Comparison of machine learning methods for predicting the recovery time of professional football players after an undiagnosed injury

Stylianos Kampakis, University College London, United Kingdom

stylianos.kampakis@gmail.com

Abstract. Injuries are a common problem in professional football. A challenge that the medical team faces is to successfully predict the recovery time of an injured player. Current medical standards can only give vague predictions as to when a player will return to play. Obviously, making an accurate prediction as soon as possible would be helpful to the coach. This research tries to answer the question of whether it is possible to predict when a player will return to play, based on information at the moment of injury, while also comparing three machine learning methods for this task: support vector machines, Gaussian processes and neural networks. The tests were conducted on data from the professional football club of Tottenham Hotspur. The results demonstrate that this task can be completed with a reasonable amount of accuracy, without any method performing significantly better than the rest. Future directions and possible improvements are discussed.

Keywords: injury prediction, football, support vector machine, neural network, Gaussian process, comparison

1 Introduction

Injuries are a common problem in every sport, including football. Professional football players get injured on average once per year [1] with 10-35 injuries occurring per 1000 game hours [2]. Injuries have been described as the main factor that prevents professional players from not being able to participate in training and playing activities [3].

The factors that cause injuries can vary. A significant percentage of injuries (9%-34%) happens due to overuse [4-5]. Most of the injuries are described as traumatic, with 29% of them being due to foul play [4]. The majority of injuries happen in play, and the most severe cases can be attributed to body contact [5].

As soon as an injury happens it is important to make an estimate of how long the player will need to recover from the injury and get back to play. This information can help the manager make appropriate changes in the squad or the tactical planning of the team. It can also help the director of the club, since new players might need to get signed in order to cover for players who are going to stay out of play for a long time. Additionally, managing the player's expectations with respect to his injury is im-

adfa, p. 1, 2011.

portant, so that the player can prepare himself mentally and psychologically. Finally, it would help the medical team by providing additional certainty in the predictions of the experts.

Currently, there is no standard method to estimate the time a player will miss from play. The time is estimated based on the experience of the physician and on recommendations by various groups and studies. The suggestions can vary quite significantly with each other, and they can also have significant variance. For example, suggestions for return to play following anterior cruciate ligament reconstruction can range from 16 to 24 weeks [6]. Similar recommendations exist for hamstring injuries [7] and concussions [10-13].

Machine learning has been used in sports for various purposes (e.g. cycling [8] and swimming [9]) including football [16-17]. The complicated and multi-factorial nature of many sports makes machine learning a natural choice for predictive tasks.

The purpose of this study is to compare different machine learning methods on predicting the recovery time of professional football athletes. The goal is to make the prediction based on information available at the time of injury, before an official diagnosis has been conducted. There are three main reasons for which the final diagnosis was left out.

First, diagnoses, in some cases, can take some time, while ideally a coach would like to know as soon as possible how long a player will stay out of play. It would be interesting to see what is the best prediction that can be obtained before an official diagnosis.

Secondly, there are many different diagnoses and different levels of abstraction that can be used. For example, in this study's dataset there were some knee injuries that were described as "knee pain, unspecified", "patellofemoral pain" and "Left knee medial meniscus". These diagnoses could be elaborated even further, or they could be abstracted, by classifying them all as "knee injuries". This is a medical problem that can influence the performance of any machine learning or statistical model that will use this information.

However, it is not entirely clear what degree of elaboration would actually help in the prediction of the response variable. For that reason it is important to know what degree of accuracy can be achieved in the prediction of the response variable before including the diagnosis, so that future research could actually tackle the problem of trying to identify the correct level of abstraction needed for this task.

Thirdly, as part of UEFA guidelines, teams in the premier league have to collect information on every injury that occurs. This information mainly concerns extrinsic factors of an injury (e.g., whether the player was running, whether there was a collision, etc.) and it is easy to collect. A proper model that tries to achieve maximum accuracy on the task of predicting the recovery time obviously requires as much information as possible, like a player's medical exams or training records. However, it would be interesting to see what is the maximum accuracy that can be achieved for this task based solely on extrinsic information. This result could be used to establish a baseline which future, more elaborate models, will improve.

