Machine Learning and Spatiotemporal Data: Overview of Data, Methods and Research in Football

Saša Tokić
Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: sasa.tokic@inf.uniri.hr

***Abstract*—The use of spatiotemporal data in football is multiplying in the last few years with the advancement of technology and the application of machine learning on this data is a growing trend both in research and practical applications in football clubs. This paper presents an overview of recent research and discusses the types and availability of data applicable for machine learning with a focus on spatiotemporal data in football. This paper provides guidelines for future research and application of new deep neural network-based approaches for sports analytics.**

***Keywords*—spatiotemporal data, sports analytics, player evaluation, tracking data, event data, deep neural network**

I. Introduction

Sports analytics has been an ever-evolving field since "Moneyball" [1] introduced the statistics-based approach for selecting baseball players. While many sports including football (soccer in the United States) have been data-rich for decades, most of this data was match sheet data or simple statistics that provides a minimal view of the game and does not help much in answering questions that are of interest to club analysts.

Rapid gathering and usage of spatiotemporal data in sports in the last decade represents a growing interest among researchers, and could also be of great practical application for (football) clubs. This paper presents an overview of recent research and discusses the types and availability of data applicable for machine learning with a special attention on spatiotemporal data in football.

While video analysis is still the de facto a standard in football analysis and scouting its application brings out two major issues:

1) Time to watch and analyze videos by humans takes a lot of time and cognitive effort, something which all but huge clubs can not afford
2) Video analysis by humans introduces cognitive biases making the analysis subjective rather than grounded in numbers

Spatiotemporal event data helps in answering many practical questions of much interest in football clubs like:

1) What is the probability of scoring a goal from a given situation?
2) What is the value of a particular pass?
3) What are common tactics opponents use against teams similar to ours?
4) Which player style is the most similar to the player club lost due to transfer or injury?

In this paper, we focus on team sports, which we define as any contact sport involving two teams and an object, usually a ball or a puck, where the game's goal is to put a ball in the opponent's goal. We will focus on football while mentioning recent research in basketball and hockey, given that the majority of the state of the art results are coming from researchers focused on these sports. Gudmundsson [2] suggests that there has been little research in football regarding spatially informed metrics and poses a question if it is possible to develop similar metrics like in other sports. While research since that publication suggests that it is indeed possible to derive new football metrics, we should note that football has some significant differences compared to other sports. Compared to basketball, football has a much lower number of points scored per game, affecting goal-related metrics and making them less reliable.

This paper is organized as follows: first, we introduce the domain of sports analytics and systemize the types of data with the emphasis on data availability where we propose possible further research in obtaining tracking data; second, we provide an

overview of recent research in the field of sports analytics with an emphasis on the spatiotemporal data analysis in football; third we show an example of applying deep learning architecture on event-based tracking data; finally we provide a conclusion of the current state and suggest further research directions possible

II. Data

Broadly speaking football related data can be grouped into three categories (see Figure 1):

1) Match sheet data
2) Event data
3) Tracking data

Data is gathered and provided by specialized providers. While the match sheet and statistics are usually provided for free by various websites and services like FBref [3] event and tracking data are, in general, not free and available. Luckily for researches, there are existing open datasets of event data provided by Wyscout [4] and Statsbomb [5].

Tracking data include Opta [6], Signality [7], SecondSpectrum[8], Metrica [9] and others.

To the best of our knowledge, Metrica Sports provides the only two publicly available games covered by tracking data. [9]. The recent addition to this is nine matches of broadcast tracking data, that is tracking data collected from commercial video broadcast provided by SkillCorner [10]

Match sheet data provides a high-level summary of the game or specific club/player during the game. Most of this data is freely available, with some providers even encouraging researchers to extract data from their websites.

Even though many sources like FBref [3] provide much more detailed statistical data (see Figure 2) than usually found in match sheet data, such as adding Expected Goals and Expected Assists, this type of data still lack much of the granularity of the other two data types typically used in football analysis nor can this type of data, in authors opinion, help clubs and scouts in answering practical tactics related questions they have.

A. Data types

Spatiotemporal data represents a type of data that consists of both time and space, as its name suggests. in sports, this data is very granular, and it usually represents 20-30Hz of data, meaning there are 20-30 data points per second representing the current positions of players and the ball or puck.

There are two types of data available for spatiotemporal research: tracking data and event data.

Tracking data can be obtained in 3 ways; static cameras on stadiums with human verification, tracking data from commercial video broadcasts and tracking data from GPS devices.

Tracking data consists of many data points. Tacking systems consisting of several cameras generate this type of data, and it is a very low level with data points usually 10-30Hz representing the player and ball positions. While this data is precious, it is also challenging for researchers to obtain the data due to its high commercial value.

Event data is a more sparse type of data, similar to tracking data; it consists of current player and ball positions, but only after a particular event happened like foul, goal, pass, or other. Although this data also has an enormous commercial value, there are already freely available datasets suitable for research.

The biggest freely available dataset is described by Pappalardo in [12] Data covers seven biggest European leagues plus World Cup 2018 and European cup 2016 as shown in Table I:

TABLE I. Competitions and corresponding data [12, Table 1.]

Competition	**#matches**	**#events**	**#players**
Spanish first division	380	628.659	619
English first division	380	643.15	603
Italian first division	380	647.372	686
German first division	306	519.407	537
French first division	380	632.807	629
World cup 2018	64	101.759	736
European cup 2016	51	78.14	552
	1.941	3,251,294	4.299

Event data consists of different types of events like pass, foul, and others with subtypes like cross-pass or simple-pass. Additionally, some providers provide tags to each event which report more details about a particular event

Event data is usually provided as JSON files or via providers' API as JSON responses (see Figure 3).

Differences in formats and data provided by various providers represent one of the engineering challenges. While also having different data structures, providers also might have different critical information. One of the providers has more human entered mistakes in their data than the other, resulting in

