Hindawi
Complexity
Volume 2021, Article ID 6676297, 11 pages
https://doi.org/10.1155/2021/6676297

Review Article

Sports Big Data: Management, Analysis, Applications, and Challenges

Zhongbo Bai[1] **and Xiaomei Bai**[2]

[1]*School of Sports Science, Anshan Normal University, Anshan, China*
[2]*Computing Center, Anshan Normal University, Anshan, China*

Correspondence should be addressed to Xiaomei Bai; xiaomeibai@outlook.com

Received 3 December 2020; Revised 11 January 2021; Accepted 18 January 2021; Published 30 January 2021

Academic Editor: chuan lin

With the rapid growth of information technology and sports, analyzing sports information has become an increasingly challenging issue. Sports big data come from the Internet and show a rapid growth trend. Sports big data contain rich information such as athletes, coaches, athletics, and swimming. Nowadays, various sports data can be easily accessed, and amazing data analysis technologies have been developed, which enable us to further explore the value behind these data. In this paper, we first introduce the background of sports big data. Secondly, we review sports big data management such as sports big data acquisition, sports big data labeling, and improvement of existing data. Thirdly, we show sports data analysis methods, including statistical analysis, sports social network analysis, and sports big data analysis service platform. Furthermore, we describe the sports big data applications such as evaluation and prediction. Finally, we investigate representative research issues in sports big data areas, including predicting the athletes' performance in the knowledge graph, finding a rising star of sports, unified sports big data platform, open sports big data, and privacy protections. This paper should help the researchers obtaining a broader understanding of sports big data and provide some potential research directions.

1. Introduction

The era of big data has brought an unprecedented impact on the development of the sports industry. Big data services closely related to it, including exercise performance, health data, training statistics, and analysis, can effectively help athletes in daily training and developing game strategies and are becoming an indispensable means for winning competitions [1–4]. Advanced big data technique has brought about the changes in the sports field. The proliferation of sports data has generated new opportunities and challenges in the field of sports big data [5, 6]. Sports big data is the product of the development of the Internet and sports. The McKinsey Global Institute gives the concept of big data, which includes four characteristics: volume, variety, velocity, and value [7]. Drawing on the definition of big data given by the McKinsey Global Institute sports big data can be defined as a sports data collection that is so large that it can acquire, store, manage, and analyze far beyond the capabilities of traditional database software tools, including five features: volume, variety, velocity, veracity, and value (see Figure 1).

Hundreds of millions of sports data are generated each day from millions of schools, various events, and communities, representing the volume feature [8–10]. The velocity feature can be reflected by the growth rate of sports data. The variety of sports big data stems from the fact that it contains various entities and relationships, which makes sports big data systems more challenging (see Figure 2). Among them, the representative processing includes name disambiguation and data duplications. The variety feature of sports big data mainly contains the following aspects: (1) physical fitness such as height, weight, vital capacity in the physical function category, as well as the 50-meter run and sitting posture in the physical fitness category; (2) physical exercise behaviors such as running, basketball, tennis, table tennis, football, archery, rowing, swimming, skipping rope, and their

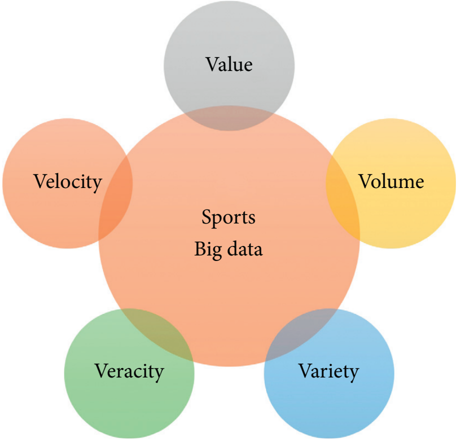

Figure 1: Five key features of sports big data.

behavior trajectory; (3) personal information (gender, age, ethnicity, birth date, language, school, province, and family); (4) various competition results. One of the most important features of sports big data is its value. At present, research related to sports big data has attracted the attention of researchers, including evaluation (evaluating player performance, evaluating student physical fitness, and evaluation of coaching results) and prediction (predicting player performance and predicting student physical fitness) [11–13].

Exploring sports big data can provide great benefits for popular sports, school sports, and competitive sports [14–17]. For example, through the management and analysis of athletes' usual physical fitness and athletic performance, it is possible to predict potential athletes. The results of these data analyses provide a favorable basis for decision-makers in the allocation of funds for athlete training. The basic motivation is to mine knowledge from sports big data to provide better sports services for athletes, coaches, competition-related decision-makers, and the public. In addition, some typical big data services such as exercise performance, health data, training statistics, and analysis can effectively help coaches and athletes in daily training and customizing game strategies and play an immeasurable role for winning competitions [18].

Sports big data analysis aims to solve the problems in sports science by relying on data mining, network science, and statistical techniques [19–23]. Sports big data analysis focuses on the discovery of the value of data and provides valuable information resources for enterprises and managers. This valuable sports information is finally displayed through visualization. For instance, the American Men's Professional Basketball League has established a complete data analysis system. They face large-scale, fast-changing, and diverse sports big data. They use sports big data to evaluate athletes and formulate new strategic plans. It is worth mentioning that they track the movement trajectories of players, referees, and the ball and then establish dynamic evaluation indicators and convert these data into valuable information. This has become a professional basketball team to win the game, evaluate players, and optimize offense and defense [24]. Different data mining methods have been applied to uncover hidden relationships, patterns, and laws in sports big data [25–27]. Due to the increasing sports data volumes and various types of sports data, sports big data is challenging.

This paper presents a review of recent developments in sports big data. To the best of our knowledge, this paper is the first effort to provide a comprehensive review of sports big data. This overview covers three aspects: sports big data management, sports big data analysis methods, and sports big data applications. In Sports Big Data Management, we introduce sports big data acquisition, sports big data labeling, and improvement of existing data. In Sports Big Data Analysis, we investigate the methods of sports big data including statistical analysis, sports social network analysis, and sports big data analysis service platform. In Sports Big Data Applications, we discuss two important applications: evaluation and prediction. In addition, we discuss several potential key issues associated with sports big data research, including predicting the athletes' performance in the knowledge graph, finding a rising star of sports, unified sports big data platform, open sports big data, and privacy protections in Open Issues and Challenges. We conclude this survey in Conclusion.

2. Sports Big Data Management

Sports big data management mainly applies data management techniques, tools, and platforms to deal with sports big data, including storage, preprocessing, processing, and security. However, big data management is a complex process, which stems from the heterogeneity and unstructured nature of data sources [28]. Sports big data management is crucial to the success of the national sports industry, teams, and individual [15, 16, 29]. The main aim of sports big data management is to mine the potential value of sports big data and to enhance data quality and accessibility for decision-making. In this section, we introduce sports big data acquisition, sports big data labeling, and improvement of existing data.

