Virtual Reality & Intelligent Hardware

• Article •

October 2022 Vol. 4 No. 5: 444–458
10.1016/j.vrih.2022.08.008

NPIPVis: A visualization system involving NBA visual analysis and integrated learning model prediction

Zhuo SHI[1,3,4], Mingrui LI[2,3*], Meng WANG[2,3], Jing SHEN[1], Wei CHEN[4], Xiaonan LUO[2,3]

1. *Department of Art and Design, University of Guilin Electronic Technology, Guilin* 541004*, China;*
2. *Department of Computer and Information Security, University of Guilin Electronic Technology, Guilin* 541004*, China;*
3. *Department of Guangxi Key Laboratory of Image and Graphic Intelligent Processing, University of Guilin Electronic Technology, Guilin* 541004*, China;*
4. *Department of State Key Laboratory of CAD&CG, University of Zhejiang, Hangzhou* 310058*, China*

Received 12 July 2022; **Revised** 25 June 2022; **Accepted** 28 August 2022

Abstract: Background Data-driven event analysis has gradually become the backbone of modern competitive sports analysis. Competitive sports data analysis tasks increasingly use computer vision and machine-learning models for intelligent data analysis. Existing sports visualization systems focus on the player–team data visualization, which is not intuitive enough for team season win–loss data and game time-series data visualization and neglects the prediction of all-star players. **Methods** This study used an interactive visualization system designed with parallel aggregated ordered hypergraph dynamic hypergraphs, Calliope visualization data story technology, and iStoryline narrative visualization technology to visualize the regular statistics and game time data of players and teams. NPIPVis includes dynamic hypergraphs of a team′s wins and losses and game plot narrative visualization components. In addition, an integrated learning-based all-star player prediction model, SRR-voting, which starts from the existing minority and majority samples, was proposed using the synthetic minority oversampling technique and RandomUnderSampler methods to generate and eliminate samples of a certain size to balance the number of all-star and average players in the datasets. Next, a random forest algorithm was introduced to extract and construct the features of players and combined with the voting integrated model to predict the all-star players, using GridSearchCV, to optimize the hyperparameters of each model in integrated learning and then combined with five-fold cross-validation to improve the generalization ability of the model. Finally, the SHapley Additive exPlanations (SHAP) model was introduced to enhance the interpretability of the model. **Results** The experimental results of comparing the SRR-voting model with six common models show that the accuracy, F1-score, and recall metrics are significantly improved, which verifies the effectiveness and practicality of the SRR-voting model. **Conclusions** This study combines data visualization and machine learning to design a National Basketball Association data visualization system to help the general audience visualize game data and predict all-star players; this can also be extended to other sports events or related fields.

Keywords: Sports visualization; Parallel aggregated ordered hypergraph; Calliope; IStoryline; Integrated learning; SHAP model

Supported by the National Natural Science Foundation of China(61862018); and the Subject of the Training Plan for Thousands of Young and Middle-aged Backbone Teachers in Guangxi Colleges and Universities(2020QGRW017).

Citation: Zhuo SHI, Mingrui LI, Meng WANG, Jing SHEN, Wei CHEN, Xiaonan LUO. NPIPVis: A visualization system involving NBA visual analysis and integrated learning model prediction. Virtual Reality & Intelligent Hardware, 2022, 4(5): 444–458

***Corresponding author**, 19032303028@mails.guet.edu

1 Introduction

The National Basketball Association (NBA) is considered the world's premier professional basketball league for men, and it records massive amounts of game data at an unprecedented rate. However, it is difficult for the average viewer to understand and analyze the NBA′s high-dimensional professional data, and data visualization[1] and machine-learning techniques[2] are difficult for the average viewer. Data visualization and machine-learning techniques can be combined to extract key information from these data, present complex data in the form of visual charts, improve the readability and operability of the data, facilitate users to observe and understand the meaning and pattern of the data, and use the data to predict the outcomes of NBA games. This paper details a study on the regular statistics and game time-series data of NBA players and teams, combined with data visualization and machine learning techniques, to explore and predict.

The NPIPVis system first uses parallel aggregate ordered hypergraph visualization (PAOHvis)[3] to create a super dynamic chart of NBA season wins and losses and then uses iStoryline[4]. Moreover, iStoryline was used to customize the NBA game plot visualization component, with many chart styles and user-friendly interaction methods. In addition, an integrated learning algorithm, SRR-voting, was proposed to predict NBA all-star players, and Calliope[5] was used to draw NBA visualization data stories. NPIPVis visualizes and interprets important NBA data, which can help general users better understand and analyze NBA teams and games, as well as understand and predict all-star players.

1.1 Sports visualization

Buono et al. analyzed soccer games with specific visualizations, highlighting players′ goal contributions and helping users understand players, team performance, and the evolution of causes of goals and passes[6]. Wu et al. designed a spatiotemporal visualization to represent team changes[7]. ForVizor, a visual analysis system, was designed and developed to enable users to track the spatiotemporal changes in formations and understand the changes and causes. Sheng et al. developed a vision-based evaluation system called GreenSea and proposed a novel recursive discriminative BLS (RDBLS) for long-term tracking[8]. Meng et al. utilized video tracking to capture the physical and tactical information of football players and proposed a football recommendation system by combining players' tracking techniques with recommendation algorithms[9]. Zhang et al. proposed an effective method to improve spatiotemporal context learning and increase accuracy by combining information from multiple views[10].

For hockey, Carsting et al. created a software system to visualize and filter game data to help the Linköping Hockey Club evaluate goalie performances[11].

Chu et al. proposed an immersive visual analysis system, TIVEE, which helps users explore and explain badminton tactics from multiple levels[12].

Wu et al. proposed iTTVis, a visual analytics system for analyzing and exploring table tennis data[13]. It provides a holistic game visualization from three perspectives: time-oriented, statistical, and tactical. Moreover, Lan et al. presented SimuExplorer, a visualization system that helps analysts explore how player behavior affects table tennis scoring rates and develop training programs that can help players improve[14].

For basketball, Goldsberry proposed a new method for quantifying the range of NBA players' shots and shooting range using spatial and visual analyses to enhance basketball expertise[15]. Lei et al. divided sports data into statistical and multidimensional data, summarized the existing work on sports data visualization, oriented fields, and common methods, and explained the basic ideas of the visual analysis of sports data[16]. In addition, Chen et al. proposed three levels of detail for NBA game visualization (i.e., season, game, and session levels) and designed and implemented a real-time system[17]. Losada et al. proposed a visual analytics system, BKViz, for basketball game analysis, focusing on individual games to better understand the game[18].