The methods that were chosen for this research were Gaussian processes, support vector machines and neural networks. The reason behind these choices is that all these methods are popular for regression tasks. While there are many other choices for solving regression problems in machine learning, these three methods have been proven and tested in a variety of applications, so they provide sensible choices for approaching this task.

The primary goal of this study was to test the degree to which this task is possible in general by reaching a level of error in the predictions that can have practical applicability, at least in some cases. Once this was established, the next goal was to see whether one of these methods is more suited for this task compared to others. The study itself is part of a greater research project that has as a final goal a fully-working predictive system that can aid football teams. Therefore, future plans, directions and suggestions for research are discussed, as well.

2 Methods

2.1 The dataset

The dataset consists of a list of injuries at Tottenham Hotspur Football Club which were recorded according to the UEFA guidelines over the period 2006-2012. For every injury, a list of variables was collected. These are presented in table 1. Note that the variable “injury” included in the dataset is not a final diagnosis, but a first general estimate such as “muscle strain” or “bone injury”.

Table 1. List of variables in the dataset

Parameter	Description
Age	The age of a player
Stage of season	The stage of season (e.g. mid-season or off-season) when the injury occurred
Where	Describes whether the injury took place in the training field or in the game
Phase of play	Describes the exact way that the injury happened (e.g. running or shooting)
Injury	Description of the injury without a specific diagnosis (e.g. bone injury or overuse)
Type	Describes whether the injury was due to overuse or it was an acute injury
Injured side	Describes whether the left or right side was injured
Position	The position of the player (e.g. forward)
Body part injured	Where the player was injured
Reoccurrence	Describes whether the same injury has happened to the same player in the past
Days unavailable	The main variable of interest in our model. It specifies how many days a player stayed out of play after his injury.

All variables, with the exception of “Age” and “Days unavailable” were categorical variables and they were converted to dummy variables in the statistical sense of the term. Therefore, for each value of a categorical variable, a binary variable was created. This gave rise to a dataset that contains 78 variables (including the response variable).

A histogram of the dataset is shown in figure 1. It is evident that most of the injuries are less than 25 days and the histogram is severely skewed. The total number of cases is 152. The mean is 15.5 and the standard deviation 36.039.

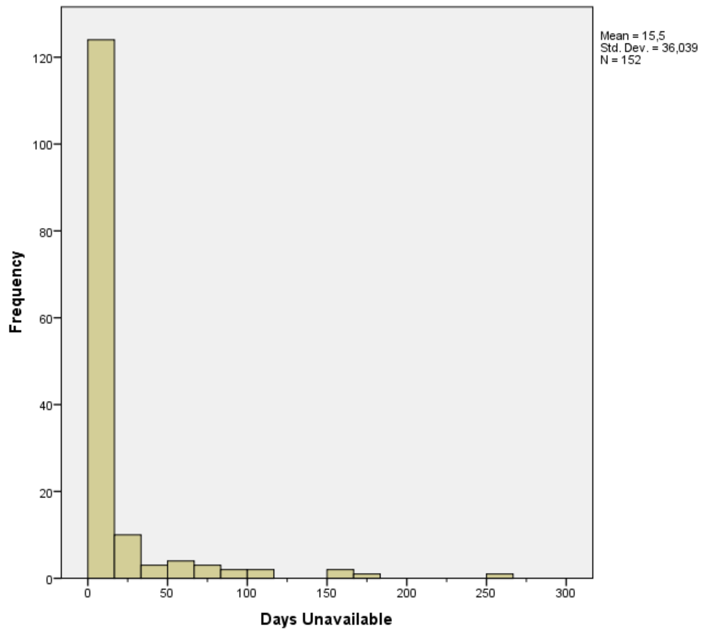

Fig. 1. Histogram of the response variable “Days unavailable”

2.2 Algorithms

Three different methods were used and evaluated: neural networks, support vector machines and Gaussian processes. Each method was executed with many different parameter sets. In order to find the best parameters, grid search was used with some additional manual tuning. Due to the number of tests (more than 50 tests for each method) conducted it is not practical to provide detailed tables and graphs for each parameter set and result.

