Digitalization in Professional Football: An Opportunity to Estimate Injury Risk

Laurent Navarro, Pierre-Eddy Dandrieux, Karsten Hollander, Pascal Edouard

Digitalization in Professional Football: An Opportunity to Estimate Injury Risk

Laurent Navarro[1] , Pierre-Eddy Dandrieux[1,2], Karsten Hollander[3] and Pascal Edouard[2]

[1] Mines Saint-Etienne, INSERM, U 1059 Sainbiose, CIS, Univ Lyon, Univ Jean Monnet, Saint-Etienne, France
[2] Univ Lyon, UJM-Saint-Etienne, Inter-university Laboratory of Human Movement Biology, EA 7424, F-42023, Saint-Etienne, France
[3] Institute of Interdisciplinary Exercise Science and Sports Medicine, MSH Medical School Hamburg, Hamburg, Germany
[4] Department of Clinical and Exercise Physiology, Sports Medicine Unit, University Hospital of Saint-Etienne, Faculty of Medicine, Saint-Etienne, France

Abstract. Digitalization in the field of sport has already been a reality for a number of years. The growing increase in the volume of data that can be acquired on athletes today makes its use possible mainly for performance enhancement and also for injury prevention. We propose in this paper to evaluate the possibility of including Artificial Intelligence (A.I.) through Machine Learning (ML) as a mean for estimating injuries in professional football, by 1) discussing the addition of ML information in the interaction between stakeholders through graph network representations, and 2) presenting the injury risk estimation through two ML techniques adapted to the characteristics of data from players. We first constructed an elementary representation for an athlete and his/her environment, and we then created a complex network of 23 professional football players. We discussed the implication of ML methods for stakeholders such as coaches, players or medical staff. Regarding injury risk estimation, we focused on methods allowing 1) to work with few data and 2) to have a certain level of explainability to avoid the well-known "black box" effect. In particular, we used decision tree and logistic regression methods to predict the occurrence of hamstring injuries in 284 professional footballers for whom baseline data, as well as sprint acceleration mechanical output measurements taken from one football season were available. The results show that the estimation of injury risk is possible to a certain extent, and that the centrality of the technical team is crucial when incorporating such methods in team sports.

Keywords: Injury risk estimation, Machine learning, Professional football, Sports science, Sports medicine, Social Networks

1 Introduction

Digital data has regularly been used to monitor, track, guide, and direct the training of athletes. A strong emphasis is placed on sports performance, but the use of this data for

injury prevention and injury risk estimation seems to represent an opportunity. At the research level, the use of databases to make predictions is increasing [1]. Indeed, medical data is often used after the injury, for diagnosis, but not for prevention or injury risk estimation. However, A.I., and more specifically Machine Learning (ML), now offers the possibility to create a form of "digital twin" [2] for athletes. Complete integration is possible thanks to sensors and questionnaire inputs. This "digital twin" theoretically makes it possible to obtain incomplete but useful models of the athlete, ahead of time, that can help reducing the occurrence of injuries.

Training rules are created from a complex and informal interdisciplinary decision-making. It is a team effort that we believe can be aided by A.I. usage with data from the athletes. This data is mainly of two types: objective and subjective. Objective data often results from sensor data, such as heart rate sensors, mechanical data from mechanical tests during training, timed data, etc. Subjective data are often provided by the athlete orally or via questionnaires on paper or via web or smartphone applications. One of the difficulties when collecting subjective data is the level of interaction that exists between the athlete and the various other stakeholders. The use of graph network representations allows to understand the different interactions between the stakeholders by formalizing interactions through mathematical rules. Consequently, a social network type approach can be envisaged [3], and the role of A.I., more precisely ML, can be understood within the decision-making process.

In addition, explainable ML is an interesting opportunity for this type of problem. Indeed, the often denounced black box effect on ML models is partly solved by models that have good explainability [4,5]. In particular, decision tree or logistic regression type models have this characteristic: the importance of the input parameters for the prediction can be quantified for example.