2.1. Sports Big Data Acquisition. A particularly important feature of sports big data is its diversity. Not only are data sources extremely broad, but data types are also extremely complex. The development of the Internet of Things, the Internet, and the sports industry has enriched the amount of sports data. Because network data are diverse, are complex in composition, and have different usage methods and utilization values for different purposes, the collection of network sports big data is also very challenging. Big data on online sports is usually collected through web crawlers. The general crawler collection process includes the following six aspects: website page analysis, link extraction, link filtering, content extraction, URL queue, and data crawling. The specific process is as follows: (1) first, write one or more sports-related target links in the URL queue as the starting point of crawling information. (2) The crawler reads the link from the URL queue and visits the sports website. (3) Crawl the corresponding sports content from the website. (4)

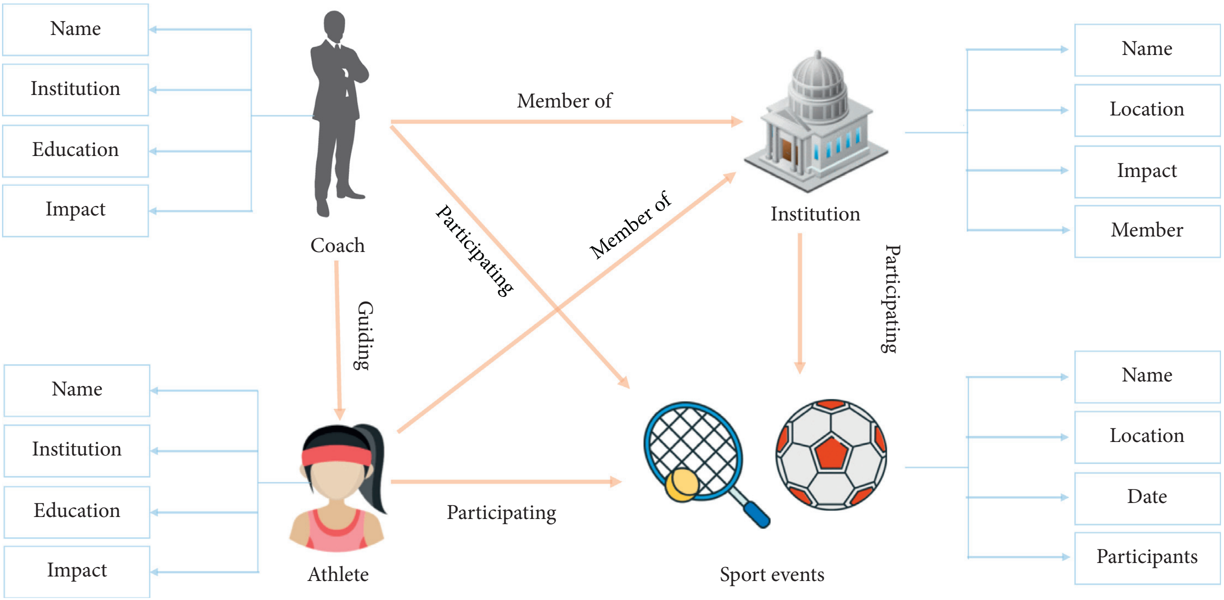

FIGURE 2: An example of entities and their relationships associated in sports big data.

Extract target data and all URL links from web content. (5) Read from the database the URL of the web page that has crawled sports content. (6) Filter URL. Compare the URLs in the current queue with the URLs that have been crawled. (7) According to the result of the comparison, decide whether to capture the content of the address. (8) The queue to be drawn. The content of the web page is written into the database, and the new links that are fetched are added to the URL queue.

Under the background of big data, to better help researchers explore the development of the smart sports industry, many open datasets can be freely downloaded on the website. Recently, a web service called Google Dataset Search (https://www.blog.google/products/search/making-it-easier-discover-datasets/) was launched for searching-related repositories of datasets on the Web. The Google Dataset Search can search for data that are not easy to search on the web. Dataset search demands dataset providers describe their datasets using various metadata (e.g., author, publication date, data content, and terms for using the data) so that they become more searchable. In addition, sports big data can be obtained by relying on the social network. For example, Open Source Sports can be accessed through the following website: http://www.seanlahman.com/open-source-sports/. It includes all kinds of sports data sources such as baseball, football, basketball, and college football. In addition, there are some other websites that provide open datasets such as the National Football League official website (http://www.NFL.com), Basketball-reference website (Basketball Reference), ACB official website (http://www.ACB.com), NBA official website (http://www.NBA.com), Equibase website (http://www.equibase.com), Basketball Federation of Serbia/Basketball Supervisor software, and Foot-Data website (http://www.football-data.co.uk).

Currently, big data applications have been popular but faced security issues and challenges. Sports big data collection is a crucial task for data processing. In addition, secure sports big data collection is a vital step for all kinds of data applications, which can provide the outcome of big data analysis. Because suspicious data sources allow data collection to explore various malicious attacks and treats, secure sports big data collection methodology is crucial for various data applications. To provide secure big data collection, researchers have made some progress in this area, for example, to provide energy-efficient data collection and security of data in a distributed environment. An effective framework is proposed to overcome these problems by relying on blockchain and deep reinforcement learning [30]. An ethereum blockchain platform can be used to provide data security when mobile terminals share the data. The framework can resolve various attacks such as majority attack, device failure, and eclipse attack [30].

2.2. Sports Big Data Labeling.

Once enough sports data have been obtained, the next processing is to label individual examples. For instance, given a basketball game dataset, label different performances for the purpose of predicting future basketball game results of the individual. Usually, data acquisition is done along with data labeling. When extracting information from the Web and constructing a knowledge base, each information is assumed to be correct and thus is implicitly labeled as true. When discussing the data labeling literature, it is easier to separate it from data acquisition because the used techniques can be different. The data labels can be divided into three categories: (1) existing labels. These existing labels can be used to learn from them to predict the rest of the labels. (2) Crowd-based. Recently, many crowdsourcing techniques can be used to help players become more effective in labeling. (3) Weak labels. Although to generate correct labels all the time is desirable, this implementation process may be very expensive. To generate less than perfect labels, weak label is an alternative method, which is used in many applications by as labeled data.

It usually takes a lot of manpower to label the data, but only a small amount of labeled data can be generated. A semisupervised learning technique explores labeled and unlabeled data for predictions [31]. A smaller branch of research is named self-labeled, which is a broad topic. One of the advantages of the self-labeled technique is that it can generate more labels by trusting one's own predictions [32]. In addition, there are graph-based label propagation techniques that are specialized in sports graph data. The semisupervised learning techniques can be used in classification, regression, and graph-based label propagation tasks. The goal of using semisupervised learning techniques in the classification task is to train the model that returns one of the multiple possible classes for each example using labeled and unlabeled datasets. The goal of using semisupervised learning techniques in the regression task is to train a model that can predict a real number given an example. The graph-based label propagation has applications in computer vision, information retrieval, social networks, and natural language processing [33–35].

