Soccer Gameplay Data Generation: Toward Integrating Computer Simulations and Human Sports Analysis

Hidehisa Akiyama[1][0000−0002−9793−9123] and
Tomoharu Nakashima[2][0000−0002−1443−0816]

[1] Okayama University of Science, Okayama, Japan
hidehisa.akiyama@ous.ac.jp
[2] Osaka Metropolitan University, Osaka, Japan
tomoharu.nakashima@omu.ac.jp

Abstract. This paper presents software that aims to generate extensive soccer gameplay data through computer simulations, addressing the scarcity of human-generated data for analysis. It discusses the challenges of analyzing human gameplay data, the need for computer simulations, and the development of software tools such as rcgamestats and rcg2data. The system uses the RoboCup Soccer Simulator to efficiently generate game data. In addition, it presents a case study on distribution analysis of ball interceptions to demonstrate the potential for various analytical purposes. The provided data includes 78,000 matches, which is equivalent to approximately 8,667 human soccer matches. Future efforts include expanding data collection and improving database system integration.

Keywords: Sports analytics · Sports data · Computer simulation

1 Introduction

RoboCup Soccer represents a form of soccer competition driven by autonomous mobile robots. Its primary objectives include the triumph of robotic teams over human champions and the collaborative engagement of robots alongside humans in the sport. To achieve these objectives, advancements in physical control mechanisms and decision-making frameworks embedded within robots (software agents) are necessary to emulate human-level teamwork.

The mechanisms underlying human decision-making in teamwork must be analyzed and comprehended in order to gain insight into human gameplay. With the advent of sophisticated measurement technologies, the analysis of human gameplay data in team sports like soccer has gained significant attraction. A multitude of data providers in the soccer domain have emerged, not limited to the World Cup but extending to leagues worldwide, attempting data collection. Analytical competitions utilizing such data have been organized, proposing various metrics for evaluating gameplay and engaging in benchmark analyses, such as trajectory prediction employing machine learning methodologies.

While the analysis of human gameplay data is becoming increasingly sophisticated, a notable challenge is the need for more data available for analysis. Although partially open data exists, gameplay data provided by data providers comes at a cost, requiring substantial resources for procurement. Moreover, due to the nature of human gameplay data, there are significant constraints on the number of matches that can be executed for the same match-up, and reproducing matches under identical conditions is unfeasible. Consequently, the collection of gameplay data under specific conditions is a practically challenging endeavor.

One approach to mitigate the data scarcity problem is to generate gameplay data through computer simulation. In this study, we provide software for generating a large amount of gameplay data, aggregating match data, and extracting event data using the RoboCup Soccer Simulator. Furthermore, we use these software tools to generate the gameplay data, resulting in the generated data being sufficiently substantial compared to data obtained from human soccer.

The structure of this paper is as follows: Section 2 presents related research and explains the positioning of the software proposed in this paper. Section 3 describes the soccer gameplay data used in the analysis and the software we provide. Section 4 introduces the data we provide. Section 5 describes the case study using the data we provide. Finally, section 6 serves as a conclusion.

2 Related Works

Analysis of human sports primarily revolves around inverse analysis, estimating models from data. Metrics such as shot and pass counts in soccer have long been the subject of data collection and analysis. Recent advancements in measurement and data analysis technologies have facilitated statistical analyses, future predictions, and player evaluations based on such data. Predictive problems in team sports like soccer can be broadly categorized into short-term predictions, spanning a few seconds, and long-term predictions, encompassing events such as shots or decisive moments. However, even short-term predictions in team sports are notoriously challenging, constituting a prominent issue in information science. For instance, numerous studies have focused on the trajectory prediction of players in sports like soccer and basketball [11,12]. Additionally, approaches have been proposed using large-scale datasets to estimate scoring probability and event occurrence probability based on machine learning predictions and evaluate "players' movements that lead to more scores" [10,4]. Algorithms designed for winning games are also employed to conduct long-term predictions. Exploration via such algorithms has been a subject of research, particularly in games such as puzzles and shooters [9]. Such approaches have been attempted even in team sports like soccer. While the RoboCup Soccer Simulator [7,1] serves as a testbed for multiagent systems, algorithm development with a focus on winning strategies is actively pursued within the competition. Conversely, simulators primarily aimed at analyzing group behaviors have also been proposed [5], suggesting a future trend toward not only winning games but also understanding and emulating human players' behaviors. Through such approaches, a deeper analysis

of human gameplay necessitates copious amounts of human-generated gameplay data.

