



**HAL**  
open science

# What data should be collected for a good handball Expected Goal model?

Alexis Mortelier, François Rioult, John Komar

## ► To cite this version:

Alexis Mortelier, François Rioult, John Komar. What data should be collected for a good handball Expected Goal model?. 10th International Workshop on Machine Learning and Data Mining for Sports Analytics (MLSA 2023), Sep 2023, Turin, Italy. pp.119-130, <10.1007/978-3-031-53833-9\_10>. <hal-04534417>

**HAL Id: hal-04534417**

**<https://hal.science/hal-04534417v1>**

Submitted on 5 Apr 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# What data should be collected for a good handball *Expected Goal* model?

Alexis Mortelier<sup>1</sup>, François Rioult<sup>1</sup>, and John Komar<sup>2</sup>

<sup>1</sup> Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

<sup>2</sup> National Institute of Education, Nanyang Technological University, Singapore, Singapore

**Abstract.** Expected goal models (xG) are of great importance as they are the most accurate predictor of future performance of teams and players in the world of soccer. This metric can be modeled by machine learning, and the models developed consider an increasing number of attributes, which increases the cost of learning it. The use of xG is not widespread in handball, so the question of learning it for this sport arose, in particular which attributes are relevant for learning. Here, we used a *wrapper* approach to determine these relevant attributes and guide teams through the data collection stage.

**Keywords:** Handball · Expected goal · Machine learning

## 1 Introduction

*Expected goal* (xG)<sup>3</sup> models are important as accurate predictors of future team and player performance. At team level, these models are better at predicting performance than current goal difference or number of shots, such as the Total Shot Ratio (TSR)<sup>4</sup>.

Models of xG are commonly used for soccer and hockey [34]. Here we considered their use for handball, which is absent from the literature. Our study also attempted to clarify what types of data are required for the design of a robust xG model for handball. For example, can we be satisfied with analyzing player positions alone, or do we need to take account of match events (passing, fouls, shooting)? Taking each type of information into account represents a cost that needs to be considered: budgets and resources vary from one sport to another, and even from one team to another within the same sport. Depending on the available budget, data acquisition technologies will provide rich, comprehensive data, or poor, summary data.

We are therefore interested here in whether a dataset must necessarily be rich to obtain relevant xG values, and how this impacts its learning cost. Indeed, data availability affects the ability to learn good models of xG [29]. In the search for young talent, a minor league team has very little data, so it's interesting to define

<sup>3</sup> <https://www.statsperform.com/opta-analytics/>

<sup>4</sup> <https://www.scisports.com/total-shots-ratio/>

an approach that allows us to judge the quality of the xG with summary data, so as not to miss out on the next recruit.

To answer the question of the relevance of the attributes needed to calculate xG, we used *machine learning* techniques in this article. These techniques make it possible to establish a model of the xG and the conditions for calculating this model. Calculation performance (accuracy, convergence speed) indicates the relevance of the attributes used: if the attribute improves learning quality, then it is relevant to consider it. This approach is called *wrapper* [20].

## 2 State of the art

### 2.1 The xG metric

In collective ball sports, goals are the most important events. Depending on the sport, these goal events may occur rarely, as in soccer, where goal events have a low probability of occurrence during a match [37]. Analysts therefore prefer to focus their attention on shots, which are more numerous than goals. This has led to the creation of measures such as the *Total Shot Ratio* or TSR [21], which evaluates the dominance of teams based on their share of shots in matches. The limitation of TSR, however, is that it does not take into account the quality of a shot. For example, a shot on goal without a goalkeeper has the same value as a shot from midfield defended by an entire team.

With the democratization of data analysis, several derivatives of this measure were created. Shot differential, Fenwick<sup>5</sup> differential (shot differential between teams) and Corsi<sup>6</sup> differential (which considers shots on target and blocked shots differently) have become popular for analyzing the performance of hockey teams and players [23]. These statistics were chosen because each has proven to be a very good indicator of performance at team level.

As not all shots are equal, a method is needed to measure the quality of a given shot or series of shots, and this is how the xG models were developed. xG is a statistical metric based on the shooter's position to calculate a probability of turning a shot into a goal. However, in sports where goal events occur most frequently, such as handball, which has around 40 goals per match for two and a half times as many shots, xG models are not used.