Hung et al. proposed BasketView, a web-based visualization system that enables basketball professionals to analyze games from three perspectives (season, game, and tournament levels)[19]. Perin et al. analyzed the current process of sports data visualization from the perspective of box-score data, tracking data, and metadata[20]. Wang et al. combined the network topology characteristics, established bypass path, pass quality, and technical statistics[21]. Thus, a novel visual evaluation model was proposed, and a visual analysis system named PlayerNetVis was developed. Ji et al. proposed a data visualization system based on the game data of NBA games from 1985 to 2019 for NBA game reports with knowledge graphs to help users quickly understand games[22]. Jin et al. proposed a player visualization system based on charts to analyze basketball players from multiple perspectives[23].

The existing sports visualization studies mentioned above, especially for NBA visualization, are biased toward expert analysis systems (e.g., Losada et al.[18] and Wang et al.[21]), and the data analyzed are often conventional player and team statistics (e.g., Jin et al.[23]). The visualization of NBA team win–loss and time-series data for the general audience deserves further study.

To address the above issues, the proposed NPIPVis system uses PAOHvis to plot a season′s win–loss hypergraphs based on an NBA team′s regular statistics from 1980 to 2022. The system uses iStoryline to process NBA game time-series data and design the game plot visualization component. In addition, NPIPVis provide rich visual charts for the general audience to better understand NBA teams and games.

1.2 Sports forecast

For tennis, Wilkens et al. applied machine-learning techniques to male and female professional singles tennis matches, analyzed data from nearly 39000 tennis matches from 2010 to 2019 and proposed a tennis match prediction and betting model to predict tennis match results and betting odds[24].

For rugby, Xia et al. proposed a network-driven approach to sports ranking and prediction, defining directed graphs of teams based on the results of the National Football League and NBA games and inferring the importance of the teams that determine their rankings using these networks[25]. South et al. used modern modeling techniques, such as neural networks and random forests, to predict the outcomes of college football games[26].

For basketball, Lam et al. proposed a pioneering modeling approach based on stacked Bayesian regression, relying on the high standard performance of sparse multiscale Gaussian process regression to predict the regular season game direction based on the data of NBA teams' assists and rebounds, linking the players' performance to the strength of the team[27]. Pai et al. proposed a hybrid model hysteresis space vector modulation decision tree (HSVMDT) based on a support vector machine (SVM) and a decision tree (DT), which provides rules to help coaches develop strategies[28]. With the help of the predicted game results and the rules generated by the HSVMDT model, coaches can easily and quickly understand the basic factors that increase the chances of winning a game. Giuliodori proposed an artificial neural network (ANN)-based model for predicting underdogs in NBA games, which trains an ANN to predict the winning team in NBA games based on information such as the number of games won and lost and the number of steals by the team[29]. Özbalta et al. predicted the salaries of new NBA contract players using a supervised machine-learning approach based on the dataset of the NBA 2K20 My Team game and the performance statistics of NBA players for the 2019-2020 season[30]. Yang et al. proposed an improved PageRank algorithm-based NBA playoff ranking prediction model to rank team strengths from the overall perspective of team game data and to predict playoff outcomes[31]. Huang et al. used a one-dimensional convolutional neural network and an SVM to predict the outcome of major league baseball games in the USA, providing some reference information for fans and team managers[32]. Kaimakamis used machine-learning algorithms such as random forest and SVM to predict the final position of NBA teams in playoffs[33]. Albert et al. proposed a hybrid machine-learning model for

predicting NBA all-star players, combining random forest, AdaBoost, and multilayer perceptron models to improve the accuracy of predicting NBA all-star players[34].

The above sports prediction studies, for example, Wilkens[24] and Giuliodori[29], often use a single machine-learning model to analyze a small amount of low-dimensional game data, with poor generalization ability, low accuracy, high memory overhead, and slow speed. In contrast, the integrated learning LightGBM[35] has the advantages of fast training, high efficiency, and good fitting ability.

To address the above problems, this paper proposes an integrated learning-based all-star player prediction model, SRR-voting, which facilitates the general audience to understand the NBA players. We used the Calliope narrative visualization technique to display the all-star player prediction results and automatically generate the visual data stories of the all-star players.

2 NBA data description

NPIPVis uses the NBA regular statistics and game time-series data provided by Basketball-Reference, a leading data site in the USA, to design a visualization view of NBA regular statistics and game time-series data.

2.1 NBA regular statistics

NPIPVis uses a crawler technique to obtain NBA regular season players and team statistics from 1980 to 2022 from the Basketball-Reference website. The NBA regular season team statistics include average steals, average turnovers, and average caps. In addition, 27 field attributes were observed, of which the playoff was the tag field. If the team entered the playoff, the playoff value was 1, and vice versa, then, 1165 data points were noted. The NBA regular season player statistics include 31 field attributes, for example, age and three-point shooting percentage, of which class is the label field. If the player was selected as an all-star, the class value was one, and vice versa, then 21161 data points were gathered.

2.2 NBA game timing data

The NBA game chronological data recorded the development of the game. Specific field attributes of NPIPVis game chronological data include the time remaining in each quarter, team A event details, team B event details, players, events, and scores(Table 1).

Table 1 Timing data field properties

Time-series data fields	Attributes
Time	12:00.0
Team B Event Details	Defensive rebound by C. Kispert
Team A Event Details	T. Young misses 2-pt jump shot from 20 ft
Score	6-0

3 Overall design of the NPIPVis visualization system

3.1 System overview

NPIPVis consists of three modules (data processing, visualization, and prediction). Figure 1 shows that the regular statistics and game time-series data are preprocessed. These data were information sources for model building, visualization, and prediction. An integrated learning-based all-star player prediction model is built and trained, and the NBA team season win–loss dynamic hypergraphs and NBA game plot narrative visualization components are then drawn. Finally, all-star player prediction results are automatically generated for visual data stories.

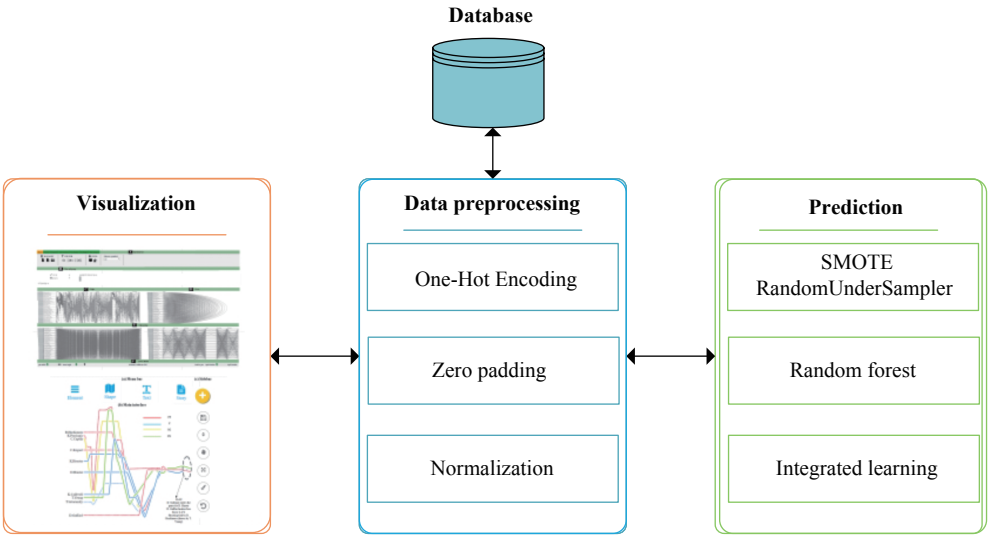

Figure 1 NPIPVis visualization system design framework.