Several open datasets are available for analyzing movement data in human-team sports. Analyzing solely positional data of players and the ball, which is challenging in team sports like soccer, is often complemented with event data, including manually recorded events like shots, passes, and fouls. Well-known data providers such as StatsBomb, WyScout, and Opta offer data for matches ranging from WorldCup to European leagues. Soccer gameplay data typically comprises events involving the ball (such as shots and passes) and the corresponding positional coordinates of players [4]. However, not all match data is publicly available, and freely accessible data is often limited.

Gameplay data in team sports is generated through mechanical extraction via image processing from competition footage, supplemented by human experts' labeling adjustments. Despite advancements in automated extraction through image processing, numerous challenges remain, including player and ball tracking, player identification, and estimation of event occurrence time and type, leading to competitions aimed at evaluating performance in data extraction [3]. Various formats have been proposed for representing gameplay data, with existing structured formats like XML, JSON, or commonly used CSV being prevalent. Efforts toward standardizing representations of soccer gameplay data [4], including player attributes, event classification during games, and organization of event-related information, have led to the development of software environments like kloppy[3]. Currently, positional data used in analyzing human soccer gameplay is represented as 2D point coordinates on a plane (the field surface), lacking posture information. Consequently, the RoboCup Soccer 2D Simulation [7] data aligns closely with this format. Being a computational simulation, it enables the execution of numerous matches under identical conditions, potentially addressing the inherent data scarcity issue in human soccer measurements. Constructing a data generation system utilizing the RoboCup Soccer 2D Simulation and aligning team behavior in the simulator with human teamwork tendencies holds promise for collaborative analysis of human soccer data. While attempts have been made to create datasets using the RoboCup Soccer 2D Simulation in the past [6], there is a demand for a more extensive play data generation system tailored to the format of human soccer data.

3 Data Generation System

3.1 Gameplay Data for Soccer

In the analysis of human soccer, information regarding the movements of players and the ball based on their positions is utilized. Positional data, obtained through external observation such as video footage, is the primary source of information. Since directly measuring players' observations or beliefs is challenging, these kinds of information is generally not employed. Currently, the coordinate

[3] `https://kloppy.pysport.org/`

system used in analyzing human soccer is two-dimensional rather than three-dimensional. That is, it represents players as single points on the field plane without posture information. Although the ball may be airborne in reality, its positional information is recorded only on the plane in the data. Therefore, these position data are highly compatible with the RoboCup Soccer 2D Simulation.

In the analysis of team sports like soccer, tracking data and event data are often analyzed together. Tracking data records the positional information of players and the ball over time, while event data records specific plays such as shots, passes, and fouls. Advancements in technology have facilitated the automatic extraction of tracking data from video footage, enabling the automated estimation of player and ball positions to some extent. However, cases where the ball's positional coordinates are recorded directly from footage are scarce. When the ball's position is not recorded, it is determined by associating player positions with events involving the ball. Automated extraction of event data from tracking data is challenging, necessitating manual annotation. Manual corrections are also required for tracking data since fully addressing occlusions and accurately identifying players pose challenges in automated extraction. While positional extraction using GPS technology is technically feasible, practical examples are rare due to operational costs.

In tracking data, positional information of players and the ball is quantified for each frame of the video footage. Therefore, it suffices to record the positional coordinates of players and the ball linked to the elapsed time since the start of the match. Event data includes information such as the occurrence time, occurrence position, end time, end position, and involved players. Various data recording formats such as CSV, XML, and JSON are commonly used for both tracking and event data.

3.2 Gameplay Data from the RoboCup Soccer Simulator

In this paper, we employ the RoboCup Soccer Simulator (rcssserver) utilized in the RoboCup Soccer Simulation 2D Competition for data generation. rcssserver is a software that realizes soccer through computer simulation. While it operates in a two-dimensional plane, rcssserver is designed as a distributed multi-agent system for 11 vs. 11 gameplay, effectively realizing competition under rules nearly identical to those of human soccer. As rcssserver employs simplified physics, it is unable to simulate human limb control. However, the tactics that occur when players compete for the ball can be imitated to some extent. Consequently, it is capable of emulating human-like group behavior, including interactions between opponents and teammates.

rcssserver is equipped with the capacity to record game logs, which can be utilized to capture the following information:

- Ball: position, velocity
- Players: Players: state (kick, failed kick, tackle, failed tackle, collision, etc.), position, velocity, body orientation, head orientation, pointing direction, field of view, focus position, stamina, player focused by auditory attention, execution count of each command

The game logs contain error-free numerical information about all objects' positions, velocities, postures, and other attributes. Therefore, there is no need for position extraction through image processing, and issues such as missing information or object identification are avoided. Essentially, it provides more comprehensive information than tracking data in human soccer without incurring the costs associated with extraction processes. Moreover, being a computational simulation, it allows for unlimited repetitions of trials given sufficient computational resources.