The better way to label examples is to do it manually. However, to a large project, it took years to complete, which most machine learning users cannot afford for their own applications. Traditionally, active learning has been an important technique in the machine learning community for carefully choosing the true examples to label and thus minimize cost. Recently, crowdsourcing techniques have been proposed in the labeling task. Therefore, there is more emphasis on how to assign tasks to ensure high-quality labels [36]. The data programming model has made progress in two aspects: accuracy and usability. Compared to training with fewer manual labels, training a discriminative model on large amounts of weak labels may result in a higher accuracy. Several systems have been developed for data programming, such as DeepDive, DDLite, and Snorkel [37–39].

2.3. Improvement of Existing Data. Machine learning technology can be used to deal with noisy data and uncorrected labels. There are a lot of literature on improving data quality [40, 41]. A representative cleaning system HoloClean constructs a probabilistic model, which uses quality rules, value relationships, and reference data, to capture how the data were generated [42]. In addition, other data cleaning tools have been developed to convert raw data into a better form for further research. Some cleaning models are designed to improve accuracy. ActiveClean model treats the training and cleaning as a form of stochastic gradient descent to improve the effect of cleaning data. TARS can be used to solve the problem of cleaning crowdsourced labels by using oracles. The TARS provides two pieces of advice. On the one hand, given test data with noisy labels, they predict how well the model may perform on the true labels by using an estimation technique. On the other hand, given training data with noisy labels, TARS can determine which examples to send to an oracle to maximize the expected model, which can improve the accuracy of each noise [43]. To obtain high-quality data labels, improving the quality of existing labels is a good solution [44]. They examine the improvement (or lack thereof) of data quality through repeated labeling and focus on the improvement of training labels for supervised induction. Their experiment results show the following advantages: (1) repeated labeling can improve the label quality and model quality; (2) for noisy labels, repeated labeling can improve the label quality; (3) and a robust technique is proposed to improve label quality.

3. Sports Big Data Analysis Methods

Big data analysis refers to the technique, which can quickly acquire valuable information from all kinds of data [45]. The big data analysis technique can use various algorithms to statistically calculate the big data and extract important analytical data to meet the actual needs. For example, in the competitive sports area, big data analysis technology can not only help coaches and athletes to analyze the previous training and competition sports behavior but also can pin the athlete's movement and physical condition and adjust the athletes' training activities to improve their competition performance. In addition, big data analysis technology can also help coaches and athletes understand the strengths and weaknesses of their opponents to achieve excellent results in large-scale events.

3.1. Statistical Analysis. Based on the statistical theory, the statistical analysis technique is proposed, which belongs to a branch of applied mathematics. The statistical analysis can provide inference for big data. In sports industry research, the statistical analysis technique usually is used to process sports datasets. Through analyzing some statistical features of sports datasets such as mean, variance, entropy, and maximum/minimum value, researchers can explore the athlete movement pattern, and based on this statistical analysis, the coaches can develop effective training plan [46].

A sports data mining tool is proposed to help improve the analysis of techniques and tactics of competitive sports [47]. In this research, the author builds two statistical databases: one is a technical dataset, and the other is a tactical dataset, including the sheets related to the badminton competition information: teams, players, coaches, technical action type, and badminton trajectory. For example, through analysis of the statistical data of technical movement used by the opponent in the competition, the opponent's action behavior can be prejudged to make an effective response plan. Based on the English Premier League, a statistically significant model is proposed by measuring the entropy of the ball passes to predict the competitive team's position [48]. Their experimental results show that the entropy can better identify the important role of defenders.

Although statistical analysis technology has played an important role in sports big data research, with the development of the sports industry and big data technology, more and more technologies such as machine learning, data mining, and predictive analysis are used in sports data research [49, 50]. These technologies usually rely on sports social networks. In the following section, we introduce sports social network analysis.

3.2. Sports Social Network Analysis. The sports social network analysis can reveal the relationship patterns in team sports. Lei et al. [51] use a questionnaire survey to investigate how the social networks of adolescent impact their sports behaviors. They conclude that the social networks of adolescents are the important factors of influencing adolescents' sports behavior. To identify the most influential Twitter accounts of major sporting events: the Track Cycling World Cups, all the tweets from the competitions from 2016 to 2018 are used, including the official hashtag of each event, mentions, and retweets [52]. They leverage the social network analysis technique to identify a part of the variables related to the influence on Twitter. The social network analysis is used to investigate the levels of cohesion occurring among recreational runners by using the running groups to prepare for the running event [53]. The social network analysis is applied to explore the goal-scoring passing networks of the 2016 European Football Championships. Their experimental results indicate that the goal-scoring passing networks have low values in terms of network density, cohesion, connections, and duration. The relationships between team performance and network measurement of the teams from the FIFA World Cup 2018 match. By comparing the performance outcomes of network measurements between winners and losers, they conclude that the general network measurements are not sensitive to the variations in the final score [54]. A passing network approach within the positioning-derived variables is leveraged to identify the contributions of individual players for the team behavior outcome during a simulated match. The results of the research indicate that the lower team passing dependency for a given player and high intrateam well-connected passing relationships are related to better outcomes [55].

3.3. Sports Big Data Analysis Service Platform. Li proposes a Hadoop-based outdoor motion sports big data analysis platform, which stores students' mass motion data and analyses these motion behaviors by the construction of a large data mining system [56]. In his research, students' physical activity information is monitored, recorded, and stored in real time by relying on wearable intelligence terminals. At the same time, these motion data will be sent to the sports big data service platform, and based on the distributed platform, each student's motion information data will be set up independently to accomplish various data analysis tasks. A beach volleyball big data analysis platform is developed to provide meaningful guidance to help coaches in developing valuable training programs and tactical decisions for beach volleyball athletes [57]. In this research, the stored big data of the beach volleyball matches can be analyzed by relying on data mining techniques, and these data information is listed as follows: the success rate of the player, technical analysis, and strategy analysis. Recently, to promote the development of big data analysis in the intelligent sports area, more researchers pay attention to distributed intelligent sensing technology. Based on the sports big data platform, the game relationship between profit and consumption intention about sports cultural hall is analyzed [58]. They leverage support vector machine technique and statistical technique to construct the pricing model, in which the dynamic pricing strategy of spare time of sports cultural hall is designed. Based on the big data cloud platform, in sunshine sports, the physical education teacher training system is developed to promote students to actively participate in physical exercise [59]. The iBall is proposed to track a ball's 3D trajectory and spin with inexpensive sensors, and the iBall integrates wireless and inertial sensory data into physics-based motion models of a ball [60]. The sports personalized content customization platform needs to improve, and the information serves as the core to build sports industry, which should receive more attention [61].