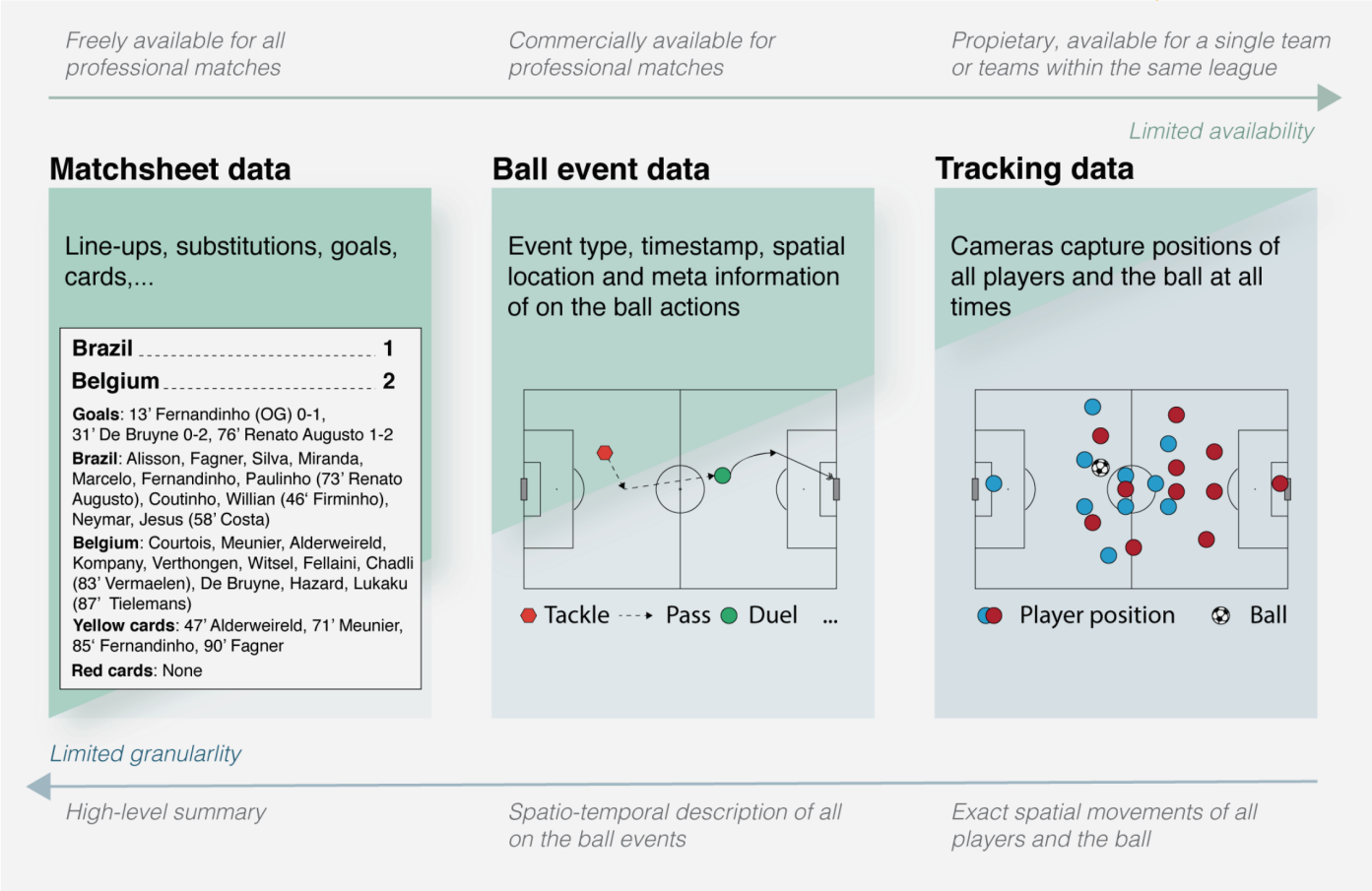

Fig. 1. Broad categorization of football data [11, page 37.]

Standard Stats 2020-2021 Liverpool: Premier League Share & more ▼ Glossary

				Playing Time			Performance						Per 90 Minutes					Expected			Per 90 Minutes					
Player	Nation	Pos	Age	MP	Starts	Min	Gls	Ast	PK	PKatt	CrdY	CrdR	Gls	Ast	G+A	G-PK	G+A-PK	xG	npxG	xA	xG	xA	xG+xA	npxG	npxG+xA	Matches
Virgil van Dijk	NED	DF	29	4	4	360	1	0	0	0	1	0	0.25	0.00	0.25	0.25	0.25	0.3	0.3	0.0	0.08	0.00	0.08	0.08	0.08	Matches
Andrew Robertson	SCO	DF	26	4	4	360	1	1	0	0	0	0	0.25	0.25	0.50	0.25	0.50	0.8	0.8	0.4	0.19	0.09	0.28	0.19	0.28	Matches
Mohamed Salah	EGY	FW	28	4	4	360	5	0	2	2	0	0	1.25	0.00	1.25	0.75	0.75	3.1	1.6	1.5	0.77	0.38	1.15	0.40	0.79	Matches
Georginio Wijnaldum	NED	MF	29	4	4	360	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	1.5	1.5	0.2	0.36	0.04	0.41	0.36	0.41	Matches
Trent Alexander-Arnold	ENG	DF	21	4	4	358	0	1	0	0	1	0	0.00	0.25	0.25	0.00	0.25	0.3	0.3	0.6	0.06	0.16	0.23	0.06	0.23	Matches
Roberto Firmino	BRA	FW	28	4	4	331	0	2	0	0	1	0	0.00	0.54	0.54	0.00	0.54	1.3	1.3	1.2	0.35	0.32	0.67	0.35	0.67	Matches
Naby Keïta	GUI	MF	25	4	4	243	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.4	0.4	0.1	0.14	0.03	0.17	0.14	0.17	Matches
Fabinho	BRA	MF,DF	26	4	3	303	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.1	0.00	0.04	0.04	0.00	0.04	Matches
Alisson	BRA	GK	27	3	3	270	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	Matches
Sadio Mané	SEN	FW	28	3	3	259	3	0	0	0	1	0	1.04	0.00	1.04	1.04	1.04	2.7	2.7	1.0	0.95	0.35	1.30	0.95	1.30	Matches
Joe Gomez	ENG	DF	23	3	3	240	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.01	0.01	0.00	0.01	Matches
Jordan Henderson	ENG	MF	30	2	2	110	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.1	0.02	0.07	0.09	0.02	0.09	Matches
Diogo Jota	POR	FW	23	2	1	101	1	0	0	0	0	0	0.89	0.00	0.89	0.89	0.89	0.5	0.5	0.0	0.41	0.03	0.44	0.41	0.44	Matches
Adrián	ESP	GK	33	1	1	90	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	Matches
James Milner	ENG	MF	34	3	0	62	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.04	0.04	0.00	0.04	Matches
Curtis Jones	ENG	MF	19	2	0	55	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.05	0.05	0.00	0.05	Matches
Takumi Minamino	JPN	FW,MF	25	3	0	51	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.09	0.00	0.09	0.09	0.09	Matches
Thiago Alcántara	ESP	MF	29	1	0	45	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.03	0.06	0.09	0.03	0.09	Matches
Joël Matip	CMR	DF	28	1	0	2	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	Matches
Caoimhin Kelleher	IRL	GK	21	0	0																					Matches
Divock Origi	BEL	FW,MF	25	0	0																					Matches
Kostas Tsimikas	GRE	DF	24	0	0																					Matches
Neco Williams	WAL	DF	19	0	0																					Matches
Squad Total				**4**	**44**	**360**	**11**	**4**	**2**	**2**	**4**	**0**	**2.75**	**1.00**	**3.75**	**2.25**	**3.25**	**10.6**	**9.2**	**5.3**	**2.66**	**1.33**	**3.99**	**2.29**	**3.62**	

Fig. 2. Standard statistical data for Liverpool, season 20/21 [3, screenshot by author]

skewed research data if the mistake is severe.

To deal with different vendors' approach to event data Decroos in [13] proposes "SPADL" Soccer Player Action Description Language) to unify and represent the data in the same way, thus abstracting away from specific vendor language and options. While commercial vendors might be interested in all the game events, research and clubs are usually focused on action events. While the end of the game is an "event," it does not carry much weight for research purposes; thus, SPADL considers only real actions like passes, which are of much interest not only to researchers but also to club stakeholders. Visualizing SPADL action can be seen in Figure 4.

