Paper : Game Analysis

Construction of Offensive Play Measurement Items and Shot Prediction Model Applying Machine Learning in Japan Professional Football League

Hirotaka Jo[*,], Hiroki Matsuoka[***], Kozue Ando[***] and Takahiko Nishijima[***]**

[*]Department of Sports Science, Shizuoka Sangyo University
1572-1 Oowara, Iwata-city, Shizuoka-pref 438-0043 Japan
[**] Doctoral Program in Health & Sport Sciences, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8574 Japan
[***] Faculty of Health & Sport Sciences, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8574 Japan
jotk.lab@gmail.com

[Received May 10, 2021; Accepted October 11, 2021]

The demand for sports analytics is increasing because it contributes to the victory of competitive sports. Although the technology for automatically measuring tracking data has improved, it is meaningful to systematize analysis using ball touch data because the stadiums that can be used are limited. Moreover, in recent years, analysis using machine learning has been increasing, and it is necessary to accumulate research. With this background, this study aimed to construct measurement items and a model for predicting to shoot in soccer offensive play by applying machine learning. Using the Delphi method, 6 items were deleted from the old measurement item group in the previous study, 11 items were newly created, and 45 items were set as the new measurement item group. A decision tree, random forest, and gradient boosting decision tree were applied to the new measurement item group, and several shot prediction models were constructed. The results indicated that the model using 23 items in the gradient boosting decision tree was the best. Furthermore, a comparison between the old and new measurement item groups revealed that the prediction accuracy of the new measurement item group was higher. In conclusion, 45 measurement items of soccer offensive play were constructed, and the shot prediction model using them was constructed by applying machine learning.

Keywords: football, offensive play, measurement items, machine learning, shot prediction

[Football Science Vol.19, 1-21, 2022]

1. Introduction

Sports analytics is the process of managing data, implementing predictive models, and using information systems for decision making, which contributes to competitive advantage in competitive sports (Alamar and Mehrotra, 2011). In other words, in competitive sports, measuring, managing, and analyzing performance during a game can provide useful insights for practice and training plans, as well as tactical and strategic planning for the next game, thereby leading to victory. This concept has been put into practical use for many years, with one of the most famous examples being Sabermetrics (James, 2003). Sabermetrics is a baseball analysis system developed in the 1970s. It is famous for the story of a team that became a regular contender for the regional championship by reinforcing its hitters based on the analysis result that the on-base percentage was a better indicator of hitting ability than the batting average. Currently, sports analytics is not limited to baseball but has become popular in many other fields. In soccer, Germany won the 2014 World Cup in Brazil by actively using data analytics (Dmonte and Dmello, 2017), and it was said that “the analytics system was the 12th player.”

In ball games such as soccer, the data used for analysis can be divided into two types: ball touch data and tracking data. Ball touch data is related to the action of the player with the ball, which cannot be

fully measured by an automatic tracking device. It is measured by an expert measurer who has undergone a certain amount of training by watching videos (Kato, 2016). Since the data are recorded every time a player takes an action, ball touch data are also called event data. Tracking data can automatically or semi-automatically measure the positions of 22 players on the pitch if automatic tracking devices are available in the match. The Japan Professional Football League (J.League) uses the TRACAB system (ChryonHego, NY, USA), which records position data at 25 frames per second (Linke et al., 2020).

Since tracking data include the positional information of players other than the ball carrier, it is thought to greatly expand the scope of analysis of collective skills, and research has been conducted on the J.League players (Matsuoka et al., 2020). However, as automatic tracking devices are specialized equipment with high installation costs, they are only available for national teams and at the professional level. In addition, such devices cannot be used in national or professional team stadiums that have not been designed for the installation of automatic tracking systems. Therefore, regardless of the type of sport, measurement and analysis using video images (Suzuki et al., 2019; Takahashi and Haseyama, 2017) are still conducted, and many applications and analysis programs (Ekin et al., 2003) have been developed for this purpose. This suggests that devising a method to analyze game performance using only ball touch data, without relying on tracking data, is considered significant enough in current competitive sports.

Data sources, analysts, and analysis methods are key elements for good sports analytics. Data sources have developed to the point where they are now referred to as “big data in sports” (Dmonte and Dmello, 2017), as the volume of information has grown enormously due to improvements in measurement accuracy and automation technology in addition to lower measurement costs (Tsuchida and Yadohisa, 2020). As for analysts, the role of sports analyst has been recognized in Japan since around 2010 (Chiba, 2020). Now, analysts are present not only in national teams and at the professional level but also in university sports, as exchanges transcending the boundaries of categories and organizations have been promoting the development of analysts (Sakaori, 2016). In the meantime, as a result of the transition of measured information into big data, conventional methods are not thought to be scalable in terms of analysis methods (Wang et al., 2015). In recent years, machine learning methods have been applied to sports data analysis (Tamura, 2020). In a study on the J.League, Matsukoka et al. (2020) applied neural networks to tracking data to develop measures for analyzing defensive and offensive plays. In addition, Jo et al. (2014) applied decision tree analysis to develop a skill rating scale for players and teams. However, since research on analysis using machine learning has a short history, it is still necessary to study, develop, and accumulate systematized analysis methods at the academic level.

Among the many machine learning algorithms, neural networks, which belong to deep learning, have been noted for their high accuracy. However, neural network algorithms have the black box problem in that the prediction process is extremely complex for human intelligence to understand. Despite many attempts at improvement (Morinaga, 2015), no clear solution exists to this problem. By contrast, decision tree analysis, classified as supervised learning and is represented by CART (Breiman et al., 1984), is characterized by the simplicity of interpretation of results due to the visual verifiability of the prediction or classification process. Since sports analytics requires the utilization of analysis results in a series of tasks (Dmonte and Dmello, 2017), it is difficult to reflect the results in practice if the prediction process is a black box in on-field feedback of analysis results in soccer. Therefore, decision tree analysis, which has a relatively clear prediction process, is considered to be useful for feedback to the sports field.

Decision tree analysis is an algorithm for prediction and classification, but the success or failure of model development largely depends on how the objective and predictor variables are established. The ultimate goal of a soccer match is to win the match, and for this purpose, strikers must aim to score, and to score goals, they must take shots (Jo et al., 2017). Based on this, “outcome,” “goal,” and “shot” may be the objective variables. Kumar (2013) reported the construction of a prediction model, with the outcome of matches as the objective variable. However, the construction of a prediction model with goals and shots as the objective variables is still unexplored. The ability to score goals highly depends on the shooting skill of the attacking player. As it is difficult to measure shooting skills with ball touch data, it is difficult to construct a prediction model of scoring

with ball touch data alone. By contrast, a model that predicts whether a player will take a shot may be constructed by processing the ball touch data from the beginning to the end of the game and by incorporating tactical and technical phenomena into the variables.

Based on the above considerations, this study intended to construct a model to predict whether a player would take a shot by listing measurement items of offensive plays based on soccer ball touch data and applying a machine learning method similar to the decision tree analysis. For this purpose, the following hypotheses were considered:

Hypothesis 1: The measurement items of offensive plays are created based on ball touch data.

Hypothesis 2: A machine learning model is constructed to predict whether a player will take a shot based on the measurement items of offensive plays by applying a method similar to the decision tree analysis.

2. Method

2.1. Definition of terms

The principal terms used in this study are defined as follows:

(1) Measurement items

Although "measurement items" in the field of measurement and evaluation, "variables" in the field of statistics and data analysis, and "features" in the field of machine learning are referred to by different terms depending on the field, they all have similar characteristics. In this study, they are considered to be synonymous and hereinafter referred to as "measurement items."

(2) Action

In ball touch data, actions refer to the individual actions that players perform during a game, such as passes, traps, dribbles, and crosses.

(3) Play

A play refers to a series of successive actions taken by a team in possession of the balluntil the said team loses the ball or takes a shot. For instance, if the sequence of actions from kick-off to trapping, passing, trapping, dribbling, passing, and ball out of play (when the ball leaves the field of play) takes place, a play comprises the actions from kick-off to ball out of play, with the six elements except for the last ball out of play being actions.

(4) Action data

A dataset in which each line consists of one action.

(5) Play data

A dataset in which action data is transformed in such a way that each row consists of one play.

2.2. Research procedure

This study was conducted in accordance with the following procedure:

(1) The new measurement items in this study (hereinafter referred to as the new set of measurement items) were established in accordance with the measurement items in the study by Jo et al. (2014) (hereinafter referred to as the old set of measurement items).

(2) The action data were converted to play data based on the definition of the new set of measurement items.

(3) Decision trees, random forests, and gradient boosting decision trees were analyzed using the play data to validate the best prediction model (hereinafter referred to as the "best model").

(4) The play data of the old set of measurement items were analyzed using the same methods and conditions as the best model, and the prediction accuracy of the old and new sets of measurement items were compared.

2.3. Target of analysis

Offensive plays in all 686 J.League Division 1 (J1) and Division 2 (J2) matches played in 2011 were analyzed. However, as set pieces are sometimes performed with special attacking patterns, plays that started with corner kicks, penalty kicks, and free kicks (both direct and indirect) were excluded. In this season, J1 comprised 18 clubs with 34 sections, J2 comprised 20 clubs with 38 sections, and 1,051 players were registered. The size of the action data used in this study was 207 columns by 1,312,117 rows, which were measured by trained staff of Data Stadium Inc. while watching the video footage of the matches. Typical measurement items included the match information such as the date of the match and the names of the teams playing the match, in addition to actions taken by the players and their positions. Every time a player took an action such as a kick-off, dribble, pass, trap, header, shot, cross, throw-in, free kick, goal kick, tackle, foul, throw-in, or goalkeeper catch was recorded.

2.4. Analysis method

2.4.1 Composition of measurement items

Jo et al. (2014) applied the Delphi method (Linstone and Turoff, 1975), a technique for gathering opinions from a group of soccer experts, to develop 40 items to measure the attacking skills demonstrated in soccer matches (referred to as the old set of measurement items in this study). The authors developed an index to measure the attacking skills of players and teams (the old set of measurement items is listed with the new measurement items in **Table 1** below). Here, decision tree analysis was used with the binary data. Whether a player took a shot (hereafter referred to as the measurement item "shot") was the objective variable and a single measurement item was the predictor variable to identify the bifurcation value that best determines whether the player would take a shot. The bifurcation value was treated as a play achievement criterion for taking a shot, and its degree of contribution for taking a shot was identified based on the shooting probability after the bifurcation. Subsequently, Jo et al. (2017) developed an algorithm to optimize offensive plays using the same measurement items. Based on these previous studies, the old set of measurement items was judged to be versatile, and thus, it was used as a reference in this study. However, some measurement items needed to be newly created or unused. Therefore, in the initial stage of this study, the measurement items were reconstructed by applying the Delphi method in accordance with the method of Jo et al. (2014). The old measurement items were carefully examined one by one, and they were not used where there was a rational reason to discard them. In addition, the factors that might affect the possibility of taking a shot were re-examined to create new measurement items that were not in the old set of measurement items. The group of experts comprised one university faculty member specializing in evaluation and measurement in health and physical education; one soccer coach with experience in the professional and the first division levels; and two active professional soccer players. The Delphi method consisted of five 90-minute meetings over a two-month period. During this period, other soccer professionals were asked for their opinions as needed when the expert group alone could not make a decision.