Historically, xG models were introduced to soccer in 2012 by Sam Green for statistics site Opta<sup>7</sup>. In soccer, the company StatsBomb<sup>8</sup> provides one of the most recognized xG<sup>9</sup> calculation services. StatsBomb uses this metric<sup>10</sup> in a variety of ways. In soccer circles, it is a benchmark calculation method on a par with that provided by Opta.

<sup>5</sup> [https://en.wikipedia.org/wiki/Fenwick\\_\(statistic\)](https://en.wikipedia.org/wiki/Fenwick_(statistic))

<sup>6</sup> [https://en.wikipedia.org/wiki/Corsi\\_\(statistic\)](https://en.wikipedia.org/wiki/Corsi_(statistic))

<sup>7</sup> <https://www.statsperform.com/opta/>

<sup>8</sup> <https://statsbomb.com/>

<sup>9</sup> <https://statsbomb.com/soccer-metrics/expected-goals-xg-explained/>

<sup>10</sup> <https://statsbomb.com/what-we-do/soccer-data/>

Models for predicting the probability of a shot turning into a goal are based on a history of similar shots. For example, for ten shots at a position, if seven are turned into goals, then the xG value for the position will be 0.7. These models can consider a wide variety of attributes, giving varied information, in order to predict the xG value of a particular shot. StatsBomb and Opta use up to 35 attributes, each advancing an xG value closer to reality. These models share a base of attributes that include the angle or distance of the shot in relation to the goal, the position of the goalkeeper, the number of defenders exerting pressure on the shooter [16,39,25,15,9,6]. In general, the addition of new attributes does not seem to improve the quality of xG models, but may be relevant for certain isolated cases.

While the xG metric indicates the most favorable positions for scoring a goal, it does not provide information on the quality of a shot: a shot with a high xG value may be stopped by the goalkeeper, the mechanics of the shot may be poor, or it may go out of frame. xG gives an indication of the position in which a player should place himself to take the shot, in order to increase his chances of turning it into a goal.

## 2.2 Use of xG

In the literature, xG can be used in several ways. Comparing the value of xG to the number of goals scored can indicate a player’s scoring ability or luck. A player who regularly scores more goals than his total xG probably has above-average shooting/finishing ability [1].

xG models are used for a variety of reasons: predicting match scores [38,40], measuring scoring efficiency by estimating goal expectancy [30], judging the efficiency of soccer shots [28] or assessing the quality of positioning without the ball, preceding shots that may lead to goals [34]. The uses of the xG value are numerous and can be adapted to measure other qualities, such as passing, possession, etc.[41,26,24].

xG is also a collective performance attribute, as the model calculates the shooting situations that a team could have converted into a goal ( $xG_{inflicted}$ ) as well as the chances that could have led to a goal being conceded ( $xG_{conceded}$ ). A team’s goal differential ( $xG_{inflicted} - xG_{conceded}$ ) can indicate how a team should behave: a negative goal differential but a positive one could indicate that a team has had little luck or has a below-average finishing ability [1].

xG can be refined to differentiate between a team’s abilities in different situations: open play, free kick, corner, etc. For example, a team that has conceded more goals from free-kicks than its xG on free-kicks is probably below average at defending these set-pieces. A team’s  $xG_{conceded}$  is an indicator of its ability to prevent goal-scoring opportunities. A team will have a lower than predicted  $xG_{conceded}$  if it protects itself from shots with a high probability of being scored.

### 3 Handball data

#### 3.1 Data acquisition

There are two types of dataset for team sports: spatio-temporal data and event data.

The *spatio-temporal* data are the trajectories of the players or the ball ; they provide location, direction and speed information for each of the entities [12].

*Event* data provides logs of player-related events (pass, shot, catch, goalkeeper stop, etc.) or technical events (red card, start of time, stop of time, foul, etc.) for a match. Each of these events can be associated with a distance, duration or direction, and is time-stamped.

These two ways of looking at data require different technologies and technical or financial resources. For example, Indoor Position System (IPS) sensors [33] are inexpensive but not very accurate, while video [27,31,32] is accurate but requires expensive processing.

#### 3.2 Qatar 2015 data

The data analyzed here comes from 88 matches of the 2015 Handball World Championship held in Qatar. These data were obtained by video analysis and manual annotation of the events, the integrity of this trajectory capture system have been measured in [4]. This dataset has been little studied [5,13,14].