SPADL defines different attributes as opposed to commercial vendors, these include [14]:

1) Time: the time in the game when the action occurred
2) StartLocation: the (x, y) location where the action started
3) EndLocation: the (x, y) location where the action ended
4) Player: the player who performed the action
5) Team: the player's team
6) ActionType: the type of the action (e.g., pass, shot, dribble)
7) BodyPart: the player's body part used for the action
8) Result: the result of the action (e.g., success or fail)

Extracting tracking data from commercial video broadcast provides another approach for obtaining valuable tracking data while considering its limitations. It would also lower the barrier of entry for the general public interested in football analytics.

TABLE II. EVENT TYPES, SUBTYPES AND TAGS [12, TABLE 2.]

type	**subtype**	**tags**
pass	cross, simple pass	accurate, not accurate, key pass, opportunity, assist, goal
foul		no card, yellow, red, 2nd yellow
shot		accurate, not accurate, block, opportunity, assist, goal
duel	air duel, dribbles, tackles, ground loose ball	accurate, not accurate
free kick	corner, shot, goal kick, throw in, penalty, simple kick	accurate, not accurate, key pass, opportunity, assist, goal
offside		
touch	acceleration, clearance, simple touch	counter attack, dangerous ball lost, missed ball, interception, opportunity, assist, goal


```
{
  "id" : "577a60d2-e469-4b7d-9b68-b0011da2f351",
  "index" : 5,
  "period" : 1,
  "timestamp" : "00:00:00.968",
  "minute" : 0,
  "second" : 0,
  "type" : { "id" : 30, "name" : "Pass"},
  "possession" : 2,
  "possession_team" : { "id" : 210, "name" : "Real Sociedad"},
  "play_pattern" : { "id" : 9, "name" : "From Kick Off" },
  "team" : { "id" : 210,"name" : "Real Sociedad"},
  "player" : {"id" : 6695,"name" : "Juan Miguel Jiménez López"},
  "position" : {"id" : 19,"name" : "Center Attacking Midfield"},
  "location" : [ 61.0, 41.0 ],
  "duration" : 2.756,
  "related_events" : [ "0f648804-fdf7-4d4f-b165-80b47eee9da7" ],
  "pass" : {
    "recipient" : {
      "id" : 6693,
      "name" : "Raúl Rodríguez Navas"
    },
    "length" : 16.643316,
    "angle" : -2.5702553,
    "height" : {
      "id" : 1,
      "name" : "Ground Pass"
    },
    "end_location" : [ 47.0, 32.0 ],
    "type" : {
      "id" : 65,
      "name" : "Kick Off"
    },
    "body_part" : {
      "id" : 40,
      "name" : "Right Foot"
    }
  }
}
```


Fig. 3. Pass event example in JSON format [5, image created by author]

Recent research by Johnson [16] in parsing player tracking data in basketball, using a video feed from a single non-stationary camera shows ninety four point five percent of placing players within a foot of their actual location. Although the research was conducted on a basketball video feed, which is different from the usual football feed, it indicates a

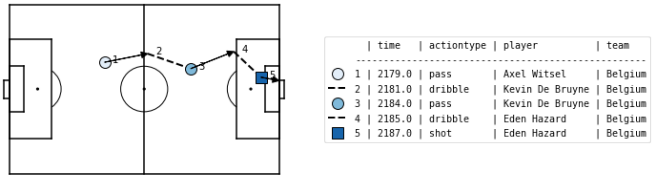

Fig. 4. Visualizing SPADL action sequence leading to a goal [15, Image 2]

promising direction for obtaining detailed football tracking data. Combined with research from Komorowski [17], where authors used a deep neural network-based detector for detecting ball in videos, we expect further advancement in this area.

On a similar note, extracting tracking data from tactical cameras might produce more reliable data in the football context. Tactical cameras are placed on high positions on the stadiums, they are stationary, and they capture the whole field and all twenty two players and the ball from a single point. While conventional methods for tracking objects in videos could provide decent tracking accuracy, researchers' problem is availability of such footage. Many stadiums are equipped with tactical cameras, but the footage is usually not freely available. However, there is public footage from tactical cameras for matches in the Men's World Cup 2018 (see Figure 5).

Fig. 5. View from a tactical camera, Brasil - Belgium World Cup 2018 - created by author

III. Related Work

A. Game analysis, strategy and related tasks

Moreover, what seems to be of pressing issue regarding current research involving tracking data is the question of reproducibility. While researchers' scientific integrity is not in question, it can pose a significant reproducibility problem and throttle future research.

Fernández et al. in [18] presents "Soccermap"; a fully CNN [19] architecture (see Figure 6) which calculates probability surfaces of potential passes (see Figure 8). The network was trained on high-frequency tracking data. By changing the output activation function, the authors conclude that the same architecture can be applied to two different problems; the estimation of pass-selection likelihood and predicting the expected value of a pass.

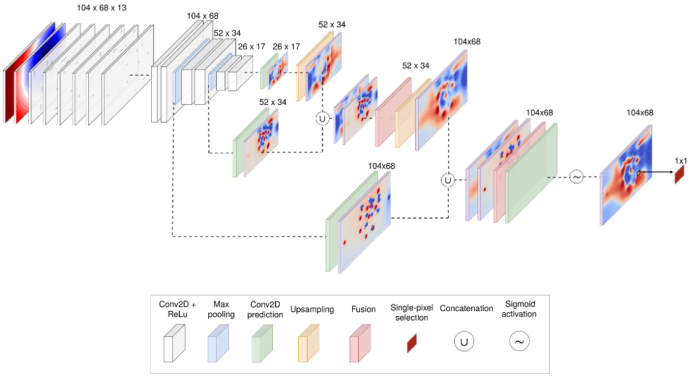

Fig. 6. "Soccermap" neural network architecture [18, Fig. 1]

Goes et al. in [20] analyzed 118 matches in the Dutch first league using positional tracking data collected at 10Hz. By applying unsupervised machine learning (KMeans), they identified dynamic formations of teams to classify successful attacks. They conclude that subgroup-level variables provide more information than team-level variables and that it is possible to identify those subgroups from positional tracking data. Practical applications of the conclusion suggest that defenders creating space for attackers are strongly dependent on those attacks' success.

Verstraete et al. in [21] analyzed SoFIFA dataset [22] as data tensor using CPD (canonical polyadic decomposition) [23] to extract interpretable latent structures. They show how grouping related skills are possible by using discovered latent structures. The authors also suggest that each player can be summarized by using a linear combination of structures. By applying Tucker decomposition of a tensor [24], authors elaborate on how a particular player's skills evolve with age. Interestingly Burzykowski in [25] demonstrates model interpretability possibilities on the same dataset.

Nunez and Dagnino in [26] applied often used metrics in football, pitch control, expected possession value and expected goals in a weighted function in order to create a competitive simulated game. Their agent was able to rank in top one point five percent in currently running Google Research Football competition [27].

Liu et al. in [28] deployed Deep Reinforcement learning [29] to extract complex dynamics from spatiotemporal data in football. The author claims

to have designed the most complex neural network architecture deployed in sports to date; a stacked two tower LSTM [30]. The network was trained on 4.5M action events from several European leagues. Authors developed a new metric called GIM, which correlates with standard success metrics in football with the possibility to fine-tune the metric to a specific league.

Decroos et al. in [31] constructed player vectors by transformation of event data using non-negative matrix factorization [32] producing a complete view of a player's style. In turn, these vectors can be used in machine learning analysis in clustering, nearest neighbor [33], or other suitable models for identifying players with a similar style, which is a crucial question in both scouting and game preparation.