2.4.2. Conversion to play data

Ball touch data are difficult to analyze in its raw form, and even when the amount of data is large, it is difficult to gather useful insights from them. Therefore, analysts need to examine complex data structures and address issues such as dealing with sparsity and scaling a large number of different data measures (Decroos, 2020). One of the objectives of this study was to build a shot prediction model. As the prediction targets play rather than act, play data in which each row consists of one play were prepared. Since the ball touch data measured by Data Stadium, Inc. were action data, the data were converted into play data based on the definition of each item of the new set of measurement items in accordance with the procedure of Jo et al. (2014).

2.4.3. Validation of the prediction model

Since players need to take shots at the goal to score in soccer, constructing a model is crucial to predict whether a player can take a shot. This study adopted decision tree analysis as a method for constructing a prediction model. There are methods called random forests (Breiman, 2001) and gradient boosting decision trees (Friedman, 2001), which are extensions of decision trees. Both these methods are a type of ensemble learning and are known for their high prediction accuracy. Random forests remove variances, and gradient boosting decision trees method eliminates biases. Few studies have analyzed sports performance data using ensemble learning methods. Thus, the possibility of improving the prediction accuracy through ensemble learning remains unknown. Therefore, this study employed random forests and gradient boosting decision trees as methods for building prediction models to determine the best method.

When constructing a machine learning model, the measurement items should be carefully examined, as the performance of the model is largely determined by the ability to prepare measurement items that contribute to the prediction results. However, more measurement items will require more processing time, causing the interpretation of the results to become more complex. Therefore, the number of measurement items should not be increased blindly. It is necessary to construct a model that retains only the items that contribute to the prediction results and excludes unnecessary measurement items. Accordingly, a prediction model was constructed

by reducing the number of measurement items step by step using three methods: decision trees, random forests, and gradient boosting decision trees. From among the multiple models generated in the process, the best model was determined based on the two perspectives of high prediction accuracy and small number of measurement items. The procedure is described below (**Figure 1**).

The analysis software used in this study was R (R Core team, 2020), a popular tool in sports analytics due to its open source development and the availability of additional features (packages) specific to many fields (Miller, 2015). First, the set of playing data was divided into training data (n=102,922) and validation data (n=44,110) in a ratio of 7:3 through random sampling. Subsequently, the training and validation data were replicated (retaining the originals) for the analysis of the three methods.

(1) Decision tree

For the decision tree method, the rpart function of R's rpart package (Therneau and Atkinson, 2019) was used. The objective variable was the measurement

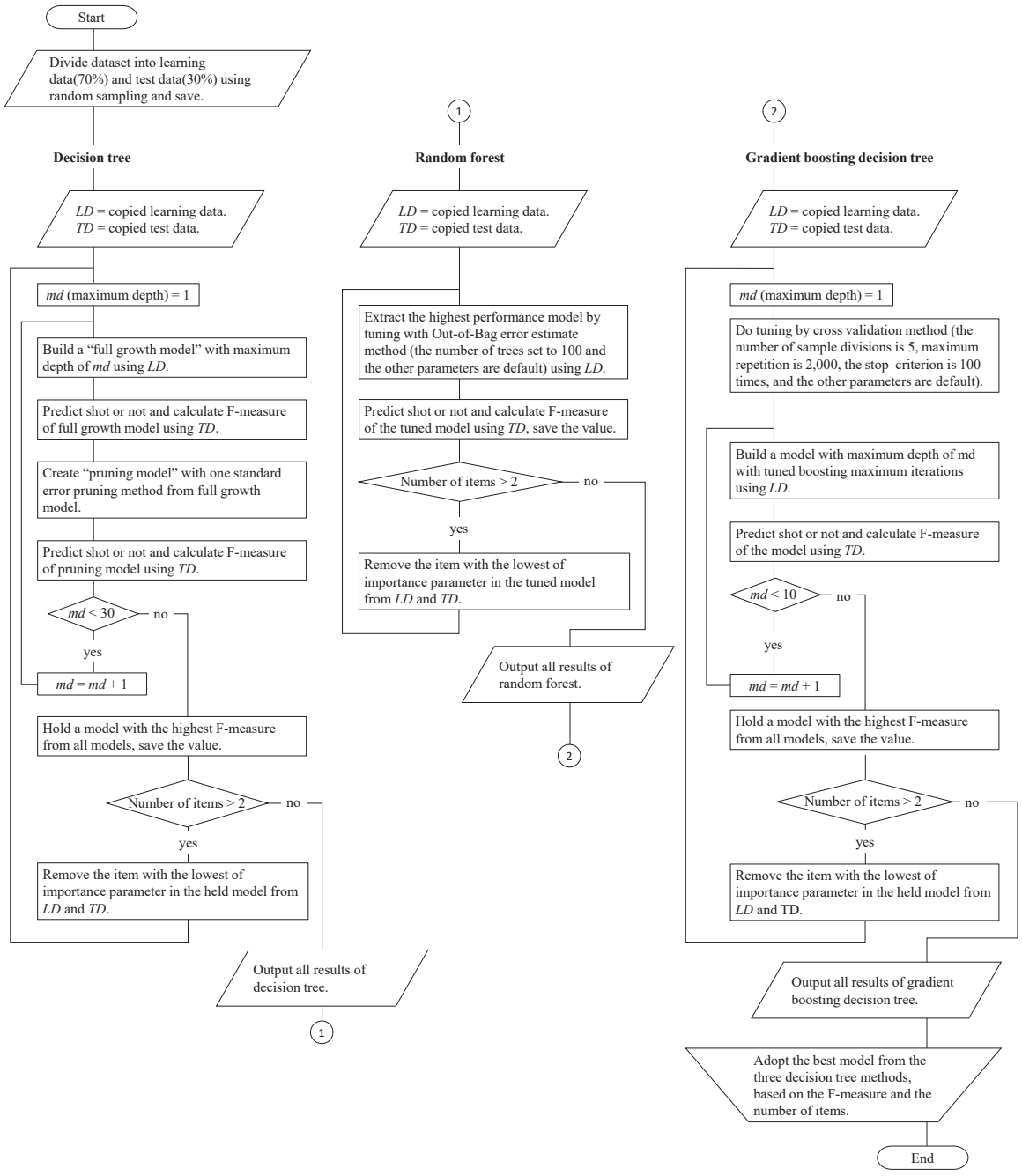

Figure 1 Flow chart of predictive model construction

item "shot" (binary data indicating whether a player scored a goal or not), and the predictor variables were all the remaining measurement items. A "full growth model" without pruning and a "pruning model" with pruning based on one standard deviation (Shimokawa et al., 2013) were constructed. In both cases, validation data were applied to predict whether a player would take a shot. A confusion matrix was developed using the measured values of the measurement item "shot" and the predicted values of whether a player would take a shot. The prediction accuracy was evaluated by calculating the f-measure based on the precision and recall rates after confirming the accuracy. The full growth and pruning models were iteratively constructed from a maximum depth of 1 to 30, and the model with the highest f-measure was reserved from a total of 60 models to store its value. To remove unnecessary measurement items that did not contribute to the prediction results, the items with the lowest importance (Breiman et al., 1984) in the reserved models were removed from the training and validation data, and the same process was repeated until two items were left.

(2) Random forest

For the random forest method, the tuneRF function of R's randomForest package (Liaw and Wiener, 2002) was used. The number of measurement items in a tree is one of the important hyperparameters in a random forest, and the tuneRF function uses the out-of-bag (OOB) error estimation method to tune the optimal number of items. In performing the tuneRF function, the number of trees to be created was set to 100. Other hyperparameters were set to default to extract the model with the best performance based on the f-measure. The measurement items used in the obtained model with the lowest importance (Breiman, 2001) were eliminated, and the analysis was repeated until only two items remained.

(3) Gradient boosting decision tree

For the gradient boosting decision tree method, R's xgboost package (Chen et al., 2020) was used. We tuned the maximum number of boosting iterations by running the cross validation method with the number of dataset partitions set to 5, the maximum number of iterations set to 2,000, the stopping criterion set to 100 times, and other hyperparameters set to default. Trees with maximum depths from 1 to 10 were created using the tuning results, and the model with the highest f-measure was reserved. As in the previous two methods, the measurement items with the lowest importance (Friedman, 2001) of the reserved models were deleted, and the analysis was repeated until only two items were left.

Eventually, models with predictor variables ranging from 44 to 2 (129 models in total) were created for the decision tree, random forest, and gradient boosting decision tree. From among these, the best model was determined based on the value of the f-measure and the number of predictor variables.

The evaluation metrics of machine learning models generally include accuracy, precision, recall, and the f-measure. A confusion matrix was created based on the measured values of the measurement item "shot" and the predicted results of the model. The true positive (TP), false positive (FP), true negative (TN), and false negative (FN) were as follows:

True positive: The model correctly predicted that the play "would end in a shot" when the actual value indicated a "shot."

False positive: The model incorrectly predicted that the play "would end in a shot" when the actual value did not indicate a "shot."

True negative: The model correctly predicted that the play "would not end in a shot" when the actual value did not indicate a "shot."

False negative: The model incorrectly predicted that the play "would not end in a shot" when the actual value indicated a "shot."

Accuracy is calculated as (TP+TN)/(TP+FP+TN+FN) and represents the probability of correctly predicting whether a player takes a shot. In soccer, the number of plays that do not result in a shot is much higher, and 91.6% of the plays in this study did not end in a shot. Thus, when the data is heavily biased toward negative (or positive), accuracy tends to be inevitably high; therefore, other evaluation metrics such as precision and recall should also be consulted.

Precision is calculated as TP/(TP+FP) and represents the probability that a player actually takes a shot in the play for which the model predicted the player would take a shot. Recall is calculated as TP/(TP+FN) and indicates the probability that the model correctly predicted that a play would end in a shot. Both fractions evaluate the predictive accuracy of true positives; however, precision has the disadvantage of not being able to judge the accuracy of a negative prediction at all, even if the model produces a negative prediction that "a player will not take a shot." Recall has the disadvantage of not being able to judge the success or failure of predictions for the

measured value that "a player did not take a shot." Therefore, the f-measure, which is the harmonic mean of precision and recall, was calculated and used as an evaluation index of the prediction accuracy of the model.

2.4.4. Comparison with the old set of measurement items

In this study, since some of the old measurement items were deleted and new items were added, it was necessary to verify which ones were superior. To this end, a model for the old set of measurement items was developed using the methods and conditions used in the best model and compared with the new set of measurement items. However, two items, "X-coordinate of the last action" and "Primary area," which were discarded from the new set of measurement items, were also removed from the old set of measurement items, and 37 old measurement items were analyzed as predictor variables.

2.5. Ethical Considerations

The data for this study were purchased from Data Stadium Inc., and usage permission was obtained. In addition, this study was conducted with the approval of the Research Ethics Committee at the Faculty of Health and Sport Sciences, University of Tsukuba (Approval No.: 30-28).

3. Results

3.1. Developed measurement items

The Delphi method was applied by a group of soccer experts. After reviewing the old set of measurement items, five items, namely "X-coordinate of the last action," "Primary area," "Mean of moving direction," "Triple speed," and "The number of tactical attackers" were deleted. Eleven new items, namely "Success rate of pass," "Success of through-ball," "Success of cross," "Flick-on," "Throw-in," "Feed," "Mean of pass distance," "Standard deviation of pass distance," "Forward propulsion," "Wide propulsion," and "Same direction" were created. The old measurement items "Turn-back of sideward" and "Turn-back" were merged into one measurement item. Finally, 45 items, including the objective variable "shot," were developed as the new measurement items (**Table 1**).