Representative research a self-powered falling point distribution statistical system is developed to give help in training guidance and real-time competition for athletes and referees [62]. Meanwhile, the edge ball judgment system has been designed. In this research, a flexible and durable high-performance wood-based triboelectric nanogenerator for self-powered sensing is developed to analyze the athletic big data. It is worth mentioning that the triboelectric and mechanical performance of the wood can improve after the treatment, including flexibly and processability.

4. Sports Big Data Application

Figure 3 shows the sports big data framework, including data source layer, data collection and exchange layer, central repository layer, data analysis layer, and application layer. The data source layer mainly includes athletes's historical data, athletes' behavior trajectory, video data, and Internet data source. The data source layer is the foundation for realizing various sports big data analysis and prediction applications. The next layer is the data collection layer, which collects data from the data source layer and performs the following processing: data collection, data storage, data exchange, manual import, and web crawler. The collected data are cleaned, and necessary processing is performed according to different application requirements. The data are classified and stored. The processed data will be stored in the central repository layer, including structured data storage, unstructured data storage, and file storage. The data analysis layer performs feature selection, relationship analysis, statistical analysis, and social network analysis according to the needs of specific applications, with the purpose of discovering potential knowledge, laws, and patterns in sports big data. Based on the above analysis results, the integration of machine learning and big data technology can promote the development of sports big data applications [63].

4.1. Evaluation. Evaluation is an important application of sports big data. Evaluation in sports circles is a valuable judgmental factor made on the outcome of an athlete's performance. In this section, factors influencing the

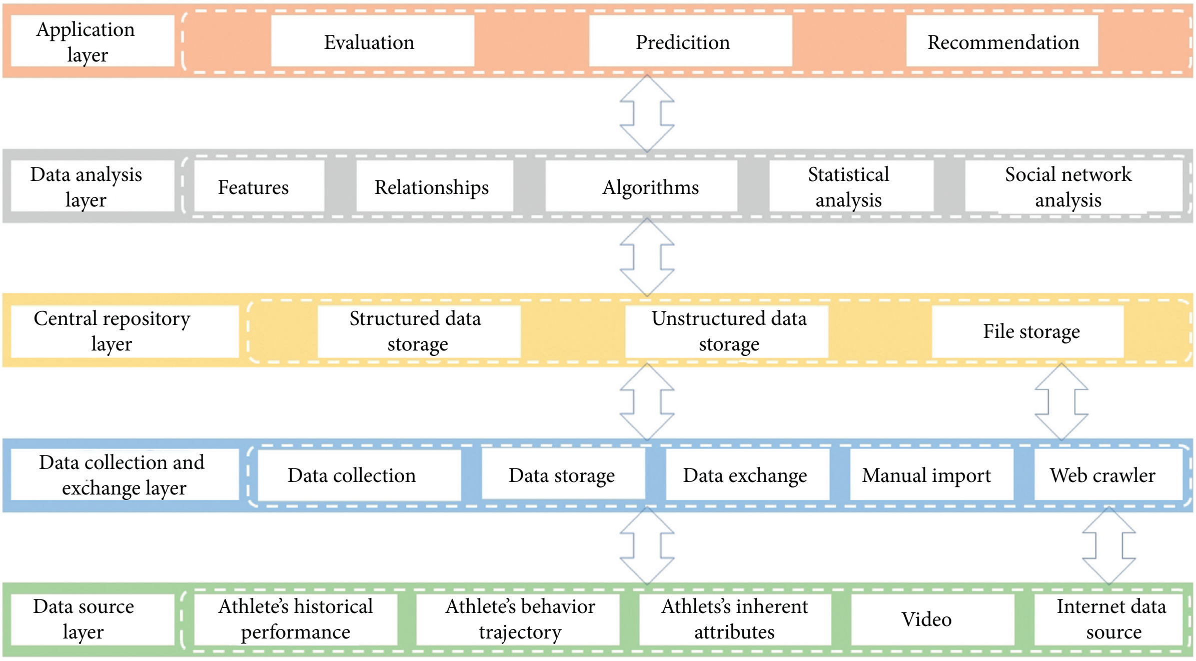

Figure 3: A framework of sports big data platform.

performance of players and data-driven evaluation models are introduced.

4.1.1. Factors Influencing the Performance of Players. Brooks et al. [64] extract feature from each possession to construct the feature vectors with the origin and destination locations in a possession. To obtain a feature vector for a possession, they average the feature vectors across all completed passes in a possession. Furthermore, each feature vector is labeled according to how the possession ended. Pappalardo et al. [65] use a database of soccer logs including 31,496,332 events, 19,619 matches, 296 clubs, and 21,361 players. Each event includes a unique event identifier, the type of the event, a timestamp, the player related to the event, the team of the player, the match, the position on the soccer field, the event subtype, and a list of tags. The type of event is composed of the pass, foul, shot, duel, free kick, offside, and touch. The foul type includes the four features: foul no card, foul yellow, foul red, and foul 2nd yellow. To compute the player performance vectors, they extract 76 features from the Wyscout soccer logs to compute the players' performance vectors. Li et al. [66] extract 22 features that are related to attacking, passing, and defending performance based on the Chinese Football Super League dataset from 2014 to 2018 to rank the teams.

4.1.2. Data-Driven Evaluation Models. The evaluation of the performance of players has attracted the interest of the scientific community and sports community, thanks to the availability of massive sports data [64–69]. Brooks et al. [64] propose a player ranking framework based on the value of passes completed, which is derived from the relationship of pass locations in the possession and shot opportunities generated. The support vector machine algorithm is used to learn the relationship, and the model has an AUROC of 0.79. Pappalardo et al. [65] design a data-driven framework which provides a principled multidimensional and role-aware evaluation for the performance of the soccer player. Based on a massive dataset of soccer logs and millions of match events from the four seasons of 18 prominent soccer competitions, they compare the proposed PlayeRank and known algorithms, and the results show that the PlayeRank outperforms the competitors. Li et al. [66] leverage a linear support vector classifier model to rank the performance of teams. Their experimental results show that the predictive accuracy of the data-driven model proposed is up to 0.83 and the ranking teams' match performance is highly correlated with their actual ranking. In addition, the rankings of different teams are highly correlated with their final league rankings. Pelechrinis et al. [67] propose a ranking algorithm based on the analysis of the teams of the corresponding leagues that capture win-lose relationships and the PageRank algorithm. The results show that the cycles in the network are significantly related to performance. Ghosh et al. [68] propose a data-driven approach to evaluate the performance of the player based on the player's stance or posture. In their experiments, they use shallow learning and deep learning algorithms to classify the strokes, which are used to analyze the stance. Based on this, they compare the stance of an intermediate or a novice player with that of a professional player. Furthermore, they learn the error between the professional player and a participant. A sensor network is used to evaluate their approach. Liu et al. [69] propose an improved evaluation method for soccer player performance. In their research, text information of postmatch reports is used, and the results indicate that the proposed method is more effective and reasonably in terms of evaluating the player