Decroos et al. in [34] introduced a variation on their previous work called Atomic-SPADL. The main difference this model introduced is treating all events as successful, meaning for example that if the pass was intercepted by the opponent team that creates a new event called "interception." While using xT model as a benchmark, which has a Pearson correlation of 0.89, their new model produced a Pearson correlation of 0.65. Although lower than the benchmark the result was better than their previous work which combined VAEP and SPADL that produced a Pearson correlation of 0.25. correlation of 0.89

Beal et al. in [35] model 2018 FIFA World Cup data as chains of interaction modeled as walks within graphs. The authors tested various network metrics to value players' contributions and sets of players based on such graphs. Authors conclude that their model can produce similar team selection as that of human coaches.

Pappalardo et al. in [36] developed "PlayeRank" a data-driven framework for role aware player performance evaluation. Authors build a three-phase approach and utilize Linear Support Vector Classifier (LSVC) [37] for the rating phase.

Groll et al. in [38] developed a hybrid model as a combination of random forests [39] and Poisson ranking. They have analyzed all matches from 4 FIFA World Cups from 2002 to 2014. The authors used the upcoming 2018 world cup as an independent test dataset and concluded that their hybrid model has greater predictive power than other methods, including betting odds.

Goes et al. in [40] analyzed a data-driven model for measuring pass effectiveness in professional football. Data used was tracking and pass data for 18 matches of 1 team in 2017–2018 Dutch premier league. The authors developed two new metrics for evaluating pass value while maintaining a goal of not overvaluing forward passes. Methods used were manual model creation and PCA (Principal component analysis) [41].

Dick et al. in [42] applied deep reinforcement learning [29] to learn valuations of multiple player positioning using positional data. The data used was tracking data at 25 Hz. The authors used a neural network to learn value function while modeling soccer matches as Markov processes similar to Yam in [43] where the pass events were modeled purely as Markov chains [44] to develop a ball progression model to detect most valuable players in Europe.

Decroos et al. in [45] proposed an improvement to VAEP model [46] which uses gradient boosting tree [47] with 151 features by developing a Generalized Additive Model (GAM) with 10 features thus improving interpretability while maintaining similar performances.

Bransen et al. in [48] introduces ECOM (Expected Contribution to the Outcome of the Match), a new metric which aims to measure players' contribution in creating goal-scoring chances while valuing they are passed. Dataset used was event data top 7 European leagues in 4 seasons, and the method used was distance-weighted k-nearest-neighbors search.

Zambom-Ferraresi et al. in [49] used the Bayesian model averaging [50] to discover determinants of sports performance in the top five European football leagues during two seasons. Their results suggest concrete attributes in sports performance analysis, suggesting that attacking actions carry more value than defensive ones.

Steiner et al. used positional data to estimate the effects of contextual features on passing decisions in football by employing first binary logistic regressions [51] to test relations between predictors and later regression models.

Pappalardo et al. in [52] used machine learning to construct a prediction model that finds team ranking in the future season by using data from previous seasons. They have analyzed 10 million events from the top 6 European leagues using OLS regression and logit classification.

McHale et al., in [53], used tracking data at 10Hz

and positional attributes of the players to identify key players in a team. Their model utilizes the probability of a successful pass and network centrality measures. For probability calculations, they show that the generalized additive mixed model produces good results. Authors suggest their findings could help trainers and scouts identify vital players in either opposition teams when recruiting new talents.

Giancola et al. in [54] focused on detecting events in football broadcast videos in order to provide a benchmark dataset for future research. They show 67,8 percent (mAP) performance in classification using ResNet-152 features and average-mAP of 49.7 percent with good annotation and 40.6 percent using weakly annotated data.

Decroos et al. in [55] developed a five-step process to discover tactics of football teams in spatiotemporal data. The dataset and approach were evaluated on the 2015/16 English Premier League event data. They approached the task by dividing the event data into phases, clustering those phases based on spatiotemporal components, ranking clusters. This proceeded with mining clusters to identify patterns and finally rank discovered patterns.

Decroos et al. in [56] developed "SoccerMix," a technique for soft clustering that enabled probabilistic representations of football actions using mixture models. They elaborate on how their approach can understand both teams' styles and recognize how one team can force opponents to change their typical style.

Decroos et al. in [57] designed a three-step approach for evaluating player performance. First, like in their other work, they split event data into phases, and by applying dynamic time warping, they rate each phase. The last step rates the actions by applying an exponential-decay-based approach. This approach enables us to find top performing players in a league or a particular match.

Steiner et al. in [58] applied regression model with four input features like the openness of passing lane, position to a ball carrier, spatial proximity, and defensive coverage which multiplied by beta coefficients calculated passing decision, that is to whom the player is most likely to pass the ball.

Horton et al. in [59] showed a model that learned to classify the quality of passes in football with an accuracy of 85.8 percent, which was compared to human observers. The model was based on computational geometry features fed into different classifiers (MLR, SVM, RUSBoost...).

Brooks et al. in [60] describe a supervised machine learning model (L2-regularized Sup- port Vector Machine (SVM) model) [37] trained in event data of the 2012/13 Spanish La Liga season to rank players based on the value of their passes alone. They create a value metric based on shot opportunities created in connection to pass locations. They conclude that predicting possession of the ball at a specific location will end up with a shot to the goal has an F-score of 0.31 and AUROC (Area Under the Receiver Operating Characteristic Curve) of 0.79.

Brooks et al. in [61] presented two approaches to finding insights in a football game by focusing on passes. The first experiment created heat-maps for each team in their event data (2012/13 Spanish La Liga). This is then used to create a unique team identification by applying the KNN model [33] with 87 percent accuracy. They also demonstrate that using supervised machine learning shots on goal can be predicted from possession information.

McHale et al. in [62] presented a model for identifying ability of football players to score goals. The model was based on event data from 2 seasons, and it was created as a mixed-effects model.

Bialkowski et al. in [63] on an entire season of tracking data with about 400,000,000 data points developed a method for analyzing both individual players and teams using minimum entropy data partitioning and expectation-maximization (EM) algorithm [64], similar to k-means [65].

Bialkowski et al. in [66] presented a method for identifying teams from spatiotemporal tracking data. The authors applied the formation of a descriptor, which was found by minimizing the entropy of role-specific maps. Authors match descriptors and multiplying it by LDA transformation with team identity predicted by applying k-NN [33].

Bunker et al. in [67] applied neural network with 10 fold Cross-Validation for predicting outcomes of football matches.

B. Game result prediction

Applying machine learning for result prediction constitutes a separate direction entirely, in our opinion, since it is not directly related to insights that club stakeholders need. While at the same time being an interesting research question for both academics and amateur gamblers. Here we briefly list

work related to result prediction using machine learning.

Tewari et al. in [68] developed a prediction system for English Premier League by applying XGboost [69], SVM [37], and Logistic regressions models with XGboost having the best F-score.

Baboota et al. in [70] applied Gaussian naive Bayes, SVM, Random forest, and gradient boosting models to predict the outcomes of EPL matches. The model with the best result was gradient boosting achieved a ranked probability score of 0.2156 while the benchmark from betting organizations was 0.2012, meaning their model could not beat the betting markets.