All measurement items were quantified on a per-play basis. "Dribble," "Pass," "Success of pass," "Success rate of pass," "Direct pass," "Consecutive direct pass," "Through-pass," "Success of through-pass," "Cross," "Success of cross," "Trap," "Rebound-ball," "Flick-on," and "Feed" measure the number of times the action occurs in a play and its success rate. "Throw-in" indicates binary data recorded as 1 when the first action of the play is a throw-in. "Total of attack actions" is the total number of attack actions in one play. "Duration of attack" represents the time taken from the beginning to the end of the play (in seconds). "Average time of attack action" indicates the average time taken per action (in seconds). "Number of attackers" is the number of attackers involved in the play.

"Total distance: TD," "Total vertical distance: TVD," "Total horizontal distance: THD," "Distance: D," "Vertical distance: VD," "Horizontal distance: HD," "Maximum vertical distance: MVD," "Maximum horizontal distance: MHD," "Mean of pass distance: MPD," and "Standard deviation of pass distance: SDPD" are measurement items related to the distance traveled by the ball (**Figure 2**). These are calculated using the following formula based on the coordinates where each action has occurred.

$$TD=\sum_{i=1}^{n-1} d_i \quad \cdots \text{Formula ①}$$

$$d_i=\sqrt{(x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2}$$

$$TVD=\sum_{i=1}^{n-1} v_i \quad \cdots \text{Formula ②}$$

$$v_i=\sqrt{(x_{i+1}-x_i)^2}$$

$$THD=\sum_{i=1}^{n-1} h_i \quad \cdots \text{Formula ③}$$

$$h_i=\sqrt{(y_{i+1}-y_i)^2}$$

$$D=\sqrt{(x_n-x_1)^2+(y_n-y_1)^2} \quad \cdots \text{Formula ④}$$

$$VD=\sqrt{(x_n-x_1)^2} \quad \cdots \text{Formula ⑤}$$

$$HD=\sqrt{(y_n-y_1)^2} \quad \cdots \text{Formula ⑥}$$

$$MVD=\sqrt{(max\{x_1,\cdots,x_n\}-min\{x_1,\cdots,x_n\})^2} \quad \cdots \text{Formula ⑦}$$

Table 1 Measurement items and definitions between new and old groups

Item no.	Measurement items (in present study)	Definitions	Old measurement items (in previous study)
1	Shot	Whether shot or not in one play.	Shot
2	Dribble	The number of dribbles in one play.	Dribble
3	Pass	The number of passes (including fail) in one play.	Pass
4	Success of pass	The number of successful passes in one play.	Success of pass
5	Success rate of pass	The rate of successful passes in one play.	(New)
6	Direct pass	The number of direct pass (pass action with no trapping) in one play.	Direct pass
7	Consecutive direct pass	The maximum number of consecutive direct pass in one play.	Consecutive direct pass
8	Through-ball	The number of successful through-ball in one play.	Through-ball
9	Success of through-ball	The number of successful through-balls in one play.	(New)
10	Cross	The number of driving a cross ball in one play.	Cross
11	Success of cross	The number of successful cross-balls in one play.	(New)
12	Trap	The number of trapping (receive the ball and set it down at foot) a ball in one play.	Trap
13	Rebound-ball	The number of getting a rebound ball in one play.	Rebound-ball
14	Flick-on	The number of flick-on (touch the ball lightly to direction shift) in one play.	(New)
15	Throw-in	Whether the play started with throw-in.	(New)
16	Feed	Whether the play started with feeding a ball from goalkeeper.	(New)
17	Total of attack action	The number of attack actions in one play.	Total of attack action
18	Duration of attack	The time (seconds) from the start to the end of the attack.	Duration of attack
19	Average time of attack action	The average time (seconds) of attack action in one play. Formula is " = duration of attack / total of attack action".	Average time of attack action
20	Number of attackers	The number of attackers involved in one play.	Number of attackers
21	Total distance	The sum of the distance (meter) between an action and the next action in one play, excluding defensive actions. a + b + c in Fig.2.	Total distance
22	Total vertical distance	The sum of the vertical distance (meter) between an action and the next action in one play, excluding defensive actions. v1 + v2 + v3 in Fig.2.	Total vertical distance
23	Total horizontal distance	The sum of the horizontal distance (meter) between an action and the next action in one play, excluding defensive actions. h1 + h2 + h3 in Fig.2.	Total horizontal distance
24	Distance	The distance (meter) between the first action and the last action in one play, excluding defensive actions. D in Fig.2.	Distance
25	Vertical distance	The vertical distance (meter) between the first action and the last action in one play, excluding defensive actions. V in Fig.2.	Vertical distance
26	Horizontal distance	The horizontal distance (meter) between the first action and the last action in one play, excluding defensive actions. V in Fig.2.	Horizontal distance
27	Maximum vertical distance	The maximum value (meter) in each vertical distances in one play. Maximum vertical distance is v1 in Fig.2, and means the depth of attack.	Maximum vertical distance
28	Maximum horizontal distance	The maximum value (meter) in each horizontal distances in one play. Maximum horizontal distance is h3 in Fig.2, and means the width of attack.	Maximum horizontal distance
29	Mean of pass distance	The mean value of pass distances (meter) in one play.	(New)
30	Standard deviation of pass distance	The standard deviation of pass distances (meter) in one play. Large SD means attack players use short passes and long passes.	(New)
31	Area of attack	The sum of triangle areas (sq. meter) created by consecutive three actions in one play, excluding defense action. In Fig.3, area of attack is triangle①②③ + triangle ②③④.	Area of attack
32	Forward propulsion	Set forward (0 degrees) to 1, rightward (90 degrees) to 0, backward (180 degrees) to -1, leftward (270 degrees) to 0, and converted the angles of ball moving into value between -1 and +1 (Fig.4). The forward propulsion is the total of the values, and the more attackers pass or dribble forward, the more the total value is larger.	(New)
33	Wide propulsion	Set rightward (90 degrees) and leftward (270 degrees) to 1, forward (0 degrees) and backward (180 degrees) to 0, and converted the angles of ball moving into value between -1 and +1 (Fig.4). The wide propulsion is the total of the values, and the more attackers pass or dribble widely (leftward or rightward), the more the total value is larger.	(New)
34	Proportion of forward	The percentage of forward moving actions in one play excluding defensive actions. The forward moving action is defined as the angle between 315 degrees and 45 degrees in Fig. 5.	Proportion of forward
35	Proportion of backward	The percentage of backward moving actions in one play excluding defensive actions. The backward moving action is defined as the angle between 135 degrees and 225 degrees in Fig. 5.	Proportion of backward
36	Proportion of rightward	The percentage of rightward moving actions in one play excluding defensive actions. The rightward moving action is defined as the angle between 45 degrees and 135 degrees in Fig. 5.	Proportion of rightward
37	Proportion of leftward	The percentage of leftward moving actions in one play excluding defensive actions. The leftward moving action is defined as the angle between 225 degrees and 315 degrees in Fig. 5.	Proportion of leftward
38	Trun back	The total of actions that the angle (internal angle) formed by three consecutive actions is under 60 degrees. The line tied 1st action and 2nd action sets to 0 degrees, the used angle is formed by 1st, 2nd, and 3rd actions (Fig.6).	Trun-back
39	Change in direction	The total of actions that the angle (internal angle) formed by three consecutive actions is over 60 degrees and under 120 degrees. The line tied 1st action and 2nd action sets to 0 degrees, the used angle is formed by 1st, 2nd, and 3rd actions (Fig.6).	Change in direction
40	Same direction	The total of actions that the angle (internal angle) formed by three consecutive actions is over 120 degrees and 180 degrees or less. The line tied 1st action and 2nd action sets to 0 degrees, the used angle is formed by 1st, 2nd, and 3rd actions (Fig.6).	(New)
41	Twice speed	The number of actions that moved the ball more than twice as fast as the previous action.	Twice speed
42	Penalty area	Whether penetrated into penalty area in one play (Fig. 7).	Penalty area
43	Side of penalty area	Whether penetrated into side of penalty area in one play (Fig. 7).	Side of penalty area
44	30m line	Whether penetrated into 30m area from goal line in one play (Fig. 7).	30m line
45	Vital area	Whether penetrated into vital area in one play (Fig. 7).	Vital area
		(Removed)	X-coordinate of the last action
		(Removed)	Primary area
		(Removed)	Mean of moving direction
	Integrated into "Turn back"	Same with "Turn back"	Turn-back of sideward
		(Removed)	Triple speed
		(Removed)	The number of tactical attackers

"(Removed)" is included in the previous study, but is not used in this study. "(New)" is a newly created measurement item in present study.

$$MHD=\sqrt{(max\{y_1,\cdots,y_n\}-min\{y_1,\cdots,y_n\})^2}$$ ・・・Formula ⑧

$$MPD=\frac{TD}{n}$$ ・・・Formula ⑨

$$SDPD=\sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n}}$$ ・・・Formula ⑩

In the above formula, i denotes the attacking action number, n denotes the number of attacking actions, and x and y are the coordinates of the location where

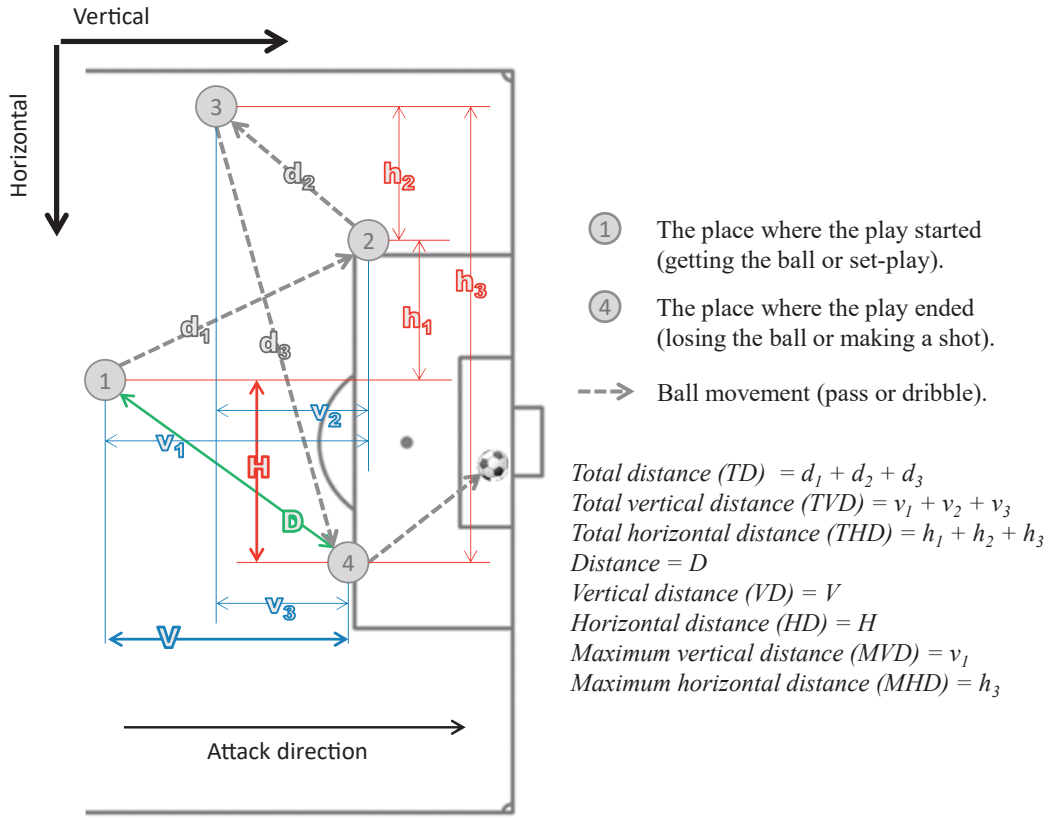

Figure 2 Definition and the example of measurement items related to distance

the action occurred. However, since a play may contain defensive actions (tackles and fouls), when converting action data into play data, a program was created to skip the loop if either or both of the actions *i* and *i+1* were defensive actions.