3.2 Software and hardware environment

The hardware configuration includes a CPU model (Intel Core i7-7700) and graphics memory capacity (4GB). The software used include Windows 10 OS and PyCharm development tool. The development languages used were Python 3.7, HTML, CSS, and JavaScript, and the main libraries included Scikit-learn, Pandas, Numpy, React, PAOHvis, iStoryline, and Calliope.

4 NBA regular statistics visualization design

PAOHvis visualizes dynamic hypergraphs. A hypergraph is a generalization of a graph in which edges can connect more than two vertices. PAOH represents vertices as parallel horizontal bars and hyperedges as vertical lines connecting two or more vertices. PAOH is the first dynamic hypergraph representation technique with high readability and no overlap and is very suitable for medium-sized dynamic hypergraph networks.

In this study, PAOHvis created a dynamic hypergraph depicting the wins and losses between teams during the 2021-2022 season. The dynamic hypergraph component is divided into seven parts: A (operation bar), B (color and group), C (origin), D (curve), E (edge patting), and F (lower sidebar). The operation bar allows the uploading of datasets and their search. Users can click color by group by sorting team data by color and by the team′s east and west divisions. The main interface uses the month of the season and the team name as the horizontal and vertical coordinates, respectively. By default, the teams are sorted by the number of games they played in the season. In each column, all games played by the corresponding team are displayed in ascending chronological order. This section describes the wins and losses for each team throughout the season. The games played by each team throughout the season were coded in ascending order from left to right. By hovering over a team such as Utah Jazz, all contract lines and teams associated with Utah Jazz will be highlighted. Hovering over team-to-team contracts, such as Utah Jazz and Miami Heat, will show all Jazz and Heat win–loss relationships as well as the home and away games for each game, with an asterisk indicating a win and a circle indicating a loss. If the space between two adjacent sessions is too narrow or too wide, users can click on the year width and house height in the sidebar to zoom in and out so that all actions are displayed without causing visual confusion (Figure 2).

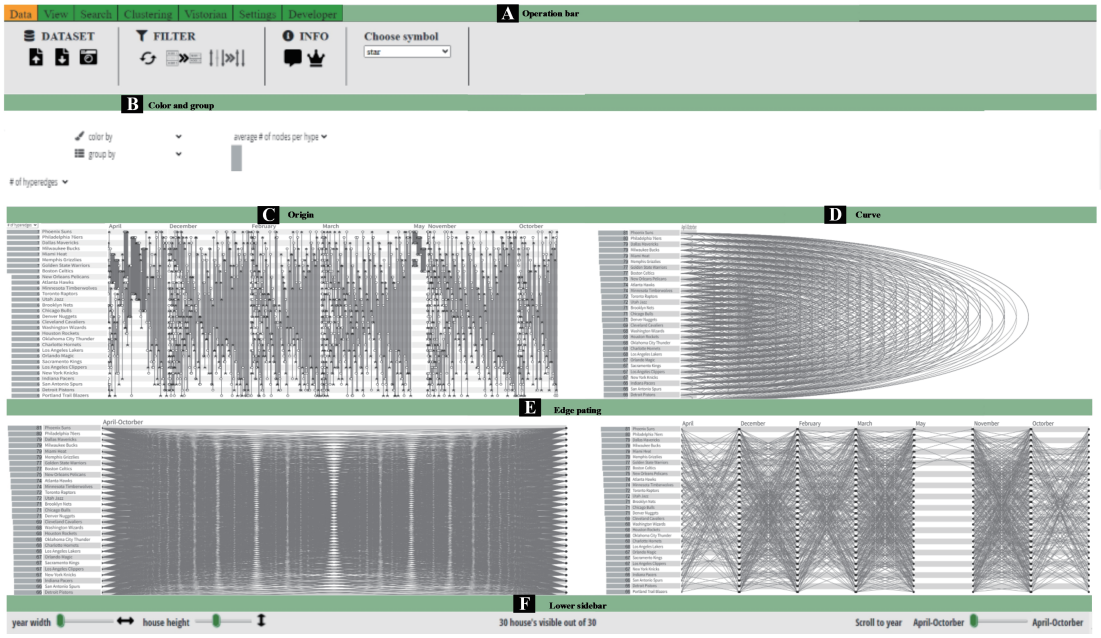

Figure 2 NBA team season win–loss dynamic super chart.

5 NBA game time-series data visualization design

After seeing the wins and losses of a team throughout the season in a super dynamic chart, users may want to take a closer look at the details of each game. This system uses narrative visualization to accomplish this. iStoryline generates storyline visualizations with various rough styles, with lines representing characters in the story. In addition, iStoryline.js uses a reinforcement learning framework to train artificial intelligence agents to help users efficiently explore the design space and generate optimized storylines. Users can visualize the game using game plot narrative visualization (Figure 3).

The contest plot narrative visualization component is divided into three parts: menu bar, main interface, and sidebar. The user can add fonts and change the line′s color through the menu bar, and move, scale, and bend the lines through the sidebar. In this article, the Washington Wizards–Atlanta Hawks game on April 6, 2022, was taken as an example. The lines in the main interface represent the players in the game: the wizard and hawk players are represented by dashed and realized lines, respectively. The players whose positions are the power forward, center, point guard, and point guard are represented by red, yellow, blue, and green, respectively. The spacing between the top and bottom indicates the switch in the players′ scenes, and the spacing between the left and right indicates the players′ playtime. Different players′ lines have contact to indicate whether the players have a confrontation or assist relationship. When the mouse is moved to the contact node, it shows the match or assists the relationship and match details (e.g., player name and event). In the picture below, the game score is 26–29; D. Gallinari replaces D. Hunter, and D. Hunter rests. D. Gafford created two free rows. Moreover, a technical foul was committed by Hachimura (T. Young).