Razali et al. in [71] used Bayesian networks to predict EPL matches in seasons 2010/11-13. By applying K-fold cross-validation, they calculated their models' accuracy to be 75.09 percent.

Danisik et al. in [72] constructed a neural network based on the LSTM regression model with input data taken from the video game "FIFA" combined with real-world matches. Their model with cross-validation produced an accuracy of 52.479 percent, slightly below bookmaker accuracy.

Cho et al. in [73] applied social network analysis of football passes and gradient boosting for predicting match outcomes. In comparisons with SVM [37], neural networks [74], decision trees, and logistic regression [51], authors conclude that their approach can provide an accurate prediction system with accuracy varying from 0.38 to 0.75 depending on the season the model was tested and the league's phase.

Ulmer et al., in [75], used seven different models that were tested against the baseline model. Models included Gaussian Naive Bayes [76], Hidden Markov Model, Multinomial Naive Bayes, RBF SVM [77], Random forest [39], Linear SVM [37], One vs. All SGD. Dataset consisted of 10 seasons of EPL from 2002 till 2012, while the test dataset was season 2013/14. Their best error rates came from Linear classifier (.48), Random Forest (.50), and SVM (.50).

IV. Deep neural network on event data

The more affordable access to event data than tracking data, an exciting research direction is applying deep neural networks [74] on said data. An example of such an experiment is shown by Pleuler in [78] and verified by the author.

Layer (type)	Output Shape	Param #	Connected to
pass_input (InputLayer)	[(None, 52, 34, 3)]	0	
conv2d (Conv2D)	(None, 50, 32, 16)	448	pass_input[0][0]
lambda (Lambda)	(None, 52, 34, 16)	0	conv2d[0][0]
conv2d_1 (Conv2D)	(None, 52, 34, 1)	17	lambda[0][0]
max_pooling2d (MaxPooling2D)	(None, 26, 17, 1)	0	conv2d_1[0][0]
conv2d_2 (Conv2D)	(None, 24, 15, 32)	320	max_pooling2d[0][0]
lambda_1 (Lambda)	(None, 26, 17, 32)	0	conv2d_2[0][0]
conv2d_3 (Conv2D)	(None, 26, 17, 1)	33	lambda_1[0][0]
up_sampling2d (UpSampling2D)	(None, 52, 34, 1)	0	conv2d_3[0][0]
conv2d_4 (Conv2D)	(None, 50, 32, 16)	160	up_sampling2d[0][0]
lambda_2 (Lambda)	(None, 52, 34, 16)	0	conv2d_4[0][0]
conv2d_5 (Conv2D)	(None, 52, 34, 1)	17	lambda_2[0][0]
dest_input (InputLayer)	[(None, 52, 34, 1)]	0	
concatenate (Concatenate)	(None, 52, 34, 2)	0	conv2d_5[0][0] dest_input[0][0]
lambda_3 (Lambda)	(None,)	0	concatenate[0][0]

Total params: 995
Trainable params: 995
Non-trainable params: 0

Fig. 7. Deep model architecture on World Cup 2018 event data [78, Fig 2.]

This experiment is inspired by [18] taking into account limitations of event data, mainly missing information about the position of other players.

This model's loss function is binary cross-entropy, which is appropriate given that the result of a pass event is either success or failure. The model was trained on 30 epochs with an "Adam" optimizer.

Compared to the original "SoccerMap," which has 401,259 parameters, this architecture consists of only 995 parameters . Even a relatively simple architecture provides good results (see Figure 9) in pass prediction probabilities (see Figure 8 as an example), which suggests that other approaches based on tracking data could be applied to event data as well. Additionally, layers in this model were reduced to 1/2 and 1/4 size of the original 104 x 68 size.

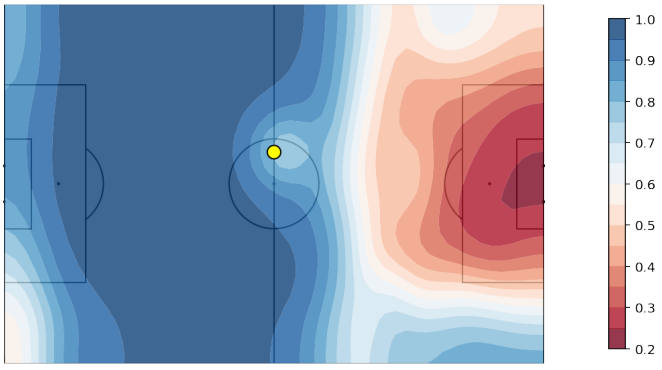

Fig. 8. Single pass probability surface [78, Fig 1.]

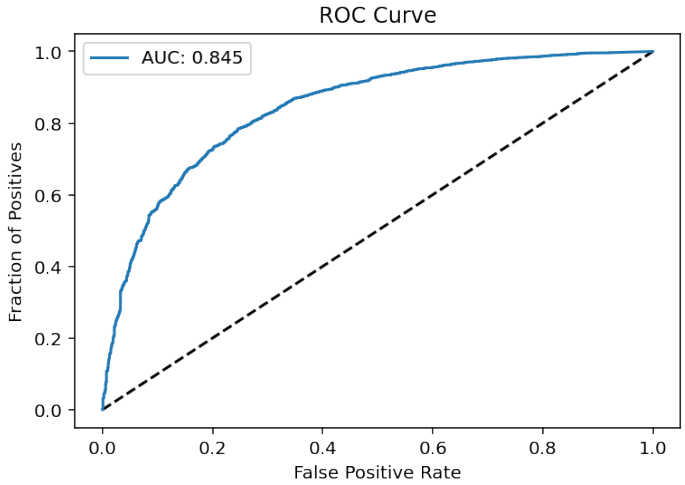

Fig. 9. AUC of architecture tested on World Cup 2018 event data [78, Fig 3.]

We propose further research based on this model in two directions. First, the model should be compared to simpler architectures and hyper-parameter tuning should be done in order to improve prediction accuracy. Second, this approach takes into account individual pass events. We suggest that the sequence of events, i.e., actions, can provide more information about probabilities of successful passes similar to the approach in [13]. With that in mind, we hypothesize that converting event data to SPADL actions and adapting architecture to accommodate "previous" information for each pass should produce higher total accuracy.

V. Future research

With recent publications, it is noticeable that the trend of event and tracking data analysis in football is shifting towards deep neural networks, while deep reinforcement learning is shown in several papers as a promising research direction. With Google and Manchester City powered Kaggle competition just recently [27] we expect more research in this area.

While there has been some research in using generative adversarial networks in basketball like [79] to the best of our knowledge, there is no similar research in football, which is another possible research direction.

Of much interest to the author is the possibility of applying knowledge gained from real-world tracking or event data in simulated environment like Google Research Football [80] with a goal of developing a realistic simulation. While the question of applicable metrics is up to a debate we suggest comparing such simulations against Google Research Football competition [27] and current reinforcement learning algorithms applied in [80]; Proximal Policy Optimization (PPO) [81], Impala [82] and Ape-X DQN [83].

Based on research in other sports like [84] and current state of research in football we expect to see much more focus on three key methods; deep neural networks [74], deep reinforcement learning [29] and generative adversarial network [85], example in basketball in [79].