"Area of attack: AA" was calculated based on Heron's formula using the following formula (**Figure 3**):

$$AA = \sum_{i=1}^{n-2} \sqrt{s(s-a)(s-b)(s-c)}$$

・・・Formula ⑪

$$s = \frac{a+b+c}{2}$$

$$a = \frac{\sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2}}{2}$$

$$b = \frac{\sqrt{(x_{i+1} - x_{i+2})^2 + (y_{i+1} - y_{i+2})^2}}{2}$$

$$c = \frac{\sqrt{(x_{i+2} - x_i)^2 + (y_{i+2} - y_i)^2}}{2}$$

In this formula *a*, *b*, and *c* indicate the distances between two points at each coordinate of the three consecutive actions. The area of the triangle formed

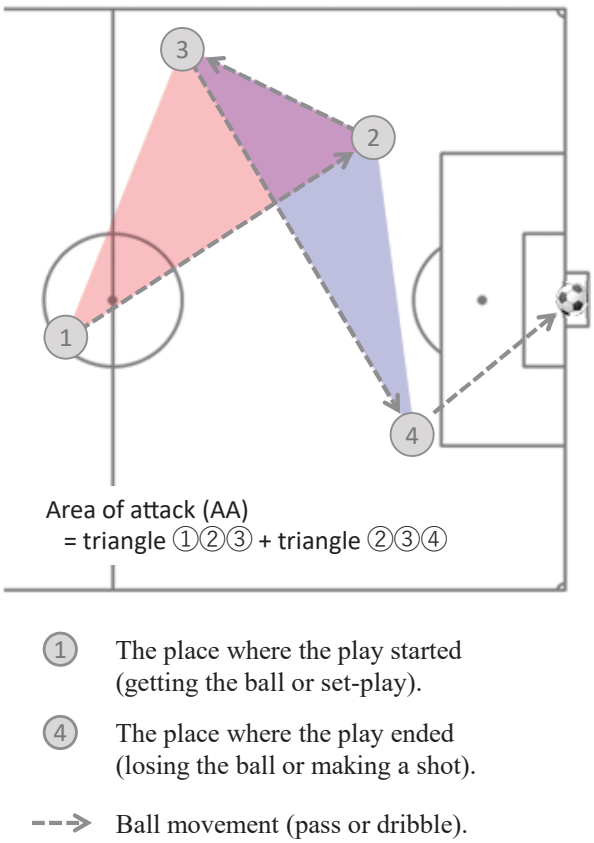

Figure 3 Definition and the example of "area of attack"

by the three consecutive actions is calculated by summing the areas of the *n-2* triangles in the play. However, a program was created to exclude actions from the calculation if any or all of the actions *i*, *i+1*, and *i+2* were defensive actions.

"Forward propulsion: FP" and "Wide propulsion: WP" were calculated by determining and summing the degree of forward propulsion, an index using the direction and distance of propulsion in each action (**Figure 4**). The angles were set to 0 degrees for forward, 90 degrees for rightward, 180 degrees for backward, and 270 degrees for leftward. The degree of forward propulsion was defined to be 1 for 0 degrees, -1 for 180 degrees, and 0 for 90 and 270 degrees. The degree of wide propulsion was defined to be 0 for 0 and 180 degrees and 1 for 90 and 270 degrees. "Forward propulsion: FP" and "Wide propulsion: WP" were calculated using the following formula:

$$FP=\sum_{i=1}^{n-1}(f_i d_i) \quad \cdots \text{Formula ⑫}$$

$$WP=\sum_{i=1}^{n-1}(w_i d_i) \quad \cdots \text{Formula ⑬}$$

In this formula, f indicates the degree of forward propulsion, w the degree of wide propulsion, and d the distance between two points.

"Proportion of forward: PF," "Proportion of backward: PB," "Proportion of rightward: PR," and "Proportion of leftward: PL" were in the old set of measurement items and continued to be used in the new set. These indices were calculated by counting the different directions (forward, back, right, and left) in which each action takes place in a play and by calculating the ratio of these directions. The angle of each direction was defined as 0 degrees in the attack direction (vertical direction), more than 315 degrees but less than 45 degrees as forward, more than 45 degrees but less than 135 degrees as rightward, more than 135 degrees but less than 225 degrees as backward, and more than 225 degrees but less than 315 degrees as leftward (**Figure 5**).

"Turn back" is a movement to return the ball to the direction it came from. "Change in direction" is a movement to change direction from vertical to horizontal or horizontal to vertical. "Same direction" refers to a movement to connect the ball in the same direction. "Turn back" is defined as the internal angle formed by three consecutive attacking actions between 0 and 60 degrees. "Change in direction" is defined as the internal angle formed between

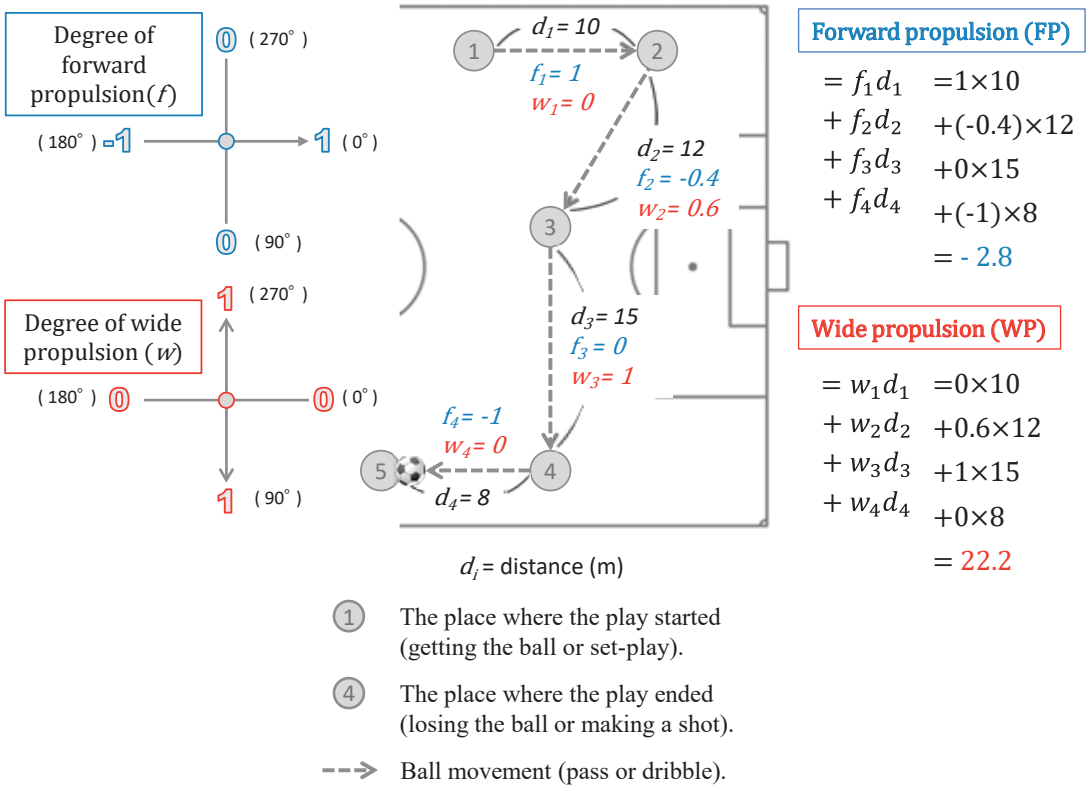

Figure 4 Definition and the example of "forward propulsion" and "wide propulsion"

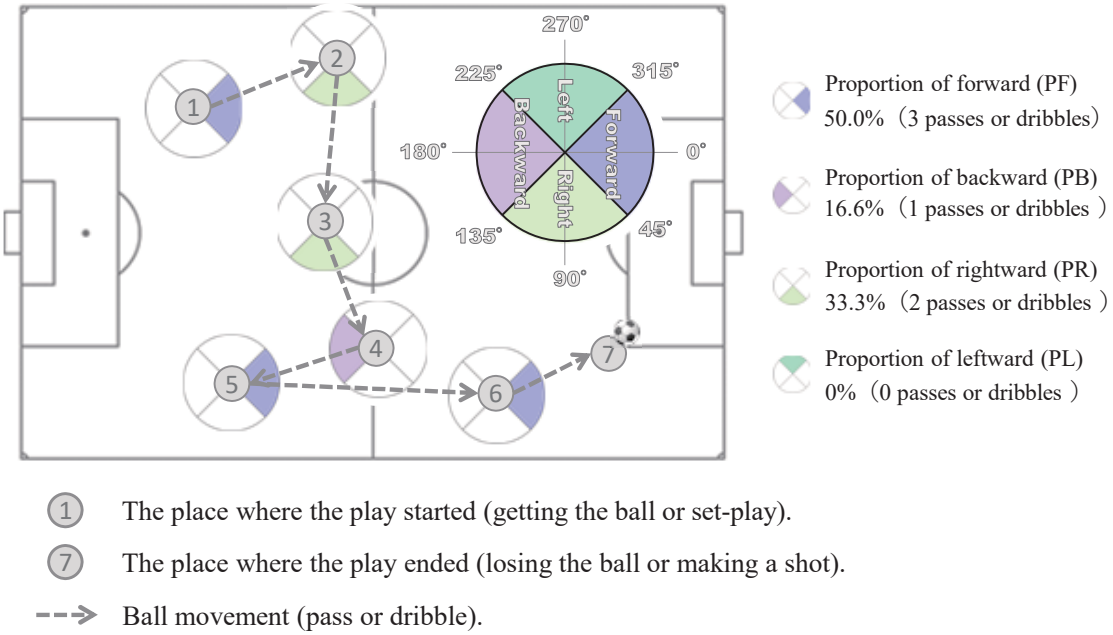

Figure 5 Definition and the example of items related to the ratio of movement directions

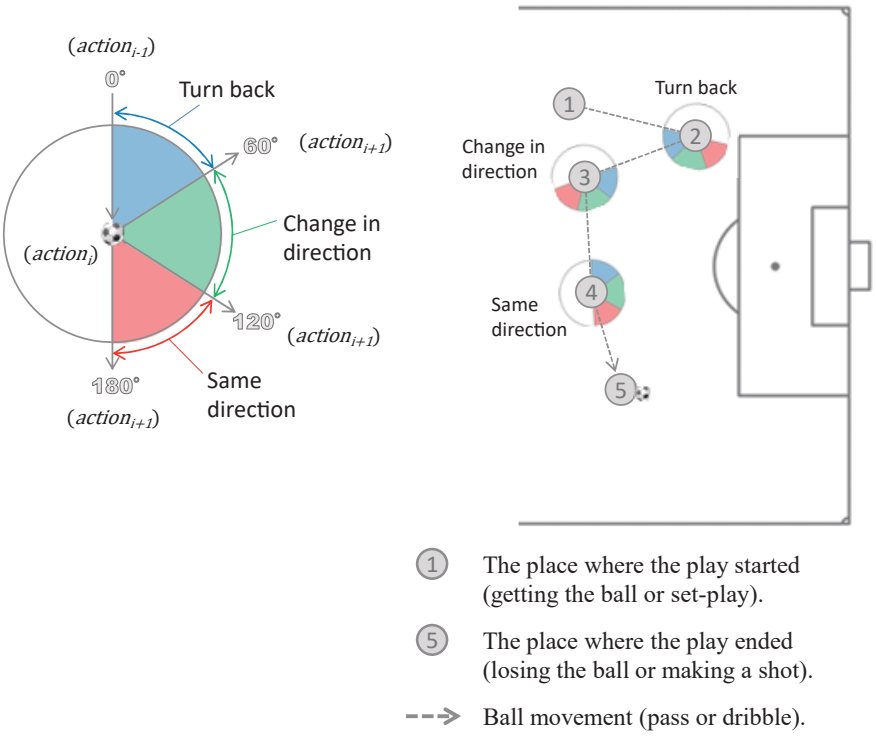

Figure 6 Definition and the example of items related to the angle of ball movement

60 and 120 degrees. “Same direction” is defined as the internal angle formed between 120 and 180 degrees (**Figure 6**). Furthermore, the number of these movements in a play was measured. However, a program was designed to exclude movements from the calculation if any or all of the actions *i*, *i+1*, and *i+2* were defensive actions.