Tables 2-6 below show the parameters that each method used and their value ranges. For each min-max pair of values 5-25 equidistant steps were used. So for example, for the momentum of the neural network the steps were [0, 0.1, 0.2,…,1]. Once the grid search was done, then some additional manual tweaking was performed.

The neural network was trained using standard backpropagation with momentum.

Table 2. Neural network parameters

	Epochs	Learning Rate	Momentum	Hidden neurons
Min	1500	0.1	0	10
Max	3000	1	1	60

Table 3. SVM parameters, kernel=RBF

	C	Sigma	Epsilon
Min	0	1	0
Max	200	20	2

Table 4. SVM parameters, kernel=polynomial

	C	Degree	Epsilon
Min	0	2	0
Max	200	7	2

Table 5. Gaussian Process parameters, kernel=RBF

	Lengthscale
Min	1
Max	50

Table 6. Gaussian Process parameters, kernel=Laplace

	Lengthscale
Min	1
Max	50

2.3 Evaluation

All the tasks were evaluated using the root mean squared error (RMSE) from the 10-fold cross validation runs of the grid search procedure. The mean of the data was used as a naïve predictor in order to compare the error of the methods against it.

Along with the RMSE the correlation was recorded as well. In the pilot experiments it was observed that, because of the distribution of the data, the error might not always carry a clear picture. In some cases the RMSE would not seem to be significantly better than using the mean as a predictor.

However, careful inspection of individual predictions showed that the RMSE could be severely affected by a few errors. The correlation between the predicted and the actual values is able to provide a scale-free measure of error. The correlation of the naïve predictor with the data is 0. Values above that can provide an additional measure of whether an algorithm can make better predictions than the mean or not.

An example can clarify this a little bit further. An SVM was trained through 10-fold cross-validation on 80% of the data and an RMSE 37.026 was achieved on the test folds. Using the mean of the data as a predictor gives an RMSE of 39.255 which is very close to that achieved through the SVM. However the correlation of SVM's predictions and the true values for the test folds is 0.49, while for the mean it is 0.

The correlation between SVM's predictions on the 20% of the data (41 points) that were not used in the training and the true values is 0.592 and the RMSE 27.74. This relationship is depicted in the scatterplot in figure 2.

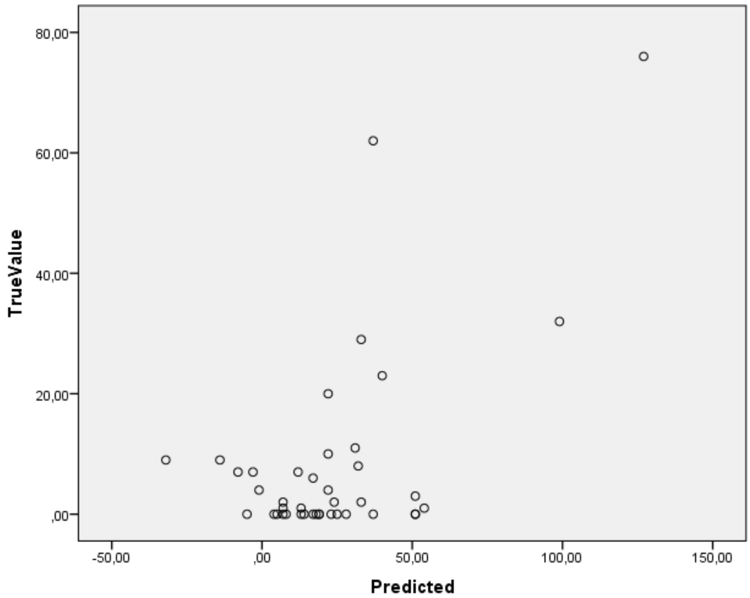

Fig. 2. Scatterplot of true values versus predicted values

For practical applications we can assume that all negative values are effectively 0. If we take that into account then the correlation rises to 0.638 and the RMSE drops to 25.6.