6 NBA game time-series data visualization design

An integrated learning-based all-star player prediction model, SRR-voting, is proposed in this study and compared with eight machine-learning models for experiments. SRR-voting uses ten models to predict 2021 all-star players. The NBA player data from 1980 to 2021 were input into the model. The NBA player data from 1980 to 2020 were used as the training set, the NBA player data from 2021 were used as the test set, and the

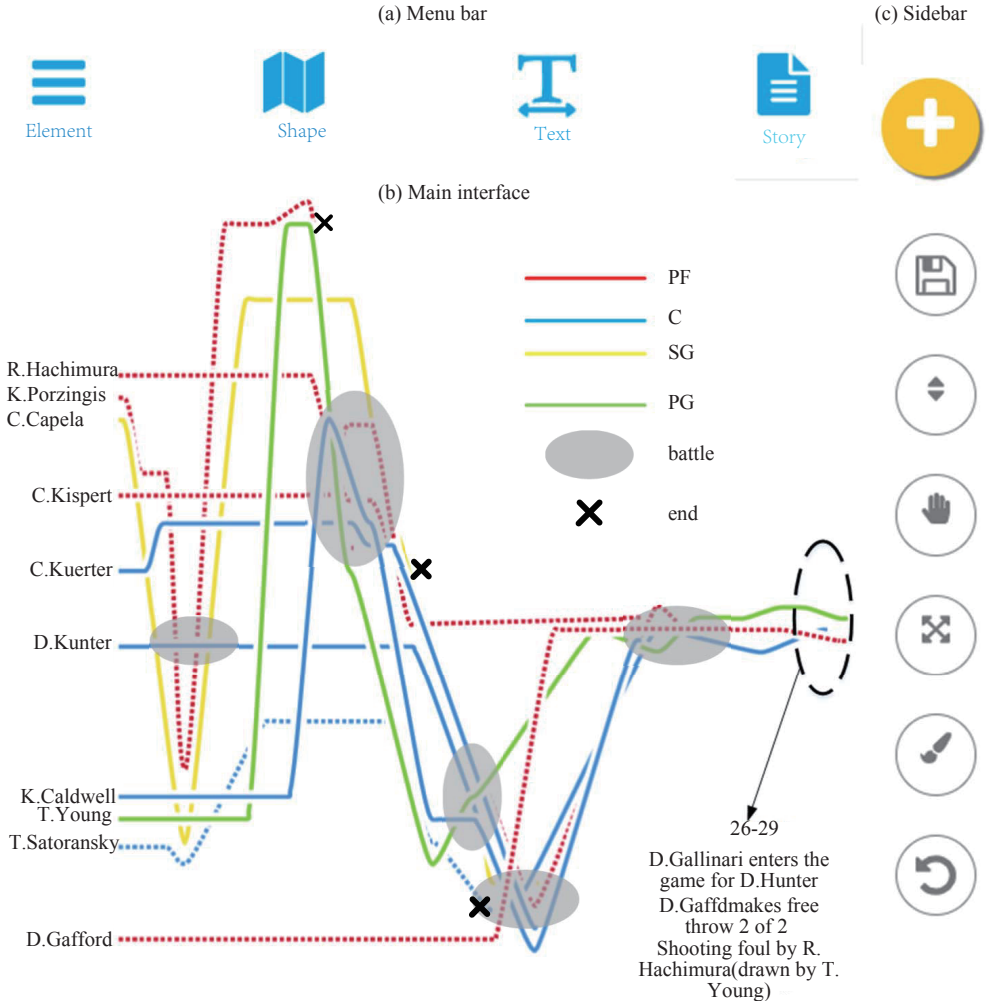

Figure 3 Visualization of the competition plot narrative.

model output the prediction results of the 2021 NBA all-star players. The experimental comparison results prove that SRR-voting prediction performed better.

6.1 Predictive evaluation indicators

In this study, precision and recall metrics were selected to evaluate the model performance. The F1-score metrics were selected to evaluate the overall ability of the model. Precision, recall, and F1-score values are between 0 and 1, and the closer they are to 1, the better the model performance. TP, TN, FP, and FN denote the true-positive, true-negative, false-positive, and false-negative cases, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

6.2 Data preprocessing

The dataset was divided into categorical and numerical features. The categorical features were processed using one-hot encoding. Some numerical features, such as assists and three-pointers, were normalized to check for outliers, and the missing values were zero-filled.

6.3 Integrated learning model

Existing NBA prediction models generally use a single or simple fusion machine-learning model, which has limitations in terms of feature extraction, rate, and accuracy, and the prediction results are not satisfactory. To solve the overfitting and slow training problems, construct effective feature information, and improve prediction accuracy, an integrated learning-based all-star player prediction model, SRR-voting, was proposed to improve the speed and accuracy while extracting effective features (Figure 4).

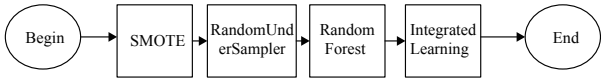

Figure 4 SRR-voting algorithm flow chart.

(1) A certain sample size was generated using the synthetic minority oversampling technique (SMOTE)[34] for the existing sample of all-star players in the minority category. A certain sample size was excluded by the RandomUnderSampler method to balance the number of average players and non-all-star players in the dataset for the existing sample of average players in the majority category.

(2) The 1980-2021 NBA regular season player data were preprocessed using the random forest algorithm, redundant features were removed, higher-order features were filtered out to construct new feature models, and the effectiveness of the random forest algorithm was verified by t-distributed stochastic neighbor embedding (t-SNE) visualization.

(3) Linear regression (LR), LightGBM, and SVM were selected for integrated learning, and each model was optimized with hyperparameters using the GridSearchCV method.

6.4 Predicting the experimental procedure

6.4.1 SMOTE and RandomUnderSampler

The SMOTE method is used to select two instances in a small class and a random point on their spatial connection line as a newly generated instance of the small class. The RandomUnderSampler method uses a certain strategy to select samples from the majority class as representative samples. Specifically, for the evaluation in this study, SMOTE and RandomUnderSampler methods were used to generate and reject samples of a certain size to obtain a balanced dataset, which was used as the input to the next step of the random forest algorithm (Figure 5).

6.4.2 Random forests

The player data and labeled features were first evaluated for feature importance. More important features were selected as the new feature models, and the feature models analyzed by random forest were used as the input for the next step of the SRR-voting algorithm (Table 2).