“Twice speed” is a binary data set where 1 is recorded when the speed of the ball is at least twice as fast as the previous action. The speed of the ball is calculated by dividing the distance between two consecutive actions by the time required. “Penalty area,” “Side of penalty area,” “30 m line,” and “Vital area” are binary data where 1 is recorded if players

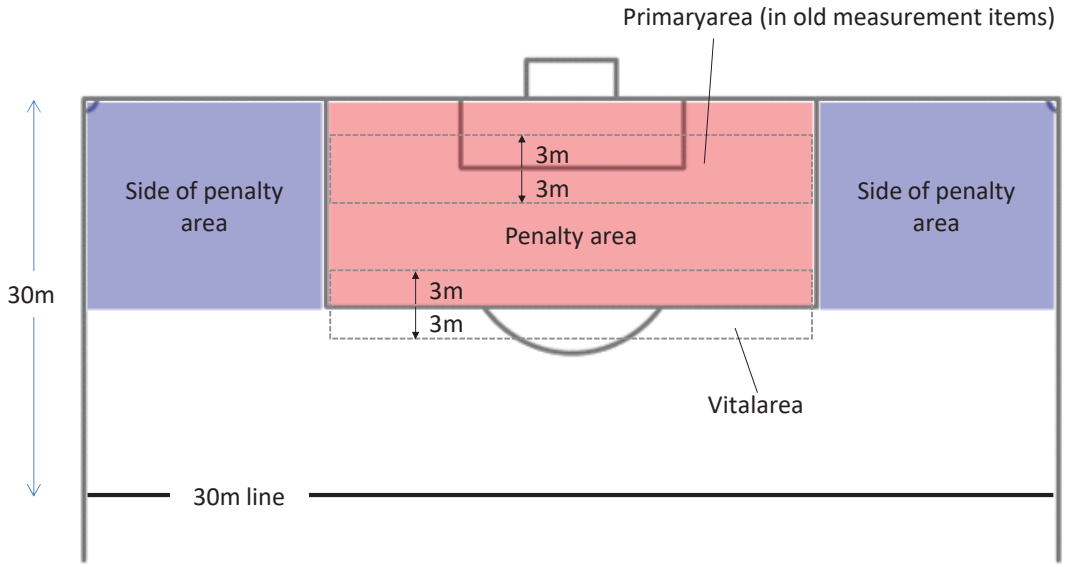

Figure 7 Definition of items related to area penetration

enter the area even once (**Figure 7**).The penalty area and vital area are partially overlapped, but this had a small effect on the analysis results. In addition, since this definition of areas is provided by Data Stadium Inc., it is desirable to keep it unchanged as much as possible (if it is changed, the process of changing the definition will have to be incorporated into the analysis flow when others analyze the data).

3.2. Conversion to play data

After converting the action data into play data based on the definition of the new set of measurement items, the converted play data (excluding plays starting with a corner kick, penalty kick, and free kick) comprised 147,032 rows. The basic statistics of the numerical measurement items are presented in **Table 2**, and the frequency distribution of the binary measurement items is indicated in **Table 3**.

3.3. Validation of the prediction model

The results of the prediction models for the number of items from 44 to 2 for the decision tree, random forest, and gradient boosting decision tree are presented in **Table 4**. The range of accuracy was 0.958 to 0.943 in the decision tree, 0.952 to 0.928 in the random forest, and 0.980 to 0.943 in the gradient boosting decision tree. At the same time, the f-measure ranged from 0.735 to 0.660 in the decision tree, 0.678 to 0.516 in the random forest, and 0.876 to 0.660 in the gradient boosting decision tree, with the gradient boosting decision tree tending to be higher than the other two methods.

In the gradient boosting decision tree, the models with 44 to 23 items indicated an f-measure of 0.85 or higher. However, when the number of items reached 22, the f-measure dropped significantly to 0.816. Therefore, the model with 23 items in the gradient boosting decision tree was selected as the best model. Predictor variables were reduced one by one based on their low importance, but the decision tree and gradient boosting decision tree tended to have eliminated similar items. By contrast, the random forest indicated a different trend. For instance, “Twice speed” was removed in the model with 43 items in the decision tree and the gradient boosting decision tree, and it was removed in the model with 38 items in the random forest.

In the best model, “Vital area” constituted 36.6% of the importance (**Figure 8**). Next, “Success rate of pass” constituted 11.3%, “Pass” constituted 9.4%, “Total of attack action” constituted 8.8%, “Trap” constituted 5.1%, “Penalty area” constituted 4.1%, and “Horizontal distance” constituted 2.3%. The other measurement items constituted less than 2%.

3.4. Comparison with the old set of measurement items

Since the gradient boosting decision tree indicated the highest prediction accuracy, this method was applied to the old set of measurement items (**Table 5**) to compare with the new set of measurement items.

Table 2 Statistics for measurement items of numerical type in play data

Item no.	Item name	Data type	Mean	Standard Deviation	Max-imum	3rd quartile	Median	1st quartile	Min-imum	Freq.
2	Dribble	Integer	0.10	0.30	3	0	0	0	0	147,032
3	Pass	Integer	3.47	3.30	44	4	2	1	0	147,032
4	Success of pass	Integer	2.68	3.16	41	4	2	1	0	147,032
5	Success rate of pass	Decimals	62.07	35.84	100.00	100.00	66.70	50.00	0.00	147,032
6	Direct pass	Integer	0.55	1.01	16	1	0	0	0	147,032
7	Consecutive direct pass	Integer	0.44	0.72	7	1	0	0	0	147,032
8	Through-ball	Integer	0.09	0.30	3	0	0	0	0	147,032
9	Success of through-ball	Integer	0.05	0.22	3	0	0	0	0	147,032
10	Cross	Integer	0.11	0.33	4	0	0	0	0	147,032
11	Success of cross	Integer	0.02	0.16	3	0	0	0	0	147,032
12	Trap	Integer	2.11	2.56	32	3	1	0	0	147,032
13	Rebound-ball	Integer	0.19	0.43	6	0	0	0	0	147,032
14	Flick-on	Integer	0.03	0.17	2	0	0	0	0	147,032
16	Feed	Integer	0.09	0.28	1	0	0	0	0	147,032
17	Total of attack action	Integer	6.06	5.81	78	8	4	2	0	147,032
18	Duration of attack	Integer	10.43	10.79	128	14	7	3	0	147,032
19	Average time of attack action	Decimals	1.66	1.00	19.00	2.00	1.56	1.00	0.00	147,032
20	Number of attackers	Integer	3.24	1.93	11	4	3	2	0	147,032
21	Total distance	Decimals	55.14	64.95	826.03	77.01	30.94	12.13	0.00	147,032
22	Total vertical distance	Decimals	34.65	38.25	499.83	50.67	21.67	7.33	0.00	147,032
23	Total horizontal distance	Decimals	34.90	46.28	604.33	47.00	16.83	5.00	0.00	147,032
24	Distance	Decimals	27.72	22.01	114.70	42.29	22.14	10.17	0.00	147,032
25	Vertical distance	Decimals	16.28	23.74	102.17	29.50	11.33	0.00	-93.50	147,032
26	Horizontal distance	Decimals	14.37	14.75	68.50	21.00	9.34	3.17	0.00	147,032
27	Maximum vertical distance	Decimals	31.56	28.31	104.67	52.67	25.83	4.33	0.00	147,032
28	Maximum horizontal distance	Decimals	29.37	24.83	69.33	52.00	26.83	3.33	0.00	147,032
29	Mean of pass distance	Decimals	13.02	10.01	92.17	18.15	13.04	6.52	0.00	147,032
30	Standard deviation of pass distance	Decimals	4.34	5.84	58.87	7.51	1.06	0.00	0.00	147,032
31	Area of attack	Decimals	1211.25	1992.04	28700.50	1626.00	262.00	0.00	0.00	147,032
32	Forward propulsion	Decimals	12.77	18.93	118.71	22.53	8.14	0.00	-74.36	147,032
33	Wide propulsion	Decimals	27.68	37.69	499.61	36.90	12.90	3.51	0.00	147,032
34	Proportion of forward	Decimals	37.23	34.21	100.00	50.00	33.33	0.00	0.00	147,032
35	Proportion of backward	Decimals	13.93	22.93	100.00	21.43	0.00	0.00	0.00	147,032
36	Proportion of rightward	Decimals	20.12	26.62	100.00	33.33	6.25	0.00	0.00	147,032
37	Proportion of leftward	Decimals	19.66	26.65	100.00	33.33	0.00	0.00	0.00	147,032
38	Trun back	Integer	0.84	1.44	21	1	0	0	0	147,032
39	Change in direction	Integer	1.57	2.43	36	2	1	0	0	147,032
40	Same direction	Integer	1.33	2.18	30	2	0	0	0	147,032
41	Twice speed	Integer	0.09	0.32	4	0	0	0	0	147,032

Table 3 Frequency distribution for measurement items of binary type in play data

			Frequency			Ratio(%)		
No.	Item name	Data type	Yes	No	Total	Yes	No	Total
1	Shot	Binary	12,346	134,686	147,032	8.40	91.60	100
15	Throw-in	Binary	37,127	109,905	147,032	25.25	74.75	100
42	Penalty area	Binary	11,042	135,990	147,032	7.51	92.49	100
43	Side of penalty area	Binary	15,498	131,534	147,032	10.54	89.46	100
44	30m line	Binary	35,766	111,266	147,032	24.33	75.67	100
45	Vital area	Binary	26,887	120,145	147,032	18.29	81.71	100

Table 4 Prediction models in decision tree, random forest, and gradient boosting decision tree methods