VI. Conclusion

Sports analytics, especially football analytics, attracted much interest in the last years with various methods employed from classical statistical approaches, through classical machine learning approach to most recent usage of deep neural networks and deep reinforcement learning models. With similar sports approaches, mainly basketball and hockey, we can expect many advances and more complicated models deployed very shortly. The problem of data availability remains to be solved. Some commercial providers like Metrica Sports are trying to make some datasets free and open for researchers. We have no reason to believe that research will have multiple tracking data games available in the near time. While tracking data is not accessible to most of the interested parties, event data is not only freely available in limited amounts, as previously shown, but it is also relatively affordable and accessible.

References

[1] M. Lewis, *Moneyball: The Art of Winning an Unfair Game*, 1st ed. New York: W. W. Norton, 2003, 288 pp., ISBN: 978-0-393-05765-2.

[2] J. Gudmundsson and M. Horton, "Spatio-Temporal Analysis of Team Sports – A Survey," *ACM Computing Surveys*, vol. 50, no. 2, pp. 1–34, Apr. 11, 2017, ISSN: 03600300. DOI: 10.1145/3054132. arXiv: 1602.06994. [Online]. Available: http://arxiv.org/abs/1602.06994 (visited on 11/16/2019).

[3] (Sep. 28, 2020). "FBref - Football Statistics and History," [Online]. Available: https://fbref.com/en/ (visited on 10/09/2020).

[4] (Sep. 28, 2020). "Wyscout - Football Professional Videos and Data Platform," [Online]. Available: https://wyscout.com/ (visited on 10/10/2020).

[5] (Sep. 28, 2020). "StatsBomb — Football Like Never Before," [Online]. Available: http://statsbomb.com/ (visited on 10/10/2020).

[6] (Sep. 28, 2020). "Opta," [Online]. Available: https://www.optasports.com (visited on 10/11/2020).

[7] (Sep. 28, 2020). "Signality," [Online]. Available: https://www.signality.com/ (visited on 11/10/2020).

[8] (Sep. 28, 2020). "Second Spectrum," [Online]. Available: https://www.secondspectrum.com/index.html (visited on 11/10/2020).

[9] (Sep. 28, 2020). "Metrica Sports Sample Data," [Online]. Available: https://github.com/metrica-sports/sample-data (visited on 10/10/2020).

[10] *SkillCorner/opendata*, SkillCorner, Sep. 28, 2020. [Online]. Available: https://github.com/SkillCorner/opendata (visited on 10/10/2020).

[11] J. Castellano, N. Evans, D. Link, L. Pappalardo, D. Memmert, S. Robertson, J. Sampaio, M. Stein, V. de Boode, C. Clemens, G. Berhalter, P. Kluivert, A. Putellas, E. Valverde, A. Zubizarreta, and R. Moreno, *Football Analytics: Now and Beyond. A Deep Dive into the Current State of Advanced Data Analytics*. 2020.

[12] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," *Scientific Data*, vol. 6, no. 1, p. 236, Dec. 2019, ISSN: 2052-4463. DOI: 10.1038/s41597-019-0247-7. [Online]. Available: http://www.nature.com/articles/s41597-019-0247-7 (visited on 02/10/2020).

[13] T. Decroos, P. Robberechts, and J. Davis. (May 5, 2020). "Introducing Atomic-SPADL: A New Way to Represent Event Stream Data," [Online]. Available: https://dtai.cs.kuleuven.be/sports/blog/introducing-atomic-spadl-a-new-way-to-represent-event-stream-data (visited on 10/11/2020).

[14] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions Speak Louder than Goals: Valuing Player Actions in Soccer," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA: ACM, Jul. 25, 2019, pp. 1851–1861, ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330758. [Online]. Available: https://dl.acm.org/doi/10.1145/3292500.3330758 (visited on 10/11/2020).

[15] *ML-KULeuven/socceraction*, KU Leuven Machine Learning Research Group, Nov. 8, 2020. [Online]. Available: https://github.com/ML-KULeuven/socceraction (visited on 11/15/2020).

[16] N. Johnson, "Extracting Player Tracking Data from Video Using Non- Stationary Cameras and a Combination of Computer Vision Techniques," p. 14, 2020.

[17] J. Komorowski, G. Kurzejamski, and G. Sarwas, "DeepBall: Deep Neural-Network Ball Detector:" in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, 2019, pp. 297–304, ISBN: 978-989-758-354-4. DOI: 10.5220/0007348902970304. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007348902970304 (visited on 11/16/2019).

[18] J. Fernández and L. Bornn, "SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer," 2020.

[19] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989, ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1989.1.4.541. [Online]. Available: https://www.mitpressjournals.org/doi/abs/10.1162/neco.1989.1.4.541 (visited on 11/15/2020).

[20] F. R. Goes, M. S. Brink, M. T. Elferink-Gemser, M. Kempe, and K. A. P. M. Lemmink, "The tactics of successful attacks in professional association football: Large-scale spatiotemporal analysis of dynamic subgroups using position tracking data," *Journal*

of Sports Sciences, vol. 0, no. 0, pp. 1–10, Oct. 27, 2020, ISSN: 0264-0414. DOI: 10.1080/02640414.2020.1834689. [Online]. Available: https://doi.org/10.1080/02640414.2020.1834689 (visited on 10/27/2020).

[21] K. Verstraete, T. Decroos, B. Coussement, N. Vannieuwenhoven, and J. Davis, "Analyzing Soccer Players' Skill Ratings Over Time Using Tensor-Based Methods," in *Machine Learning and Knowledge Discovery in Databases*, ser. Communications in Computer and Information Science, P. Cellier and K. Driessens, Eds., vol. 1168, Cham: Springer International Publishing, 2020, pp. 225–234, ISBN: 978-3-030-43886-9 978-3-030-43887-6. DOI: 10.1007/978-3-030-43887-6_17. [Online]. Available: http://link.springer.com/10.1007/978-3-030-43887-6_17 (visited on 10/04/2020).

[22] (Sep. 28, 2020). "Players FIFA 21 Oct 9, 2020 SoFIFA," [Online]. Available: https://sofifa.com/ (visited on 10/11/2020).

[23] L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-Based Algorithms for Tensor Decompositions: Canonical Polyadic Decomposition, Decomposition in Rank-$(L_r,L_r,1)$ Terms, and a New Generalization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 695–720, Jan. 2013, ISSN: 1052-6234, 1095-7189. DOI: 10.1137/120868323. [Online]. Available: http://epubs.siam.org/doi/10.1137/120868323 (visited on 11/15/2020).

[24] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, Sep. 1966, ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02289464. [Online]. Available: http://link.springer.com/10.1007/BF02289464 (visited on 11/15/2020).

[25] P. B. a. T. Burzykowski, *Explanatory Model Analysis*. 2020. [Online]. Available: https://pbiecek.github.io/ema/ (visited on 10/04/2020).

[26] J. C. Nunez and B. Dagnino, "Exploring the application of soccer mathematical models to game generation on a simulated environment," presented at the Sports Tomorrow, Barca innovation hub, 2020, p. 10.

[27] (2020). "Google Research Football League," [Online]. Available: https://research-football.dev/ (visited on 11/14/2020).