However, rcssserver does not support the recording of event data. Consequently, a separate process is required to extract events from game logs. The game logs include flags indicating whether each player has touched the ball. Leveraging this information enables the automated extraction of most event data from game logs with minimal manual labeling efforts, bypassing the need for extensive human intervention.

The game logs recorded by rcssserver are in text format, but Lisp-like format is the default. Starting with rcssserver version 18.0.0, game logs are also available in JSON format. In each case, a separate process is required to extract the necessary information from the game logs.

3.3 Distributed Execution System: rcgamestats

To address the need for a large number of matches to generate extensive gameplay data, efficient execution of these matches is essential. In response to this challenge, we have developed a system called rcgamestats for facilitating distributed execution of simulations [4]. rcgamestats offers a mechanism for distributing simulator execution, which significantly reduces the time required for match execution, given the availability of computational resources in parallel. rcggamestats makes it possible to efficiently generate gameplay data on a large scale.

Key features of rcgamestats include:

- Distributed execution with any number of hosts: Matches can be distributed across any number of computational hosts within the same LAN.
- Match scheduling: Users can specify arbitrary combinations of teams from the available pool and designate the number of matches to be executed.
- Aggregation and display of match results: Results from each match are aggregated into a database, and the summary of these results is displayed through a web interface(Fig.1 and 2). The code currently available offers functionality for aggregation and display using Google Spreadsheet as a simplified version(Fig.3).

[4] Available at: `https://github.com/hidehisaakiyama/rcgamestats`

Fig. 1: An example of the aggregated results display. The dedicated web interface allows not only to view the results in table format but also to select the display content using filters and to perform analyses using various indicators.

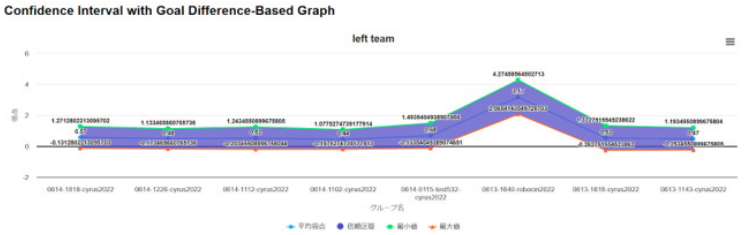


Fig. 2: An example of the visualization. The dedicated web interface provides functions that facilitate graphical visualization and simple statistical analysis.

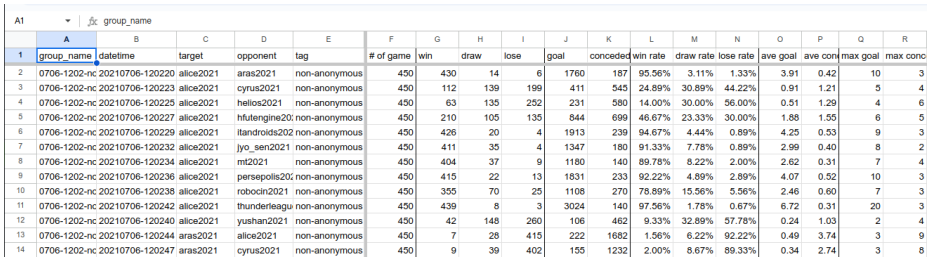
3.4 Data Converter and Extractor: rcg2data

To facilitate the analysis of the game logs recorded by rcssserver and to extract event information in a format suitable for analysis, we have developed a software tool called rcg2data [5]. rcg2data provides two main functionalities:

- Conversion of game logs recorded by rcssserver into CSV format.
- Extracting event information from the game logs recorded by rcssserver. In the current implementation, pass, interception, and shot events can be extracted from the game log and saved in CSV format.