Number of items	Decision tree							Random forest						Gradient boosting decision tree						
	Max depth	One S.E. pruning	Item to be removed		Evaluation indices				Numer of items in a tree	Item to be removed		Evaluation indices				Max depth	number of boosting repetion	Item to be removed		Evaluation indices			
			Item name	importance	Accuracy	Precision	Recall	F-measure		Item name	importance	Accuracy	Precision	Recall	F-measure			Item name	importance	Accuracy	Precision	Recall	F-measure
44	16	Not	Flick-on	1.6	.954	.702	.743	.722	6	Flick-on	44.9	.952	.594	.780	.675	3	1454	Flick-on	0.000	.980	.850	.905	.876
43	16	Not	Twice speed	5.0	.954	.702	.743	.722	6	Proportion of backward	49.1	.952	.596	.782	.676	4	805	Twice speed	0.001	.980	.849	.903	.875
42	16	Not	Feed	21.5	.954	.702	.743	.722	6	Feed	74.0	.952	.597	.777	.675	4	399	Consecutive direct pass	0.001	.979	.843	.907	.874
41	16	Not	Trun back	22.3	.954	.703	.743	.722	6	Maximum horizontal distance	84.1	.952	.595	.778	.674	4	1340	Feed	0.001	.980	.850	.901	.875
40	16	Not	Proportion of backward	57.4	.954	.702	.744	.722	6	Total distance	79.4	.951	.595	.775	.673	3	836	Success of through-ball	0.001	.979	.848	.902	.874
39	16	Not	Same direction	59.4	.954	.701	.743	.722	6	Proportion of forward	91.7	.951	.596	.775	.674	3	1029	Through-ball	0.002	.980	.848	.905	.875
38	16	Not	Proportion of rightward	70.0	.954	.703	.741	.721	6	Twice speed	102.5	.951	.597	.773	.674	3	1345	Side of penalty area	0.002	.979	.848	.902	.874
37	16	Not	Change in direction	75.4	.954	.704	.742	.722	6	Total vertical distance	107.6	.951	.598	.772	.674	3	1223	Trun back	0.002	.979	.847	.903	.874
36	14	Not	Proportion of leftward	69.4	.954	.702	.743	.722	6	Wide propulsion	107.6	.952	.603	.775	.678	4	1002	Direct pass	0.003	.979	.851	.900	.874
35	14	Not	Through-ball	85.2	.954	.701	.744	.722	5	Direct pass	103.1	.952	.590	.782	.673	4	821	Same direction	0.003	.979	.844	.900	.871
34	14	Not	Throw-in	88.4	.954	.702	.742	.721	5	Success of pass	108.9	.951	.590	.780	.672	4	774	30m line	0.003	.979	.845	.897	.870
33	14	Not	Side of penalty area	107.8	.954	.705	.741	.723	5	Total of attack action	129.6	.952	.591	.782	.673	5	514	Change in direction	0.004	.978	.835	.892	.863
32	14	Not	Proportion of forward	105.7	.954	.700	.742	.721	5	Success of through-ball	128.4	.951	.592	.774	.671	3	1539	Proportion of rightward	0.004	.979	.845	.899	.871
31	14	Not	Consecutive direct pass	131.3	.954	.698	.740	.719	5	S.D. of pass distance	142.6	.951	.593	.772	.671	3	883	Proportion of leftward	0.003	.979	.844	.902	.872
30	14	Not	Direct pass	134.2	.954	.699	.740	.719	5	Total horizontal distance	146.5	.951	.594	.769	.670	4	1037	Throw-in	0.005	.978	.846	.892	.868
29	14	Not	Rebound-ball	161.5	.954	.699	.741	.719	5	Duration of attack	154.8	.950	.593	.766	.669	3	1282	Cross	0.004	.978	.834	.895	.863
28	16	Not	Maximum horizontal distance	193.1	.953	.695	.733	.713	5	Proportion of rightward	161.7	.950	.594	.761	.667	4	941	Success of pass	0.005	.977	.833	.886	.859
27	21	Not	Dribble	214.7	.953	.697	.731	.714	5	Maximum vertical distance	164.7	.950	.595	.765	.669	4	1147	Proportion of backward	0.006	.977	.835	.888	.861
26	13	Not	S.D. of pass distance	203.3	.954	.677	.750	.711	5	Proportion of leftward	162.0	.950	.596	.762	.669	3	632	Proportion of forward	0.003	.977	.829	.895	.861
25	13	Not	Horizontal distance	210.9	.954	.682	.750	.714	5	Trap	164.6	.950	.597	.760	.669	3	898	Total horizontal distance	0.006	.978	.832	.894	.862
24	14	Not	Cross	248.6	.955	.680	.757	.716	4	Consecutive direct pass	136.2	.949	.572	.762	.654	4	884	Success of cross	0.008	.977	.833	.891	.861
23	14	Not	Maximum vertical distance	262.4	.954	.675	.749	.710	4	Average time of attack action	133.1	.950	.582	.766	.661	4	616	Dribble	0.010	.977	.827	.890	.857
22	14	Not	Forward propulsion	271.0	.953	.678	.745	.710	4	Trun back	127.2	.949	.580	.755	.656	5	267	Total vertical distance	0.011	.970	.786	.847	.816
21	3	Not	Total vertical distance	1.7	.945	.669	.672	.671	4	Same direction	130.4	.948	.561	.759	.645	4	356	Rebound-ball	0.010	.970	.788	.850	.818
20	18	Not	Area of attack	417.5	.951	.670	.731	.699	4	Forward propulsion	131.8	.947	.552	.760	.639	4	291	S.D. of pass distance	0.010	.970	.781	.853	.815
19	18	Not	Distance	429.3	.954	.682	.747	.713	4	Horizontal distance	125.4	.947	.552	.760	.639	5	831	Number of attackers	0.014	.970	.787	.844	.815
18	14	Not	Wide propulsion	349.4	.954	.686	.749	.716	4	Side of penalty area	117.8	.948	.551	.761	.640	4	524	Maximum vertical distance	0.016	.969	.780	.843	.810
17	14	Not	Mean of pass distance	384.9	.955	.687	.757	.720	4	Dribble	111.0	.947	.549	.760	.637	8	41	Wide propulsion	0.012	.968	.765	.843	.802
16	28	Run	Total horizontal distance	283.9	.958	.683	.794	.735	4	Change in direction	103.7	.947	.542	.760	.633	6	131	Maximum horizontal distance	0.017	.969	.779	.846	.812
15	20	Run	Average time of attack action	325.7	.958	.683	.794	.734	6	Vertical distance	125.8	.947	.554	.747	.636	8	43	Mean of pass distance	0.020	.968	.762	.844	.801
14	18	Not	Success of through-ball	540.6	.955	.686	.755	.719	6	Throw-in	131.7	.946	.548	.744	.631	8	44	Distance	0.021	.968	.766	.841	.801
13	18	Not	Number of attackers	694.8	.956	.696	.758	.725	6	Number of attackers	123.0	.946	.541	.751	.629	8	61	Average time of attack action	0.024	.968	.768	.839	.802
12	25	Not	Trap	720.0	.956	.689	.762	.724	6	Area of attack	113.9	.946	.529	.752	.621	7	91	Forward propulsion	0.026	.968	.768	.836	.801
11	6	Not	Total of attack action	168.0	.949	.696	.700	.698	3	Through-ball	73.4	.944	.508	.745	.604	7	87	Area of attack	0.032	.968	.768	.835	.801
10	6	Not	Success of pass	173.9	.949	.700	.696	.698	6	Distance	128.8	.944	.532	.735	.617	7	55	Vertical distance	0.034	.968	.772	.833	.801
9	6	Not	Total distance	199.7	.949	.700	.696	.698	3	Rebound-ball	76.3	.944	.551	.716	.623	8	45	Horizontal distance	0.042	.967	.771	.829	.799
8	6	Not	Duration of attack	388.5	.949	.698	.697	.697	4	Mean of pass distance	111.7	.944	.568	.710	.631	10	40	Penalty area	0.050	.966	.765	.818	.790
7	3	Not	Vertical distance	337.7	.945	.669	.672	.671	4	Pass	127.5	.943	.559	.708	.625	10	32	Duration of attack	0.068	.964	.754	.811	.781
6	3	Not	Pass	604.1	.945	.669	.672	.671	4	Cross	142.3	.940	.420	.760	.541	9	32	Total distance	0.070	.964	.744	.818	.779
5	5	Not	Success of cross	896.5	.944	.685	.661	.673	4	30m line	58.4	.939	.422	.751	.540	10	33	Trap	0.096	.965	.757	.816	.785
4	3	Not	30m line	1358.9	.943	.682	.655	.668	2	Success of cross	486.4	.939	.384	.788	.516	6	14	Pass	0.083	.949	.674	.709	.691
3	6	Not	Penalty area	1871.3	.943	.698	.650	.673	2	Success rate of pass	1222.2	.939	.392	.767	.519	10	21	Total of attack action	0.077	.945	.682	.674	.678
2	3	Not	Success rate of pass	1877.0	.943	.661	.659	.660	1	Penalty area	2270.5	.928	.475	.595	.528	2	1	Success rate of pass	0.293	.943	.661	.659	.660

The range of accuracy for the old set of measurement items was 0.958 to 0.897, and the f-measure ranged from 0.799 to 0.528. By the number of items, the f-measure was 0.795 for the old set of measurement items, 0.874 for the new set of measurement items in the model with 37 items, and 0.785 for the old set of measurement items and 0.857 for the new set of measurement items in the model with 23 items. These results indicate that the prediction accuracy is higher in the new set of measurement items.

4. Discussion

4.1. Measurement items

This study was based on the measurement items of Jo et al. (2014), from which six items were deleted (including the integrated items) and 11 new items were created. This section first discusses the reasons why the six items were removed. The "X-coordinate of the last action" indicates where the last action of the play ended. If it ends in the front, this indicates how far into the opponent's territory the player has advanced; thus, it is considered to be a measurement item that helps, to a great extent, to determine whether a player takes a shot. However, this study was intended to predict whether the play would end in a shot. Therefore, It is inappropriate as a predictor variable because the result shows whether the player was able to take a shot once the play is over.

The "Primary area" is the area extending from the 5.5-meter line of the goal area to 3 meters in front of and behind the penalty area. It is a measurement item that indicates the entry into the area closest to the opponent's goal post among the old measurement items. When a player enters the primary area, the player is highly likely to take a shot. In fact, the shooting rate for the plays in which the player entered the primary area was 79%. In other words, the model can predict whether a player will take a shot 79% of the time based merely on the item "Primary area." This approach reduces the need to apply machine learning methods. In addition, since only 7.51% of the plays entered the primary area, it would be ideal to develop a model with high prediction accuracy without this item; thus, it was removed from the new set of measurement items. Since this also applies to the old set of measurement items, this item was not used in the comparison of prediction accuracy between the old and new sets of measurement items.

The "Mean of moving direction" is the average of the moving directions of each action, but it was judged inappropriate as it would be close to zero when the opponent was propelled from side to side (it involves a calculation process of summing positive and negative values). Instead, new measurement items "Wide propulsion," "Turn back," "Change in direction," and "Same direction" were added.