In this paper we propose the use of digital data to help estimate injury risk in football players. Therefore, we implemented a micro/macro type approach based on the use of an elementary social network type model of the interactions between the athlete and his/her environment that we replicated to create complex networks. We then analyzed these complex networks using graph representations and betweenness centrality measures [6,7]. In particular, the importance of stakeholders is shown and discussed. We then used as an example of the potential of ML techniques to estimate injury risk, a database of 284 players from 16 professional football teams for which training data and hamstring injury occurrence data are recorded. We first implemented two ML algorithms on the data to predict hamstring injury occurrence: decision tree and logistic regression. Then we addressed the questions of the place of this tool through the analysis of different graphs, and the relationships between social network analysis and importance of ML parameters.

2 Graph Description of the Problem

In this section, we used a graph network approach. The idea was to create a micro-scale model of the athlete, which described all interactions between an athlete and his immediate environment. Then, a macro model connected a number of micro systems to observe the importance of the different stakeholders as well as the place of the athletes

in a network. In this model, the nodes correspond to the different stakeholders, and the edges correspond to the interactions between them. An interaction corresponds to a relationship between two stakeholders, through which the behaviors of these individuals influence each other and change accordingly. The edges are therefore all considered to be bidirectional. We considered two types of interactions: human/human interactions and human/machine interactions (A.I.).

We describe an example of an elementary system for one athlete (Figure 1) for illustration purpose, which is not exhaustive. However, it is still fairly representative to the athlete's environment. The different stakeholders that we chose to describe in this system are: athlete, technical team, medical staff, social life, family, work (which can correspond to school or higher education depending on the age of the athlete), administrative leaders, and finally A.I. and Data Scientists. Some interactions ("research only") appear in red color. These interactions are provisional, they are active in the research phase for the construction of the A.I. system, but are intended to be deactivated in routine operation.

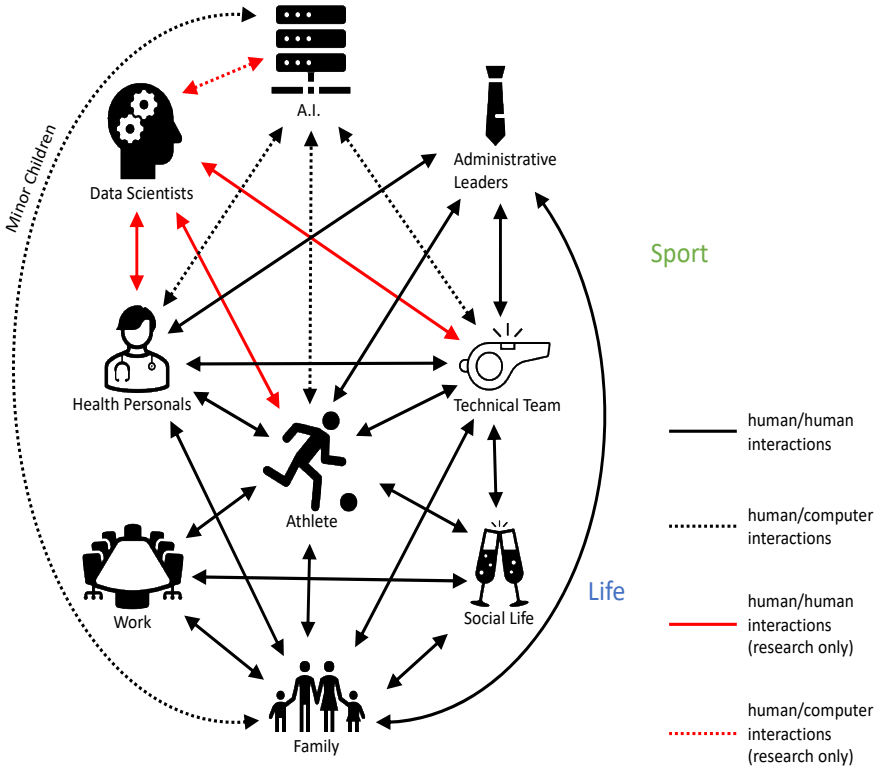

Fig. 1. Proposed elementary system representing the athlete in his/her environment, and the interactions with the stakeholders.