6.4.3 t-SNE

The t-SNE[36] was used to visualize the nonlinear dimensionality reduction of high-dimensional data. In this

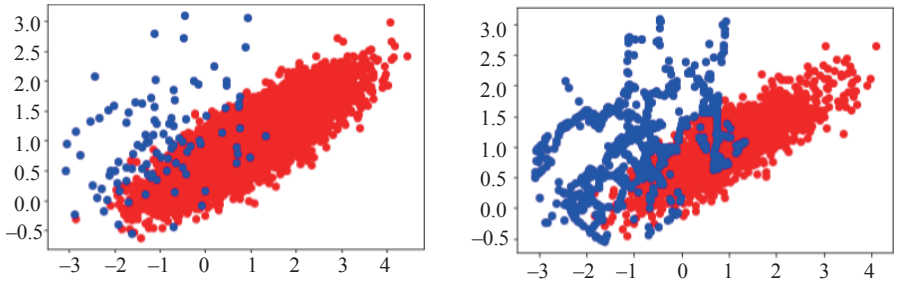

Figure 5 SRR-voting algorithm flow chart.

study, t-SNE reduced the multidimensional data of NBA players to two-dimensional data and visualized them to verify the effectiveness of the random forest method. The red and green nodes represent the non-star and all-star players, respectively. The more distant the nodes, the sparser the distribution, indicating that the random forest method effectively reduces dimensionality (Figure 6).

Table 2 Timing data field properties

Data features	Features importance
FG	0.143
2P	0.119
BLK	0.049
STL	0.030
AST	0.023
…	…

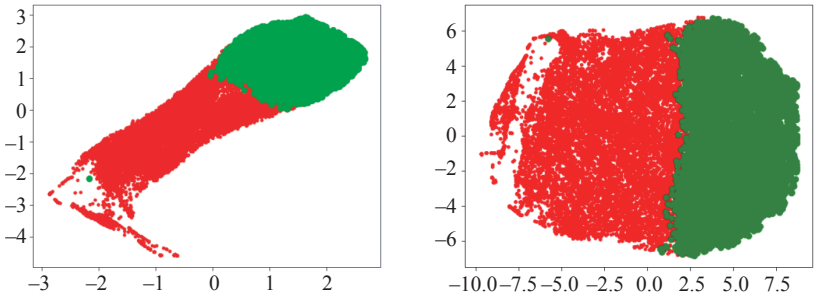

Figure 6 SRR-voting algorithm flow chart.

6.4.4 Integrated learning

The basic principle of integrated learning is to construct several weak classifiers with low classification accuracy and form a strong classifier based on the results of each weak classifier, following a certain strategy to solve the classification problem. In this study, LR, SVM, K-nearest neighbor (KNN), DT, random forest (RF), bagging, stochastic gradient descent (SGD), gradient learning, and classification of the weak classifiers are used. SGD, gradient boosting, XGBoost, local cascade ensemble, and GridSearchCV were used to optimize each model with hyperparameters. Furthermore, three classifiers with the best results–LR, SVM, and LightGBM–were selected for voting integration learning.

6.4.5 GridSearchCV

GridSearchCV is a tuning technique that automatically calculates the best combination of parameters to improve the prediction scores in the current study. The new feature model obtained in the previous step was fed into the integrated learning model optimized by GridSearchCV, and the final prediction results were output. The integrated learning algorithm uses multiple parameters, and tuning is necessary to improve the model′s training, which can directly affect prediction results.

6.4.6 K-Fold cross-validation

In this study, the K-fold was used to validate the results of each model. Because this study used a small dataset, five cross-validations were performed on the NBA player dataset to mitigate overfitting from repeated learning. Thus, 80% and 20% of the data were used as the training and test data, respectively. In each iteration, the scores were recorded and averaged at the end, and the average was used to compare the effects of each model (Table 3).

Table 3 Algorithm core hyperparameter parameter settings

Algorithm	Parameter
LR	'penalty':['l1', 'l2'], 'C':[0.001, 0.01, 0.1, 1, 10]
SVM	'n_neighbors':list(range(2, 5, 1)), 'algorithm':['auto','ball_tree','kd_tree','brute']
KNN	'C': list(range(0.5, 1, 0.1)), 'kernel':['rfb', 'poly', 'sigmod', 'linear']
DT	'criterion':['gini', 'entropy'], 'max_depth':list(range(2, 5, 1)), 'min_samples_leaf':list(range(3, 7, 1))
RF	'n_estimators': list(range(100, 200, 50)), 'criterion':['gini', 'entropy'], 'max_depth':list(range(2,5,1))
Bagging	'n_estimators': list(range(10, 20, 5))
SGD	'penalty':['l2','l1'], 'max_iter': list(range(1000, 2000, 500))

6.5 Comparison of prediction model effects

To validate the effectiveness of the SRR-voting model in predicting NBA players on the NBA player dataset, the SRR-voting model was compared with LightGBM[35], Kaimakamis[33], XGBoost[35], and Wilkens[24]. Comparison experiments were conducted with six models, such as LightGBM and XGBoost in Kaimakamis C as well as KNN in Wilkens S. Comparison results are shown in the table below. In the table, ①, ②, ③, ④, ⑤, and ⑥ represent the SMOTE +RandomUnderSampler+RF, XGBoost, bagging, LR +SVM+LGB, KNN, and LightGBM models, respectively (Table 4).

Table 4 Comparison of ten NBA prediction algorithms on the player dataset

Model	Precision	Recall	F_1
②	0.9238	0.9412	0.9324
①③⑤	0.9263	0.9431	0.9351
①⑤	0.9250	0.9423	0.9335
①④	**0.9305**	**0.9592**	**0.9446**
⑤	0.9176	0.9229	0.9202
⑥	0.9251	0.9433	0.9345
①②③	0.9289	0.9115	0.9200

In summary, all seven models performed well on the NBA player dataset; however, the SRR-voting model outperformed each model in aggregate. The SRR-voting model is the best in precision, recall, and F1 measure metrics when predicting players under similar conditions compared with the remaining six models, which verifies the effectiveness of SMOTE, RandomUnderSampler, random forest, LightGBM, SVM, and RF algorithms, and is at most 1.29, 4.77, and 2.46 higher than the remaining eight algorithms. This is because the SRR-voting model first uses the SMOTE and RandomUnderSampler algorithms to generate and reject a certain sample size to balance the number of all-stars and regular players in the dataset to improve the classification effect and then uses the RF algorithm to extract and construct new features for the player dataset. The RF algorithm was then used to extract and construct features from the player dataset, retain the effective features to build new feature models, and select three algorithms (SVM, LightGBM, and RF) to integrate voting learning. Moreover, GridSearchCV was used to optimize each model with hyperparameters, combined with five-fold cross-validation to improve model generalization. The SRR-voting algorithm is an effective NBA all-star player prediction algorithm(Figure 7).

6.6 Comparison of SHAP-based model interpretation analysis

In this study, the results of the NBA all-star player prediction model were analyzed using the SHapley Additive exPlanations (SHAP) framework. Figure 8 shows the factors that influence the selection of all-star players based on the importance of their characteristics. The higher the value of these characteristics, the higher the probability of being selected as an all-star. In addition, Figure 8 shows the important characteristics of NBA players′ general statistics, from top to bottom, from the most important to the least important characteristics. All NBA players′ regular data features contribute equally to both categories of being selected as an all-star (playoff = 1) or not (playoff = 0) because the number of all-stars and regular players in the NBA players′ regular data is balanced after SMOTE and random downsampling methods. The Figure 8 shows that the differences in the characteristics of the number of points (PTS), number of free throws (FT), and number of 2-pointers have a significant impact on the model, while the number of shots and number of rebounds are the least important characteristics.