These data combine spatio-temporal and event-based information: each player's trajectory at 10 Hz and details of match events such as passes, receptions, shots and goals. For each match, a statistical summary table is available for each player. This table communicates individual information, such as walking distance, running distance, shooting distance at 7m and 9m, as well as the player's role for each match.

#### 3.3 Dataset enrichment

From this dataset, it is possible to pre-calculate various attributes providing additional information on the state of play, such as the distance of the shot from the goal, the angle of opening, etc. More complex attributes provide information on the pressure exerted on the shooter. For example, Voronoï [2] regions are used to determine the area a player can reach before all others (assuming all players have the same speed). If individual speeds are taken into account, the dominant region [36] or zone of influence [35] can be calculated. These tools are widely used to analyze team sports such as football [18].

Each player has a dominant zone. The boundaries of these zones are delimited by the bisectors between the players: the closer the neighbors, the smaller the player's zone. The area of the zone is ultimately indicative of the pressure exerted on a player.

The Delaunay triangulation [8] of a discrete set  $P$  of points is the dual graph of the Voronoï diagram associated with  $P$ . Each cell of the Voronoï diagram

is associated with a vertex in the Delaunay triangulation. These vertices are connected by an edge if the cells are adjacent. The number of edges connecting a player to defenders is a relevant indicator of the pressure exerted on that player.

## 4 Method for calculating xG

### 4.1 Sample calculation

The first method for calculating the xG metric involves sampling the data to determine the ratio between the number of goals scored and the number of shots taken for each shooting position. Since shots are always taken from different positions, the field is divided into a grid with cells of varying size, depending on the precision required. The ratio between the number of goals scored and the number of shots taken is then calculated for each grid cell. The ratio obtained reflects *the ground truth* provided by the data. These values can be visualized using a heat map. The advantage of this method is that only the shooter's position is required to determine this ground truth.

### 4.2 Model calculation

The second method for obtaining the xG metric is to design a model. A model is a function that takes as input various arguments such as the angle of the goal opening, the distance of the shot, etc., and gives as output the value of the xG.

For example, if we denote  $xG(d, \theta)$  the function of arguments  $d$  the shooting distance from the goal and  $\theta$  the opening angle of the goal, we could define this function as equal to  $\frac{1}{d\theta}$ . Indeed, this function has good properties because the goal probability is inversely proportional to distance and opening angle. However, the exact expression of this function is not known, not least because it depends on many other arguments, such as the size of the shooter.

What's more, each handball expert has his or her own unique personal experience and intuitions. It would be difficult to reach a consensus on the precise definition of the *expected goal*.

Determining the right model and therefore the exact expression of the function is a task taken on by machine learning, based on ground truth. For the remainder of this presentation, we have used seven of the most widely used learning algorithms: RANDOMFOREST [3], logistic regression [22], decision trees [19] and gradient boosting [7,17]: XGBOOST, CATBOOST and ADABOOST. We don't need to go into the technical details of how these algorithms work, but the essential point is to understand that, from a series of matches between shooting situations and the outcome of the shot, the algorithm calculates the most accurate model that functionally relates a shot to its outcome.

The models under study operate in a *supervised classification* mode. Their role is namely to classify shooting opportunities according to their outcome, positive or negative. In practice, these models calculate the probability of shot success and use an activation threshold to make the classification decision. The performance of a model is evaluated by its recall, precision, F1-score and area under the ROC curve [10].

### 4.3 *wrapper*: Attribute selection approach

In order to determine the handball attributes needed to learn a good handball model, a technique known as attribute selection is required. These techniques can be very costly if they have to exhaustively test all potential attribute combinations: for  $n$  attributes, there are in fact  $2^n - 1$  possible combinations. In practice, we distinguish between the *filter* [11] and *wrapper* [20] approaches. The *filter* approach examines the statistical characteristics of attributes, such as correlation coefficients. The *wrapper* approach iteratively selects the attributes that induce good performance of the learning algorithm employed (see Figure 1).

The *wrapper* approach thus performs a search in the space of attribute subsets, guided by the performance of the model learning algorithm. The evaluation criterion is simply the algorithm’s performance measure, which depends on the calculated model. The *wrapper* approach enables finer selection than the *filter* approach, but at the cost of a longer computation time.