We constructed two types of graphs: a first type of graph comprising only one (minor) athlete (Figure 2) and a second containing 23 adult players (Figure 3), which corresponds to the standards for the number of players in a professional football team before 2020. These graphs are simulated ones, and edges values have all been set to 1, as we consider binary interactions between nodes for the sake of simplicity, but also because objective data about the importance of these interactions do not exist at this time.

For the professional adult players, we removed the node "Work", and the edges between the nodes "Family" and "A.I.". We also created recursive links between social lives, families and athletes to illustrate the dense social interactions existing in professional sports environments. Each type of graph is declined in three cases: the classic case (classic) where there is no A.I., the case with A.I. and the research case with A.I. and data scientist (Data Sc.).

We chose to analyze the graphs in terms of betweenness centrality, both for nodes [6] and edges [7]. Betweenness centrality is equal to the number of times one node (or edge) is on the shortest path between any two other nodes in the graph. On the graphs, the greater the diameter of a node, the greater its betweenness centrality, and the thicker an edge, the greater its betweenness centrality.

Colors correspond to the type of nodes: blue for athletes, yellow for sports professionals or A.I. stakeholders, and red for life environment.

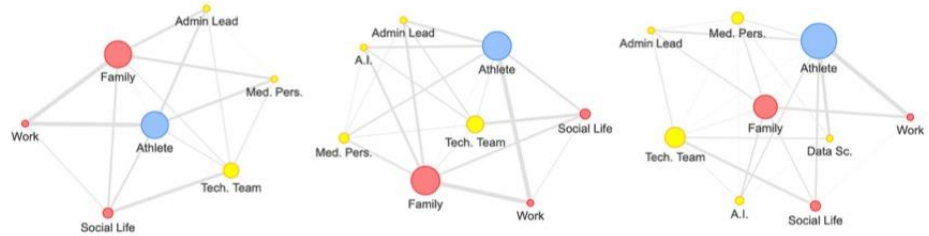

Fig. 2. Graphs of the three cases for one minor athlete: classic (left), with A.I. (middle) and research with A.I. and Data Scientist (right).

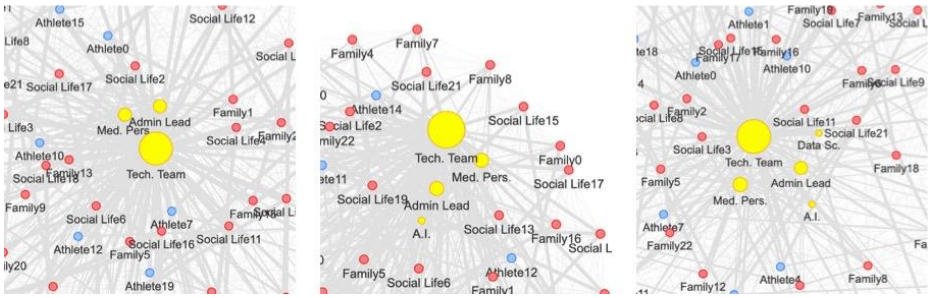

Fig. 3. Graphs of the three cases for 23 professional adult football players: classic (left), with A.I. (middle) and research with A.I. and Data Scientist (right).

Table 1. normalized betweenness values of nodes of the two different graphs. Max values are indicated in **bold**.

	classic	A.I.	A.I. and D.S.	classic	A.I.	A.I. and D.S.
Stakeholder	**Single athlete**			**23 football players**		
Athlete	**0.1333**	**0.1468**	**0.1845**	0.0050	0.0052	0.0054
Family	**0.1333**	**0.1468**	0.1101	0.0050	0.0049	0.0047
Tech. Team	0.0444	0.0595	0.0833	**0.1826**	**0.1851**	**0.1873**
Med. Pers.	0.0000	0.0119	0.0298	0.0407	0.0426	0.0444
Social Life	0.0222	0.0159	0.0119	0.0059	0.0057	0.0056
A.I.	N/A	0.0000	0.0089	N/A	0.0000	0.0000
Admin Lead	0.0000	0.0000	0.0000	0.0407	0.0396	0.0385
Work	0.0000	0.0000	0.0000	N/A	N/A	N/A
Data Sc.	N/A	N/A	0.0000	N/A	N/A	0.0000

Table 1 gives the normalized (on a [0,1] range) betweenness centralities of the different types of nodes in the graphs. In the 23 players case, the betweenness is considered for one athlete, one family and one social life, as these are the same for the 23. Technical team, medical personal, A.I., administrative leaders and data scientists are unique for each graph. Work has been removed for the 23 players graphs (Figure 3).

