]. Barra et al. observed that FM analysis relied on annotations of individual players, team actions, and player performance, which made annotating FM at a fine-grained level a very expensive and error prone task. Therefore, an AI-based FM annotation system was designed, demonstrating its effectiveness in real-world application scenarios through experiments [8]. Chen proposed an optimal trajectory prediction method based on the sliding window for football long-range shooting to improve the accuracy of football shooting trajectory prediction. The experimental results showed that the deviation of this method was less than 0.2, and the trajectory prediction accuracy was as high as 99.9%, which met the practical needs of football trajectory prediction [9]. Huang and Bai designed an intelligent sport prediction and analysis system based on particle swarm optimization algorithm. Combining the prediction performance of particle swarm optimization algorithm edge computing in artificial intelligence with traditional prediction analysis methods, an intelligent sport prediction analysis system was designed. The research results showed that the prediction accuracy of

the system reached 89.6% [10]. The above results indicated that the prediction accuracy of sports event prediction analysis still had certain limitations. The characteristics of FM, such as data being prone to errors and ambiguity, may result in lower prediction accuracy. Therefore, this study introduced feature engineering to optimize the indicators used for event prediction, to obtain more accurate prediction results. Hu et al. introduced the inherent flaws of tensor decomposition models when potential factors were represented as the edge information function. The Friedman tensor was proposed as a large-scale prediction tool for high-dimensional data. Compared with the Light Gradient Boosting Machine (LightGBM) algorithm and the Field Perception Factor Decomposition Machine, the designed method had good tracking records, which was widely used in large-scale prediction [11]. Traditional line loss methods are not suitable for situations where data quality is lacking in actual production environments. Therefore, Tang et al. designed a short-term line loss prediction method for low-voltage distribution networks based on K-means algorithm and LightGBM algorithm. The prediction accuracy of the research method was significantly higher than that of back propagation neural network and support vector regression methods [12]. Masood et al. proposed a machine learning model to address the limited ability of experience propagation models to capture various propagation environment characteristics. Even in sparse training data, the overall performance of LightGBM was still superior to other machine learning algorithms. Compared with the empirical model, its prediction accuracy increased by 65% [13]. The above results showed that the overall performance of LightGBM algorithm was better than that of other machine learning algorithms in the case of various training data. The prediction accuracy was significantly improved. However, a single ensemble learning tree model only had a good effect on a certain type of training data, and had a poor effect on other data. Aiming at the above problems, this paper introduced Random Forest (RF) algorithm to realize multi-aspect prediction learning of the model. Based on the above literature summary, Table 1 is compiled.

Table 1: Summary of literature results

Author	Research contents	Performance index result
S. Barra et al.[8]	A football match annotation system based on machine learning algorithm is designed	/
Z. Chen[9]	An optimal trajectory prediction method based on sliding window is designed	The deviation of the research method is always below 0.2, and the accuracy of trajectory prediction is as high as 99.9%
Y. Huang et al.[10]	In this paper, an intelligent sport prediction analysis system is designed based on the prediction performance of edge computing of particle swarm optimization algorithm and the traditional sports event prediction method	The prediction accuracy of this method is about 89.6%
R. Hu et al.[11]	A Friedman tensor method is proposed, which is used as a tool for large-scale and high dimensional data prediction, and the performance of LightGBM algorithm is compared	/
Z. Tang et al.[12]	A short-term line loss prediction algorithm for low voltage distribution network based on K-means-LightGBM is designed	The prediction accuracy of the proposed method is higher than that of the back propagation neural network and the traditional LightGBM algorithm
U. Masood et al.[13]	A model based on machine learning is proposed and path loss is estimated by a key predictor	In the case of sparse training data, LighrGBM algorithm is generally better than other algorithms, and the prediction accuracy is improved by 65%. Compared with ray tracing, the prediction time is reduced by 13 times

Based on the above content, it can be seen that the existing method with the best comprehensive performance at present is based on K-means-LightGBM algorithm, but the accuracy of this method in FM prediction is relatively limited. In addition, great research achievements have been made in sports event analysis and the application of LightGBM algorithm. However, the feature analysis of large-scale sports event data and the accuracy of FM prediction are limited. Therefore, firstly, the dataset is obtained through a crawler algorithm and processed using a series of data processing techniques. Secondly, FM features are constructed and the LightGBM decision algorithm is optimized using RF algorithm to obtain the LightGBM Decision Optimization (LDO) algorithm. Then, Fuzzy Grey Relational Analysis (FGRA) is designed. Finally, the LDO algorithm and FGRA are fused to obtain FM analysis and prediction methods. The research aims to provide more accurate FM

analysis and prediction results, optimize tactics, improve athlete training methods, assist coaches in formulating game strategies and selecting starting lineups, and lay a solid data foundation for fans to make bets. There are two main points of research innovation. The first is to propose a LightGBM decision algorithm based on RF, namely LDO, to solve the single LightGBM decision algorithm having better experimental results only in a certain type of data and may have poorer experimental results in other models. The second is to design a new analysis method that combines fuzzy theory with grey relational model, and apply it to FM result analysis.

2 Methods and materials

To design a scientific and reasonable method for analyzing and evaluating FM indicators, and provide more effective guidance for athlete training, the collection and processing methods of FM data are first introduced. Then, based on the LightGBM decision algorithm, the RF is fused to obtain the LMO for prediction. Finally, a FGRA method combining fuzzy theory with grey relational model is designed to obtain FM analysis and prediction methods that integrate LDO and FGRA.

2.1 The processing method for football match data

With the flourishing development of artificial intelligence technology, big data technology plays an extremely important role in the modern football. At the same time, the continuous improvement of data makes the football data utilization more visual, thereby quantifying the data and facilitating the subsequent use of intelligent methods to replace subjective judgments and decisions [14-16]. Therefore, the study first explores data based on feature engineering, which has two stages: data capture and data processing. In the data capture stage, considering the particularity of FM, its results are usually influenced by several factors and cannot be expressed accurately based on mathematical equations. The FM data acquisition is the most fundamental and crucial part, and its quantity and quality are related to subsequent analysis and prediction. Therefore, the scouting net is selected for data extraction and analysis, and crawler algorithms are used to capture data. In addition, considering the large amount of data that needs to be collected by crawlers, the Python and asynchronous Scrapy framework are used to build a data scraping project, and saved it through a non relational database. Therefore, the data scraping stage can be obtained, as shown in Figure 1.

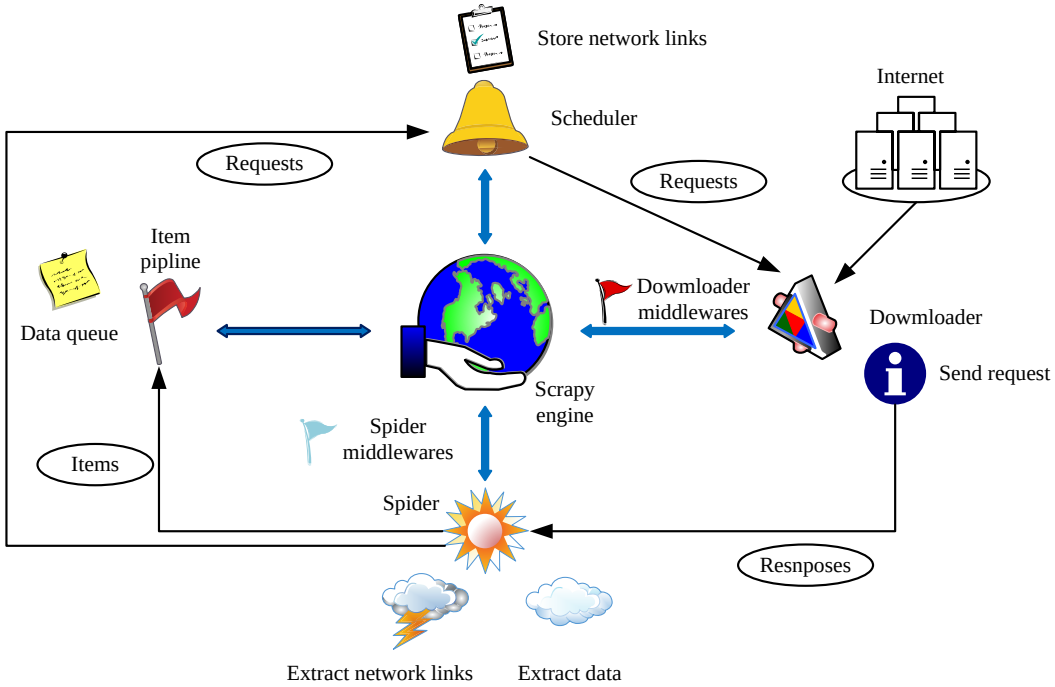


Figure 1: Process diagram of data collection stage

Based on the special structure of the scouting website, the asynchronous Scrapy framework ensures the safe and smooth operation of the website while collecting specific data. Each specific process set by it is the foundation for ensuring accurate data acquisition. The specific operation process is as follows. Firstly, the domain of the crawler starts the scouting website. The Scrapy engine processes the website while directly transmitting the link of the first crawled website to Schedler. After simple processing, the generated request is transmitted to the engine. The scheduler feeds back the network links in the list to the engine, and the engine

transmits it to the down-loader based on the download middleware. Secondly, after the network download is completed, the response content is returned to the engine. After receiving the response, the engine can transmit it to the spider program for processing. Then the crawler processes the corresponding items, returns them to the crawled items, and transmits new requests to the engine. Finally, the above operation is repeated until there are no new requests in the schedule. The relationship between the engine and the domain can be disconnected. The unique structural framework diagram of the scouting net is shown in Figure 2.