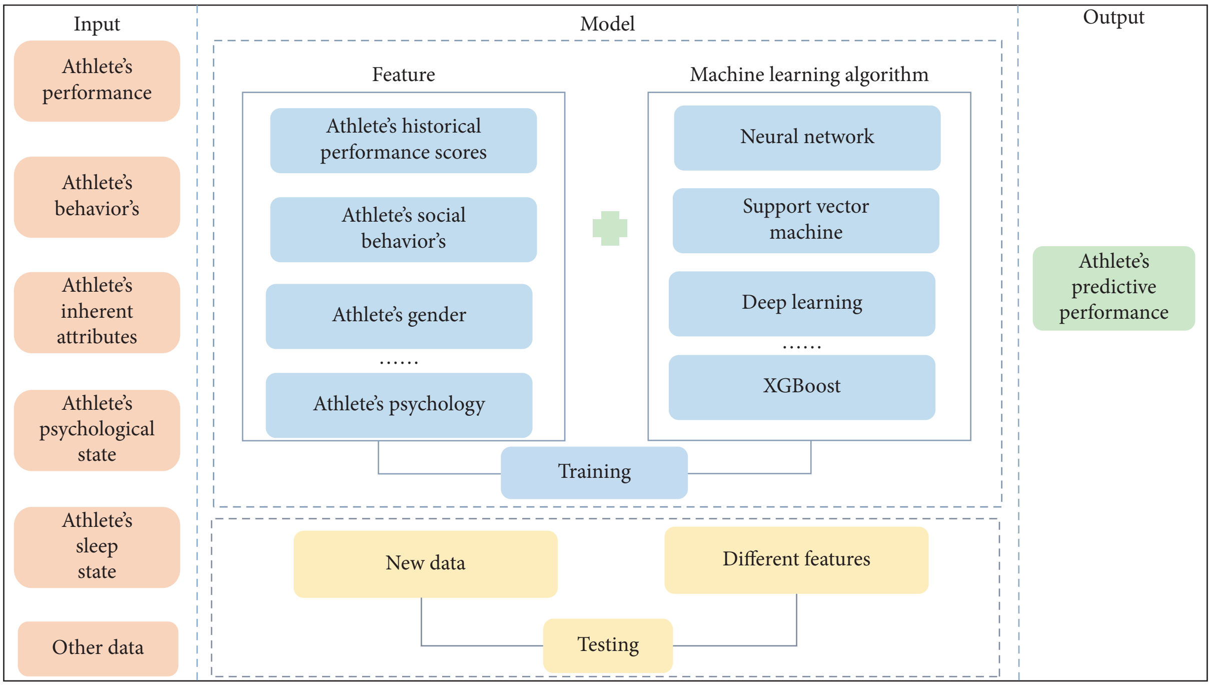

Figure 4: A model for predicting athlete's performance.

performance. Sales and Tjortjis offer a critical analysis of performance metrics to understand the strengths and weaknesses of different teams [70].

4.2. Prediction. As we all know, sports big data can bring unprecedented changes in the sports industry. Recognizing and mining valuable sports big data can not only improve the competitive level of individuals and teams but also promote the development of national fitness. Prediction is an important research aspect of sports big data applications. It is very meaningful to tracking and predicting sports performance, which can bring the following advantages: (1) it can help coaches find a rising star in sports; (2) it can help coaches and athletes to develop effective training plans; (3) and it can help coaches and athletes master the opponent's habits and specialties in the game to make value judgments in the game. For example, by analyzing each athlete's training state and recent game performance, the coach can make the right decisions in selecting players for the game. There is a lot of literature on the prediction of athlete performance.

Figure 4 shows a predictive model for predicting an athlete's performance. The predictive model mainly includes three parts: input, model, and output. In the input part, some raw sports big data is obtained such as athlete's historical performance, athlete's training and competing behavior, athlete's inherent attributes, athlete's psychological state, and sleep state. The model part includes two steps: training and testing. Usually, some important features are used to train the predictive model, including athlete's historical performance scores, athlete's social behavior, athlete's gender, and other features. Based on these important features, some machine learning algorithms are leveraged to train the model, including neural network [71], support vector machine [72], deep learning [73], and XGboost [74]. After training, new data and the selected features are used to test and predict the athlete's performance. In the following sections, we focus on the factors and models that affect athletes' performance.

4.2.1. Factors Influencing Athlete Performance. A predictive model of athlete performance takes most factors that influence his/her performance change into account [75–78]. These important features are divided into the following categories: (1) sports historical scores; (2) sports behavior attributes, including athlete movement process, game performance, and training performance; (3) inherent attributes, including the static attributes of athlete's gender, family, and education; (4) psychologic health state. Bunker et al. [79] use match-related features and external features such as the results of historical matches, player performance indicators, opposition information, recent form, and player available for the match. Constantinou and Fenton [80] leverage the following features: EU competition (EU competition (PS), qualified for EU competition (NS), team stress and fatigue (PS), and team stress and fatigue (PS)), EU matches (EU involvement experience (PS) and EU involvement experience (NS)), managerial changes (new manager, longevity of previous, managerial instability, and managerial ability), newly promoted, injury level (PS), player MotM (squad ability to deal with injuries (NS) and squad ability to deal with injuries (PS)), league points (league points (PS), league points (NS), and league points difference between seasons), team wages (team wages (PS), team wages (NS), adversaries' average team wage (PS), adversaries' average team wage (NS), team wages % difference from average adversary (PS), and team wages % difference from average adversary (NS)), and net transfer spending (net transfer spending, adversaries' average net transfer spending, and net transfer spending relative to average opponent) to predict the accuracy of the long-term football team performance. Thabtah et al. [81] predict NBA game results by using influential features such as defensive rebounds feature, three-point percentage, free throws made, and total rebounds, which can increase the accuracy of the predictive model. Li et al. [29] propose a predictive method to predict sports team performance by using players' and team's historical data, including game count, playing time, two points, three points, free point, free throw, defensive rebound, offensive rebound, assist, steal, block, turnover, and personal foul. These features mentioned above and machine learning techniques are used to realize the predictive model.

4.2.2. Data-Driven Predictive Models. As the volume of structured and unstructured data in the sports industry, the prediction of the performance of athletes is an important application of sports big data, and machine learning technology is usually used to predict the performance of athletes [75, 78, 82–86]. However, the prediction of the results of sports events is hard task [87]. To learn about the skills of teams, Aoki et al. [87] propose a probabilistic graphical model, and based on this model, they decompose the relative weights of luck and skill in each game. Their experimental results indicate that luck is substantially present in the most competitive championships and give an explanation of why the complex feature-based models hardly beat simple models for predicting the outcome of sports. Bunker et al. [79] propose a predictive framework for predicting the sports results by applying the artificial neural network. The predictive accuracy of their model obtained is higher than the predictive results of traditional mathematical and statistical models. Constantinou and Fenton [80] improve predictive accuracy in the long-term football team performance. Their model can predict the total league points that a team is expected to obtain throughout the season. In their model, a Bayesian network model is used. To predict the results of the NBA game, Thabtah et al. [81] propose an intelligent machine learning framework, framework naive Bayes, artificial neural network, and decision tree algorithms are used. Li et al. [29] leverage a multivariate logistic regression analysis to identify the relationship between the winning probability and the results of the game at the team level. Their experimental results show the efficacy of the predictive model based on the National Basketball Association and Golden State Warriors datasets. Zhu and Sun [77] predict the athletes' performance by using a supporting vector machine algorithm. Their experimental results show that, compared with the current predictive model of athlete performance, the predictive accuracy of their model is more reliable.