Figure 9 shows that the differences in characteristics, such as PTS, assists, and defensive rebounds, have a more significant effect on the model, with the probability of being selected as an all-star increasing as that value increases. In contrast, the number of starts (GS) and the number of appearances (G) were the least important characteristics.

The first sample data in the NBA dataset were used for the analysis. The y-axis shows the different NBA data feature values, and the x-axis indicates the SHAP values. Red and blue indicate the positive and negative gains, respectively. Figure 10 shows that G, MP, PF, and age have positive gains for the first sample data, and

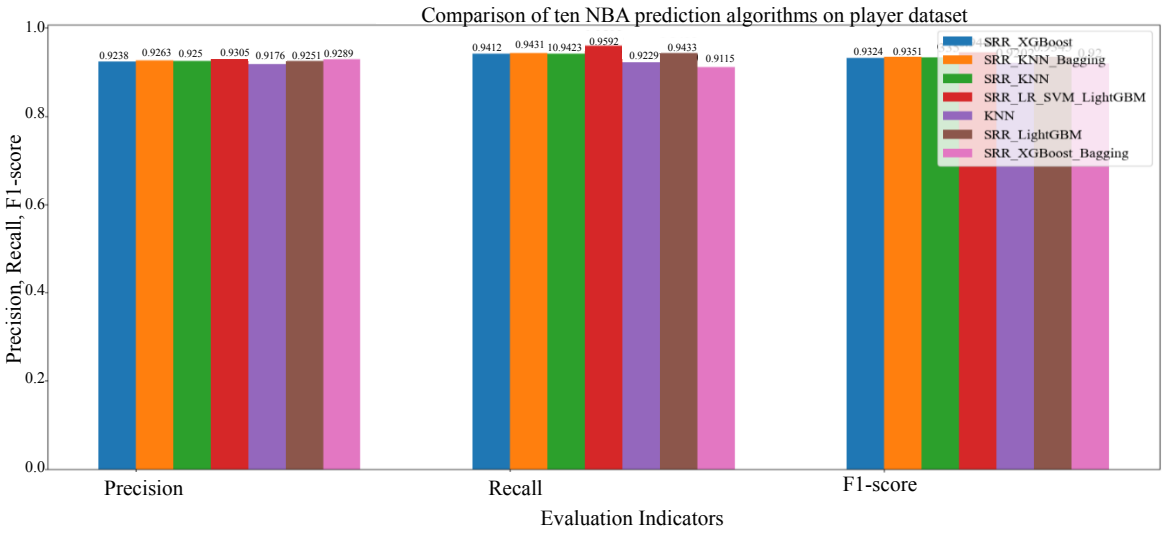

Figure 7 Comparison of 7 algorithms on NBA player dataset.

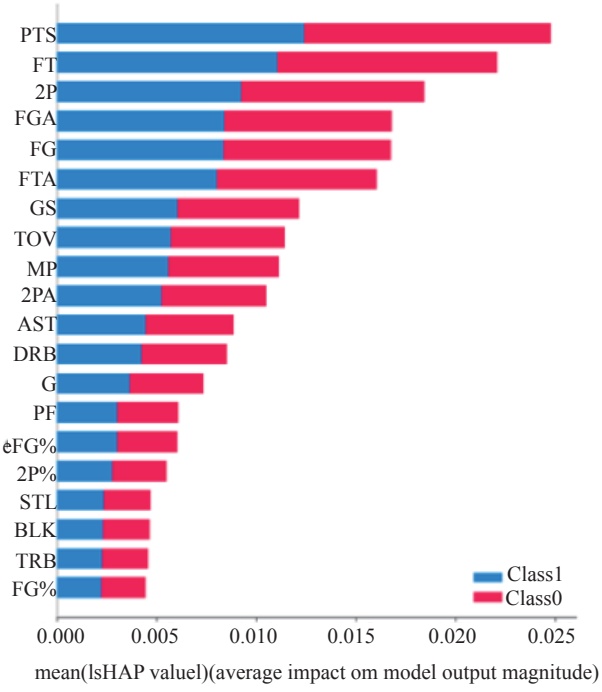

Figure 8 SRR-voting algorithm flow chart.

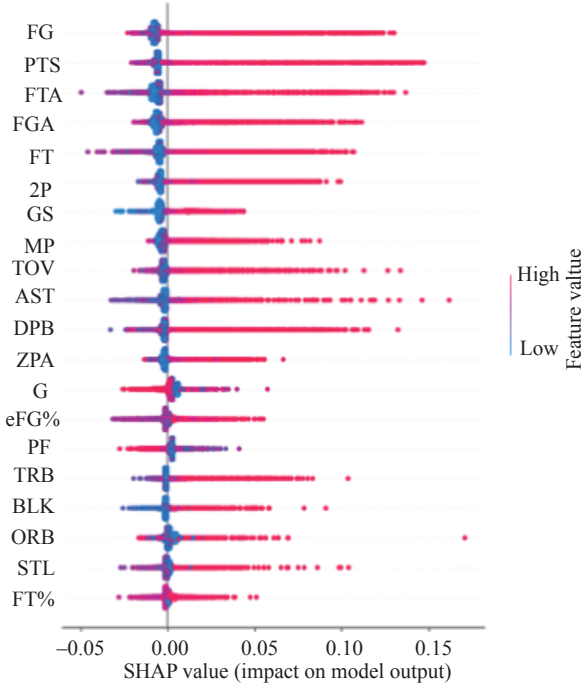

Figure 9 SHAP feature density scatter plot.

PTS, FTA, GS, TOV, BLK, and 17 other features have positive gains for the first sample data. The first sample exhibited a negative gain.

6.7 NBA all-star player visualization data stories

Player visualization data stories are automatically generated using Calliope narrative visualization technology to process regular statistics of players over the same period. Calliope created visual data stories from input spreadsheets and easily modified the generated stories based on an online story editor. It uses a logic-oriented Monte Carlo tree search algorithm that explores the data space provided by the input spreadsheet to progressively generate and organize story segments in a logical order. Calliope automatically transforms data files into a variety of visual data insight presentations, including fact sheets, big visual screens, data cartoons, and

data videos. Finally, Calliope Sheets were used to visualize the results of the NBA all-star prediction experiment and generate a visual data story of NBA players. As shown in Figure 11, from left to right, the colors of the chart are yellow, blue, and green, representing the chart to be explained. Extreme value, comparison, and distribution. The first picture is a bar chart, which shows that the SRR-voting algorithm in this study is the best in the evaluation index recall compared with other algorithms. The second picture is a text chart, which shows that the SRR-voting algorithm in this study is the best. Compared with the KNN algorithm, Voting has a difference of 0.01 in the evaluation index precision. The third image is a tree diagram, and the size of each module represents the level of each algorithm in the precision of the evaluation index.

