Machine Learning
https://doi.org/10.1007/s10994-018-5763-8

Guest editorial: special issue on machine learning for soccer

Daniel Berrar[1] · **Philippe Lopes**[2,3] · **Jesse Davis**[4] · **Werner Dubitzky**[5]

Received: 12 September 2018 / Accepted: 18 September 2018

Soccer[1] is the biggest global sport and a fast-growing, multi-billion dollar industry. Advanced data analytics are being more frequently employed on both the club and national levels to improve performance, equipment, marketing, scouting, etc. Soccer therefore offers interesting challenges for the machine learning community. This special issue solicited articles on all aspects of data analysis and machine learning for soccer.

As part of the special issue, we posed the 2017 Soccer Prediction Challenge that revolved around predicting the outcomes of *future* soccer matches. This is an interesting task for the general public, researchers, clubs, media, news and advertising companies, and professional odds setters. Soccer outcome prediction has been the subject of research since at least the 1960s (Reep and Benjamin 1968; Hill 1974; Maher 1982; Dixon and Coles 1997; Angelini and Angelis 2017). Various statistical techniques have been used for outcome prediction, including Poisson models (Karlis and Ntzoufras 2003), Bayesian models (Baio and Blangiardo 2010; Rue and Salvesen 2000), rating systems (Hvattum and Arntzen 2010), and more recently also machine learning methods, such as kernel-based relational learning (Van Haaren and Van den Broeck 2011). O'Donoghue et al. (2004) used machine learning and statistical methods to predict the results of the 2002 FIFA World Cup but achieved the best prediction with a simulation on a commercial game console.

✉ Daniel Berrar
daniel.berrar@ict.e.titech.ac.jp

Philippe Lopes
philippe.lopes@univ-evry.fr

Jesse Davis
jesse.davis@kuleuven.be

Werner Dubitzky
werner.dubitzky@helmholtz-muenchen.de

1 Data Science Lab, Department of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan

2 Sport and Exercise Science Department, University of Evry-Val d'Essonne, Evry, France

3 INSERM 1124, Paris Descartes University, Paris, France

4 Department of Computer Science, KU Leuven, Leuven, Belgium

5 Research Unit Scientific Computing, German Research Center for Environmental Health, Helmholtz Zentrum München, Neuherberg, Germany

[1] We use the term *soccer* instead of *football* because in an international context, it is less ambiguous than the term *football*, which also relates to *American football*.

Published online: 05 October 2018

The fundamental research question of the 2017 Soccer Prediction Challenge was the following: "To what extent is it possible to predict the outcome of a soccer match, given commonly available match data?" The competition's task was to use machine learning to predict the outcome of *future* soccer matches. To do so, the participants received v1.0 of the International Open Soccer Database, which has been under development since 2001. The database contains the match reports of 216,743 past matches from 52 soccer leagues in 35 countries covering the years 2000–2017. Each match report specifies the name of the home and away team, respectively, the goals scored by each team, the date on which the match was played, as well as the corresponding soccer league and season. Such match reports represent the most commonly available data about soccer matches around the world. Thus, models learned from this data can be applied to future soccer matches without requiring special arrangements with commercial entities that collect and sell more sophisticated data about soccer matches.

Using only the provided data, the Challenge participants had to develop machine learning models in order to predict the outcome of 206 *future matches* that took place after the submission deadline. Thus, when the participants submitted their predictions, the outcomes for these matches were not known to anyone. The goal was to minimize the average ranked probability score (RPS_{avg}) (Epstein 1969) of the predictions. These conditions highlight two goals of the challenge, which were (1) to gauge the limits of predictability with these commonly available data, and (2) to pose a real-world machine learning challenge with a fixed time line involving the prediction of *real future events*. The last point is a key factor that distinguishes the 2017 Soccer Prediction Challenge from other data mining challenges.

Table 1 summarizes the results. Usually, data mining competitions prohibit the organizers from participating. Because this competition involved predicting the outcomes of real future events that were unknown to us, too, we adhered to the same rules and submitted our predictions as Team DBL.[2] Nonetheless, we considered our predictions to be out-of-competition.

We congratulate the winners of the 2017 Soccer Prediction Challenge:

1. First place: Team OH (Hubáček et al. 2018).
2. Second place: Team ACC (Constantinou 2018).
3. Third place: Team FK (Tsokos et al. 2018).