By using the mean as a predictor the correlation is obviously still 0 for this data, but the RMSE is 17.5. If we based our conclusions solely on the RMSE, then we'd assume that the SVM is not performing significantly better. However, the difference between the amended predictions for the SVM and the predictions used for the mean can be clearly seen in the scatterplot in figure 3.

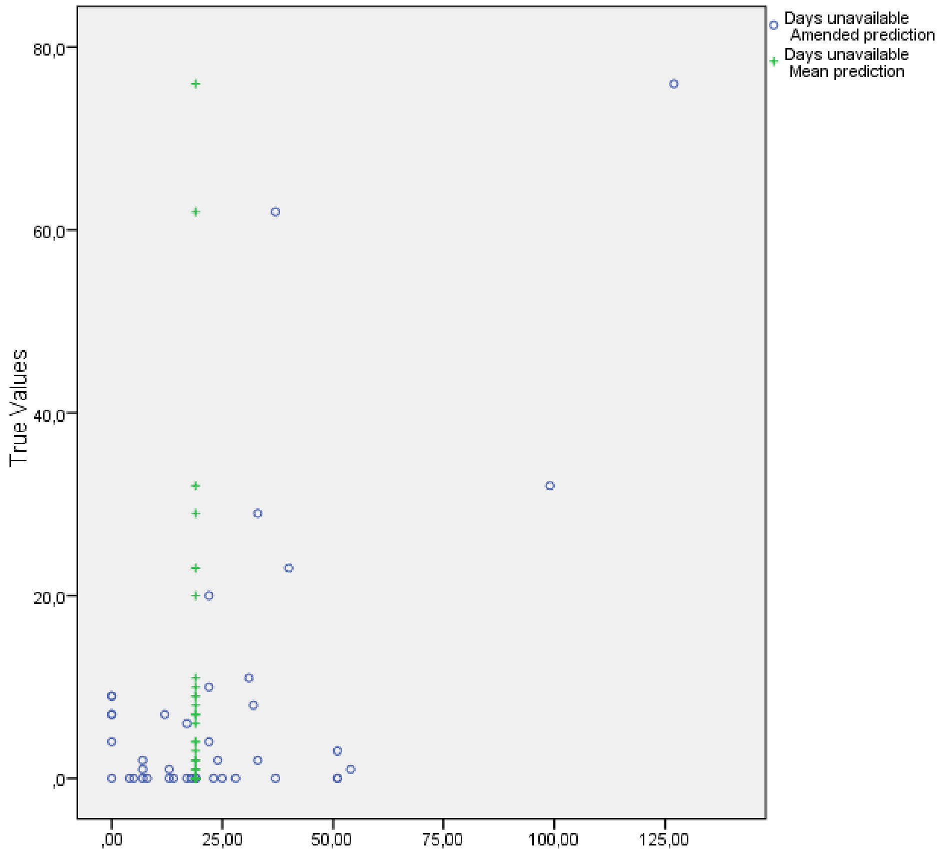

Fig. 3. Scatterplot of true values versus amended predicted values and predictions through the mean

All methods were evaluated using 10-fold cross validation and all tests were executed using RapidMiner version 5.3. The main criterion for choosing the best parameters for each algorithm was the RMSE achieved on the test folds of the cross-validation. The correlation was taken into account when some manual tweaking of the parameters was performed, once the grid search was over.

3 Results

The best results parameter settings achieved for each method are shown in table 7. These parameter settings where evaluated again for each classifier by running 20 rounds of 10-fold cross validation. The RMSEs and correlations for all runs were averaged and they are reported along with their corresponding standard deviations in table 7.