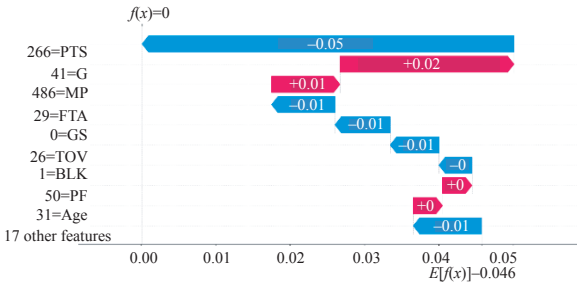

Figure 10 Single-sample feature impact map

7 Customer feedback

To evaluate the effect of the visualization system NPIPVis on the completion of the goal of the basketball game data analysis, this study conducted a user evaluation of the visualization system NPIPVis. During the evaluation process, a total of 10 evaluators were invited, including eight NBA basketball fans, of which two ordinary viewers had only heard about the NBA. After the guidance, the evaluators quickly understood how to use the NPIPVis visualization system. The game time-series data visualization design (T2) and NBA player visualization data stories (T3) are quantitatively evaluated, with a score of 1.00–5.00 points (1.00 = bad effect, 5.00 = very good effect). The scoring of each visualization scheme of the NPIPVis system is shown in Table 5. It can be seen from the table that most of the evaluators believe that the visualization schemes of the visualization system NPIPVis can efficiently and directly analyze the data of basketball events and obtain event information quickly, which is better than the traditional competition sports data analysis tool.

Among the visualization schemes, the evaluation users were the most satisfied with the visualization design

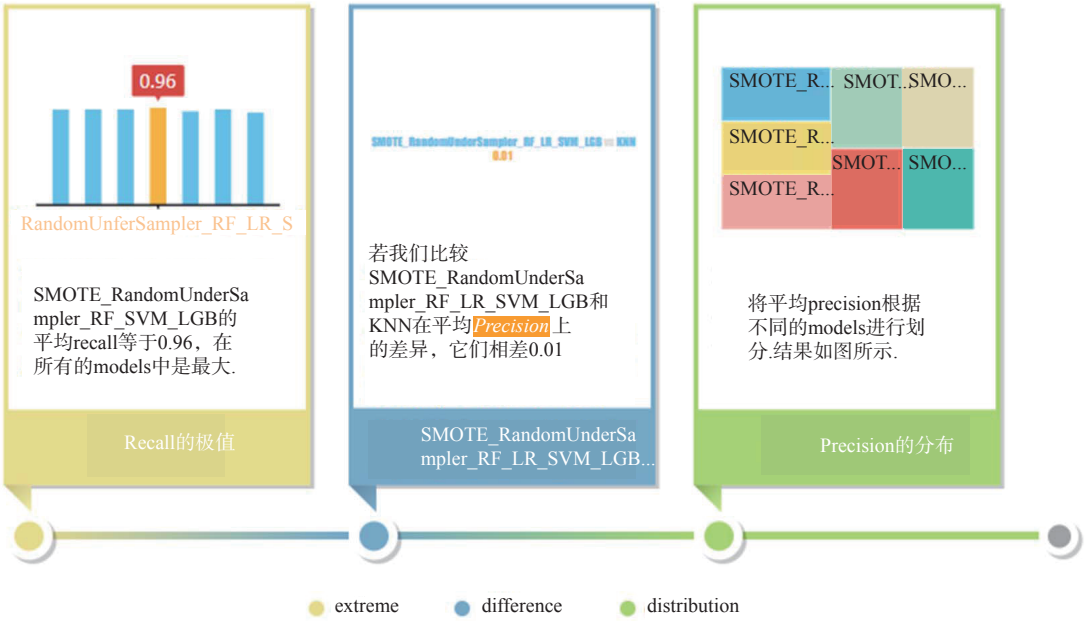

Figure 11 NBA player visualization data story.

of NBA regular statistical data and gave an average score of 4.02. Most evaluation users are satisfied with the visualization design of NBA game time-series data and consider it to be a novel narrative visualization method. Most of the evaluation users were generally satisfied with the visual data stories of NBA players, and their scores were the lowest.

Table 5 User ratings for visualization solutions

Client	T1	T2	T3
NBA fans	3.94	3.65	3.37
General user	4.10	4.26	3.91
Summary	4.02	3.96	3.64

8 Conclusion

In this study, a visualization system, NPIPVis, was designed for the general audience, which includes dynamic hypergraphs of NBA team wins and losses, and visualization views of game plot narratives to help users understand NBA teams and games. In addition, an integrated learning model called SRR-voting has been proposed to predict all-star players. In addition, an integrated learning model, SRR-voting, was proposed for predicting all-star players using case studies of 2021 player data from the Basketball-Reference website and comparing it with six prediction models. In the future, NPIPVis will support more views, such as box line charts and other basic charts, and further optimize the game plot visualization view to analyze players, teams, and game data from multiple perspectives, providing rich visualization components that allow users to intuitively and easily interpret NBA players, teams, and games.

Declaration of competing interest

We declare that we have no conflict of interest.

References

1 Chen W, Shen Z, Tao Y. Big data series: Data visualization. Beijing: Electronic Industry Press, 2013
DOI: CNKI:SUN:IGXN.0.2014-01-040

2 Ji S, Li J, Du T. A review of machine learning model interpretability methods, applications and security research. Computer Research and Development, 2019, 56(10): 2071–2096
DOI: 10.7544/issn1000-1239.2019.20190540

3 Valdivia P, Buono P, Plaisant C, Dufournaud N, Fekete J D. Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(1): 1–13
DOI: 10.1109/tvcg.2019.2933196

4 Tang T, Rubab S, Lai J, Cui W, Yu L, Wu Y. iStoryline: effective convergence to hand-drawn storylines. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 769–778
DOI: 10.1109/tvcg.2018.2864899

5 Shi D, Xu X, Sun F, Shi Y, Cao N. Calliope: automatic visual data story generation from a spreadsheet. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 453–463
DOI: 10.1109/tvcg.2020.3030403

6 Buono P, Ceriani M, Costabile M F, Valdivia P. Visual analysis of goal-leading phases in soccer. 14th Biannual Conference of the Italian SIGCHI Chapter. Bolzano, Italy. New York, ACM, 2021, 1–5
DOI: 10.1145/3464385.3464740

7 Wu Y, Xie X, Wang J, Deng D, Liang H, Zhang H, Cheng S, Chen W. ForVizor: visualizing spatio-temporal team formations in soccer. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 65–75
DOI: 10.1109/tvcg.2018.2865041

8 Sheng B, Li P, Zhang Y, Mao L, Chen C L P. GreenSea: visual soccer analysis using broad learning system. IEEE Transactions on Cybernetics, 2021, 51(3): 1463–1477
DOI: 10.1109/tcyb.2020.2988792

9 Meng X, Li Z, Wang S, Karambakhsh A, Sheng B, Yang P, Li P, Mao L. A video information driven football recommendation system. Computers & Electrical Engineering, 2020, 85106699
DOI: 10.1016/j.compeleceng.2020.106699

10 Zhang P, Zheng L, Jiang Y, Mao L, Li Z, Sheng B. Tracking soccer players using spatio-temporal context learning under multiple views.