The Database, the 2017 Soccer Prediction Challenge and its results are described in Dubitzky et al.'s article entitled "The Open International Soccer Database for Machine Learning" (Dubitzky et al. 2018). All materials related to the Database and Challenge are publicly available under the CC0 1.0 Universal license through the Open Science Framework project sites.[3]

This special issue features selected papers of the top-performing teams that participated in the Challenge. In total, the special issue received ten submissions from participating teams, and four of these submissions were accepted. Seven further submissions reporting on machine learning methods for soccer were unrelated to the Challenge. Of these seven papers, one was accepted.

This special issue consists of six papers that are briefly discussed as follows. The article "Learning to predict soccer results from relational data with gradient boosted trees" by Hubáček et al. describes the winning approach for the Challenge. Their model is based on manually engineered features and extreme gradient boosted trees.

[2] Team members included Werner Dubitzky, Daniel Berrar, and Philippe Lopes.

[3] Open International Soccer Database (Dubitzky et al. 2017), available at https://osf.io/kqcye/, and the 2017 Soccer Prediction Challenge (Berrar et al. 2017), available at https://osf.io/ftuva/.

Table 1 Summary of the results for the 2017 Soccer Prediction Challenge

Rank	Team	RPS_{avg}	Accuracy	Method
1	Team DBL*	0.2054	0.5194	Berrar et al. (2018)
2	Team OH	0.2063	0.5243	Hubáček et al. (2018)
3	Team ACC	0.2083	0.5146	Constantinou (2018)
4	Team FK	0.2087	0.5388	Tsokos et al. (2018)
5	Team DBL*	0.2149	0.5049	Berrar et al. (2018)
6	Team HEM	0.2177	0.4660	N/A
7	League priors	0.2255	0.4515	Prior information based on leagues
8	Team EB	0.2258	0.4854	N/A
9	Global priors	0.2261	0.4515	Global priors of win, draw, lose
10	Team LJ	0.2313	0.4126	N/A
11	Team AT	0.3981	0.3883	N/A
12	Team LHE	0.4515	0.3398	N/A
13	Team EDS	0.4515	0.3592	N/A

Participating teams are ranked based on increasing values of the average ranked probability score, calculated from the submitted predictions for the 206 games from the prediction set. The accuracy, i.e., the proportion of correctly predicted games, is also shown. Submissions by the organizers (Team DBL) are out-of-competition and marked by *

In "*Dolores*: A model that predicts football match outcomes from all over the world", Constantinou presents a model for soccer outcome prediction based on hybrid Bayesian networks and dynamic performance rating that placed second in the Challenge. A comparison with bookmakers' odds revealed that *Dolores* could also increase profitability in terms of return of investment, albeit only marginally.

The article "Modeling outcomes of soccer matches" by Tsokos et al. compares various extensions of Bradley–Terry models and a hierarchical log-linear Poisson model for the prediction of soccer outcomes. Their best model achieved third place in the Challenge.

The article titled "Incorporating domain knowledge in machine learning for soccer outcome prediction" by Berrar, Lopes, and Dubitzky presents two new feature engineering methods for match outcome prediction: *recency feature extraction* and *rating feature learning*. With the latter method, we constructed a learning set and trained a k-nearest neighbor model, which achieved the best performance among all models submitted to the Challenge. We conclude that the key challenge in soccer prediction lies in domain knowledge integration.

The article "Probabilistic movement models and zones of control" by Brefeld et al. is a submitted paper not directly related to the Challenge. The authors present a probabilistic, data-driven movement model to estimate positions, directions, and velocities of players at observed timestamps. Using their model, the authors derive *zones of control*, also known as *dominant regions*. If the ball falls into the zone of control of a player, then this player is most likely to gain control over the ball; consequently, the more space a team controls, the more dominant the team is. A comparison with existing movement models suggests that this model leads to a more realistic estimation of zones of control. This model might give useful insights into game tactics and team performance, not only for soccer but also other, similar team sports.