We observed that in the single athlete's case, the athlete has the most important betweenness centrality, which means that his relative "power" or influence is high. Its value is shared with family for the classic and A.I. cases. Concerning the 23 players case, the technical team has in any case a way more important betweenness centrality, followed by the medical team and the administrative leaders.

3 Machine Learning on Real-world Data

This section aims to highlight the potential of ML techniques to estimate injury risk in sports. We chose an approach mostly presented in the current literature: estimation of injury risk during the season based on baseline data (i.e., data at the start of the season). As mentioned above, explainability is crucial when it comes to including A.I. algorithms in a network composed mainly of humans who are used to interacting in a system that already works. Also, we chose two methods for their relative explainability: decision trees and logistic regressions. For these two methods, we can compute the weight of each parameter in the final prediction, so feedback other than the final prediction can be provided to the stakeholders concerned.

For this example, we used a dataset from 284 male football players from 16 professional football teams from three countries (Japan, France and Finland) over one season. More details, as well as statistics about these data can be found in [8]. The outcome was the occurrence of hamstring injuries during the season. At the end of the season, 47 hamstring injuries affected 38 players. At the start of the season, all players performed a 30m sprint to measure sprint acceleration. Data for each athlete included: binary coded country group (country), age, height, body mass, training volume, history

of hamstring injuries (previous season), and data recorded during training sessions: horizontal force production capacity (FH0 and V0), the maximum power, the force-speed profile, the time at 5 m, the time at 10 m, the time at 20 m, and the maximum speed [8]. Thus, all these data constitute the inputs of the model, and the output is the outcome, i.e. the occurrence of hamstring injuries during the season.

For the hyperparameters tuning, we followed the principle of nested cross-validation, which limits the bias on small datasets [9]. For each of the two models (i.e., decision trees and logistic regressions), 200 nested cross validation iterations were carried out. Here, the dataset was divided into 10 equal parts, nine of which (train) are used for hyperparameter selection. Then, this 90% of the dataset were further divided into 10 parts and 9 of them were used to predict the 10th. Hyperparameter optimization was performed during this operation, with a grid-search algorithm. Then, the performance evaluation of the model was carried out on the first of the 10 initial parts which have been left out. There are therefore 10 scores for each iteration, i.e., 200×10=2000 sets of hyperparameters.

Table 2 presents the performance results for the two methods, namely decision tree and logistic regression. The mean and standard deviation of the recall, specificity, precision, accuracy and ROC-AUC parameters are specified. It is interesting to observe that these parameters have a very precise explanation with respect to the estimation of the risk of injury occurrence. The recall parameter indicates the ability of the model to detect injuries. The specificity parameter indicates the ability of the model to detect non-injured. The precision parameter indicates the proportion of injured among positive predictions. The accuracy parameter indicates the ability of the model to make good predictions of injured and non-injured. Finally, the ROC-AUC parameter represents the overall measure of the model's performance.

Table 2. Performance parameters for the two methods tested 200 times wit mean and std.

Model	Recall (mean ± std)		Specificity (mean ± std)		Precision (mean ± std)		Accuracy (mean ± std)		ROC-AUC (mean ± std)	
Decision Tree	0.58	0.07	0.67	0.04	0.22	0.03	0.66	0.03	0.65	0.04
Logistic Regr.	0.69	0.03	0.76	0.01	0.32	0.02	0.75	0.01	0.77	0.01

Figure 4 shows the importance of features for both models in injury risk estimation. Since logistic regression has a better performance, we base our analysis on the latter in the following paragraphs. We noted that the most influential parameters on the injury were height, followed by the fact of belonging to the Finland country group, followed by the time at 20 m, followed by the fact of belonging to the France country group, followed by the fact of belonging to the Japan country group.