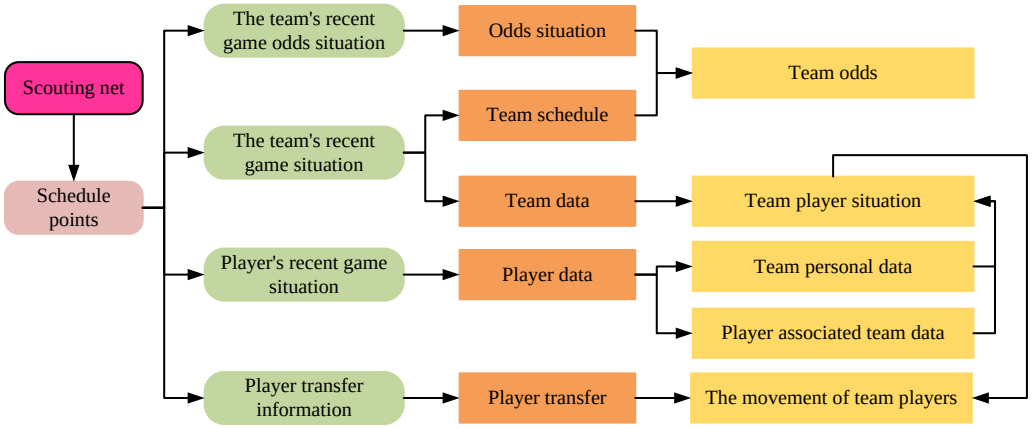


Figure 2: The unique structural framework diagram of the scouting net

After the crawler completes data extraction, the collected massive and unordered data needs to be stored in a database for subsequent data analysis. Usually, database selection needs to meet the characteristics of convenient use, long storage time, certain stability, and ensure data integrity. The data collection processes this time involves 100 requests per second. It is also necessary to analyze the existing data as much as possible through the obtained data. At present, storing a hypertext tag file on a website requires 1MB of storage space. 100MB of data content can be stored per second. Therefore, a non-relational database Not Only SQL (NoSQL) is chosen for research. This database does not require strong requirements for the relationships and structures between data, which is beneficial for FM data storage and frequent data reading [17-19]. This study ensures data accuracy through several repeated crawls. The large amount of FM historical data obtained mainly includes the mainstream European and American leagues from 2018 to 2023, namely La Liga, Bundesliga, Serie A and Ligue 1, Serie B, Bundesliga, Japanese professional leagues, Major League Soccer in the United States, English Championship, Football Association Challenge Cup, and Portuguese Super League.

In the data processing stage, tasks such as data cleaning and transformation are included. For abnormal data, it is deleted or modified to improve the quality of the data sample. The data cleaning task mainly includes the recent match status data of athletes, the recent match status data of the team, the recent match odds situation, and the athlete transfer information table. For different collected data, a one-on-one approach is used for processing, as follows. The scores of athletes and teams are scored using deleted variables, while control rate, shooting, passing, and success rate are filled in with statistics. The relevant competition results are filled with dummy variables. The median or mean is used to replace outliers in competition information. The missing transfer time for athletes is filled with re-collection. The match time, athlete and team names, and scores are treated with noise. The odds and loss of home and away game information are filled in and regressed based on the loss situation and the last five games, respectively. In the data integration stage, which includes data integration, simplification, and transformation, the obtained data is first stored in the MongoDB database. Then 564829 athlete transfer information, 1256984 athlete recent game data, and 65142 team recent game data are obtained. Based on the number of matches, and team names, data integration is conducted to obtain the FM historical data analysis library. Then, in data reduction, each attribute is normalized, and the calculation is shown in equation (1).

$$g_s = \frac{g - g_{\min}}{g_{\max} - g_{\min}} \tag{1}$$

In equation (1), g and g_s represent the match data before and after normalization, respectively. $g_{\min}$ and $g_{\max}$ are the minimum and maximum values of a certain attribute, respectively. Finally, in data conversion, the match results correspond to -1, 0, and 1 for wins, draws, and losses, respectively. The ball control rates of 0% -30%, 40% -50%, 50% -60%, 60% -70%, and 70% -100% correspond to K1-K5, respectively. In the classification of lost attributes for important athletes, if there are no red cards, less than two yellow cards, and important athletes are not lost, the corresponding attribute

value is D1. If there is a red card and the red card athlete is lost, it is D2. If there are two yellow cards and the yellow card player is lost, it is D3. Athletes who are injured and lost are D4. This athlete scores high but does not participate in the competition due to injury, with a score of D5.

2.2 FM Data analysis based on LightGBM decision algorithm

After the above FM data processing is completed, the features can be constructed. Then the importance of the features obtained from the whole process analysis to predict the FM historical data can be predicted. Finally, the LightGBM decision algorithm is used to predict. Firstly, the characteristic engineering by mastering the principle of FM and the rules of sports competition is established. Because FM has the correlation of human, natural and realistic factors, the research can complete the whole process of feature engineering by deeply mining the features from seven dimensions, including Home and Away Offense and Defense (HAOD), Recent Status of Home and Away Games (RAHAG), Performance Status (PS), History of Two Teams Fighting (HTTF), Odds Situation (OS), Point Difference (PD), and Round Information (RI), as shown in Figure 3.

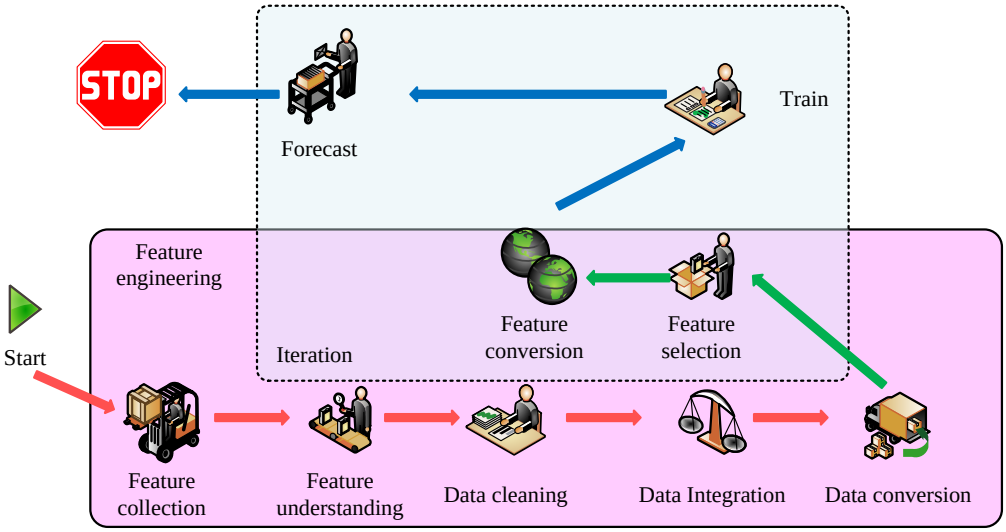


Figure 3: Schematic diagram of the entire process of feature engineering

In Figure 3, feature engineering is related to whether the data can effectively mine the hidden information of the data, which affects whether the subsequent model can achieve better training results. Feature construction is the core part of feature engineering, which generates new feature variables by combining single feature with other features. The mental generative variable is a multi-variable feature constructed on the basis of the original single feature. A single feature may have a weak correlation with the FM prediction result, but this does not mean that the feature is insignificant or invalid, because the influencing factors of the FM prediction result are often complex rather than single. It is more likely that the FM prediction results show a change rule under the joint action of many factors. Therefore, the new regenerative feature will be more effective than the original feature, contain more information, and usually better reflect the change of the FM prediction results. The main purpose of the subsequent model is to balance the over-fitting and under-fitting cases, improve the generalization ability of the model while ensuring the accuracy, and generally maintain the balance by controlling the number of input features. If the number of features is too small, under-fitting problems may occur in model training due to the lack of information, and retaining more features can enable the model to fully learn the feature information. Therefore, the feature engineering to construct features is not only considered as obtaining more dimensional features, but more importantly, to minimize the impact of certain variable features through feature intersection and the combination of other features. A new feature which can better reflect the change rule of FM prediction results is obtained. After that, the original features obtained after data cleaning are set as the Control Group (CG). The features processed by feature engineering are set as the Experimental Group (EG) [20-22]. All the features of the 64 items in the CG are shown in Figure 4.