As "Triple speed" and "Twice speed" refer to the

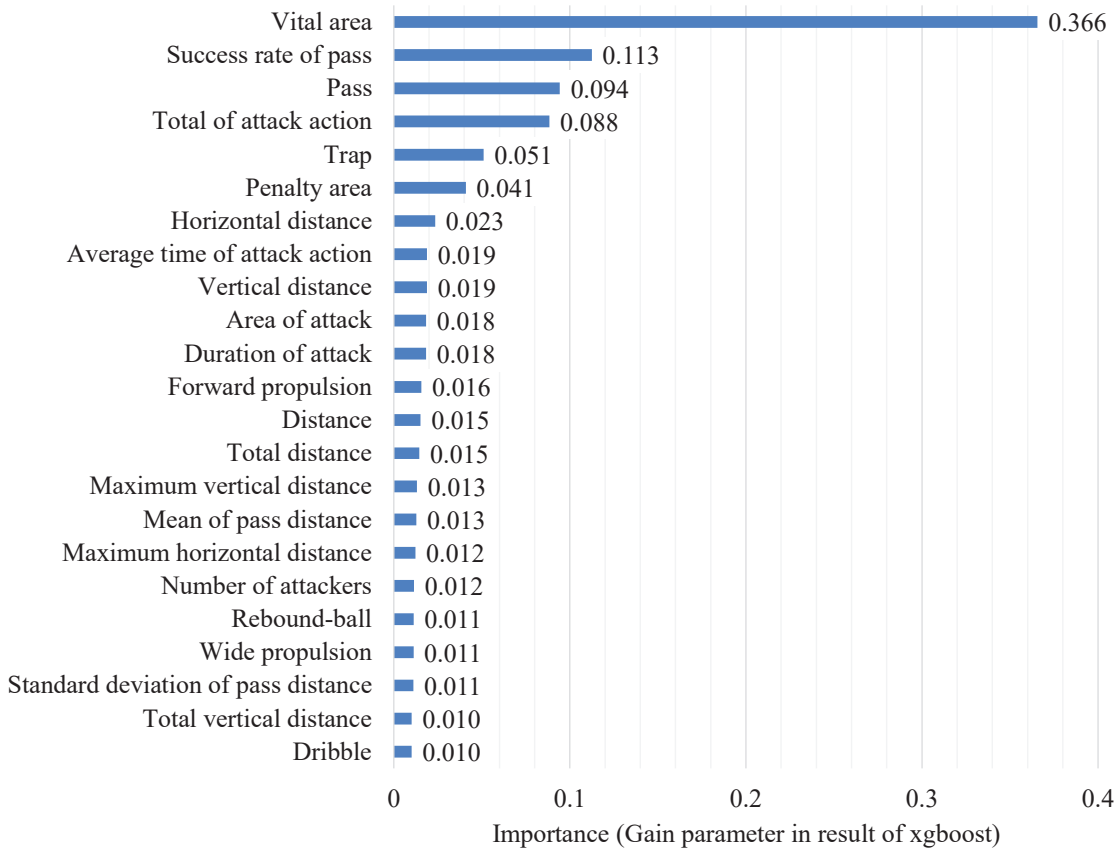

Figure 8 Importance of the best model (23 measurement items in the gradient boosting decision tree)

Table 5 The shot prediction models in gradient boosting decision tree using old measurement items group

Number of items	Max depth	number of boosting repetition	Item to be removed		Evaluation indices			
			Item name	Impor-tance	Accuracy	Precision	Recall	F-measure
37	4	411	Triple speed	0.002	.957	.754	.840	.795
36	3	595	Trun-back	0.001	.958	.751	.846	.796
35	2	676	Consecutive direct pass	0.000	.958	.755	.845	.798
34	4	337	Turn-back of sideward	0.002	.958	.757	.846	.799
33	4	359	Side of penalty area	0.002	.958	.751	.847	.796
32	4	330	Twice speed	0.002	.957	.756	.838	.795
31	3	481	Change in direction	0.002	.957	.755	.837	.794
30	3	445	Dribble	0.002	.958	.757	.842	.797
29	4	331	Number of attackers	0.003	.956	.747	.835	.788
28	3	688	Through-ball	0.003	.958	.762	.837	.798
27	3	406	Mean of moving direction	0.005	.957	.750	.837	.791
26	5	318	30m line	0.008	.956	.747	.831	.787
25	3	415	Total horizontal distance	0.006	.956	.744	.838	.788
24	3	596	Proportion of backward	0.008	.957	.756	.840	.796
23	4	559	Trap	0.010	.955	.751	.824	.785
22	6	137	Total vertical distance	0.013	.955	.740	.831	.783
21	4	354	Rebound-ball	0.012	.956	.749	.831	.788
20	4	346	Proportion of rightward	0.011	.955	.740	.831	.783
19	4	315	Cross	0.012	.955	.743	.833	.785
18	4	382	The number of tactical attackers	0.014	.954	.738	.828	.781
17	7	142	Proportion of forward	0.023	.952	.721	.818	.766
16	6	181	Proportion of leftward	0.023	.953	.728	.821	.771
15	8	28	Area of attack	0.016	.950	.707	.817	.758
14	8	30	Vertical distance	0.019	.950	.708	.817	.759
13	10	32	Duration of attack	0.028	.951	.717	.813	.762
12	8	32	Maximum horizontal distance	0.023	.951	.713	.816	.761
11	10	30	Success of pass	0.029	.950	.714	.809	.759
10	8	22	Maximum vertical distance	0.020	.948	.687	.812	.745
9	8	26	Total distance	0.029	.949	.691	.812	.747
8	10	29	Horizontal distance	0.070	.948	.700	.806	.749
7	9	19	Distance	0.059	.946	.680	.797	.734
6	9	46	Direct pass	0.099	.947	.705	.790	.745
5	7	18	Pass	0.103	.927	.583	.700	.636
4	2	1	Penalty area	0.239	.907	.538	.582	.559
3	7	1	Total of attack action	0.058	.897	.526	.530	.528

same action, they were deleted from the perspective of linear dependence. Although Jo et al. (2014) mentioned technology and tactics in their analysis, "The number of tactical attackers" was also removed as this study did not use the concept of technology and tactics.

The reason why some new items were created is discussed next. There are three types of passes: normal pass, through-pass, and cross. Although the old set of measurement items included "Success of pass," "Success of through-pass" and "Success of cross" were not included. "Pass" includes passes that were not received by a teammate. "Success of pass" counts only passes received by a teammate. As both qualitatively differ, "Through-pass" and "Cross" were unified and a measurement item for the number of successes of each was added.

The number of plays with no through passes was 132,487 (91.4% of the total), and the number of plays with no crosses was 132,487 (90.1% of the total). When these numbers are zero, the denominator is zero; therefore, the success rate cannot be calculated and is a missing value in the data set. The items "Success rate of through-pass" and "Success rate of cross" were not created as it would be difficult to handle items with a missing number exceeding 90%. However, as there were 7,711 plays (5.2% of the total) with no normal passes, the item "Success rate of pass" was added.

"Flick-on" is the action of lightly touching the

ball to change its pass course. Flick-on was added as a new item because it can surprise the opposing defender if successful. "Feed" is the pass after the goalkeeper has caught the ball. In modern soccer, there is often a strategy called "build-up" in which the goalkeeper and the defender build up the attack. "Feed" was added as a new item to distinguish it. "Throw-in" was another new entry added to distinguish between attacks that begin with a throw-in and attacks after the ball has been taken from the opposing defender.

"Mean of pass distance" is the average distance of passes in a play. A larger value indicates that more long passes were used, and a smaller value indicates that more short passes were used. "Standard deviation of pass distance" is the standard deviation of the pass distance in a play. A large value indicates that both long and short passes were used.

"Forward propulsion" indicates the amount of force that carries the ball forward. "Wide propulsion" indicates the amount of force that propels the ball from side to side. As it is an index that combines the angle at which the ball moves forward and the distance, it has characteristics different from the measurement items that only measure distance or angle.

The old set of measurement items contained items to measure the forward-backward turn and the left-right turn, but in the new set of measurement items, these two were combined into "Turn back." "Turn back" refers to the movement of returning the ball to the direction it came from, such as returning a received ball to the same player or sending a lowered ball back to the front.

"Change in direction" is a measurement item included in the old set of measurement items. It refers to a movement that converts a horizontal pass into a vertical pass, and vice versa. If, in addition to "Turn back" and "Change in direction," there was an item to measure the movement of a horizontal pass into a horizontal direction and a vertical pass into a vertical direction, it would be possible to measure all the angles the ball travels. Therefore, "Same direction" was created.

4.2. Conversion to play data

Before analyzing large-scale data, it is necessary to analyze the data structure and preprocess it to deal with sparsity and scaling of different data scales (unification of scales) (Decroos, 2020). In this study, the action data were converted into play data to create the data structure necessary for developing the shot prediction model. As the decision tree method can be used for both quantitative and qualitative variables, scaling was determined to be unnecessary at this point. The "Penalty area" and the "Vital area," which are closest to the goal, were considered to contribute greatly to the number of shots. However, as the attacks leading up to them are also important, although differences in the size of the contribution would be apparent, sparsity was also considered to not be a problem.

After converting 1,312,117 lines of action data, 147,032 lines of play data were generated, with an average of 8.9 actions per play. "Dribble" indicated a lower average of 0.1 times per play (once in 10 plays), but this may be due to the measurement criteria. When the ball was carried for a certain length of time, the action was measured as a dribble. It was not measured as a dribble when it was moved slightly with a fine touch.

Among others, the average number of times of "Through-ball," "Cross," "Feed," and "Flick-on" was low. The mean value of "Total of attack actions," which was 6.06, greatly differed from the mean value of "Number of attackers," which was 3.24. This is because the same players took multiple actions, most of which were a combination of trapping and passing as well as receiving and delivering the ball.

The mean values of the measurement items concerning the distance traveled were 34.65 m for "Total vertical distance" and 34.90 m for "Total horizontal distance," showing very little difference. The differences between "Vertical distance," which was 16.28 m, and "Horizontal distance," which was 14.37 m, and between "Maximum vertical distance," which was 31.56 m, and "Maximum horizontal distance," which was 29.37 m, were also small. However, the mean values of "Forward propulsion" and "Wide propulsion" were 12.77 and 27.68, respectively, with the former being more than twice as large as the latter. The sum of the distances between the actions tended to be larger in the vertical direction because the distances were calculated as positive values, even for backward passes. At the same time, "Forward propulsion" is designed to take a negative value for backward passes, which reflects the movement away from the goal post and thus correctly reflects the vertical movement in actual soccer play.

4.3. Validation of the prediction model

Many analytical methods have been designed for prediction. Not only traditional statistical methods but also machine learning methods have attracted attention in recent years. Machine learning methods can be broadly classified into "supervised learning," "unsupervised learning," and "reinforcement learning." Since the data in this study contain items that measure whether a shot was taken (i.e., answers exist), supervised learning was appropriate. However, since the study is not at the stage of constructing a system that automatically improves learning accuracy, reinforcement learning cannot be used yet. Typical methods for supervised learning include neural networks, support vector machines, and decision trees; however, it is desirable that the prediction process be as clear as possible as the prediction model leads to the clarification of the movements necessary for taking a shot. Given this point, decision tree analysis is a useful method because it can visualize the prediction process. In addition, decision trees are considered to be the best method when prediction accuracy is important for 100 items or less (Microsoft, 2020). For these reasons, decision tree analysis was adopted as the method for model construction in this study. It was extended to ensemble learning by testing the random forest and gradient boosting decision tree methods.

For the verification of prediction accuracy, three methods were compared: decision tree, random forest, and gradient boosting decision tree. Accuracy was above 0.9 for all methods, which may be due to the fact that 91.6% of the total number of plays did not end in a shot. If the expected value of the prediction results derived by the model is 50% for both "ending in a shot" and "not ending in a shot," the probability of accuracy for "not ending in a shot," i.e., the expected value of true negative, is $0.5 \times 0.916 = 0.458$. This indicates that even if the prediction model has low performance, 45.8% of the predictions will theoretically be accurate. By contrast, the probability of accuracy for "ending in a shot," i.e., the expected value of true positive, is $0.5 \times 0.084 = 0.042$ (4.2%). This means that the prediction accuracy will not improve unless the performance of the prediction model is high. Thus, since the data were biased toward negatives (not ending in a shot), it was better to use the f-measure, which is the harmonic mean of precision and recall, to verify the accuracy of the prediction model.