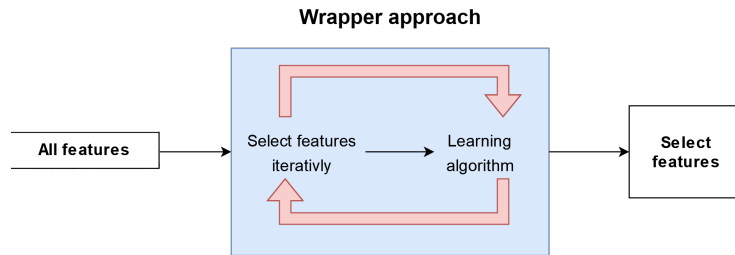


Fig. 1. *wrapper* approach.

In the experiments in section 5, we’re not looking for the optimal attribute configuration. We simply compare several emblematic families (see Table 1):

1. the first *pos* configuration considers only the shooter’s position ;
2. the second ( $pos + \theta + d$ ) adds the opening angle  $\theta$  of the goal and the distance to the goal  $d$  ;
3. *event data* contains position data plus information on match events (time elapsed since previous shot, goal differential, scores, etc.).
4. *spatial data* focus on the geometric characteristics of the game situation at the moment of shooting (goalkeeper position, shooter speed, number of defenders, etc.) ;
5. *combined data* consider all attributes.

The attributes chosen come from the literature on different xG models for football [16,39,25,15,9,6]. All these models use a base consisting of the shooter’s position, the opening angle of the goal and the distance to the goal.

However, distance and angle are calculated from the shooter’s  $x$  and  $y$  positions. They are therefore a priori redundant with the shooter’s position. This is

data	$pos$	$pos + \theta + d$	event	spatial	combined
shooter's $(x, y)$ position	×	×	×	×	×
angle $\theta$ of goal opening		×	×	×	×
distance from goal		×	×	×	×
time elapsed since previous shot			×		×
goal difference			×		×
shooter's team score			×		×
opposing team's score			×		×
does the shooter's team lead?			×		×
ball possession time			×		×
goalkeeper's position $(x, y)$				×	×
distance from goalkeeper to goal				×	×
shooter's speed				×	×
game situation (counter-attack or placed attack)				×	×
dominant region area				×	×
number of connections in the Delaunay graph				×	×
number of defenders				×	×

**Table 1.** Attribute families for calculating xG.

why the model will be trained, on the one hand with the positions alone, and on the other with the addition of distance and angle. This distinction will enable us to decide whether the position is sufficient to calculate a good model.

For event data, the attributes are directly derived or calculated from the raw data. The characterization of the game situation (placed attack or transition game) is determined from the positions of the team's players: after a change in the team in possession of the ball, a transition situation is characterized by a surplus of players, indicating the delay of one or more defenders.

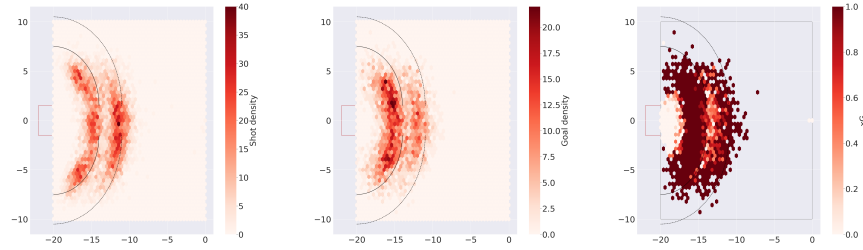
For spatial data, the area of the shooter's dominant Voronoï region and the number of neighbors in the Delaunay graph are considered. This information is thought to be indicative of the pressure exerted on the shooter.

## 5 Results

### 5.1 Sample calculation

On the Qatar 2015 data, shot density, goal density and xG are represented by heat maps in Figure 2. Although this article asks which attribute families should be considered for good xG learning, it seems relevant to visualize and analyze ground truth.

Most shots are taken from the 6m and 9m lines. On the other hand, their success seems to depend on the distance from the goal: the further away the shooter is, the more the angle of his shot seems likely to face the goal. The density suggests a shooting cone, with the wide base closest to the goal, and its apex closer to midfield. We can assume that there is a correlation between the



**Fig. 2.** Shot density (left), goal density (middle) and xG (right) in handball.

chance of success on goal and the angle and distance of the shooter. Due to the atypical topography of a handball pitch, which prohibits players from entering the 6m zone, we observe a higher success rate for shots in this area. Indeed, an attacker can enter the zone for a brief moment while in the air, and thus be in a duel with the goalkeeper, with no defender to get in the way.