Using rcg2data makes it possible to generate data equivalent to those commonly used in human soccer analysis. Figure 4 shows an example of event data extracted from a game log file. The figure shows a list of plays where the ball was passed, or the opposing team intercepted the ball. The figure shows a list of plays in which the ball was passed to a teammate or intercepted by an opponent player.

[5] Available at: `https://github.com/hidehisaakiyama/rcg2data`



group_name	datetime	target	opponent	tag	# of game	win	draw	lose	goal	conceded	win rate	draw rate	lose rate	ave goal	ave conc	max goal	max conc
0706-1202-nc	20210706-120220	alice2021	aras2021	non-anonymous	450	430	14	6	1760	187	95.56%	3.11%	1.33%	3.91	0.42	10	3
0706-1202-nc	20210706-120223	alice2021	cyrus2021	non-anonymous	450	112	139	199	411	545	24.89%	30.89%	44.22%	0.91	1.21	5	4
0706-1202-nc	20210706-120225	alice2021	helios2021	non-anonymous	450	63	135	252	231	580	14.00%	30.00%	56.00%	0.51	1.29	4	6
0706-1202-nc	20210706-120227	alice2021	hfutengine20	non-anonymous	450	210	105	135	844	699	46.67%	23.33%	30.00%	1.88	1.55	6	5
0706-1202-nc	20210706-120229	alice2021	itandroids202	non-anonymous	450	426	20	4	1913	239	94.67%	4.44%	0.89%	4.25	0.53	9	3
0706-1202-nc	20210706-120232	alice2021	jyo_sen2021	non-anonymous	450	411	35	4	1347	180	91.33%	7.78%	0.89%	2.99	0.40	8	2
0706-1202-nc	20210706-120234	alice2021	mt2021	non-anonymous	450	404	37	9	1180	140	89.78%	8.22%	2.00%	2.62	0.31	7	4
0706-1202-nc	20210706-120236	alice2021	persepolis20	non-anonymous	450	415	22	13	1831	233	92.22%	4.89%	2.89%	4.07	0.52	10	3
0706-1202-nc	20210706-120238	alice2021	robocin2021	non-anonymous	450	355	70	25	1108	270	78.89%	15.56%	5.56%	2.46	0.60	7	3
0706-1202-nc	20210706-120242	alice2021	thunderleagu	non-anonymous	450	439	8	3	3024	140	97.56%	1.78%	0.67%	6.72	0.31	20	3
0706-1202-nc	20210706-120240	alice2021	yushan2021	non-anonymous	450	42	148	260	106	462	9.33%	32.89%	57.78%	0.24	1.03	2	4
0706-1202-nc	20210706-120244	aras2021	alice2021	non-anonymous	450	7	28	415	222	1682	1.56%	6.22%	92.22%	0.49	3.74	3	9
0706-1202-nc	20210706-120247	aras2021	cyrus2021	non-anonymous	450	9	39	402	155	1232	2.00%	8.67%	89.33%	0.34	2.74	3	8

Fig. 3: An example of displaying the aggregated results using Google Spreadsheet.

While it is difficult to analyze team play using tracking data alone, it is possible to analyze it using information from such events.

Type	Side1	Unum1	Time1	X1	Y1	Side2	Unum2	Time2	X2	Y2	Success
Pass	left	10	105	11.0433	32.3346	left	5	113	-2.7	30.9469	true
Pass	left	5	125	4.4467	29.5977	left	10	128	9.2386	31.489	true
Pass	left	10	143	16.6419	31.1778	left	5	152	0.0918	26.3957	true
Pass	left	5	161	4.2961	25.8656	left	6	169	2.7422	12.9022	true
Interception	left	6	173	4.7946	15.0694	right	7	174	6.834	15.8583	true
Pass	right	7	176	6.5191	16.1724	right	9	182	3.9038	24.3288	true
Pass	right	9	183	3.6503	24.2566	right	7	187	3.1349	19.623	true
Pass	right	7	188	3.1646	19.682	right	9	191	3.0246	25.26	true
Pass	right	9	204	-7.1198	24.3199	right	11	208	-10.6632	17.1693	true
Pass	right	11	209	-10.9034	17.1474	right	9	212	-8.9923	22.7204	true
Pass	right	9	212	-8.9923	22.7204	right	11	216	-12.6582	18.0909	true
Pass	right	11	227	-16.2424	16.4082	right	6	234	-15.0141	2.6168	true
Pass	right	6	234	-15.0141	2.6168	right	8	243	-13.1596	-17.0355	true
Pass	left	7	297	-24.0591	-22.1905	left	6	301	-19.5139	-14.1352	true
Pass	left	6	310	-14.5264	-14.1491	left	7	312	-16.7063	-18.5533	true

Fig. 4: An example of event data extracted from a game log file using rcg2data.