The total RMSE is the error on the whole dataset once the full model has been built.

Table 7. Results for each method

Method	Parameters	RMSE (test)	Correlation (test)	RMSE (total)
SVM	Polynomial kernel, degree=3, C=71, epsilon=1	31.8717 +/- 0.882	0.3565+/- 0.066	4.899
Gaussian Process	RBF kernel, lengthscale=7	32.098+/- 0.846	0.3934+/- 0.0411	5.795
Neural Network	Neurons=45, epochs=2500 learning rate=0.4, momentum=0.2	32.585 +/- 2.02	0.3607+/- 0.0672	1.303

Using the mean as a predictor the RMSE obtained was 35.92. The Wilcoxon signed rank test was conducted for the test RMSE of each classifier. The goal was to check if the error is significantly lower than the error achieved by just using the mean as a predictor. The p-value was less than 0.01 for all three classifiers.

The p-value for a test that the correlation of the classifiers is 0 had a p-value less than 0.01 for all three classifiers.

4 Discussion

It is evident that this task can be predicted with some degree of accuracy, albeit small. The means of the errors and their variances indicate that no single method seems to perform significantly better to others. However, the important point is that some estimate can be obtained. The extrinsic information that is collected for an injury seems to be useful, even to a small extent, when predicting the recovery time of players after an injury.

The results become more important when considering that the size of the dataset is small for this task and it concerns only a single football club. There are many types of injuries in football that can occur under different circumstances. Future research should use datasets from other football clubs in order to verify and expand the current results. Ideally, datasets from football clubs from different countries should be obtained, since the style of play in each country, along with other factors (e.g. a country's climate), could influence the response variable.

Obviously, the end goal is the practical applicability of the results. An issue with the evaluation of the results is the desired degree of accuracy that is required for a method in this task to be considered successful from the perspective of practical applicability. Football teams play a certain amount of games within a season. Usually this is 4 league games per month, and maybe some more cup games and games in European competitions. If a player is injured in a game, it might not matter so much whether he will be back in play in 3 or 5 days, as long as the coach knows that in 7 days, when the next game starts, he will be ready to play.

Furthermore, the dataset contains many transient injuries. The meaning of the word transient is vague and its interpretation is better left to a medical professional, but in general it describes injuries where the recovery time was 0 days or close to that value (e.g 2 days). Many of these cases do not require the execution of a predictive algorithm, because the medical professionals of the team can very quickly classify the injury as transient. Predictions are more helpful for injuries that have longer lasting effects, for example, more than a couple of weeks. This means, that the margin of error can be higher. If the medical staff's opinion is that a player will miss 5 to 10 weeks, then a prediction that manages to narrow down this margin to, for example, 6 to 7 weeks, can help the coach make better decisions and plan for the future.

An interesting feature of this task is that the models could be included in a diagnostic protocol. After each injury, the medical staff will conduct detailed medical tests in order to diagnose the injury. Models like the ones described in this paper could accompany a diagnosis, providing some additional support for the experts' estimates.

Furthermore, additional information that could be available at the moment of injury includes anthropometric and medical information such as the height, weight or medical blood tests of players. This information could improve the accuracy of the model, while also staying true to its original goal of making predictions right after an injury has occurred.

Finally, future research could also solve the problem of how additional official diagnostic information could be used alongside this model in order to make more accurate predictions.

5 Conclusion

This research dealt with the question of whether it is possible to predict the recovery time after an injury in professional football without an official diagnosis, while it also tests 3 methods against each other for this task. The results demonstrate that it is possible to reach some degree of accuracy in this task, but the size of the dataset, and maybe the variables themselves, limit the accuracy that can be reached. No single method was deemed to be significantly better than any of the other methods that were used.

However, this work paves the way for future research that can include bigger and more complicated datasets and can also be extended by protocols that can combine experts' opinions. Future research will built on top of the current results in order to provide a functional system for assessing injuries in professional football.