## 5.2 wrapper approach

Table 2 shows performance (accuracy, precision, recall, etc.) for each attribute family depending on model chosen. Learning times are in seconds for the reference attribute family (*pos* only), and in percentage relatively to the reference.

In our experiments, we found that the RANDOMFOREST algorithm performed best. For RANDOMFOREST, the F1-score is between 0.75 and 0.78 – recall and precision are almost identical – and the AUC varies between 0.71 and 0.75. This is the sign of a high-performance model that predicts rather correctly whether there is a goal in a given situation. The other models have an F1-score between 0.66 and 0.72, and an AUC between 0.63 and 0.67: RANDOMFOREST performs much better than the others.

As attribute richness increases, the improvement in learning quality is not very significant (F1-score of 0.75 for positions alone, 0.78 for all attributes). An increase in learning time is observed between the simplest configuration (*pos*) and the richest configuration (*combined data*).

We can also see that there is very little difference between event and spatial attribute configurations. The extra cost involved in calculating spatial attributes, such as those associated with the Delaunay graph, which requires the position of all players, is not always justified. This is characteristic of handball, whereas in soccer, the richness of attributes is important and the consideration of spatial attributes is necessary.

## 6 Conclusion

Adapting the calculation of the xG metric to handball is a crucial step in analyzing team performance in this sport. An *wrapper* approach was used to determine

input	model	accuracy	precision	recall	F1-score	AUC	learning time
pos	RANDOMFOREST	<b>0.72</b>	<b>0.77</b>	<b>0.74</b>	<b>0.75</b>	<b>0.71</b>	258s
	LOGISTIC REGRESSION	0.64	0.64	0.69	0.66	0.63	16s
	DECISION TREE	0.67	0.75	0.69	0.72	0.66	2s
	XGBOOST	0.66	0.75	0.68	0.71	0.66	29s
	CATBOOST	0.65	0.71	0.68	0.70	0.65	106s
	ADABOOST	0.65	0.72	0.68	0.70	0.64	34s
pos+ $\theta$ +d	RANDOMFOREST	<b>0.73</b>	<b>0.78</b>	<b>0.74</b>	<b>0.76</b>	<b>0.72</b>	+62%
	LOGISTIC REGRESSION	0.63	0.64	0.69	0.66	0.63	+21%
	DECISION TREE	0.65	0.71	0.68	0.70	0.65	+28%
	XGBOOST	0.66	0.73	0.68	0.71	0.65	+33%
	CATBOOST	0.65	0.73	0.68	0.70	0.65	+60%
	ADABOOST	0.65	0.72	0.68	0.70	0.64	+70%
event data	RANDOMFOREST	<b>0.74</b>	<b>0.79</b>	<b>0.76</b>	<b>0.77</b>	<b>0.73</b>	+75%
	LOGISTIC REGRESSION	0.63	0.63	0.69	0.66	0.63	+1631%
	DECISION TREE	0.65	0.74	0.67	0.70	0.65	+700%
	XGBOOST	0.66	0.73	0.69	0.71	0.66	+177%
	CATBOOST	0.66	0.74	0.68	0.71	0.65	+54%
	ADABOOST	0.65	0.72	0.68	0.70	0.64	+70%
spatial data	RANDOMFOREST	<b>0.75</b>	<b>0.77</b>	<b>0.78</b>	<b>0.78</b>	<b>0.75</b>	+84%
	LOGISTIC REGRESSION	0.64	0.64	0.69	0.67	0.64	+768%
	DECISION TREE	0.65	0.73	0.67	0.70	0.65	+2000%
	XGBOOST	0.70	0.74	0.73	0.73	0.69	+134%
	CATBOOST	0.66	0.72	0.69	0.70	0.65	+148%
	ADABOOST	0.67	0.73	0.69	0.71	0.66	+114%
combined data	RANDOMFOREST	<b>0.75</b>	<b>0.77</b>	<b>0.78</b>	<b>0.78</b>	<b>0.75</b>	+49%
	LOGISTIC REGRESSION	0.64	0.64	0.69	0.67	0.64	+2406%
	DECISION TREE	0.65	0.67	0.69	0.68	0.64	+4450%
	XGBOOST	0.68	0.73	0.70	0.72	0.67	+144%
	CATBOOST	0.67	0.72	0.70	0.71	0.66	+102%
	ADABOOST	0.66	0.72	0.68	0.70	0.65	+238%

**Table 2.** Models performances on different attribute families.

the relevant attributes for learning xG. In soccer, the position of the shooter, the distance to the goal and the position of the defenders are relevant. However, adapting the soccer models for handball, which include a large number of attributes, did not significantly improve model accuracy.