4 Provided Data

We generated actual gameplay data using the developed system. A total of 1,000 round-robin matches were played by teams participating in the RoboCup 2021 Soccer Simulation 2D Competition. Thirteen teams were involved in the matches: Alice2021, CYRUS, FRA-UNIted, HELIOS2021, HfutEngine2021, ITAndroids, Jyo_sen2021, MT2021, Oxsy, Persepolis, RoboCin, ThunderLeague, and YuShan2021. In 2022, the specifications of the simulator were modified to prohibit players from performing backward dashes. Consequently, some teams were unable to move as intended in RoboCup 2022 and RoboCup 2023. Therefore, we used teams from RoboCup 2021 for this study. This resulted in a total of 78,000 matches.

The data volume for the game log files of 78,000 matches was approximately 500GB after zip-compressed. The match results and team binary information for RoboCup 2021 are available on the simulator's official website

[6]. The generated data is available at `https://github.com/hidehisaakiyama/RoboCup2D-data/`.

A match played by rcssserver lasts about 10 minutes, which is one-ninth the length of a human soccer match. Therefore, the generated data is gameplay data equivalent to about 8,667 human soccer matches. It can be said that the gameplay data is sufficiently large compared to that obtained from human soccer data providers.

5 Case Study: Distribution Analysis of Ball Interception

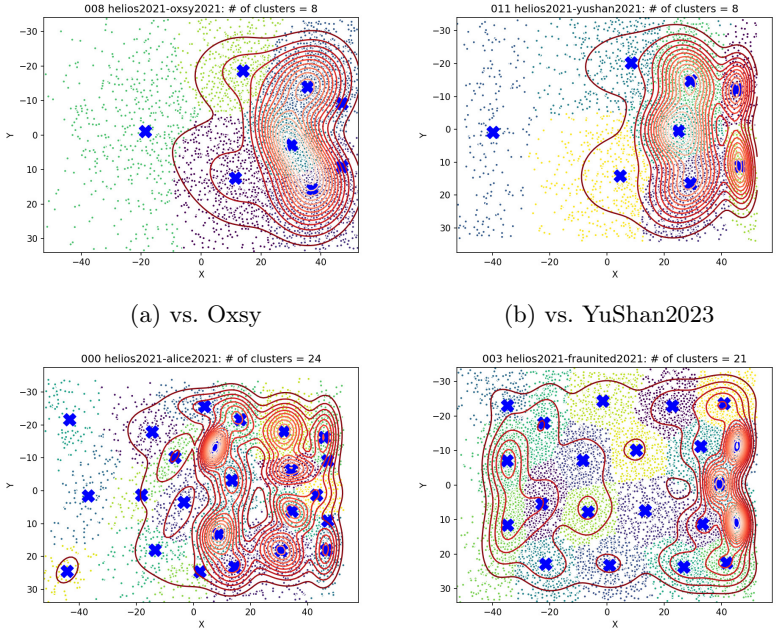
We use the interception information from the event data to estimate the probability distribution of ball interceptions for each matchup combination. While kernel density estimation is a common method for estimating such distributions, we opt for a Gaussian Mixture Model in this study. Since the goal is to analyze the similarity of distributions, we chose a Gaussian Mixture Model, which tends to have clustering characteristics, making it suitable for analyzing the similarity of distributions. Furthermore, to analyze the similarity of the estimated distributions, we perform hierarchical clustering based on distribution distances. We use the Earth Mover's Distance (EMD) [8] as the distance metric between distributions.

Here, we present partial results of the clustering performed for games between HELIOS2021, the runner-up of RoboCup2021, and other teams. Figure 5 shows contour plots of the probability distribution estimation. From these figures, it can be observed that the distribution of ball interception positions exhibits distinctive characteristics for each team, suggesting a relationship between the distribution features and the team's defensive strategies. Figure 6 shows the dendrogram resulting from hierarchical clustering. The results suggest that the team at the bottom (FRA-UNIted) employs a unique defensive strategy, while the top two teams (YuShan2023 and Oxsy), though also distinctive, likely use relatively similar defensive strategies. Although no quantitative evaluation has been made, it is suggested that the characteristics appearing in the distribution and the clustering results may be similar to human observations.