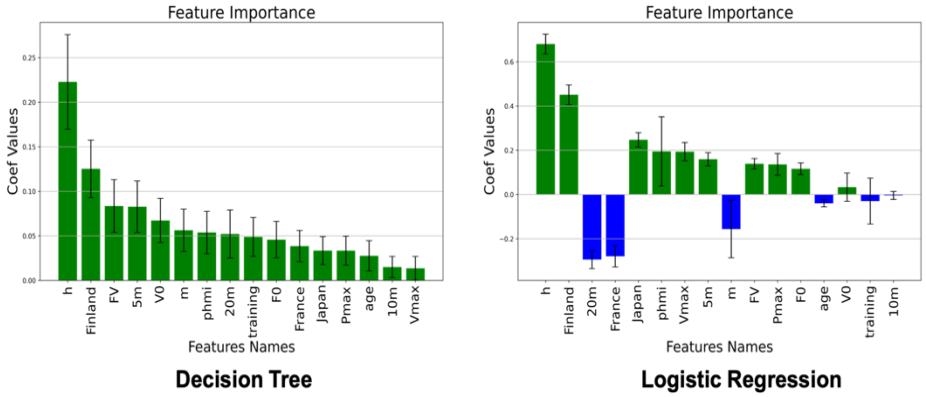

Fig. 4. Importance of features for the two methods tested.

Height is an intrinsic non-modifiable parameter of an adult athlete and is not controllable but must be taken into account by all stakeholders. Indeed, although it is non-modifiable, better management of the other modifiable parameters should be proposed. The time at 20 m parameter is related to sprint performance, and this implies that the technical team must take this parameter into account when building their training programs. A discussion between the technical team and the medical staff is thus recommended. Belonging to a specific country group seems to be a very important parameter. With all the precautions that must obviously be taken, this data raises the legitimate question of the training methods used by each country group.

4 Discussion

In all network representations with multiples athletes, the technical team occupies a central position, with a node of the highest betweenness centrality. This centrality is not a surprising phenomenon. Coaches in particular have a central position in the lives of athletes. However, the athlete must be at the center of the process, as shown in the graph Figure 1.

Sections 2 and 3 raised the question of the place of artificial intelligence in sport, by first proposing a network model, and an example of machine learning on real-world data. The confrontation of these two approaches shows that the technical team, in the case of a professional football team, carries a great responsibility in the overall process, and therefore the occurrence of injuries. ML seems capable, to a certain extent, of predicting the possible occurrence of injuries. It is then the role of the technical team to take the prediction results into consideration when creating training programs.

If one normalizes the importance of the parameters by the betweenness centralities (by simple multiplication, like done in Table 3) in the classic case, one can see what are the real means of action on the occurrence of injuries. Indeed, since one isolated athlete has a very low betweenness centrality compared to the doctor and the technical team, all the parameters specific to him (age, height, body mass, FH0 and V0,

maximum power, force-speed profile, time at 5 m, time at 10 m, time at 20 m, and maximum speed) are not controllable. The means of action that are relegated to the technical and medical teams, and therefore preponderant, are: the country, which is linked to the training programs, the history of hamstring injuries, and the volume of training.

Table 3. Example of correction of feature importance with betweenness centralities. The last column corresponds to the product of first column and fourth column.

Feature importance	Feature	Most infuencal Stakeholder	Betweenness centrality	Corrected Value
0.680472408	h	Athlete	0.0050	0.003395516
0.451034408	Finland	Tech. Team	0.1826	**0.082378062**
-0.294332569	20m	Athlete	0.0050	-0.001468702
-0.279330584	France	Tech. Team	0.1826	**-0.051017642**
0.247351867	Japan	Tech. Team	0.1826	**0.045176969**
0.194580992	phmi	Med. Pers.	0.0407	**0.007924184**
0.1935081	Vmax	Athlete	0.0050	0.000965594
0.159192767	5m	Athlete	0.0050	0.000794362
-0.15626462	m	Athlete	0.0050	-0.000779751
0.139131464	FV	Athlete	0.0050	0.000694258
0.135991412	Pmax	Athlete	0.0050	0.000678589
0.116354602	F0	Athlete	0.0050	0.000580602
-0.039871139	age	Athlete	0.0050	-0.000198955
0.033002423	V0	Athlete	0.0050	0.00016468
-0.029759597	training	Tech. Team	0.1826	**-0.005435368**
-0.004197034	10m	Athlete	0.0050	-2.09429E-05