Soccer provides many fascinating challenges for machine learning, and we hope that this special issue will spur further research. Particularly interesting new data and challenges are the following:

Event streams This type of data annotates specific events that occur in a soccer match (Opta Sports 2018; Wyscout 2018; STATS' SportVU 2018). The precise number of events, each event's definition, and what information is available about the event varies according to the provider. Typically, there are around 40 different types of events, and each event lists the type of the event, the players who are involved, a timestamp, the start location of the event, and the end location of the event (if applicable). Sometimes additional information may be available, for example, if a shot was made by the head or foot. Typical events include passes, clearances, fouls, shots, cards, and substitutions.
Optical tracking A variety of companies, such as ChyronHego (2018), Stats LLC (2018), SciSports (2018), and Second Spectrum (2018) record the locations of the players and the ball at a high frequency using optical tracking systems during matches.
Player monitoring Players are often outfitted with sensor systems (STATSports 2018; Catapult 2018) including accelerometers, gyroscopes, heart rate monitors, and GPS during training sessions and games (if permitted). Furthermore, the data generated by these devices may be augmented with additional data, such as fatigue ratings (e.g., the rating of perceived exertion) or general wellness scores (e.g., muscle soreness). Additionally, clubs record and store information from physical testings (e.g., flexibility, speed, maximum rate of oxygen consumption during exercise, etc.).

These types of data are of interest to a variety of different parties. Clubs and national teams are continually trying to exploit these types of data to improve performance, equipment, marketing, scouting, etc. Fans may be interested in analyzing and debating the performances of their favorite teams. Bettors and oddsmakers are interested in how these data can be exploited to turn a profit. This has lead to an explosion of interest in data science and analytics, specifically for the following tasks:

Evaluating actions One of the most prominent new metrics is known as *expected goals* (Eastwood 2015; Eggels 2016; Lucey et al. 2015; Ijtsma 2015; Caley 2015). The objective is to quantify the quality of a shot by training a statistical model that predicts the probability of scoring based on the features of the shot (e.g., location, angle to the goal, etc.). More recently, there have been attempts to move beyond simply evaluating shots by assigning values to other actions on the pitch, such as shots or even individual movements based on event streams and/or optical tracking data (Decroos et al. 2018; Spearman 2018; Gyarmati and Stanojevic 2016; Bransen and Van Haaren 2018; Pappalardo et al. 2018). By evaluating all actions, it is possible to derive rankings or overall ratings of players.
Identifying tactics and strategy One line of work looks at trajectory data produced by optical tracking to try to understand tactics, such as how play is built up from the back (Knauf et al. 2016), analyzing how effective a team is at creating scoring opportunities (Fernando et al. 2015), or using data-driven ghosting methods to understand how a team should have addressed certain situations (Le et al. 2017). Researchers have also analyzed tactics from event stream data to find commonly occuring sequences of events that lead to attempts on goal (Van Haaren et al. 2015) or identify whether an attempt is likely in the near future (Decroos et al. 2017). Substantial attention has been devoted to understanding passing behaviors, particularly in terms of finding different types of recurrent passing patterns (Gyarmati et al. 2014; Gyarmati and Anguera 2015; Bekkers and Dabadghao 2017). Other tasks include predicting if a pass will succeed (Spearman et al. 2017), classifying different types of passes (Chawla et al. 2017), and predicting whom a player

may pass to (Vercruyssen et al. 2016). Finally, researchers have built occupancy maps based on ball movements (Lucey et al. 2013) and attempted to recognize team formations (Bialkowski et al. 2014).

Monitoring players' health Currently, professional soccer clubs monitor training and match load of all players. From a sports science perspective, both the external load and internal load are of interest. Intuitively, the external load captures the level of activity (e.g., amount, intensity, etc.) performed by players, and it is often measured by having players wear sensors (e.g., GPS and accelerometer). The internal load captures the body's physiological response to the activity, and it is measured by having the players report the rating of perceived exertion. Researchers have explored using machine learning techniques to investigate the relations between these two loads as well as perceived wellness (Rossi et al. 2017; Vandewiele et al. 2017; Jaspers et al. 2018a, b), which could help optimize training routines. Another promising but very challenging task is to build models to assess a player's risk of a non-contact injury based on physical and testing data collected from the players (Kampakis 2016; Rossi et al. 2018).

We would like to thank everyone who was involved in this special issue, particularly all contributing authors and the reviewers. We also thank the editorial and publishing staff at Springer for their support. Special thanks also goes to Peter A. Flach, Editor-in-Chief of *Machine Learning*, and Dragos D. Margineantu, Editor for Special Issues.