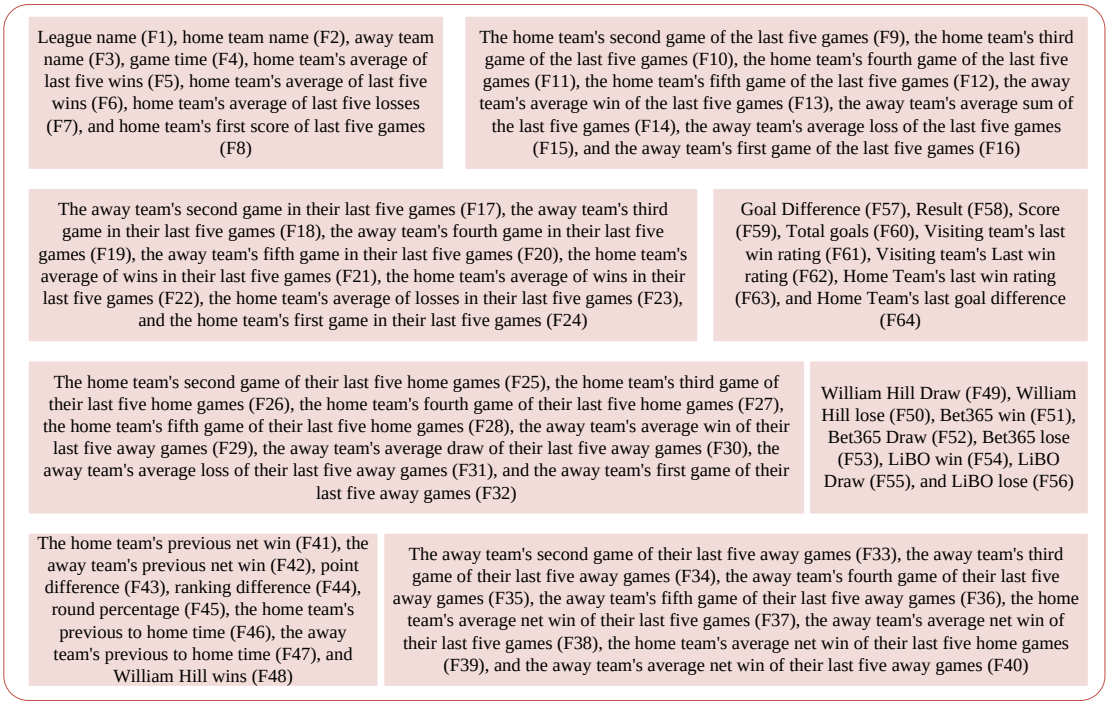


Figure 4: Feature set of the CG

In Figure 4, due to the feature data name being too long, F1-F64 is used for replacement. The dataset partitioning of the EG is optimized through feature engineering. The specific features are as follows. The home advantage and home team advantage are recorded as S1 and S2. The average values of the total field and sub field are denoted as S4 and S5, respectively. The scores of the home team, away team, home team home field, and away team away field in the last five games are recorded as S6-S9. The integral difference is (0, 3) and (3, 6), denoted as S7-S8, respectively. The two features with integrals of (0, 3) and rounds exceeding 0.8 and 0.9 are respectively denoted as S10 and S11. The rest ratio and its intensity are denoted as S12 and S13, respectively. The winning loss rate and draw rate are recorded as S14 and S15, respectively. The credibility corresponding to the odds of the home and the away is recorded as S16 and S17, respectively. The home and away teams of the two teams, excluding the two teams in this game, and the corresponding strength ratio of the two teams are set as S18-S20. The recent abnormal matches and the corresponding home and away matches between the two teams are set to S21 and S22, respectively. The goal ratios for home and away teams are recorded as S23 and S24, respectively. The home and away, the last five home and away games, as well as the last home and away games of the home and away teams, are recorded as S25-S28. Except for the last home and away game and the last five games, as well as the home and away teams of the home team and the last five games of home and away teams, the corresponding goal difference is recorded as S29-S32. The characteristics of S33-S41 are consistent with those of F49-F57 in CG, totaling 41. Finally, the LightGBM decision algorithm can be constructed for analysis and prediction, which is learning framework based on the gradient boosting decision tree. It has efficient training speed, low memory usage and expenses, high accuracy, and suitability for large-scale data processing, which has extremely high application value in the industrial field [23-25]. However, a single LightGBM decision algorithm may have good performance on specific types of training data, but it may also have poor experimental results on other types of training data. Therefore, this study investigates the ensemble tree approach with different ensemble methods, namely the RF algorithm for optimization, so that the LDO algorithm has serial and parallel learning advantages [26-27]. The process framework of LDO algorithm is shown in Figure 5.

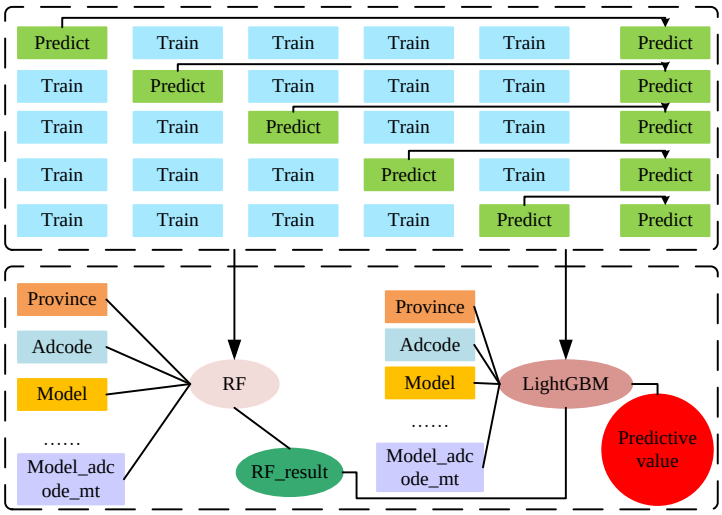


Figure 5: The flow framework of LDO algorithm

In Figure 5, the prediction result of RF is introduced into LightGBM decision algorithm as a new feature. Then LDO algorithm is modeled and predicted, so as to realize the multifaceted learning and prediction of the algorithm. RF belongs to bagging framework, which uses the sampling method with return to get different training subsets, and each base learner carries out parallel training based on the above [28-30]. The base model has low dependence. Finally, the prediction results of all the base learners are integrated. The mean method is adopted, as shown in equation (2).

$$B_N(x)=\frac{1}{N\sum_n b_{n(x)}} \tag{2}$$

In equation (2), $b(x)$ and N are the base learners and their number, respectively. The LightGBM decision algorithm belongs to the Boosting framework. The base learners are trained sequentially in a ladder like manner, and the error rate is shown in equation (3).

$$\zeta_n=\sum_{i=1}^{m}\omega_n(i)\left[y_i\neq b(x_i)\right] \tag{3}$$

In equation (3), m and $\omega_n(i)$ respectively represent the number and weight of training samples. The weight ψ_n of the next classifier is shown in equation (4).

$$\psi_n=\frac{1}{2}\ln\frac{1-\zeta_n}{\zeta_n} \tag{4}$$

The weight of the training sample is updated, as shown in equation (5).

$$\omega_{n+1}(i)=\frac{\omega_n(i)}{C}e^{-\psi^n b^n(x)y(x)} \tag{5}$$

In equation (5), C stands for the constant, which is used for weight normalization. The final strong learner can be obtained by continuously training the next classifier. The expression is shown in equation (6).

$$B(x)=sign\left[\psi^1 b^1(x)+\psi^2 b^2(x)+...+\psi^N b^N(x)+\right] \tag{6}$$

Finally, the training results of all the base learners of LightGBM decision algorithm adopt the weighted method to get the final prediction results. From the above, RF pays attention to variance and improves the prediction accuracy by reducing the training variance, while LightGBM decision algorithm pays more attention to deviation and improves the prediction accuracy by reducing the deviation. Fusing the two integrated tree models can effectively avoid conflicts caused by the fusion of different types of methods. Meanwhile, more valuable information can be learned from more angles to balance the prediction results of LDO.

2.3 Analysis and prediction methods for football matches integrating FGRA

Based on the 64 FM features obtained above, it can be concluded that there are many factors that affect FM. At the same time, there is uncertainty, complexity, diversity, and ambiguity among different factors. Therefore, it is difficult to distinguish the correlation between influencing factors and results. In response to the above issues, the study first uses a grey relational analysis to mine the known information obtained at a deeper level, accurately describing the real objects. In addition, fuzzy theory can determine possible state combinations, which is applicable to the nonlinear problems proposed in this study. The membership function is a key measurement method for fuzzy evaluation and an important component of the grey relational analysis method. Therefore, the study chooses Pearson correlation analysis to analyze the influence of different factors. The specific process of FGRA method is as follows. Firstly, the reference sequence is compared with the comparison sequence, which is a feature sequence that can reflect the results, i.e.