5. Open Issues and Challenges

5.1. Predicting the Athlete's Performance in Knowledge Graph. Although researchers have achieved unprecedented results in predicting athlete performance, most of their prediction models focus on feature extraction and machine learning algorithms. In the current research, an issue is that the knowledge relationship between athlete performance and athletes, coaches, and events is ignored. Existing researchers pay more attention to the statistical relationship between them. How to predict athletes' performance more accurately? One possible solution is to predict the performance of athletes based on the knowledge graph of sports big data. Therefore, how to construct a knowledge graph of sports performance and performance-related entities is a crucial task. In addition, how to use the constructed knowledge graph of sports big data to predict the performance of athletes is very challenging.

5.2. Finding Rising Star of Sports. The success of a sports career not only depends on the athlete's personal ability but also is related to the athlete's team and country. For a team or a country, cultivating an outstanding athlete requires a lot of manpower and material resources. The rising star of sports refers to the athletes who are not outstanding among the peers, and they are at the beginning stage of their sports career, but they have a trend of becoming sports stars in the future. Finding a rising star of sports not only provides constructive guidance on the investment of national funds but also provides necessary help for athletes to show excellent performance earlier. However, little is known about how to find the rising star of sports. Current research mostly uses statistical methods for the evaluation of the athletes. How to construct the knowledge graph for finding a rising star is a challenging task.

5.3. Unified Sports Big Data Platform. In the traditional sports system, different sports institutions construct independent sports data platforms according to the needs of their clubs or teams, resulting in data island. It is necessary and crucial to integrate different sports systems and build a unified big service platform for sports big data. Based on the platform, researchers can analyze various relationships between sports entities. Furthermore, via the platform, coaches, athletes, and teams can obtain valuable information on games. Based on this platform, a multidimensional portrait of athletes, coaches, teams, and countries can be made, and accurate services can be provided to them, such as recommending coaches and clubs and identifying rising stars of sports.

5.4. Open Sports Big Data. Sports big data has attracted the attention of domestic and foreign researchers and sports industry builders. Sports big data not only provides sports enthusiasts with various precise services but also guides sports decision-makers to allocate the funds for improving the performance of athletes. However, there are very few shared sports big data for researchers, unlike scholarly data resources such as Google Scholar, Mendeley, and Web of Science. In addition, although different sports institutions and clubs currently have different sports data, these data are mainly in isolation. Therefore, sports big data is open for researchers to help and to promote the vigorous development of sports and achieve the purpose of providing convenient, efficient, and precise service for sportsmen and sports enthusiasts.

5.5. Privacy Protections. In the era of big data, while sports big data brings great value, it also brings some problems in athletes' privacy protection. How to protect athletes' privacy and prevent sensitive information leakage in the process of sports big data development and application has become a new challenge. On the one hand, the personal privacy of athletes requires international sports organizations to establish an independent privacy protection agency; on the other hand, it is necessary to create a special privacy system, the purpose of which is to ensure the priority of privacy of athletes. For athletes' privacy protection, it is necessary to implement fine-grained authority control and to cooperate with relevant data desensitization strategies to better protect athletes' privacy.

6. Conclusion

In this paper, we have provided a comprehensive review of sports big data, focusing on sports big data management, sports big data analysis methods, and sports big data applications. There are several changes in the sports big data field: (1) from simple statistic evaluation to model-based evaluation; (2) from simple statistic analysis to data-driven performance prediction of athletes; (3) from social network analysis to knowledge graph analysis; (4) from explicit sports features to implicit sports features. However, the analysis of the literature on sports big data has led to the conclusion that although researchers have proposed some methods to resolve the problems in sports big data area, the solutions of some crucial issues remain unknown, such as predicting the athletes' performance in the knowledge graph, finding a rising star of sports, unified sports big data platform, open sports big data, and privacy protections.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was partially supported by the Liaoning Provincial Key R&D Guidance Project (2018104021).

References

[1] G. Liu, Y. Luo, O. Schulte, and T. Kharrat, "Deep soccer analytics: learning an action-value function for evaluating soccer players," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1531–1559, 2020.

[2] Z. Haiyun and X. Yizhe, "Sports performance prediction model based on integrated learning algorithm and cloud computing hadoop platform," *Microprocessors and Microsystems*, vol. 79, p. 103322, 2020.

[3] P. Power, H. Ruiz, X. Wei, and P. Lucey, "Not all passes are created equal: objectively measuring the risk and reward of passes in soccer from tracking data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1605–1613, London, UK, 2017.

[4] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports," *ACM Computing Surveys*, vol. 50, no. 2, p. 22, 2017.

[5] L. Pappalardo, P. Cintia, A. Rossi et al., "A public data set of spatio-temporal match events in soccer competitions," *Scientific Data*, vol. 6, no. 1, pp. 1–15, 2019.

[6] D. Patel, D. Shah, and M. Shah, "The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports," *Annals of Data Science*, vol. 7, no. 1, pp. 1–16, 2020.

[7] M. M. Gobble, "Big data: the next big thing in innovation," *Research-technology Management*, vol. 56, no. 1, pp. 64–67, 2013.

[8] M. Du and X. Yuan, "A survey of competitive sports data visualization and visual analysis," *Journal of Visualization*, vol. 56, pp. 1–21, 2020.

[9] Y. Zhang, Y. Zhang, X. Zhao, Z. Zhang, and H. Chen, "Design and data analysis of sports information acquisition system based on internet of medical things," *IEEE Access*, vol. 8, pp. 84792–84805, 2020.

[10] Z. Yin and W. Cui, "Outlier data mining model for sports data analysis," *Journal of Intelligent and Fuzzy Systems*, vol. 22, no. 1–10, 2020.

[11] J. H. Dugdale, D. Sanders, T. Myers, A. M. Williams, and A. M. Hunter, "A case study comparison of objective and subjective evaluation methods of physical qualities in youth soccer players," *Journal of Sports Sciences*, vol. 38, no. 11-12, pp. 1304–1312, 2020.

[12] Q. Yi, M.-Á. Gómez-Ruano, H. Liu et al., "Evaluation of the technical performance of football players in the uefa champions league," *International Journal of Environmental Research and Public Health*, vol. 17, no. 2, p. 604, 2020.