References

Angelini, G., & De Angelis, L. (2017). PARX model for football match predictions. *Journal of Forecasting*, *36*(7), 795–807.

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, *37*(2), 253–264.

Bekkers, J., & Dabadghao, S. (2017). Flow motifs in soccer: What can passing behavior tell us? In *Proceedings of the 11th MIT sloan sports analytics conference* (pp. 1–31).

Berrar, D., Lopes, P., Davis, J., & Dubitzky, W. (2017). The 2017 soccer prediction challenge. https://doi.org/10.17605/OSF.IO/FTUVA. Accessed 7 Sep 2018.

Berrar, D., Lopes, P., & Dubitzky, W. (2018). Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine learning. https://doi.org/10.1007/s10994-018-5747-8.

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., & Matthews, I. (2014). Identifying team style in soccer using formations learned from spatiotemporal tracking data. In *IEEE international conference on data mining workshop* (pp. 9–14).

Bransen, L., & Van Haaren, J. (2018). Measuring football players' on-the-ball contributions from passes during games. In *Proceedings of the 5th workshop on machine learning and data mining for sports analytics, ECML/PKDD* 2018 (pp. 1–13).

Caley, M. (2015). Premier League projections and new expected goals. http://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/premier-league-projections-and-new-expected-goals. Accessed 6 Sep 2018.

Catapult. (2018). https://www.catapultsports.com/. Accessed 3 Aug 2018.

Chawla, S., Estephan, J., Gudmundsson, J., & Horton, M. (2017). Classification of passes in football matches using spatiotemporal data. *ACM Transactions Spatial Algorithms and Systems*, *3*(2), 6:1–6:30.

ChyronHego. (2018). http://www.chyronhego.com. Accessed 3 March 2018.

Constantinou, A. (2018). *Dolores*: a model that predicts football match outcomes from all over the world. Machine Learning. https://doi.org/10.1007/s10994-018-5703-7.

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2018). Actions speak louder than goals: Valuing player actions in soccer. arXiv:1802.07127.

Decroos, T., Dzyuba, V., Van Haaren, J., & Davis, J. (2017). Predicting soccer highlights from spatio-temporal match event streams. In *Proceedings of the 31st AAAI conference on artificial intelligence* (pp. 1302–1308).

Dixon, M., & Coles, S. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, *46*(2), 265–280.

Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2017). The open international soccer database. https://doi.org/10.17605/OSF.IO/KQCYE.
Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2018). The open international soccer database. Machine Learning. https://doi.org/10.1007/s10994-018-5726-0.
Eastwood, M. (2015). Expected goals and support vector machines. http://pena.lt/y/2015/07/13/expected-goals-svm/. Accessed 6 Sep 2018.
Eggels, H. (2016). Expected goals in soccer: Explaining match results using predictive analytics. MSc thesis, Department of Mathematics and Computer Science, Eindhoven University of Technology.
Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*(6), 985–987.
Fernando, T., Wei, X., Fookes, C., Sridharan, S., & Lucey, P. (2015). Discovering methods of scoring in soccer using tracking data. In *Proceedings of the KDD workshop on large-scale sports analytics* (pp. 1–4).
Gyarmati, L., & Anguera, X. (2015). Automatic extraction of the passing strategies of soccer teams. arXiv:1508.02171.
Gyarmati, L., Kwak, H., & Rodriguez, P. (2014). Searching for a unique style in soccer. arXiv:1409.0308.
Gyarmati, L., & Stanojevic, R. (2016). Qpass: A merit-based evaluation of soccer passes. CoRR abs/1608.03532, arXiv:1608.03532.
Hill, I. (1974). Association football and statistical inference. *Applied Statistics*, *23*(2), 203–208.
Hubáček, O., Šourek, G., & Železný, F. (2018). Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning. https://doi.org/10.1007/s10994-018-5704-6.
Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470.
Ijtsma, S. (2015). A close look at my new Expected Goals Model. http://11tegen11.net/2015/08/14/a-close-look-at-my-new-expected-goals-model/. Accessed 6 Sep 2018.
Jaspers, A., De Beéck, T. O., Brink, M. S., Frencken, W. G., Staes, F., Davis, J. J., et al. (2018a). Relationships between the external and internal training load in professional soccer: What can we learn from machine learning? *International Journal of Sports Physiology and Performance*, *13*(5), 625–630.
Jaspers, A., Op De Beéck, T., Brink, M.S., Frencken, W.G., Staes, F., Davis, J., & Helsen, W. (2018b). Predicting future perceived wellness in professional soccer: the role of preceding load and wellness. International Journal of Sports Physiology and Performance (to appear)
Kampakis, S. (2016). Predictive modeling of football injuries. Ph.D. thesis, Department of Computer Science, University College London.
Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 381–393.
Knauf, K., Memmert, D., & Brefeld, U. (2016). Spatio-temporal convolution kernels. *Machine Learning*, *102*(2), 247–273.
Le, H., Carr, P., Yue, Y., & Lucey, P. (2017). Data-driven ghosting using deep imitation learning. In *Proceedings of the 11th MIT sloan sports analytics conference* 2017 (pp. 1–15).
Lucey, P., Bialkowski, A., Monfort, M., Carr, P., & Matthews, I. (2015). Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *Proceedings of the 9th MIT sloan sports analytics conference* (pp. 1–9).
Lucey, P., Oliver, D., Carr, P., Roth, J., & Matthews, I. (2013). Assessing team strategy using spatiotemporal data. In *Proceedings of the 19th international conference on knowledge discovery and data mining* (pp. 1366–1374).
Maher, M. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118.
O'Donoghue, P., Dubitzky, W., Lopes, P., Berrar, D., Lagan, K., Hassan, D., et al. (2004). An evaluation of quantitative and qualitative methods of predicting the 2002 FIFA World Cup. *Journal of Sports Sciences*, *22*(6), 513–514.
Opta Sports. (2018). http://www.optasports.com. Accessed 3 Aug 2018.
Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2018). PlayeRank: Multi-dimensional and role-aware rating of soccer player performance. arXiv:1802.04987.
Reep, C., & Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society, Series A (General)*, *131*(4), 581–585.
Rossi, A., Pappalardo, L., Cintia, P., Iaia, F., Fernandez, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS ONE*, *13*(7), e0201264.
Rossi, A., Perri, E., Trecroci, A., Savino, M., Alberti, G., & Iaia, M.F. (2017). GPS data reflect players' internal load in soccer. In *Proceedings of the 2017 IEEE international conference on data mining workshops* (pp. 890–893).
Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 399–418.

SciSports. (2018). http://www.scisports.com. Accessed 3 Aug 2018.
Second spectrum. (2018). http://www.secondspectrum.com. Accessed 3 Aug 2018.
Spearman, W. (2018). Beyond expected goals. In *Proceedings of the 12th MIT sloan sports analytics conference* (pp. 1–17).
Spearman, W., Pop, P., Basye, A., Hotovy, R., & Dick, G. (2017). Physics-based modeling of pass probabilities in soccer. In *Proceedings of the 11th MIT sloan sports analytics conference* (pp. 1–14).
STATS' SportVU. (2018). http://www.stats.com/sportvu. Accessed 3 Aug 2018.
STATSports. (2018). https://statsports.com/. Accessed 3 Aug 2018.
Tsokos, A., Narayanan, S., Kosmidis, G, I, Baio., Cucuringu, M., Whitaker, G., & Király, F. (2018). Modeling outcomes of soccer matches. Machine Learning. https://doi.org/10.1007/s10994-018-5741-1.
Van Haaren, J., Dzyuba, V., Hannosset, S., & Davis, J. (2015). Automatically discovering offensive patterns in soccer match data. In *Proceedings of advances in intelligent data analysis XIV* (pp. 286–297).
Van Haaren, J., & Van den Broeck, G. (2011). Relational learning for football-related predictions. In *Proceedings of the 21st international conference on inductive logic programming* (pp. 1–6).
Vandewiele, G., Geurkink, Y., Lievens, M., Ongenae, F., Turck, F.D., & Boone, J. (2017). Enabling training personalization by predicting the session rate of perceived exertion. In *Proceedings of the 4th workshop on machine learning and data mining for sports analytics, ECML/PKDD* 2018 (pp. 31–40).
Vercruyssen, V., De Raedt, L., & Davis, J. (2016). Qualitative spatial reasoning for soccer pass prediction. In *Proceedings of the 3rd workshop on machine learning and data mining for sports analytics, ECML/PKDD* 2016 (pp. 1–10).
Wyscout. (2018). https://wyscout.com/. Accessed 3 Aug 2018.