$Y_i = (y_{i1}, y_{i2}, ..., y_{in},)$. The latter is a data sequence that can affect system behavior, set as matrix U, and the expression is shown in equation (7).

$$U_N = \begin{bmatrix} U_{n1}(1) & U_{n1}(2) & \cdots & U_{n1}(a) \\ U_{n2}(1) & U_{n2}(2) & \cdots & U_{n2}(a) \\ \vdots & \vdots & \vdots & \vdots \\ U_{nk}(1) & U_{nk}(2) & \cdots & U_{nk}(a) \end{bmatrix} \quad (7)$$

In equation (7), a represents the number of Y_i. Each Y_i object contains k comparative indicators. Secondly, the data is standardized, and the study uses non-dimensional conversion of data intervals, as calculated in equation (8).

$$U_{ij} = \frac{u(d) - {}^{\min}_{d}u(d)}{{}^{\max}_{d}u(d) - {}^{\min}_{d}u(d)}, d = 1,2,...,a \quad (8)$$

In equation (8), ${}^{\min}_{d}u(d)$ and ${}^{\max}_{d}u(d)$ correspond to the minimum and maximum values of the data in U. Then the Spearman correlation method is used to calculate the fuzzy membership degree r, as shown in equation (9).

$$r = \frac{\sum_{i=1}^{N_t}(u_i - \bar{u})(y_i - \bar{y})}{\left[\sum_{i=1}^{N_t}(u_i - \bar{u})^2 \sum_{i=1}^{N_t}(y_i - \bar{y})^2\right]^{\frac{1}{2}}} \quad (9)$$

In equation (9), u_i and y_i correspond to the level of the observed value i. $\bar{u}$ and $\bar{y}$ are the evaluation levels of u and y. N_t represents the total amount of observations. Then the grey relational coefficient $\chi_{ij}(d)$ is calculated, as shown in equation (10).

$$\chi_{ij}(d) = \frac{\Delta_{\min} + \alpha\Delta_{\max}}{\Delta_{ij}(d) + \alpha\Delta_{\max}} \quad (10)$$

In equation (10), $\Delta_{\max}$ and $\Delta_{\min}$ correspond to the maximum and minimum absolute differences of elements in Y_i and U_i. α represents the resolution coefficient. $\Delta_{ij}(d)$ represents the absolute difference between Y_i and U_i at the U_i-th point. Finally, the weighted grey relational degree $\gamma_{ij}(d)$ is calculated, as expressed in equation (11).

$$\gamma_{ij}(d) = \frac{1}{a}\sum_{d=1}^{a}\omega(d)\zeta_{ij}(d) \quad (11)$$

A comprehensive indicator of FM influencing factors, namely the fuzzy grey relational degree R_{ij}, can be obtained through r and γ_{ij}. The calculation is shown in equation (12).

$$R_{ij} = \frac{r + \gamma_{ij}}{2} \quad (12)$$

Therefore, the construction process of the FGRA multi-factor analysis model can be obtained, as shown in Figure 6.

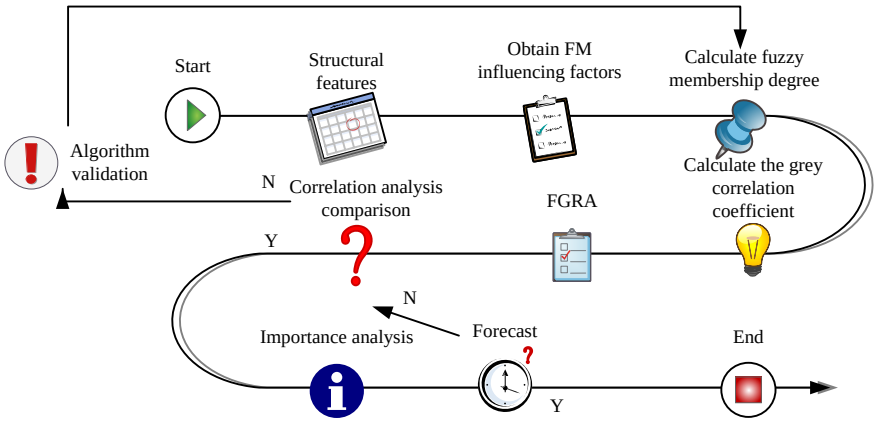


Figure 6: Construction process of FGRA multi-factor analysis model

In Figure 6, the processed FM data is first used to calculate the corresponding fuzzy function and grey relational analysis. Then, the dataset of feature importance is obtained by substituting it into FGRA. Then, through feature importance screening, it is transmitted to the LDO algorithm for prediction comparison. The prediction results are filtered to improve the features. Finally, the best prediction result is obtained based on importance analysis. Based on the above content, the framework of FM historical data analysis and prediction process that integrates LDO algorithm and FGRA can be obtained, as shown in Figure 7.

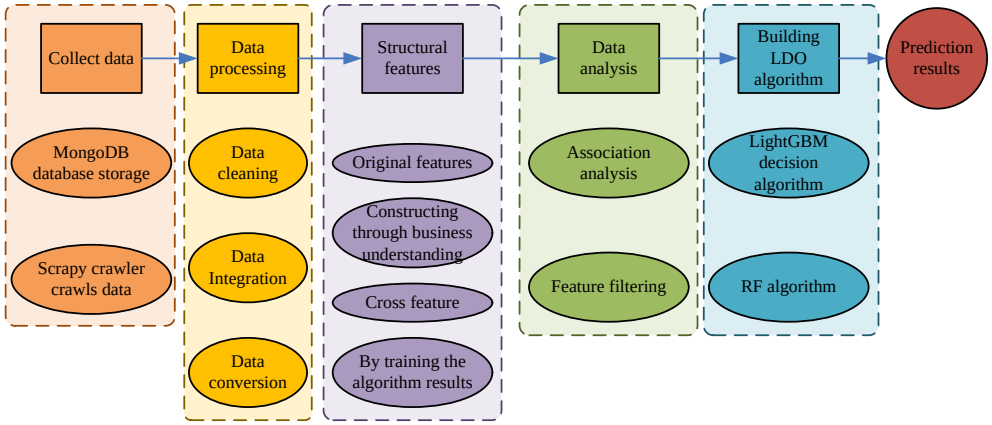


Figure 7: Framework for FM historical data analysis and prediction process integrating LDO algorithm and FGRA

In Figure 7, it is mainly divided into five parts: data acquisition, data processing, structural characteristics, data analysis and model building. Finally, the prediction results can be obtained.

3 Results and discussion

To verify the stability and accuracy of the proposed method, the FM data prediction based on LDO is tested for multiple times. Then the FM results combining LDO and FGRA are analyzed. Finally, the results are comprehensively discussed.

3.1 Analysis of FM data prediction results based on LDO

To verify the performance of LDO, the experimental scenario is a computer with Windows10, 2.60GHz CPU and 8GB RAM. The algorithm is implemented on the software Python. In addition, the commonly used accuracy, F1 value, Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) are selected for evaluation. The obtained data set is divided into training set and verification set according to 9:1. The average value of 10-fold cross validation results is compared. Considering the uniqueness of FM, the competition system and rules of each league are different. The study conducts experiment on 10714 complete game data extracted from the six European leagues from 2018 to 2023. The match results correspond to -1, 0, and 1 for wins, draws, and losses, respectively. The LDO algorithm needs to focus on and adjust the following parameters. To ensure the consistency of the random numbers generated each time, the value range of leaves should be less than 2max_depth. If the total number of leaves exceeds this range, over-fitting will be caused. Therefore, this parameter is set to 25_1 in the model, the depth limit of the tree is -1, and the learning rate of the tree in each iteration is set to 0.05. The L1 regularization parameter and L2 policy parameter are set to 0.25. In order to prevent over-fitting, the minimum sample number of a leaf is set to 5, the number of base learners is set to 2000, the random seed parameter is set to 2020, and the data sampling ratio and feature sampling ratio are set to 0.9 and 0.7, respectively. In addition, to more scientifically evaluate the effectiveness of the research method, the current mainstream algorithms are used for comparative experiments, namely, the Extreme Learning Machine Based on Drosophila Algorithm Optimization (DAO-ELM), the Backpropagation Neural Network and Kriging Algorithm (BPNN-KA) and Multivariate Analysis Algorithm (MAA). The experiment is carried out in the CG and the EG, and the corresponding influencing factors are 41 and 64, respectively. Thus, the prediction accuracy change results of different algorithms can be obtained, as shown in Figure 8.