[13] L. V. D. Broucke and S. Baert, "And at the end, the germans always win, don't they? an evaluation of country-specific scoring behaviour in the dying seconds of international club soccer games," *PLoS One*, vol. 14, no. 4, 2019.

[14] S.-U. Park, H. Ahn, D.-K. Kim, and W.-Y. So, "Big data analysis of sports and physical activities among Korean adolescents," *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, p. 5577, 2020.

[15] Y. Liu, "Teaching effect and improvement model of college basketball sports based on big data analysis," *Journal of Physics Conference Series*, vol. 1533, Article ID 042056, 2020.

[16] H. Liu, ""Opportunities, challenges and countermeasures for the development of China's sports industry in the era of big data," *Journal of Physics Conference*, vol. 1237, Article ID 022012, 2019.

[17] E. Tian, "A prospect for the geographical research of sport in the age of big data," *Sport in Society*, vol. 23, no. 1, pp. 159–169, 2020.

[18] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of "big data" on cloud computing: review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2015.

[19] V. S. Tseng, C.-H. Chou, K.-Q. Yang, and J. C. Tseng, "A big data analytical framework for sports behavior mining and personalized health services," in *Proceedings of the 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 178–183, New York, NY, USA, 2017.

[20] K. Fujimoto, T. A. B. Snijders, and T. W. Valente, "Multivariate dynamics of one-mode and two-mode networks: explaining similarity in sports participation among friends," *Network Science*, vol. 6, no. 3, pp. 370–395, 2018.

[21] W. Wang, F. Xia, H. Nie et al., "Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, pp. 1–10, 2020.

[22] W. Wang, J. Liu, Z. Yang, X. Kong, and F. Xia, "Sustainable collaborator recommendation based on conference closure," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 311–322, 2019.

[23] W. Wang, X. Zhao, Z. Gong, Z. Chen, N. Zhang, and W. Wei, "An attention-based deep learning framework for trip destination prediction of sharing bike," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1–10, 2020.

[24] Y. Zhen-Xing, Y. Jun, B. Jie, L. Lin-Xing, C. N. University, and B. University, "Research on the data analysis of NBA based on big data technologies," *China Sport Ence and Technology*, vol. 52, no. 1, pp. 96–104, 2016.

[25] R. Rein and D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," *Springerplus*, vol. 5, no. 1, p. 1410, 2016.

[26] E. Morgulev, O. H. Azar, and R. Lidor, "Sports analytics and the big-data era," *International Journal of Data Science and Analytics*, vol. 5, no. 4, pp. 213–222, 2018.

[27] J. Wang and B. Lv, "Big data analysis and research on consumption demand of sports fitness leisure activities," *Cluster Computing*, vol. 22, no. 2, pp. 3573–3582, 2019.

[28] A. Siddiqa, I. A. T. Hashem, I. Yaqoob et al., "A survey of big data management: taxonomy and state-of-the-art," *Journal of Network and Computer Applications*, vol. 71, no. 1, pp. 151–166, 2016.

[29] Y. Li, L. Wang, and F. Li, "A data-driven prediction approach for sports team performance and its application to national basketball association," *Omega*, vol. 98, 2019.

[30] C. H. Liu, Q. Lin, and S. Wen, "Blockchain-enabled data collection and sharing for industrial iot with deep reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3516–3526, 2019.

[31] X. J. Zhu, *Semi-supervised Learning Literature Survey*, University of Wisconsin-Madison Department of Computer Sciences, London, UK, 2005.

[32] I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study," *Knowledge and Information Systems*, vol. 42, no. 2, pp. 245–284, 2015.

[33] P. P. Talukdar and W. W. An, "Improved evaluation method for soccer player performance using affective computing cohen, scaling graph-based semi supervised learning to large number of labels using count-min sketch," *AISTATS*, vol. 33, pp. 940–947, 2014.

[34] G. Yang, W. Zheng, C. Che, and W. Wang, "Graph-based label propagation algorithm for community detection," *International Journal of Machine Learning & Cybernetics*, vol. 5439, pp. 1–11, 2020.

[35] Y. Ni, J. Chai, Y. Wang, and W. Fang, "A fast radio map construction method merging self-adaptive local linear embedding (lle) and graph-based label propagation in wlan fingerprint localization systems," *Sensors*, vol. 20, no. 3, p. 767, 2020.

[36] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data-AI integration perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 1, no. 1, p. 1, 2019.

[37] C. Zhang, *Deepdive: A Data Management System for Automatic Knowledge Base Construction*, Gradworks, London, UK, 2015.

[38] H. R. Ehrenberg, J. Shin, A. J. Ratner, J. A. Fries, and C. Ré, *Data Programming with Ddlite: Putting Humans in a Different Part of the Loop*, Workshop, London, UK, 2016.

[39] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, and C. Ré, "Snorkel: rapid training data creation with weak supervision," *Proceedings of the Vldb Endowment*, vol. 11, no. 3, 2017.

[40] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, and A. Liu, "Big data cleaning based on mobile edge computing in industrial sensor-cloud," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1321–1329, 2020.

[41] K. Dhinakaran and G. Geetharamani, "A review on big data cleaning and analytical tools," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 9, no. 4, pp. 90–96, 2019.

[42] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, "Holoclean: Holistic data repairs with probabilistic inference," *PVLDB*, vol. 1011 pages, 2017.

[43] M. Dolatshah, M. Teoh, J. Wang, and J. Pei, "Cleaning crowdsourced labels using oracles for statistical classification," *Proceedings of the VLDB Endowment*, vol. 12, no. 4, 2018.

[44] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*, pp. 1–9, Berlin, Germany, 2008.

[45] V. S. Tseng, C.-H. Chou, K.-Q. Yang, and J. C. Tseng, "A big data analytical framework for sports behavior mining and personalized health services," in *Proceedings of the 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 178–183, IEEE, Berlin, Germany, 2017.

[46] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: a survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[47] L. Pan, "A big data-based data mining tool for physical education and technical and tactical analysis," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 22, pp. 220–231, 2019.

[48] Y. Neuman, N. Israeli, D. Vilenchik, and Y. Cohen, "The adaptive behavior of a soccer team: an entropy-based analysis," *Entropy*, vol. 20, no. 10, p. 758, 2018.

[49] W. Fan and A. Bifet, "Mining big data: current status, and forecast to the future," *ACM SIGKDD Explorations Newsletter*, vol. 16, no. 1, pp. 1–5, 2013.

[50] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient machine learning for big data: a review," *Big Data Research*, vol. 2, no. 3, pp. 87–93, 2015.

[51] L. Lei, H. Zhang, and X. Wang, "Adolescent sports behavior and social networks: the role of social efficacy and self-presentation in sports behavior," *Complexity*, vol. 2020, no. 6, pp. 1–10, 2020.