Among the three methods, the gradient boosting decision tree had the best f-measure. The random forest had the lowest accuracy. The random forest method, like the gradient boosting decision tree, is an ensemble learning method, but its accuracy was lower than that of its base method, the decision tree. This may be due to the mechanism of random forests. In random forests, some measurement items are randomly selected to construct multiple trees. If highly important measurement items are selected, the prediction accuracy increases; however, if the trees are constructed with only low importance measurement items, the prediction accuracy will remain low. Therefore, when the average value is finally calculated, the prediction accuracy will have decreased. For this reason, random forests are not suitable for data sets where the strength of the relationship between each measurement item and the objective variable is biased (i.e., sparse), and this seems to have been the case with the soccer ball touch data in this study.

The highest f-measure in the gradient boosting decision tree was 0.876 with 44 items; however, the model with 23 items, or 21 items less than the model with 44 items, also showed a high f-measure of 0.857. A greater number of measurement items does not necessarily produce a better result. It is desirable not to use unnecessary measurement items from the perspective of computational load and labor of the measurer. These reasons led to the conclusion that the model with 23 items in the gradient boosting decision tree was optimal within the scope of this study. The next model with 22 items had dribbles removed, and the f-value dropped significantly to 0.816. This suggests that dribbling is a relatively influential item in the prediction of shots because movements such as cutting into the penalty area and running up the side to use a cross are observed in actual games. At the same time, among the new items, "Flick-on" and "Feed" had low importance and were listed among the items to be deleted at an early stage, indicating that their contribution to shot prediction was small.

The importance of "Vital area" in the best model constituted 36.6% of the total. The fact that the players had a 40% chance of shooting when entering the vital area may have contributed to its high importance. Although the importance of the remaining items was not very high, the precision of the best model was 82.7%, suggesting that items other

than "Vital area" increased the prediction accuracy by 42.7%. This indicates that items other than "Vital area" are also important measurement items. By contrast, the importance of "Penalty area," which is closer to the goal post and wider than the vital area, was 4.1%. In actual matches, since the penalty area is often crowded with defenders, shots are often taken from outside the penalty area. "Vital area" is more important than "Penalty area" in determining whether a player takes a shot.

In this study, decision tree analysis, which has a relatively clear prediction process, and its ensemble learning methods, random forest and gradient boosting, were applied to address the black box problem. Since the output of a decision tree has a single tree structure, the bifurcation of each node can be identified directly. By contrast, the gradient boosting decision tree, in which the best model was developed, has an algorithm that creates multiple small trees in series and improves the error of the previous tree in the next tree. Therefore, the same measurement items are used many times in different trees, each time producing different bifurcation values and nodes after bifurcation. Although individual tree structures can be interpreted in a similar way to decision trees, it is relatively difficult to capture the overall trend. Thus, when analysts interpret the results, they should rely on the importance of the gradient boosting decision tree or obtain the average of the bifurcation values of each measurement item across all tree structures. Further research is required in this regard and is a research topic for the future.

4.4. Comparison with the old set of measurement items

In this study, 11 items were newly created and 5 items of the old set of measurement items were removed. The best model did not include six of the newly measurement items: "Success of through-pass," "Success of cross," "Flick-on," "Throw-in," "Feed," and "Same direction." This means that 5 items were newly added and 5 items were deleted from the old set of items, resulting in a total of 10 items being replaced. The maximum f-measure of the new set of items was 0.876 and that of the old set was 0.799. The change in the measurement items improved the f-measure by 0.077. In particular, "Success rate of pass" is a measurement item of high importance and is considered to have contributed to the shot prediction. The above result indicates that the new measurement items are better than the old ones in the study by Jo et al. (2014).

4.5. Extensibility

The measurement items and prediction model in this study have the potential to be applied to further research and to the field of sport competition. There have been very few studies on the development of measurement items for game performance using machine learning in addition to the Delphi method, and this method is thought to be applicable to other sports. In addition, this study revealed the measurement items related to shots in offensive plays in soccer, which will enable players to construct and evaluate their plays based on the results of measurement.

5. Research limitations

The limitations of this study by sample, by measurement items, and by the data set used limit the generalizability of the conclusions. This study utilized ball touch data from all matches and all plays in the J.League in 2011. Although the amount of information in the ball touch data is sufficient, it does not include any tracking data of players not in possession of the ball and thus cannot measure the spatial movement of the group. Therefore, in interpreting the measurement items related to offensive actions in soccer (new measurement items) defined in this study, it is important to keep in mind that the information is limited to ball touch data.

At the same time, there are also limitations due to the data analysis methods used. Among the many machine learning methods, three types were applied in this study: decision trees, random forests, and gradient boosting decision trees. However, other methods were not discussed. Although the gradient boosting decision tree had the highest prediction accuracy, it was not compared with other methods.

6. Conclusion

Ball touch data in soccer matches are useful; in academia, it is necessary to accumulate studies on analyses applying machine learning. In light of this, this study aimed to construct a machine learning model capable of predicting whether a player would

take a shot by developing measurement items for offensive plays based on the soccer ball touch data. For this purpose, the following hypotheses were discussed:

Hypothesis 1: The measurement items of offensive plays are created based on ball touch data.

Hypothesis 2: A machine learning model is constructed that applies decision tree analysis to predict whether a player will take a shot based on the measurement items of offensive plays.

Although there are various machine learning methods, the study was conducted using decision tree analysis, which has a relatively clear prediction process, and its ensemble learning methods, random forests and gradient boosting decision trees. Through these methods, the study aimed to address the black box problem and identify the model with the highest prediction accuracy. Six measurement items were deleted, and 11 items were newly created based on the 40 items (old set of measurement items) in the study by Jo et al. (2014). As a result of comparing the prediction accuracy of the new measurement items and the old measurement items, the following conclusions were obtained:

(1) Based on the ball touch data, 45 offensive play measurement items, including the objective variable "shot," were developed.

(2) By applying gradient boosting decision tree analysis, a machine learning model was built to predict whether a player would take a shot based on the 23 offensive play measurement items.

7. Acknowledgements

This research was funded by the Shizuoka Sangyo University special research support grant, with data provided by Data Stadium Inc. I would like to express my gratitude to both institutions.

References

Alamar, B. and Mehrotra, V. (2011). Beyond moneyball: The rapidly evolving world of sports analytics, Part I. https://pubsonline.informs.org/do/10.1287/LYTX.2011.05.05/full/. (accessed 2021-04-17).

Breiman, L., Friedman, J. H., Olsehn, R. A., and Stone, C. J. (1984). Classification and regression trees. Wadsworth Inc.

Breiman, L. (2001). Random forests. Machine Learning, 45: 5-32.

Chen, T., He, T., Benesty, M., Khotilovich, B., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2020). xgboost: Extreme gradient boosting. R package version 1.2.0.1. https://CRAN.R-project.org/package=xgboost. (accessed 2021-04-25).

Chiba, Y. (2020). Fencing and sports analytics: Initiatives for the Tokyo Olympics. Monthly Statistics, 71: 10-16. (in Japanese).

Decroos, T. (2020). Soccer analytics meets artificial intelligence: Learning value and style from soccer event stream data. Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science in University of Liverpool.

Dmonte, R., and Dmello, A. (2017). Big data in sports: Leverage big data in sports: An insight using SAP HANA. IJERT, 6: 380-383.

Ekin, A., Takelp, M., and Mehrotra, R. (2003). Automatic soccer video analysis and summarization. IEEE transactions on image processing, 12: 796-807.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5): 1189-1232.

James, B. (2003). The new Bill James baseball abstract. Free Press.

Jo, H., Yokoo, T., Ando, K., Nishijima, T., Kumagai, S., Naomoto, H., Suzuki, K., Yamada, H., Nakano, T., and Saito, K. (2014). Attack power index of players and teams in J League. Research on Sports Data Analytics: Theory, Methodology, and Applications, 1:21-26. (in Japanese).

Jo, H., Oosawa, K., Mishio, S., Ando, K., Suzuki, K., and Nishijima, T. (2017). Development of optimization algorithm for attack play in football. Proceedings of the Institute of Statistical Mathematics, 64: 309-321. (in Japanese).

Kato, K. (2016). Data analysis and team reinforcement in soccer. IEICE Communications Society Magazine, 10: 29-34. (in Japanese).

Kumar, G. (2013). Machine learning for soccer analytics. Doi:10.13140/RG.2.1.4628.3761.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R News 2(3), pp.18-22.

Linke, D., Link, D., and Lames, M. (2020). Football-specific validity of TRACAB's optical video tracking systems. Plos ONE, 15: e0230179.

Linstone, A. H., and Turoff, M. (1975). The Delphi method. Massachusetts: Addison-wesley.

Matsuoka, H., Tahara, Y., Ando, K., and Nishijima, T. (2020). Development of defence and offence play items for deep learning model of offence play analysis in soccer game. Football Science, 17: 69-85.

Microsoft (2020). Machine learning algorithm cheat sheet. https://docs.microsoft.com/ja-jp/azure/machine-learning/algorithm-cheat-sheet. (accessed 2021-04-29).

Miller, W. (2015). Sports analytics and data science: Winning the game with methods and models. Paul Boger, New Jersey.

Morinaga, S. (2015). Practical application of large-scale forecasting system based on big data analysis: Simultaneous achievement of "high accuracy", "white box" and "low cost". Monthly Statistics, 66: 14-19. (in Japanese).

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. (accessed 2021-05-5).

Sakaori, F. (2016). Sports and statistical science. Monthly Statistics, 67: 41-44. (in Japanese).

Shimokawa, T., Sugimoto, T., and Goto, M. (2013). Classification and regression tree (CART) method and the surroundings. In Kim, M. (eds.), Data science learned in R vol. 9 tree structured

analysis (pp.1-33). Tokyo: Kyoritsu Publishing. (in Japanese).
Suzuki, K., Nagai, S., Ogaki, R., Iwai, K., Furukawa, T., Miyakawa, S., and Takemura, M. (2019). Video analysis of tackling situations leading to concussion in collegiate rugby union. J. Phys. Fitness Sports Med., 8: 79-88.
Takahashi, S., Haseyama, S. (2017). A method of important player extraction based on link analysis in soccer videos. ITE Trans. on MTA, 5: 42-48.
Tamura, Y. (2020). Special feature: About sports analytics. Monthly Statistics, 71: 2-3. (in Japanese).
Therneau, T., and Atkinson, B. (2019). Rpart: Recursive partitioning and regression trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart. (accessed 2021-04-22).
Tsuchida, J., and Yadohisa, H. (2020), Recent trends in soccer data analysis: Tracking data and the surroundings. Monthly Statistics, 71: 4-9. (in Japanese).
Wang, Q., Zhu, H., Hu, W., Shen, Z., and Yao, Y. (2015). Discerning tactical patterns for professional soccer teams: An enhanced topic model with applications. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2197-2206.

Name:
Hirotaka Jo

Affiliation:
*Department of Sports Science, Shizuoka Sangyo University
**Doctoral Program in Health & Sport Sciences, University of Tsukuba

Address:
*1572-1 Oowara, Iwata-city, Shizuoka-pref 438-0043 Japan
**1-1-1 Tennodai, Tsukuba, Ibaraki 305-8574 Japan

Brief Biographical History:
2014- Health and physical education teacher, Junior & Senior High School at Komaba, University of Tsukuba.
2019- Associate lecturer, Shizuoka Sangyo University.

Main Works:
• Jo, H., Oosawa, K., Mishio, S., Ando, K., Suzuki, K., and Nishijima, T. (2017). Development of optimization algorithm for attack play in football., Proceedings of the Institute of Statistical Mathematics, 64: 309-321. (in Japanese).

Membership in Learned Societies:
• Japanese Society of Science and Football
• Japan Society of Physical education, Health and Sport Sciences
• Japanese Society of Test and Measurement in Health and Physical Education