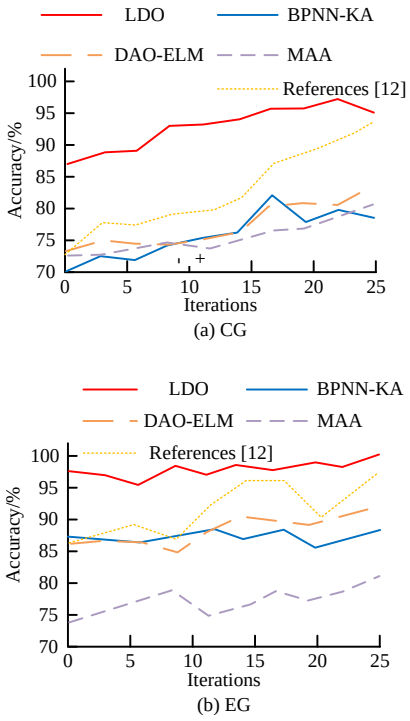
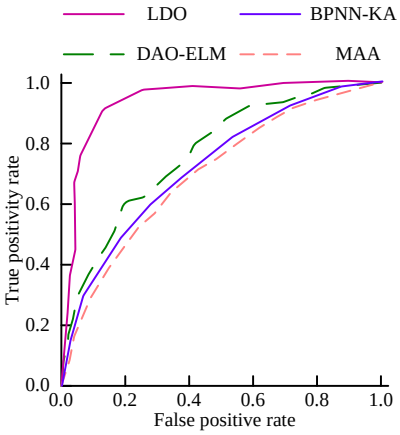
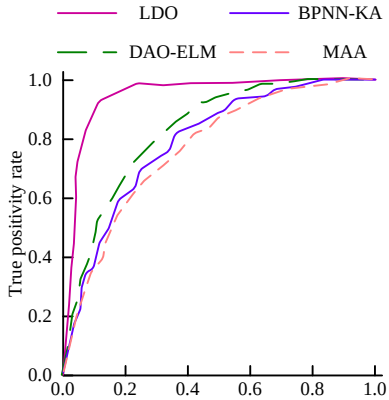


Figure 8: Changes in prediction accuracy of different algorithms

Figure 8 (a) and 8 (b) respectively show the accuracy change curves of different algorithms in the CG and the EG. In the CG and the EG, the accuracy change curves of the LDO showed a stable and smooth trend, and the accuracy was the highest in the whole process, with an average accuracy of 95.31% and 86.74% respectively. The accuracy of MAA in the CG and the EG was the lowest, with the value of 71.36% and 81.32%, respectively, but the corresponding curve change range was small. In the CG, the accuracy of DAO-ELM and BPNN-KA was 72.13% and 73.51%, respectively; In the EG, the accuracy of DAO-ELM and BPNN-KA was 82.16% and 83.35%, respectively. The above results showed that although the feature count of EG reduced by 48.8%, the prediction accuracy of LDO increased by 8.57%. In addition, the accuracy of K-means-LightGBM algorithm in literature [12] was relatively good in the two groups, slightly worse than the performance of the research method. The accuracy results of CG and EG were 86.72% and 92.75%, respectively. It means that FM feature engineering can effectively and deeply mine relevant influencing factors, which confirms the importance of FM for structural features. The ROC results of different algorithms in the two groups can be obtained through multiple experiments on CG and EG, as shown in Figure 9.



(a) CG



(b) EG

Figure 9: ROC results for two groups under different algorithms

Figure 9 (a) and Figure 9 (b) correspond to the ROC changes of different algorithms in CG and EG. In the CG, the AUC values of LDO, DAO-ELM, BPNN-KA and MAA were 0.9124, 0.7796, 0.7662 and 0.7213 respectively. In the EG, the AUC values of LDO, DAO-ELM, BPNN-KA and MAA corresponded to 0.9767, 0.8695, 0.8452 and 0.7963. In addition, the LDO could wrap the other algorithms in both groups, which verified the superiority of the algorithm in FM data association analysis and prediction performance.

3.2 Analysis of football match results based on LDO and FGRA

To explore the feasibility of FM analysis and prediction method combining LDO and FGRA, the results of each actual game are restored in FM prediction as much as possible. The research first calculates the influencing factor R_{ij} of FM results and the weighted grey relational degree α. The results are shown in Table 2.

Table 2: Two tables of correlation results

Number	α	R_{ij}	Number	α	R_{ij}	Number	α	R_{ij}	Number	α	R_{ij}
F1	0.6104	0.4119	F12	0.5878	0.3042	F23	0.6569	0.4022	F34	0.5927	0.2977
F2	0.5674	0.2993	F13	0.5789	0.2904	F24	0.6592	0.3966	F35	0.5382	0.4832
F3	0.5923	0.3021	F14	0.5764	0.2877	F25	0.5937	0.3278	F36	0.5465	0.3391
F4	0.5746	0.2946	F15	0.5548	0.2911	F26	0.6031	0.3483	F37	0.5676	0.4966
F5	0.6016	0.3891	F16	0.5857	0.3026	F27	0.5637	0.3491	F38	0.5344	0.4792
F6	0.5999	0.3849	F17	0.5719	0.4996	F28	0.5856	0.4122	F39	0.5585	0.3452
F7	0.6096	0.3869	F18	0.5848	0.3621	F29	0.5686	0.3491	F40	0.5742	0.4984
F8	0.5563	0.3592	F19	0.6036	0.3847	F30	0.5757	0.3971	F41	0.5396	0.4830
F9	0.5879	0.2981	F20	0.6101	0.4236	F31	0.6016	0.3895	F42	0.5563	0.3442
F10	0.6167	0.4283	F21	0.6059	0.3172	F32	0.6152	0.4531	F43	0.5715	0.4981
F11	0.6282	0.3162	F22	0.6382	0.3672	F33	0.5967	0.3069	/	/	/

From Table 2, taking 0.4 as the grey relational degree standard, the thirteen features F1, F10, F17, F20, F23, F28, F32, F35, F37, F38, F40, F41 and F43 had a high relational degree with the result. From the perspective of feature extraction, there were four strongly correlated feature vectors, namely HAOD, OS, RAHAG and HTTF, which showed that the team's recent game status was very important for future games and had a positive role in promoting the team's adjustment. The OS can extract nine feature vectors, which means that although the initial compensation is a pre-judgment built before the competition, it brings together the power of professional gambling companies and has a certain correlation with FM results. At the same time, it shows that the odds agency has a strong analysis on FM, which can effectively help relevant fans or teams to carry out subsequent game analysis. HTTF can extract F23, which means that the recent abnormal match between the two teams has reference value for the next FM between the two teams. HAOD extracts F28 and F32, which shows that the team's winning rate in the last five games has a great impact on the next FM. It can help the team optimize the tactical analysis in time. The above results show that FGRA can better quantify the data with different characteristics, which is also convenient for subsequent analysis and statistics of influencing factors, laying a solid foundation for the actual result prediction. To further test the application effect of FGRA, regression analysis (method 1), maximum information coefficient (method 2), Pearson correlation coefficient (method 3) and Kendall correlation coefficient (method 4) were used for comparison. The importance comparison results of different correlation analysis methods are obtained, as shown in Figure 10.

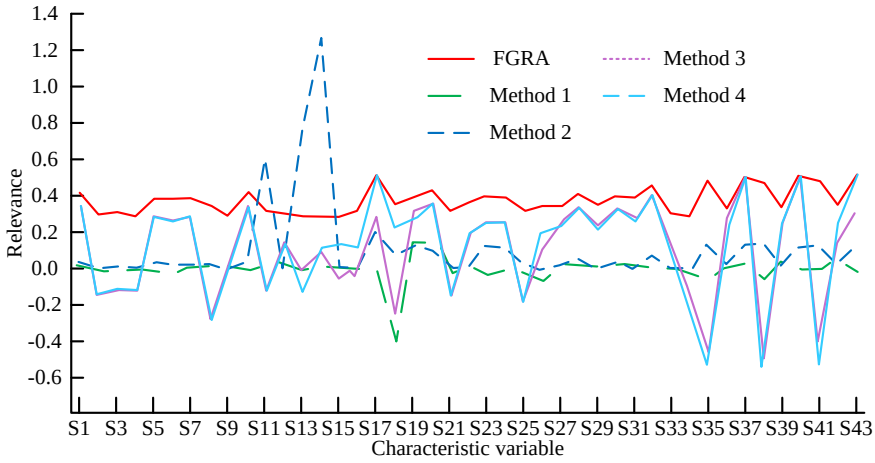


Figure 10: Comparison of importance results of different correlation analysis methods

From Figure 10, in the result of method 3, the correlation coefficient of the three eigenvectors F15, F16 and F26 was less than 0.10, which indicated that the above eigenvectors were statistically uncorrelated. In method 4, F9, F17 and F34 were also not statistically correlated. In method 1, most of the eigenvectors were not statistically correlated. On the whole, different correlation analysis methods had high contribution characteristics in the HAO, OS and DHTTF. This meant that HAO and OS had a correlation effect on the prediction of FM results and provided strong support. From the correlation analysis, the correlation analysis method related to the correlation coefficient couldn't clearly distinguish the contribution of features to FM results. The FGRA method for feature extraction has higher ambiguity, so it has better application effects. Based on the importance results of different correlation analysis methods, the corresponding eigenvalue table is obtained, and the results are shown in Table 3.