There can exist some tension between the technical team and the medical staff when it comes to injury prevention [10]. The technical team is looking for performance, which influences the risk of injury. The medical staff tries to avoid injuries, that is their main goal. We are therefore in the presence of a dual performance/injury prevention objective. As the example on our athlete database shows, performance tends to be correlated with the risk of injury [11]. Thus, the overall problem can be seen as an optimization problem, under constraints of increasing performance and decreasing injury risk.

In our models, the importance of the interactions was not specified, this is one of the limitations. It is obvious that the interactions do not have the same importance between the different stakeholders. In addition, its importance can vary according to the situation and the athlete's personality. A model integrating the importance of the interactions

would make it possible to calculate the importance of the nodes differently. We could therefore understand more precisely what actions could be put in place to put the athlete back at the center of the process and partially reduce the centrality of the technical team. The place of A.I. in this operation could be decisive. Indeed, the transmission of information and the restriction according to the positions of the stakeholders could enable to control the importance of the interactions, via a mediation between the different stakeholders.

5. Conclusion

We proposed in this paper to study two approaches related to digitalization in sport for the estimation of injury risk: graph social networks, and ML. These two approaches are complementary, the first allowing to understand the importance of stakeholders in a sports context, and the second allowing to assess the possibility of predicting the risk of injury using athlete's data. It is the combination of these two approaches that is interesting, as it shows how the integration of artificial intelligence in sport can influence the risk of injury. In particular, the role of the technical team is crucial, and it appears to be its responsibility to integrate the results of artificial intelligence predictions for the construction of training programs.

Future work will focus on the specification and improvement of the two proposed approaches. Networks embedding the weight of the nodes and the importance of the edges will be developed through the use of questionnaires. This will allow a finer understanding of the interactions. The explainable machine learning approach will also be developed through the construction of algorithms truly adapted to the world of sport. These specific models will be developed in a multidisciplinary framework involving sports scientists, sports doctors, sports scientists, and engineers.

References

1. Van Eetvelde, Hans, et al. "Machine learning methods in sport injury prediction and prevention: a systematic review." Journal of experimental orthopaedics 8.1 (2021): 1-15.
2. Barricelli, Barbara Rita, et al. "Human digital twin for fitness management." Ieee Access 8 (2020): 26637-26664.
3. Wäsche, Hagen, et al. "Social network analysis in sport research: an emerging paradigm." European Journal for Sport and Society 14.2 (2017): 138-165
4. Burkart, Nadia, and Marco F. Huber. "A survey on the explainability of supervised machine learning." Journal of Artificial Intelligence Research 70 (2021): 245-317.
5. Lundberg, Scott M., et al. "From local explanations to global understanding with explainable AI for trees." Nature machine intelligence 2.1 (2020): 56-67.
6. Freeman, Linton C. "A set of measures of centrality based on betweenness." Sociometry (1977): 35-41.

7. Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the national academy of sciences 99.12 (2002): 7821-7826.
8. Edouard, Pascal, et al. "Low horizontal force production capacity during sprinting as a potential risk factor of hamstring injury in football." International journal of environmental research and public health 18.15 (2021): 7827.
9. Vabalas, Andrius, et al. "Machine learning algorithm validation with a limited sample size." PloS one 14.11 (2019): e0224365.
10. Ekstrand, Jan, et al. "Communication quality between the medical team and the head coach/manager is associated with injury burden and player availability in elite football clubs." British Journal of Sports Medicine 53.5 (2019): 304-308.
11. Chapon, Joris, Laurent Navarro, and Pascal Edouard. "Relationships between performance and injury occurrence in athletics (track and field): A pilot study on 8 national-level athletes from sprints, jumps and combined events followed during at least five consecutive seasons." Frontiers in Sports and Active Living: 176.