[52] J. M. Lamirán-Palomares, T. Baviera, and A. Baviera-Puig, "Sports influencers on twitter. analysis and comparative study of track cycling world cups 2016 and 2018," *Social Sciences*, vol. 9, no. 10, p. 169, 2020.

[53] M. E. Hambrick, S. H. Schmidt, and A. M. Cintron, "Cohesion and leadership in individual sports: a social network analysis of participation in recreational running groups," *Managing Sport and Leisure*, vol. 23, no. 3, pp. 225–239, 2018.

[54] S. Mclean, P. M. Salmon, A. D. Gorman, N. J. Stevens, and C. Solomon, "A social network analysis of the goal scoring passing networks of the 2016 european football championships," *Human Movement Science*, vol. 57, pp. 400–408, 2018.

[55] B. Gonçalves, D. Coutinho, S. Santos, C. Lago-Penas, S. Jiménez, and J. Sampaio, "Exploring team passing networks and player movement dynamics in youth association football," *PLoS One*, vol. 12, no. 1, Article ID e0171156, 2017.

[56] M. Li, "Big data based outdoor sports monitor and analysis system design of university," in *Proceedings of the 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pp. 245–248, IEEE, London, UK, 2019.

[57] W. Song, M. Xu, and Y. Dolma, "Design and implementation of beach sports big data analysis system based on computer technology," *Journal of Coastal Research*, vol. 94, no. 1, pp. 327–331, 2019.

[58] R. Jiang and Y. Li, "Dynamic pricing analysis of redundant time of sports culture hall based on big data platform," *Personal and Ubiquitous Computing*, vol. 24, no. 1, pp. 19–31, 2020.

[59] Q. Ma, L. Liu, Y. Xie, and F. Lei, "Research on the physical education teacher training system in sunshine sports based on big data platform," *Revista de la Facultad de Ingeniería*, vol. 32, no. 4, pp. 780–787, 2017.

[60] M. Gowda, A. Dhekne, S. Shen et al., "Iot platform for sports analytics," *GetMobile: Mobile Computing and Communications*, vol. 21, no. 4, pp. 8–14, 2018.

[61] C. Deng, Z. Tang, and Z. Zhao, "Countermeasures for the innovative development of China's sports industry under the background of big data," in *Proceedings of the International Conference on Big Data Analytics for Cyber-Physical-Systems*, pp. 1223–1229, New York, NY, USA, 2019.

[62] J. Luo, Z. Wang, L. Xu et al., "Flexible and durable wood-based triboelectric nanogenerators for self-powered sensing in athletic big data analytics," *Nature Communications*, vol. 10, no. 1, p. 5147, 2019.

[63] C. Cuevas, D. Quilón, and N. García, "Techniques and applications for soccer video analysis: A survey," *Multimedia Tools and Applications*, vol. 79, pp. 1–37, 2020.

[64] J. Brooks, M. Kerr, and J. Guttag, "Developing a data-driven player ranking in soccer using predictive model weights," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 49–55, London, UK, 2016.

[65] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, "PlayeRank," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 5, pp. 1–27, 2019.

[66] Y. Li, R. Ma, B. Gonçalves, B. Gong, Y. Cui, and Y. Shen, "Data-driven team ranking and match performance analysis in Chinese football super league," *Chaos, Solitons & Fractals*, vol. 141, 2020.

[67] K. Pelechrinis, E. Papalexakis, and C. Faloutsos, "Sportsnetrank: network-based sports team ranking," 2016.

[68] I. Ghosh, S. R. Ramamurthy, and N. Roy, "Stancescorer: A data driven approach to score badminton player," in *Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1–6, New York, NY, USA, 2020.

[69] W. Liu, X. Xie, S. Ma, and Y. Wang, "Artificial Intelligence and Big Data," *ICAIBD*, vol. 9, 2020.

[70] V. Sarlis and C. Tjortjis, "Sports analytics-evaluation of basketball players and team performance," *Information Systems*, vol. 93, p. 101562, 2020.

[71] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[72] M. M. Adankon and M. Cheriet, "Support vector machine," in *Proceedings of the International Conference on Intelligent Networks and Intelligent Systems*, pp. 418–421, Berlin, Germany, 2010.

[73] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[74] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Berlin, Germany, 2016.

[75] K. Dong-Wook, J. W. Park, and Jae-Hyun'Choi, "A study for the prediction of winning of e-sports using machine learning," *Jounal of the Korea Society of Information Technology Policy & Management*, vol. 9, no. 1, pp. 319–325, 2017.

[76] M. K. Langaroudi and M. Yamaghani, "Sports result prediction based on machine learning and computational intelligence approaches: a survey," *Journal of Advances in Computer Engineering and Technology*, vol. 5, no. 1, pp. 27–36, 2019.

[77] P. Zhu and F. Sun, "Sports athletes' performance prediction model based on machine learning algorithm," in *Proceedings of the International Conference on Applications and Techniques in Cyber Security and Intelligence*, pp. 498–505, London, UK, 2019.

[78] T. Horvat and J. Job, "The use of machine learning in sport outcome prediction: a review," *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, vol. 10, no. 5, 2020.

[79] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied Computing and Informatics*, vol. 15, no. 1, pp. 27–33, 2019.

[80] A. Constantinou and N. Fenton, "Towards smart-data: improving predictive accuracy in long-term football team performance," *Knowledge-Based Systems*, vol. 124, pp. 93–104, 2017.

[81] F. Thabtah, L. Zhang, and N. Abdelhamid, "Nba game result prediction using feature analysis and machine learning," *Annals of Data Science*, vol. 6, no. 1, pp. 103–116, 2019.

[82] B. Xu, "Prediction of sports performance based on genetic algorithm and artificial neural network," *International Journal of Digital Content Technology and Its Applications*, vol. 6, no. 22, pp. 141–149, 2012.

[83] Y. Yue, P. Lucey, P. Carr, A. Bialkowski, and I. Matthews, "Learning fine-grained spatial models for dynamic sports play prediction," in *Proceedings of the 2014 IEEE International Conference on Data Mining*, pp. 670–679, Berlin, Germany, 2014.

[84] H. Kim, "Study on the prediction of the number of spectators and it's factors in pro sports by machine learning method," *The Korean Data Analysis Society*, vol. 21, no. 4, pp. 1867–1880, 2019.

[85] H. Yoon, "The study on the prediction of insolvency of Korean sports industry using machine learning," *The Korean Journal of Physical Education*, vol. 58, no. 6, pp. 165–176, 2019.

[86] T. L. G. Bergkamp, R. J. R. den Hartigh, W. G. P. Frencken, A. S. M. Niessen, and R. R. Meijer, "The validity of small-sided games in predicting 11-vs-11 soccer game performance," *PLoS One*, vol. 15, no. 9, 2020.

[87] R. Y. Aoki, R. M. Assuncao, and P. O. V. de Melo, "Luck is hard to beat: the difficulty of sports prediction," 2017.