Table 3: The result table of eigenvalues extracted by different association analysis methods

Correlation analysis method	Eigenvalue set
Method 1	F7, F8, F11, F12, F18, F19, F20, F22, F26, F35, F37, F38, F39 and F42
Method 2	F11, F13, F14, F17, F18, F19, F20, F32, F35, F37, F38, F40, F41 and F43
Method 3	F1, F5, F13, F14, F20, F23, F26, F28, F32, F35, F37, F38, F40, F41 and F43
Method 4	F1, F8, F10, F13, F14, F17, F20, F28,

FGRA	F30, F37, F38, F40, F41 and F43 F1, F10, F17, F20, F23, F28, F32, F35, F37, F38, F40, F41 and F43

In Table 3, in the two dimensions of PI and rounds, the importance of method 1 was prominent, while the PS dimension showed the importance, which indicated the protection of the team for the ability of athletes in a certain range. The FGRA method proposed in this study played a better role in feature selection. To further analyze the prediction effect of different prediction algorithms combined with various correlation analysis methods, the accuracy comparison experiment was conducted. The results are shown in Figure 11.

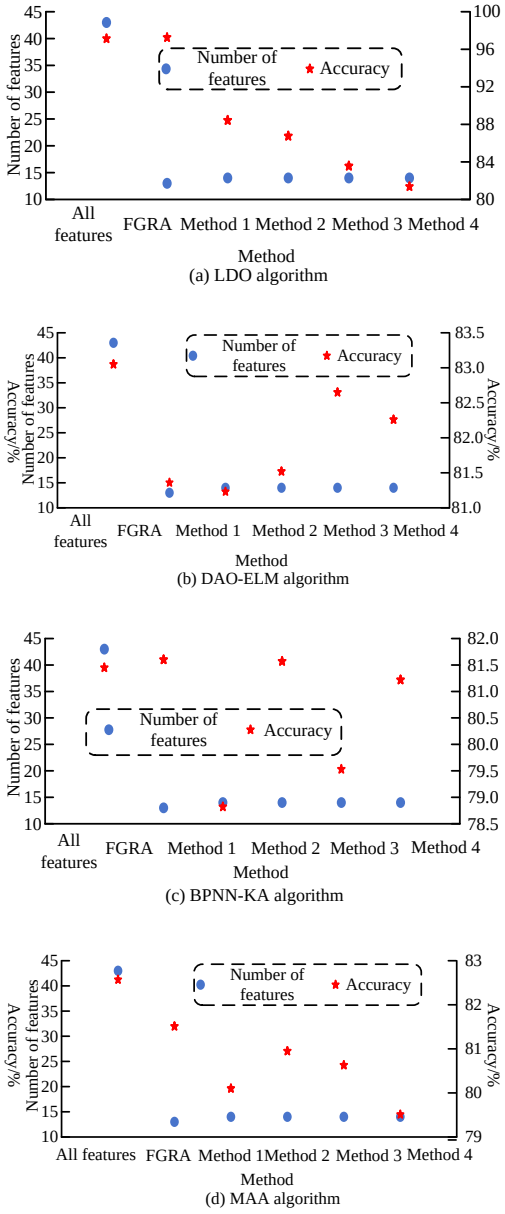
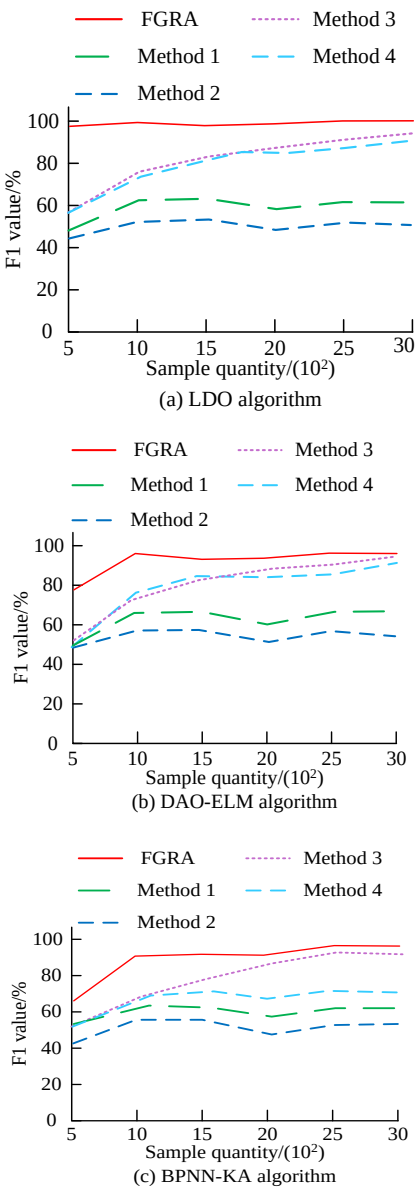
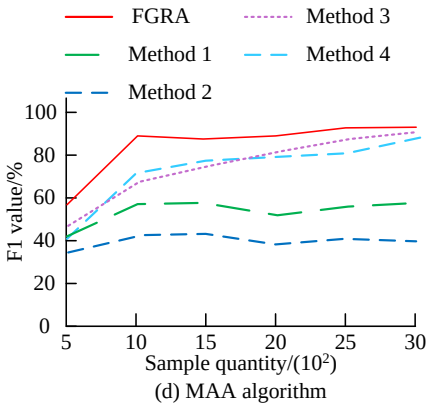


Figure 11: Comparison of prediction accuracy of different prediction algorithms combined with correlation analysis methods

Figure 11 (a)-Figure 11 (d) respectively correspond to the prediction accuracy results of LDO, DAO-ELM, BPNN-KA and MAA with different correlation analysis methods. From Figure 11, the accuracy of FM analysis and prediction method combining LDO and FGRA was the highest, at 97.26%. The accuracy was 0.12% higher than that of the research method, while the accuracy of the method 4 was 15.89% lower than that of the research method. Compared with other algorithms, the research method still showed excellent results. In order to further explore the performance of the research method, the study conducts experiments using FI values. The effects of different prediction algorithms are compared combined with association analysis methods. The results are shown in Figure 12.





(d) MAA algorithm

Figure 12: Comparison of F1 values of different prediction algorithms combined with correlation analysis methods

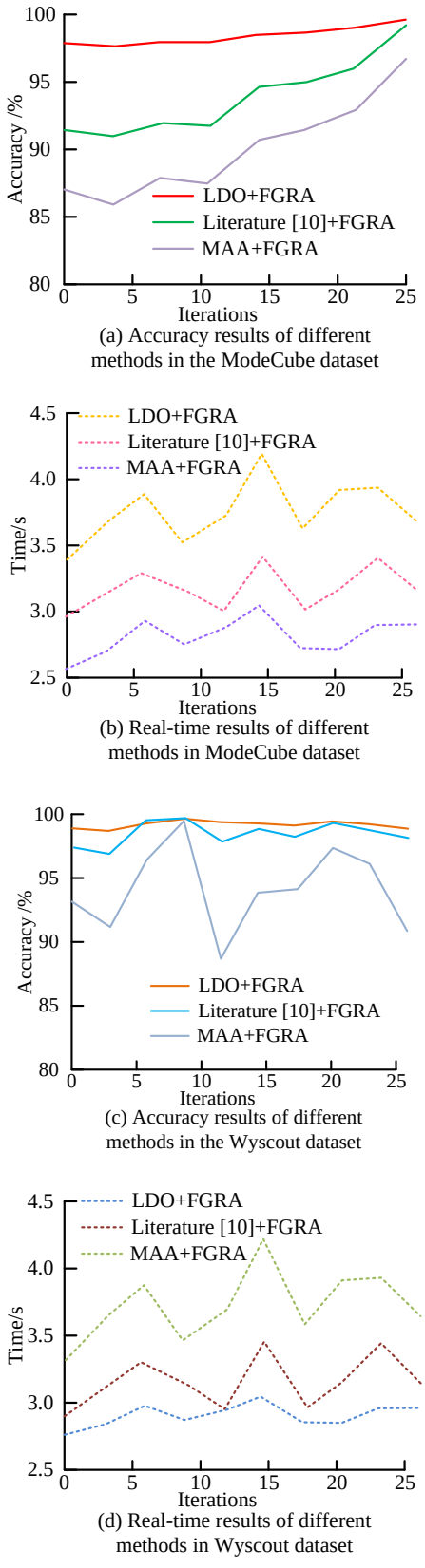
Figure 12 (a)-Figure 12 (d) respectively show the F1 value results of LDO, DAO-ELM, BPNN-KA and MAA with different correlation analysis methods. In Figure 12, the FM analysis and prediction method combining LDO and FGRA still maintained a stable and high F1 value throughout the whole process with the increase of the number of samples, and the average F1 value was 93.71%. The F1 value curves of other algorithms combined with the correlation analysis method showed large fluctuations. When the number of samples exceeded 2000, the F1 value fluctuated around 60%. The combination of MAA and method 2 had the worst effect, with an average F1 value of 41.26%. Finally, the ROC results of different prediction algorithms combined with correlation analysis methods are compared and analyzed, as shown in Table 4.