Multimedia Tools and Applications, 2018, 77(15): 18935–18955
DOI: 10.1007/s11042-017-5316-3

11 Carsting T, Gummesson J. GoalMate: an application for visualization of ice hockey statistics. 2021

12 Chu X, Xie X, Ye S, Lu H, Xiao H, Yuan Z, Chen Z, Zhang H, Wu Y. TIVEE: visual exploration and explanation of badminton tactics in immersive visualizations. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 118–128
DOI: 10.1109/tvcg.2021.3114861

13 Wu Y, Lan J, Shu X, Ji C, Zhao K, Wang J, Zhang H. iTTVis: interactive visualization of table tennis data. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 709–718
DOI: 10.1109/tvcg.2017.2744218

14 Lan J, Zhou Z, Wang J, Zhang H, Xie X, Wu Y. SimuExplorer: visual exploration of game simulation in table tennis. IEEE Transactions on Visualization and Computer Graphics, 2021, 991
DOI: 10.1109/tvcg.2021.3130422

15 Goldsberry K. Courtvision: new visual and spatial analytics for the NBA. http://www.sloansportsconference.com/wp-content/uploads/2012/02/GoldsberrySloanSubmission.pdf. 2015

16 Lei H, Lao T, Liu Z. Review of sports data visualization. Journal of Computer Aided Design and Graphics, 2015, 27(9): 1605–1616
DOI: 10.3969/j.issn.1003-9775.2015.09.003

17 Chen W, Lao T, Xia J, Huang X, Zhu B, Hu W, Guan H. GameFlow: narrative visualization of NBA basketball games. IEEE Transactions on Multimedia, 2016, 18(11): 2247–2256
DOI: 10.1109/tmm.2016.2614221

18 Losada A G, Theron R, Benito A. BKViz: a basketball visual analysis tool. IEEE Computer Graphics and Applications, 2016, 36(6): 58–68
DOI: 10.1109/mcg.2016.124

19 Hung S, Xu J Y, Gao X F, Chen G H. Basketview: interactive visualization of NBA games. In: Proceedings of the 2018 International Conference on Data Science and Information Technology. Singapore, Singapore. New York, ACM, 2018, 11–17
DOI: 10.1145/3239283.3239298

20 Perin C, Vuillemot R, Stolper C D, Stasko J T, Wood J, Carpendale S. State of the art of sports data visualization. Computer Graphics Forum, 2018, 37(3): 663–686
DOI: 10.1111/cgf.13447

21 Wang S, Zhou S Y. PlayerNetVis: a visual analytics system for evaluating NBA player performance based on network topology. In: Proeedings of 2020 7th International Conference on Information Science and Control Engineering (ICISCE). Changsha, China, IEEE, 2020, 1006–1010
DOI: 10.1109/icisce50968.2020.00206

22 Ji N, Gao Y, Zhao Y, Yu D, Chu S. Basket news visualization combined with knowledge graph. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(6): 837–846
DOI: 10.3724/SP.J.1089.2021.18590

23 Jin Y, Jia J, Hong M. Data analysis and visualization for NBA players. Computer Applications and Software, 2021, 38(8): 84-91
DOI: 10.3969/j.issn.1000-386x.2021.08.015

24 Wilkens S. Sports prediction and betting models in the machine learning age: the case of tennis. Journal of Sports Analytics, 2021, 7(2): 99–117
DOI: 10.3233/jsa-200463

25 Xia V, Jain K, Krishna A, Brinton C G. A network-driven methodology for sports ranking and prediction. In: Proeedings of 2018 52nd Annual Conference on Information Sciences and Systems (CISS). Princeton, NJ, USA, IEEE, 2018, 1–6
DOI: 10.1109/ciss.2018.8362324

26 South C, Egros E. Forecasting college football game outcomes using modern modeling techniques. Journal of Sports Analytics, 2020, 6(1): 25–33
DOI: 10.3233/jsa-190314

27 Lam M W Y. One-match-ahead forecasting in two-team sports with stacked Bayesian regressions. Journal of Artificial Intelligence and Soft Computing Research, 2018, 8(3): 159–171
DOI: 10.1515/jaiscr-2018-0011

28 Pai P F, ChangLiao L H, Lin K P. Analyzing basketball games by a support vector machines with decision tree model. Neural Computing and Applications, 2017, 28(12): 4159–4167
DOI: 10.1007/s00521-016-2321-9

29 Giuliodori P. An artificial neural network-based prediction model for underdog teams in NBA matche. MLSA@ PKDD/ECML. 2017

30 Özbalta E, Yavuz M, Kaya T. National basketball association player salary prediction using supervised machine learning method. International Conference on Intelligent and Fuzzy Systems. Springer, 2021, 189–196
DOI: 10.1007/978-3-030-85577-2_22

31 Yang F, Zhang J. The ranking prediction of nba playoffs based on improved pagerank algorithm. Complexity, 2021
DOI: 10.1155/2021/6641242

32 Huang M L, Li Y Z. Use of machine learning and deep learning to predict the outcomes of major league baseball matches. Applied Sciences, 2021, 11(10): 4499
DOI: 10.3390/app11104499

33 Kaimakamis C. Sports analytics algorithm for NBA champion prediction. 2021

34 Albert A A, de Mingo López L F, Allbright K, Gómez Blas N. A hybrid machine learning model for predicting USA NBA all-stars. Electronics, 2022, 11(1): 97
DOI: 10.3390/electronics11010097

35 Daoud E A. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. International Journal of Computer and Information Engineering, 2019, 13(1): 6–10
DOI: 10.5281/zenodo.3607805

36 Wu J L, Wang J X, Xiao H, Ling J L. Visualization of high dimensional turbulence simulation data using t-SNE. In: 19th AIAA Non-Deterministic Approaches Conference. Grapevine, Texas, Reston, Virginia: AIAA, 2017, 1770
DOI: 10.2514/6.2017-1770