The way in which the calculation of the xG metric is adjusted is influenced by the availability of data, as well as by the cost that the team is willing to incur to collect this data. In handball, our experiments have shown that it is not mandatory to invest huge resources in data collection, and that only the shooting positions could be sufficient. This is a significant advance in the analysis of player and team handball performance, which could have practical applications in the field of training and strategy.

## Acknowledgements

This work was partially funded by the ANR and the Normandy Region as part of the HAISCoDe project.

## References

1. Anzer, G., Bauer, P.: A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living* **3**, 624475 (2021)
2. Aurenhammer, F.: Voronoi diagrams : a survey of a fundamental geometric data structure. *ACM Computing Surveys* **23**(3), 345–405 (september 1991)
3. Breiman, L.: Random forests. *Machine learning* **45**, 5–32 (2001)
4. Buchheit, M., Allen, A., Poon, T.K., Modonutti, M., Gregson, W., Di Salvo, V.: Integrating different tracking systems in football: multiple camera semi-automatic system, local position measurement and GPS technologies. *Journal of sports sciences* **32**(20), 1844–1857 (2014)
5. Cardinale, M., Whiteley, R., Hosny, A.A., Popovic, N.: Activity profiles and positional differences of handball players during the world championships in qatar 2015. *International journal of sports physiology and performance* **12**(7), 908–915 (2017)
6. Cavus, M., Biecek, P.: Explainable expected goal models for performance analysis in football analytics (2022). <https://doi.org/10.48550/ARXIV.2206.07212>, <https://arxiv.org/abs/2206.07212>
7. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al.: Xgboost: extreme gradient boosting. *R package version 0.4-2* **1**(4), 1–4 (2015)
8. Delaunay, B.: Sur la sphère vide. *Proceedings du Congrès international des mathématiciens de 1924* pp. 695–700 (1924)
9. Fairchild, A., Pelechrinis, K., Kokkodis, M.: Spatial analysis of shots in mls: A model for expected goals and fractal dimensionality. *Journal of Sports Analytics* **4**(3), 165–174 (2018)
10. Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. *Technical Report HPL-2003-4*, HP Laboratories (2003)

11. Germano, M.: Turbulence: the filtering approach. *Journal of Fluid Mechanics* **238**, 325–336 (1992)
12. Gudmundsson, J., Horton, M.: Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR)* **50**(2), 1–34 (2017)
13. Hansen, C., Sanz-Lopez, F., Whiteley, R., Popovic, N., Ahmed, H.A., Cardinale, M.: Performance analysis of male handball goalkeepers at the world handball championship 2015. *Biology of sport* **34**(4), 393 (2017)
14. Hansen, C., Whiteley, R., Wilhelm, A., Popovic, N., Ahmed, H., Cardinale, M., et al.: A video-based analysis to classify shoulder injuries during the handball world championships 2015. *Sportverletzung Sportschaden: Organ der Gesellschaft für Orthopädisch-traumatologische Sportmedizin* **33**(1), 30–35 (2019)
15. Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., Meyer, T.: Machine learning in men’s professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching* **14**(6), 798–817 (2019). <https://doi.org/10.1177/1747954119879350>, <https://doi.org/10.1177/1747954119879350>
16. Hewitt, J.H., Karakuş, O.: A machine learning approach for player and position adjusted expected goals in football (soccer) (2023). <https://doi.org/10.48550/ARXIV.2301.13052>, <https://arxiv.org/abs/2301.13052>
17. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30** (2017)
18. Kim, S.: Voronoi analysis of a soccer game. *Nonlinear Analysis: Modelling and Control* **9**(3), 233–240 (Jul 2004). <https://doi.org/10.15388/NA.2004.9.3.15154>, <https://www.journals.vu.lt/nonlinear-analysis/article/view/15154>
19. Kingsford, C., Salzberg, S.L.: What are decision trees? *Nature biotechnology* **26**(9), 1011–1013 (2008)
20. Kohavi, R., John, G.H.: The wrapper approach. In: *Feature extraction, construction and selection*, pp. 33–50. Springer (1998)
21. Lago-Ballesteros, J., Lago-Peñas, C.: Performance in team sports: Identifying the keys to success in soccer. *Journal of Human kinetics* **25**(2010), 85–91 (2010)
22. LaValley, M.P.: Logistic regression. *Circulation* **117**(18), 2395–2399 (2008)
23. Macdonald, B.: Adjusted plus-minus for NHL players using ridge regression with goals, shots, fenwick, and corsi. *Journal of Quantitative Analysis in Sports* **8**(3) (2012). <https://doi.org/doi:10.1515/1559-0410.1447>, <https://doi.org/10.1515/1559-0410.1447>
24. Macdonald, B.: An expected goals model for evaluating NHL teams and players. In: *Proceedings of the 2012 MIT Sloan Sports Analytics Conference* (2012)
25. Madrero Pardo, P.: Creating a model for expected Goals in football using qualitative player information. Ph.D. thesis, UPC, Facultat d’Informàtica de Barcelona, Departament de Ciències de la Computació (Jun 2020), <http://hdl.handle.net/2117/328922>
26. Madrero Pardo, P.: Creating a model for expected goals in football using qualitative player information. Master’s thesis, Universitat Politècnica de Catalunya (2020)
27. Pettersen, S.A., Johansen, D., Johansen, H., Berg-Johansen, V., Gaddam, V.R., Mortensen, A., Langseth, R., Griwodz, C., Stensland, H.K., Halvorsen, P.: Soccer video and player position dataset. In: *Proceedings of the 5th ACM Multimedia Systems Conference*. pp. 18–23 (2014)
28. Rathke, A.: An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise* **12**(2), 514–529 (2017)