[28] G. Liu, Y. Luo, O. Schulte, and T. Kharrat, "Deep soccer analytics: Learning an action-value function for evaluating soccer players," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1531–1559, Sep. 2020, ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-020-00705-9. [Online]. Available: http://link.springer.com/10.1007/s10618-020-00705-9 (visited on 10/04/2020).

[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, ser. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press, 1998, 322 pp., ISBN: 978-0-262-19398-6.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1, 1997, ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735 (visited on 11/15/2020).

[31] T. Decroos and J. Davis, "Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, Eds., vol. 11908, Cham: Springer International Publishing, 2020, pp. 569–584, ISBN: 978-3-030-46132-4 978-3-030-46133-1. DOI: 10.1007/978-3-030-46133-1_34. [Online]. Available: http://link.springer.com/10.1007/978-3-030-46133-1_34 (visited on 10/04/2020).

[32] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," presented at the NIPS, 2001, p. 7.

[33] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 4, pp. 580–585, Jul. 1985, ISSN: 0018-9472, 2168-2909. DOI: 10.1109/TSMC.1985.6313426. [Online]. Available: http://ieeexplore.ieee.org/document/6313426/ (visited on 11/15/2020).

[34] (May 5, 2020). "Introducing Atomic-SPADL: A New Way to Represent Event Stream Data," [Online]. Available: https://dtai.cs.kuleuven.be/sports/blog/introducing-atomic-spadl-a-new-way-to-represent-event-stream-data (visited on 10/04/2020).

[35] R. Beal, N. Changder, T. Norman, and S. Ramchurn, "Learning the Value of Teamwork to Form Efficient Teams," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 7063–7070, Apr. 3, 2020, ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v34i05.6192. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6192 (visited on 10/04/2020).

[36] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, "PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 5, pp. 1–27, Sep. 12, 2019, ISSN: 21576904. DOI: 10.1145/3343172. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3360733.3343172 (visited on 02/10/2020).

[37] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00994018. [Online]. Available: http://link.springer.com/10.1007/BF00994018 (visited on 11/15/2020).

[38] A. Groll, C. Ley, G. Schauberger, and H. Van Eetvelde, "A hybrid random forest to predict soccer matches in international tournaments," *Journal of Quantitative Analysis in Sports*, vol. 15, no. 4, pp. 271–287, Oct. 25, 2019, ISSN: 1559-0410, 2194-6388. DOI: 10.1515/jqas-2018-0060. [Online]. Available: http://www.degruyter.com/view/j/jqas.2019.15.issue-4/jqas-2018-0060/jqas-2018-0060.xml (visited on 01/14/2020).

[39] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, ISSN: 08856125. DOI: 10.1023/A:1010933404324. [Online]. Available: http://link.springer.com/10.1023/A:1010933404324 (visited on 11/15/2020).

[40] F. R. Goes, M. Kempe, L. A. Meerhoff, and K. A. Lemmink, "Not Every Pass Can Be an Assist: A Data-Driven Model to Measure Pass Effectiveness in Professional Soccer Matches," *Big Data*, vol. 7, no. 1, pp. 57–70, Mar. 2019, ISSN: 2167-6461, 2167-647X. DOI: 10.1089/big.2018.0067. [Online]. Available: https://www.liebertpub.com/doi/10.1089/big.2018.0067 (visited on 10/04/2020).

[41] K. Pearson, "LIII. *On lines and planes of closest fit to systems of points in space*," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, ISSN: 1941-5982, 1941-5990. DOI: 10.1080/14786440109462720. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/14786440109462720 (visited on 11/15/2020).

[42] U. Dick and U. Brefeld, "Learning to Rate Player Positioning in Soccer," *Big Data*, vol. 7, no. 1, pp. 71–82, Mar. 2019, ISSN: 2167-6461, 2167-647X. DOI: 10.1089/big.2018.0054. [Online]. Available: https://www.liebertpub.com/doi/10.1089/big.2018.0054 (visited on 10/04/2020).

[43] (Feb. 21, 2019). "Attacking Contributions: Markov Models for Football," [Online]. Available: https://statsbomb.com/2019/02/attacking-contributions-markov-models-for-football/ (visited on 10/04/2020).

[44] L. E. Baum and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, Dec. 1966, ISSN: 0003-4851. DOI: 10.1214/aoms/1177699147. [Online]. Available: http://projecteuclid.org/euclid.aoms/1177699147 (visited on 11/15/2020).

[45] T. Decroos and J. Davis, "Interpretable Prediction of Goals in Soccer," p. 17, 2019.

[46] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "VAEP: An Objective Approach to Valuing On-the-Ball Actions in Soccer (Extended Abstract)," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, Jul. 2020, pp. 4696–4700, ISBN: 978-0-9992411-6-5. DOI: 10.24963/ijcai.2020/648. [Online].

Available: https://www.ijcai.org/proceedings/2020/648 (visited on 10/12/2020).

[47] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, ISSN: 0090-5364. DOI: 10.1214/aos/1013203451. [Online]. Available: http://projecteuclid.org/euclid.aos/1013203451 (visited on 11/15/2020).

[48] L. Bransen, J. Van Haaren, and M. van de Velden, "Measuring soccer players' contributions to chance creation by valuing their passes," *Journal of Quantitative Analysis in Sports*, vol. 15, no. 2, pp. 97–116, Jun. 26, 2019, ISSN: 1559-0410, 2194-6388. DOI: 10.1515/jqas-2018-0020. [Online]. Available: https://www.degruyter.com/doi/10.1515/jqas-2018-0020 (visited on 10/04/2020).

[49] F. Zambom-Ferraresi, V. Rios, and F. Lera-Lopez, "Determinants of sport performance in European football: What can we learn from the data?" *Decision Support Systems*, vol. 114, pp. 18–28, Oct. 2018, ISSN: 01679236. DOI: 10.1016/j.dss.2018.08.006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167923618301350 (visited on 02/04/2020).

[50] C. T. Volinsky, A. E. Raftery, D. Madigan, and J. A. Hoeting, "Bayesian model averaging: A tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–417, Nov. 1999, ISSN: 0883-4237. DOI: 10.1214/ss/1009212519. [Online]. Available: http://projecteuclid.org/euclid.ss/1009212519 (visited on 11/15/2020).

[51] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958, ISSN: 0035-9246. JSTOR: 2983890.

[52] L. Pappalardo and P. Cintia, "Quantifying the relation between performance and success in soccer," *Advances in Complex Systems*, vol. 21, p. 1750014, 03n04 May 2018, ISSN: 0219-5259, 1793-6802. DOI: 10.1142/S021952591750014X. arXiv: 1705.00885. [Online]. Available: http://arxiv.org/abs/1705.00885 (visited on 11/16/2019).

[53] I. G. McHale and S. D. Relton, "Identifying key players in soccer teams using network analysis and pass difficulty," *European Journal of Operational Research*, vol. 268, no. 1, pp. 339–347, Jul. 2018, ISSN: 03772217. DOI: 10.1016/j.ejor.2018.01.018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0377221718300365 (visited on 02/04/2020).

[54] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1792–179 210, Jun. 2018. DOI: 10.1109/CVPRW.2018.00223. arXiv: 1804.04527. [Online]. Available: http://arxiv.org/abs/1804.04527 (visited on 10/08/2020).