6 Conclusion

This paper proposes and implements software for generating a large volume of soccer gameplay data using computer simulation. Furthermore, in the pursuit of merging human soccer with computer simulation, software has been developed to organize data compatible with human soccer. Using the implemented software and team binaries from RoboCup2021, actual data generation was performed and made openly available. Through simple experiments with the generated data, the potential for various analytical purposes was demonstrated.

[6] `https://rcsoccersim.github.io/`



(a) vs. Oxsy (b) vs. YuShan2023

(c) vs. Alice2023 (d) vs. FRA-UNIted

Fig. 5: Results of estimating the probability distribution using interception information obtained from event data between HELIOS2021 and each team. Each uses event information extracted from data for 500 matches in which HELIOS2021 was on the left side.

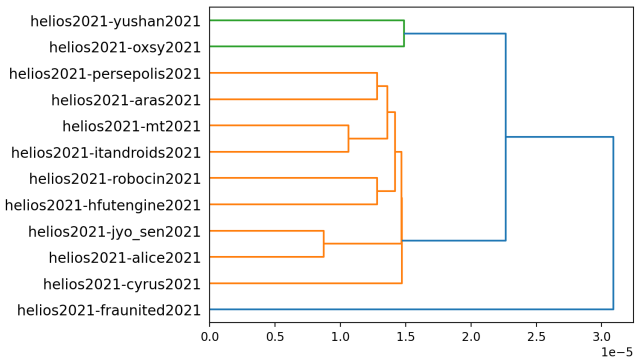


Fig. 6: Dendrogram showing the results of hierarchical clustering using EMD as a distance metric.

Future challenges include preparing agents that mimic human play, collecting data from a broader range of match-ups, and improving integration with database systems.

References

1. Akiyama, H., Dorer, K., Lau, N.: On the Progress of Soccer Simulation Leagues, RoboCup 2014: Robot World Cup XVIII. RoboCup 2014. Lecture Notes in Computer Science, vol 8992. pp. 599–610. (2015)
2. Akiyama, H., Nakashima, T.: HELIOS Base: An Open Source Package for the RoboCup Soccer 2D Simulation, RoboCup 2013: Robot World Cup XVII. Lecture Notes in Computer Science, vol 8371. pp. 528–535, (2014)
3. Cioppa, A., Giancola, S., Deliege, A. and Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M.: SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos, Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3491–3502, (2022)
4. Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: Actions speak louder than goals: Valuing player actions in soccer, in KDD, pp. 1851–1861, (2019)
5. Kurach, K., Raichuk, A., Stańczyk, P., Zając, M., Bachem, O., Espeholt, L., Riquelme, C., Vincent, D., Michalski, M., Bousquet, O., Gelly, S., "Google Research Football: A Novel Reinforcement Learning Environment", Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), pp. 4501–4510. (2020)
6. Michael, O., Obst, O., Schmidsberger, F., Stolzenburg, F.: RoboCupSimData: Software and Data for Machine Learning from RoboCup Simulation League, RoboCup 2018: Robot World Cup XXII. Lecture Notes in Computer Science, vol 11374. (2019)
7. Noda, I., Matsubara, H., Hiraki, K, Frank, I.: Soccer Server: A Tool for Research on Multiagent Systems, Applied Artificial Intelligence 12, no. 2–3, pp.233—250. (1998)
8. Rubner, Y., Tomasi, C., Guibas, L. J.: A metric for distributions with applications to image databases In IEEE International Conference on Computer Vision, pp. 59–66. (1998)
9. Silver, D. et al.: Mastering the game of Go with deep neural networks and tree search. Nature, vol. 529, no. 7587, pp. 484-489. (2016)
10. Spearman, W.: Beyond expected goals, In Proceedings of the 12th MIT Sloan Sports Analytics Conference, pp. 1–17. (2018)
11. Teranishi, M., Tsutsui, K., Takeda, K., Fujii, K.: Evaluation of Creating Scoring Opportunities for Teammates in Soccer via Trajectory Prediction, Machine Learning and Data Mining for Sports Analytics. MLSA 2022. Communications in Computer and Information Science, vol 1783. (2023)
12. Yeh, R. A., Schwing, A. G., Huang, J., Murphy, K.: Diverse Generation for Multi-Agent Sports Games, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4605–4614 (2019)