29. Robberechts, P., Davis, J.: How data availability affects the ability to learn good xG models. In: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (eds.) *Machine Learning and Data Mining for Sports Analytics*. pp. 17–27. Springer International Publishing, Cham (2020)
30. Ruiz, H., Lisboa, P., Neilson, P., Gregson, W.: Measuring scoring efficiency through goal expectancy estimation. In: *ESANN 2015 proceedings of the European symposium on artificial neural networks, computational intelligence and machine learning*. pp. 149–154 (2015)
31. Sanford, R., Gorji, S., Hafemann, L.G., Pourbabae, B., Javan, M.: Group activity detection from trajectory and video data in soccer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 898–899 (2020)
32. Scott, A., Uchida, I., Onishi, M., Kameda, Y., Fukui, K., Fujii, K.: Soccertrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3569–3579 (2022)
33. Serpiello, F., Hopkins, W., Barnes, S., Tavrou, J., Duthie, G., Aughey, R., Ball, K.: Validity of an ultra-wideband local positioning system to measure locomotion in indoor sports. *Journal of sports sciences* **36**(15), 1727–1733 (2018)
34. Spearman, W.: Beyond expected goals. In: *Proceedings of the 12th MIT sloan sports analytics conference*. pp. 1–17 (2018)
35. Taki, T., Hasegawa, J.: Visualization of dominant region in team games and its application to teamwork analysis. In: *Proceedings Computer Graphics International 2000*. pp. 227–235 (2000). <https://doi.org/10.1109/CGI.2000.852338>
36. Taki, T., Hasegawa, J.i., Fukumura, T.: Development of motion analysis system for quantitative evaluation of teamwork in soccer games. In: *Proceedings of 3rd IEEE international conference on image processing*. vol. 3, pp. 815–818. IEEE (1996)
37. Tenga, A., Ronglan, L.T., Bahr, R.: Measuring the effectiveness of offensive match-play in professional soccer. *European journal of sport science* **10**(4), 269–277 (2010)
38. Tiippana, T., et al.: How accurately does the expected goals model reflect goalscoring and success in football? (2020)
39. Umami, I., Gautama, D., Hatta, H.: implementing the expected goal (xG) model to predict scores in soccer matches. *International Journal of Informatics and Information Systems* **4**(1), 38–54 (2021). <https://doi.org/10.47738/ijjis.v4i1.76>, <http://ijjis.org/index.php/IJIS/article/view/76>
40. Umami, I., Gautama, D.H., Hatta, H.R.: Implementing the expected goal (xG) model to predict scores in soccer matches. *International Journal of Informatics and Information Systems* **4**(1), 38–54 (2021)
41. Van Haaren, J.: Why Would I Trust Your Numbers? On the Explainability of Expected Values in Soccer. arXiv preprint arXiv:2105.13778 (2021)