[55] T. Decroos, J. Van Haaren, and J. Davis, "Automatic Discovery of Tactics in Spatio-Temporal Soccer Match Data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London United Kingdom: ACM, Jul. 19, 2018, pp. 223–232, ISBN: 978-1-4503-5552-0. DOI: 10.1145/3219819.3219832. [Online]. Available: https://dl.acm.org/doi/10.1145/3219819.3219832 (visited on 10/04/2020).

[56] T. Decroos, M. V. Roy, and J. Davis, "SoccerMix: Representing Soccer Actions with Mixture Models," p. 16, 2018.

[57] T. Decroos, "STARSS: A Spatio-Temporal Action Rating System for Soccer," p. 10, 2018.

[58] S. Steiner, S. Rauh, M. Rumo, K. Sonderegger, and R. Seiler, "Using position data to estimate effects of contextual features on passing decisions in football," p. 9, 2018. DOI: 10.15203/CISS_2018.009.

[59] M. Horton, J. Gudmundsson, S. Chawla, and J. Estephan, "Classification of Passes in Football Matches using Spatiotemporal Data," *ACM Transactions on Spatial Algorithms and Systems*, vol. 3, no. 2, pp. 1–30, Aug. 29, 2017, ISSN: 2374-0353, 2374-0361. DOI: 10.1145/3105576. arXiv: 1407.5093. [Online]. Available: http://arxiv.org/abs/1407.5093 (visited on 10/04/2020).

[60] J. Brooks, M. Kerr, and J. Guttag, "Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights," in *Proceedings of the 22nd ACM SIGKDD Inter-*

national Conference on Knowledge Discovery and Data Mining - KDD '16, San Francisco, California, USA: ACM Press, 2016, pp. 49–55, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939695. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2939672.2939695 (visited on 01/26/2020).

[61] ——, "Using machine learning to draw inferences from pass location data in soccer: Drawing Inferences from Pass Location Data Using Machine Learning," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 9, no. 5, pp. 338–349, Oct. 2016, ISSN: 19321864. DOI: 10.1002/sam.11318. [Online]. Available: http://doi.wiley.com/10.1002/sam.11318 (visited on 10/04/2020).

[62] I. G. McHale and L. Szczepanski, "A mixed effects model for identifying goal scoring ability of footballers," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 177, no. 2, pp. 397–417, Feb. 2014, ISSN: 09641998. DOI: 10.1111/rssa.12015. [Online]. Available: http://doi.wiley.com/10.1111/rssa.12015 (visited on 02/02/2020).

[63] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data," in *2014 IEEE International Conference on Data Mining*, Shenzhen, China: IEEE, Dec. 2014, pp. 725–730, ISBN: 978-1-4799-4302-9 978-1-4799-4303-6. DOI: 10.1109/ICDM.2014.133. [Online]. Available: http://ieeexplore.ieee.org/document/7023391/ (visited on 11/16/2019).

[64] T. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, Nov./1996, ISSN: 10535888. DOI: 10.1109/79.543975. [Online]. Available: http://ieeexplore.ieee.org/document/543975/ (visited on 11/15/2020).

[65] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489. [Online]. Available: http://ieeexplore.ieee.org/document/1056489/ (visited on 11/15/2020).

[66] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data," in *2014 IEEE International Conference on Data Mining Workshop*, Shenzhen, China: IEEE, Dec. 2014, pp. 9–14, ISBN: 978-1-4799-4274-9 978-1-4799-4275-6. DOI: 10.1109/ICDMW.2014.167. [Online]. Available: http://ieeexplore.ieee.org/document/7022571/ (visited on 11/16/2019).

[67] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied Computing and Informatics*, vol. 15, no. 1, pp. 27–33, Jan. 2019, ISSN: 22108327. DOI: 10.1016/j.aci.2017.09.005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2210832717301485 (visited on 10/12/2020).

[68] A. Tewari, T. Parwani, A. Phanse, A. Sharma, and A. Shetty, "Soccer Analytics using Machine Learning," 2019. DOI: 10.5120/IJCA2019918773.

[69] *Dmlc/xgboost*, Distributed (Deep) Machine Learning Community, Nov. 15, 2020. [Online]. Available: https://github.com/dmlc/xgboost (visited on 11/15/2020).

[70] R. Baboota and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for English Premier League," 2019. DOI: 10.1016/J.IJFORECAST.2018.01.003.

[71] N. Razali, A. Mustapha, F. A. Yatim, and R. A. Aziz, "Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)," 2017. DOI: 10.1088/1757-899X/226/1/012099.

[72] N. Danisik, P. Lacko, and M. Farkas, "Football Match Prediction Using Players Attributes," *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 2018. DOI: 10.1109/DISA.2018.8490613.

[73] Y. Cho, J. Yoon, and S. Lee, "Using social network analysis and gradient boosting to develop a soccer win-lose prediction model," *Eng. Appl. Artif. Intell.*, 2018. DOI: 10.1016/j.engappai.2018.04.010.

[74] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, Mas-

sachusetts: The MIT Press, 2016, 775 pp., ISBN: 978-0-262-03561-3.

[75] B. Ulmer and M. Fernandez, "Predicting Soccer Match Results in the English Premier League," p. 5, 2014.

[76] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'95, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Aug. 18, 1995, pp. 338–345, ISBN: 978-1-55860-385-1.

[77] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin, "Training and Testing Low-degree Polynomial Data Mappings via Linear SVM," *Journal of Machine Learning Research*, vol. 11, pp. 1471–1490, 2010.

[78] D. Pleuler. (2020). "Google Colaboratory," [Online]. Available: https://colab.research.google.com/github/devinpleuler/analytics-handbook/blob/master/notebooks/nn_pass_difficulty.ipynb (visited on 11/10/2020).

[79] H.-Y. Hsieh, C.-Y. Chen, Y.-S. Wang, and J.-H. Chuang. (Oct. 7, 2019). "BasketballGAN: Generating Basketball Play Simulation Through Sketching." arXiv: 1909.07088 [cs], [Online]. Available: http://arxiv.org/abs/1909.07088 (visited on 02/12/2020).

[80] K. Kurach, A. Raichuk, P. Stanczyk, M. Zajac, O. Bachem, L. Espeholt, C. Riquelme, D. Vincent, M. Michalski, O. Bousquet, and S. Gelly. (Apr. 14, 2020). "Google Research Football: A Novel Reinforcement Learning Environment." arXiv: 1907.11180 [cs, stat], [Online]. Available: http://arxiv.org/abs/1907.11180 (visited on 10/04/2020).

[81] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. (Aug. 28, 2017). "Proximal Policy Optimization Algorithms." arXiv: 1707.06347 [cs], [Online]. Available: http://arxiv.org/abs/1707.06347 (visited on 11/15/2020).

[82] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu. (Jun. 28, 2018). "IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures." arXiv: 1802.01561 [cs], [Online]. Available: http://arxiv.org/abs/1802.01561 (visited on 11/15/2020).

[83] D. Horgan, J. Quan, G. Barth-Maron, and M. Hessel, "DISTRIBUTED PRIORITIZED EXPERIENCE REPLAY," p. 19, 2018.

[84] T. Seidl, A. Cherukumudi, A. Hartnett, P. Carr, and P. Lucey, "Bhostgusters: Realtime Interactive Play Sketching with Synthesized NBA Defenses," p. 13, 2018.

[85] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," p. 9, 2014.