Table 4: Comparison of ROC results using different prediction algorithms combined with correlation analysis methods

Prediction algorithm	Correlation analysis method	Number of features	AUC
LDO	Method 1	14	0.8068
	Method 2	14	0.8097
	Method 3	14	0.8036
	Method 4	14	0.8025
	FGRA	13	0.9885
	All features	41	0.9879
DAO-ELM	Method 1	14	0.7758
	Method 2	14	0.7902
	Method 3	14	0.6805
	Method 4	14	0.7864
	FGRA	13	0.8452
	All features	41	0.8563
BPNN-KA	Method 1	14	0.7536
	Method 2	14	0.7816
	Method 3	14	0.6726
	Method 4	14	0.7753
	FGRA	13	0.8406
	All features	41	0.8477
MAA	Method 1	14	0.7928
	Method 2	14	0.7732
	Method 3	14	0.7945
	Method 4	14	0.7614
	FGRA	13	0.8389
	All features	41	0.8421

In Table 4, the AUC value of FM analysis and prediction method integrating LDO and FGRA was 0.9881, which was 0.06% higher than that of all features. This shows that this method has a good correlation analysis effect on FM with many influencing factors. It can effectively and deeply excavate the valuable features in the actual game, with excellent effects on FM prediction. The combination of each algorithm with FGRA method had higher AUC value. The AUC values combined with DAO-ELM, BPNN-KA and MAA were 0.8452, 0.8406 and 0.8389, respectively. In order to verify the robustness of the research method, the research method is evaluated from two perspectives of accuracy and real-time performance. The public ModeCube dataset (http://modelcube.cn/) and Wyscout dataset (https://figshare.com/collections/Soccer_match_event_dataset/4415000/ 5) are used to verify. The former contains about 25,000 matches in 11 European countries and their top leagues from 2008-2016, as well as detailed events of more than 10,000 matches. The latter covers 1,941 games and about 3 million events. The robustness results of different methods under ModeCube and Wyscout datasets

can be obtained, as shown in Figure 13.



(a) Accuracy results of different methods in the ModeCube dataset

(b) Real-time results of different methods in ModeCube dataset

(c) Accuracy results of different methods in the Wyscout dataset

(d) Real-time results of different methods in Wyscout dataset

Figure 13: Robustness results of different methods on ModeCube dataset and Wyscout dataset

Figure 13 (a) and Figure 13 (b) respectively show the accuracy and real-time results of different methods in the ModeCube dataset. Figure 13 (c) and Figure 13 (d) respectively correspond to the accuracy and real-time results of different methods in the Wyscout dataset. From Figure 13, in the two datasets, the proposed FM match analysis and prediction method combining LDO algorithm and FGRA was the most robust. The accuracy and real-time results of ModeCube dataset were 99.13% and 2.93s, respectively. The accuracy and real-time results of Wyscout dataset were 99.56% and 3.16s, respectively. The prediction method based on particle swarm optimization algorithm in literature [10] had the second-best performance. The accuracy and real-time results of ModeCube dataset were 98.25% and 3.05s, and the accuracy and real-time results of Wyscout dataset were 98.11% and 3.36%, respectively. The above results show that the proposed method has good robustness. The accuracy, F1 value, and AUC index all have good results, and the robustness of the proposed method is significantly improved compared with the existing method.

3.3 Discussion

The digital development of football in China is currently slow due to the high difficulty in collecting and using FM data. The reference indicators also need to be based on the results of previous matches. Therefore, conducting FM analysis and prediction is extremely challenging. The study first focused on the characteristics of FM data, selected Python and asynchronous Scrapy framework to build a data capture project, and saved it through a non-relational database. Secondly, the study used RF to optimize the LightGBM decision algorithm and obtained the LDO. Then the FGRA method was designed. Finally, the LDO algorithm was combined to establish FM analysis and prediction method. The innovation of the research is mainly in the following two aspects. Firstly, the optimized LightGBM decision algorithm based on RF algorithm, namely LDO algorithm, is proposed to solve the problem that a single LightGBM decision algorithm has good experimental effect only in certain types of data, but may have poor experimental effect in other models. Secondly, a new analysis method combining fuzzy theory and grey relational analysis is designed and applied to FM result analysis. In addition, the research method introduces feature engineering to construct accurately predicted feature indexes from multiple angles and aspects. The FGRA method is used to extract high-contribution features, so as to show the importance of screening features from the actual results.

The following results can be obtained. Firstly, the prediction accuracy and ROC experiments of different prediction algorithms were carried out. In the EG and the CG, the accuracy curve of the LDO algorithm shows a stable and smooth trend, with an average accuracy and AUC value of 95.31% and 86.74%, respectively. In

addition, the accuracy of the K-means-LightGBM algorithm in literature [12] in the two groups was only slightly worse than the performance of the research method, and the accuracy results in the CG and EG were 86.72% and 92.75%, respectively. The above results may be due to the fact that the research method uses RF algorithm for optimization to cope with massive and complex FM data, which can significantly improve the prediction performance of a single Light algorithm. Secondly, experiments are conducted on different prediction algorithms combined with the accuracy, F1 value and AUC of various correlation analysis methods. It was found that the FM analysis and prediction method combining LDO algorithm and FGRA had the highest accuracy of 97.26%, and the accuracy of this method was 0.12% higher than that of all characteristic influencing factors. The F1 value change curves of other algorithms combined with the correlation analysis method showed large fluctuations. When the number of samples exceeded 2000, the F1 value fluctuated around 60%. Moreover, all the algorithms combined with FGRA obtained higher AUC values. The AUC values combined with DAO-ELM, BPNN-KA and MAA were 0.8452, 0.8406 and 0.8389, respectively. Thirdly, robustness experiments were conducted on different methods in different datasets. The results showed that the robustness of the research method was the best. The accuracy and real-time results of the ModeCube dataset were 99.13% and 2.93s, respectively. The accuracy and real-time results of the Wyscout dataset were 99.56% and 3.16s, respectively. However, the performance of the method in reference [10] was second. The accuracy and real-time results of ModeCube dataset were 98.25% and 3.05s, respectively, and the accuracy and real-time results of Wyscout dataset were 98.11% and 3.36%, respectively. The above results indicate that the robustness of the research method is significantly improved. It may be due to the fact that the research method, combining LightGBM algorithm and RF algorithm, can solve the adverse impact of a single model in a certain aspect on the final result, improve the impact of better performance methods on the final result, and achieve the balance between over-fitting and under-fitting.

Based on the above content, it can be concluded that the method proposed in the study can effectively solve the ambiguity and complexity of factors affecting FM data that traditional analysis methods cannot effectively analyze. It can also quantitatively and deeply mine competition data, providing a solid and powerful foundation for FM prediction. The contribution of the research is to solve the problem that the massive data and complex multi-level data in FM cannot be comprehensively analyzed, provide technical support for the iteration of the sports industry, promote the innovation and development of the sports industry, and accelerate the process of science and technology of the sports industry.

4 Conclusion

In response to the large volume, multiple levels, and high complexity of FM data, as well as the diversity and ambiguity among different factors that affect the results, the required dataset was first obtained through crawler algorithms. The collected data was processed using methods such as data clarity and data integration. Next, the feature construction was carried out, followed by proposing the LDO algorithm for prediction. Finally, the FGRA method was designed and combined with the LDO algorithm to obtain FM analysis and prediction methods. From the experimental results, in the EG and CG, the feature amount decreased by 48.8%, but the LDO had the highest accuracy and AUC value, with an average accuracy of 95.31% and 86.74%, respectively. The AUC values were 0.9124 and 0.9767, respectively. The accuracy and F1 value of the FM analysis and prediction method that integrated LDO and FGRA were the highest, with 97.26% and 93.71%, respectively. Compared with all influencing factors with the same features, the accuracy of this method was 0.12% higher, while the accuracy of method 4 was 15.89% lower than that of the research method. The AUC value of the FM analysis and prediction method that integrated LDO and FGRA was 9885, which was 0.06% higher than the AUC of all features. In summary, the method proposed in the study can effectively deal with complex FM data, while conducting analysis and prediction, providing reference for the analysis and prediction of related competitions. However, there are still shortcomings in the research due to the uniqueness of FM. The extraction of relevant influencing factors is not yet complete. In actual competitions, there are also non textual data influencing factors, such as weather conditions, venue environment, and athlete injuries. Therefore, in future research, these factors that affect FM can be further improved.