6 Acknowledgments

The author of this research would like to thank Wayne Diesel of Tottenham Hotspur Football Club for kindly sharing the data for conducting this research.

Bibliography

1. A. Junge and J. Dvorak, "Soccer injuries: a review on incidence and prevention," *Sports Medicine*, vol. 34, no. 13, pp. 929-938, 2004.
2. J. Dvorak and A. Junge, "Football injuries and physical symptoms: a review of the literature," *The American Journal of Sports and Medicine*, vol. 28, no. 5, 2000.
3. L. Parry and B. Drust, "Is injury the major cause of elite soccer players being unavailable to train and play during the competitive season?," *Physical Therapy in Sport*, vol. 7, no. 2, pp. 58-64, 2006.
4. R. D. Hawkins and C. W. Fuller, "Risk assessment in professional football: an examination of accidents and incidents in the 1994 World Cup finals," *British Journal of Sports Medicine*, vol. 30, no. 2, pp. 165-170, 1996.
5. L. Peterson, A. Junge, J. Chomiak, T. Graf-Baumann and J. Dvorak, "Incidence of football injuries and complaints in different age groups and skill-level groups," *The American Journal of Sports Medicine*, vol. 28, no. 5, 2000.
6. M. Bizzini, D. Hancock and F. Impellizzeri, "Suggestions from the field for return to sports participation following anterior cruciate ligament reconstruction: soccer," *The Journal of Orthepaedic and Sports Physical Therapy*, vol. 42, no. 4, 2012.
7. J. Mendiguchia and M. Brughelli, "A return-to-sport algorithm for acute hamstring injuries," *Physical Therapy in Sport*, vol. 12, no. 1, 2011.
8. B. Ofoghi, J. Zeleznikow, C. MacMahon and D. Dwyer, "Supporting athlete selection and strategic planning in track cycling omnium: A statistical and machine learning approach," *Information Sciences*, vol. 233, pp. 200-213, 2013.
9. E. Meżyk and O. Unold, "Machine learning approach to model sport training," *Computers in Human Behavior*, vol. 27, no. 5, p. 1499–1506, 2011.
10. A. Arnason, A. Gudmundsson, H. A. Dahl and E. Jóhannsson, "Soccer injuries in Iceland," *Scandinavian Journal of Sports & Medicine in Sports*, vol. 6, no. 1, pp. 40-45, 1996.
11. A. B. Nielsen and J. Yde, "Epidemiology and traumatology of injuries in soccer," *The American Journal of Sports Medicine*, vol. 17, no. 6, pp. 803-807, 1989.
12. R. C. Cantu, "Guidelines for return to contact sports after a cerebral concussion," *Physician and Sports Medicine*, vol. 14, no. 10, pp. 75-83, 1986.
13. G. Dicker, "A sports doctor's dilemma in concussion," *Sports Medicine Training and Rehabilitation*, vol. 2, pp. 203-209, 1991.
14. M. Lovell, M. Collins and J. Bradley, "Return to play following sports-related concussion," *Clinics in Sports Medicine*, vol. 23, pp. 421-441, 2004.
15. M. W. Collins, M. R. Lovell and D. B. McKeag, "Current issues in managing

sports concussion," *The Journal of the American Medical Association,* vol. 282, pp. 2283 - 2285, 1999.

16. A. Joseph, N. E. Fenton and M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques," *Knowledge-Based Systems*, vol. 19, no. 7, pp. 544-553.

17. B. Min, J. Kim, C. Choe, H. Eom and B. R. I. McKay, "A compound framework for sports results prediction: A football case study," *Knowledge-Based Systems*, vol. 21, no. 7, 2008.

18. S. Marsland, Machine learning: an algorithmic perspective, CRC Press, 2009.

19. A. Junge and J. Dvorak, "Influece of definition and data collection on the incidence of injuries in in football," *The American Journal of Sports Medicine,* vol. 28, no. 5, 2000.