References

[1] F. Kong, and S. Ren, “Analysis on the influence of nanotechnology development on sports health industry,” International Journal of Nanotechnology, vol. 19, no. 6/11, pp. 1034-1044, 2022. https://doi.org/10.1504/ijnt.2022.10054009

[2] M. Li, H. Dong, F. Zhang, and X, Liu, “A method for top view pedestrian flow detection based on small target tracking,” Informatica, vol. 48, no. 11, pp. 59-70, 2024. https://doi.org/10.31449/inf.v48i11.6033

[3] M. Serazio, “The irreverent life and uncompromising death of Deadspin: Sports blogging as punk journalism,” Journalism, vol. 23, no. 2, pp. 461-478, 2022. https://doi.org/10.1177/1464884920987690

[4] M. B. ermansen, A. V. Christiansen, and S. Frische, “SARS-CoV-2 prevalence and transmission in swimming activities: Results from a retrospective cohort study,” Scandinavian Journal of Medicine &

Science in Sports, vol. 32, no. 1, pp. 242-254, 2022. https://doi.org/10.1111/sms.14071

[5] B. A. A. White, and J. Quinn, "Personal growth and emotional intelligence: foundational skills for the leader," Clinics in Sports Medicine, vol. 42, no. 2, pp. 261-267, 2023. https://doi.org/10.1016/j.csm.2022.11.008

[6] N. Liu, and P. Liu, "Goaling recognition based on intelligent analysis of real-time basketball image of Internet of Things," The Journal of Supercomputing, vol. 78, no. 1, pp. 123-143, 2022. https://doi.org/10.1007/s11227-021-03877-3

[7] M. Ahin, and M. Uar, "Prediction of sports attendance: A comparative analysis," Journal of Sports Engineering and Technology, vol. 236, no. 2, pp. 106-123, 2022. https://doi.org/10.1177/1754337120983135

[8] S. Barra, S. M. Carta, A. Giuliani, A. Pisu, A. S. Podda, and D. Riboni, "FootApp: An AI-powered system for football match annotation," Multimedia Tools and Applications, vol. 82, no. 4, pp. 5547-5567, 2023. https://doi.org/10.1007/s11042-022-13359-0

[9] Z. Chen, "Prediction method of best running track of long distance shooting based on sliding window," International Journal of Reasoning-based Intelligent Systems, vol. 14, no. 2/3, pp. 84-90, 2022. https://doi.org/10.1504/ijris.2022.10048822

[10] Y. Huang, and Y. Bai, "Intelligent sports prediction analysis system based on edge computing of particle swarm optimization algorithm," IEEE Consumer Electronics Magazine, vol. 12, no. 2, pp. 73-82, 2023. https://doi.org/10.1109/MCE.2021.3139837

[11] R. Hu, G. K. Nicholls, and D. Sejdinovic, "Large scale tensor regression using kernels and variational inference," Machine Learning, vol. 111, no. 7, pp. 2663-2713, 2022. https://doi.org/10.1007/s10994-021-06067-7

[12] Z. Tang, Y. Xiao, Y. Jiao, X. Li, C. Zhang, J. Sun, and P. Wang. "Research on short-term low-voltage distribution network line loss prediction based on Kmeans-LightGBM," Journal of Circuits, Systems and Computers, vol. 31, no. 13, pp. 135-146, 2022. https://doi.org/10.1142/S0218126622502280

[13] U. Masood, H. Farooq, A. Imran, and A. Abu-Dayya, "Interpretable AI-based large-scale 3D pathloss prediction model for enabling emerging self-driving networks," IEEE Transactions on Mobile Computing, vol. 22, no. 7, pp. 3967-3984, 2023. https://doi.org/10.1109/TMC.2022.3147191

[14] M. Hasanvand, M. Nooshyar, E. Moharamkhani, and A. Selyari, "Machine learning methodology for identifying vehicles using image processing," Artificial Intelligence and Applications, vol. 1, no. 3, pp. 170-178, 2023. https://doi.org/10.47852/bonviewAIA3202833

[15] G. Mehdi, H. Hooman, Y. Liu, S. Peyman, and R. Arif, "Data mining techniques for web mining: A survey," Artificial Intelligence and Applications, vol. 1, no. 1, pp. 3-10, 2022. https://doi.org/10.47852/bonviewAIA2202290

[16] D. Charles, "The lead-lag relationship between international food prices, freight rates, and Trinidad and Tobago's food inflation: A support vector regression analysis," Green and Low-Carbon Economy, vol. 1, no. 2, pp. 94-103, 2023. https://doi.org/10.47852/bonviewGLCE3202797

[17] V. Leonidaki, and M. P. Constantinou, "A comparison of completion and recovery rates between first-line protocol-based cognitive behavioural therapy and non-manualized relational therapies within a UK psychological service," Clinical Psychology & Psychotherapy, vol. 29, no. 2, pp. 754-766, 2022. https://doi.org/10.1002/cpp.2669

[18] L. Ma, and Z. Zhang, "The contribution of databases towards understanding the universe of long non-coding RNAs," Nature Reviews. Molecular Cell Biology, vol. 27, no. 9, pp. 601-602, 2023. https://doi.org/10.1038/s41580-023-00612-z

[19] S. A. Obead, H. Y. Lin, E. Rosnes, and J. Kliewer, "Private linear computation for noncolluding coded databases," IEEE Journal on Selected Areas in Communications, vol. 40, no. 3, pp. 847-861, 2022. https://doi.org/10.1109/JSAC.2022.3142362

[20] H. Xia, Z. Qiu, and Y. Liu, "A stress-influence-function with adaptive strength feature approach for stress constrained continuum topology optimization via small-loop sequential strategy," International Journal for Numerical Methods in Engineering, vol. 123, no. 1, pp. 41-68, 2022. https://doi.org/10.1002/nme.6840

[21] T. Zhang, J. Chen, S. He, and Z. Zhou, "Prior knowledge-augmented self-supervised feature learning for few-shot intelligent fault diagnosis of machines," IEEE Transactions on Industrial Electronics, vol. 69, no. 10, pp. 10573-10584, 2022. https://doi.org/10.1109/TIE.2022.3140403

[22] H. Alqahtani, and A. Ray, "Feature selection of surface topography parameters for fatigue-damage detection using pearson correlation method and neural network analysis," Fatigue & Fracture of Engineering Materials & Structures, vol. 46, no. 5, pp. 1810-1820, 2023. https://doi.org/10.1111/ffe.13963

[23] Imai N, Takeyoshi M, Aizawa S, M. Tsurumaki, M. Kurosawa, A. Toyoda, M. Sugiyama, K. Kasahara, S. Ogata, T. Omori, and M. Hirota, "Improved performance of the SH test as an in vitro skin sensitization test with a new predictive model and decision tree," Journal of Applied Toxicology, vol. 69, no. 5/6, pp. 1029-1043, 2022. https://doi.org/10.1002/jat.4275

[24] O. Mehrpour, F. Saeedi, and C. Hoyte, "Decision tree outcome prediction of acute acetaminophen exposure in the United States: A study of 30,000 cases from the national poison data system," Basic

& Clinical Pharmacology & Toxicology, vol. 130, no. 1, pp. 191-199, 2022. https://doi.org/10.1111/bcpt.13674

[25] H. Xu, X. Rui, Z. Wang, Z. Qiu, L. Cai, Z. Zhang, and M. Zhao, "The effects of physical properties on the compression modulus of coastal fully-weathered red sandstone: Consolidation compressibility and the settlement calculation of foundation," Journal of Coastal Research, vol. 38, no. 4, pp. 870-884, 2022. https://doi.org/10.2112/JCOASTRES-D-21-00112.1

[26] W. Zhang, Y. He, L. Wang, S. Liu, and X. Meng, "Landslide susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing," Geological Journal, vol. 58, no. 6, pp. 2372-2387, 2023. https://doi.org/10.1002/gj.4683

[27] R. S. W. Chang, P. R. Tacon, M. Abiragi, L. Cao, G. Hong, J. Le, P. Ricchiuto, C. Daluwatte, D. Ouyang, and I. M. Chiu, "Random forest machine learning to detect cardiac amyloidosis," Journal of the American College of Cardiology, vol. 83, no. 13, pp. 311-311, 2024. https://doi.org/10.1016/S0735-1097(24)02301-5

[28] Q. Li, J. Han, W. Wang, W. Cui, D. L. Federico, X. Yang, Y. Zhou, and R. Shi, "What to expect from dynamical modelling of cluster haloes-II. Investigating dynamical state indicators with Random forest," Monthly Notices of the Royal Astronomical Society, vol. 514, no. 4, pp. 5890-5904, 2022. https://doi.org/10.1093/mnras/stac1739

[29] Y. Liu, and B. Pan, "Profit estimation model and financial risk prediction combining multi-scale convolutional feature extractor and BGRU model," Informatica, vol. 48, no. 11, pp. 15-32, March, https://doi.org/10.31449/inf.v48i11.5941

[30] C. E. Searle, A. Kaszta, D. T. Bauer, K. Kesch, J. E. Hunt, R. Mandisodza-Chikerema, M. V. Flyman, D. W. Macdonald, A. J. Dickman, and A. J. Loveridge, "Random forest modelling of multi-scale, multi-species habitat associations within KAZA transfrontier conservation area using spoor data," Journal of Applied Ecology, vol. 59, no. 9, pp. 2346-2359, 2022. https://doi.org/10.1111/1365-2664.14234