Ayala, Francisco ORCID logoORCID: https://orcid.org/0000-0003-2210-7389, López-Valenciano, Alejandro, Jose, Antonio, De Ste Croix, Mark B ORCID logoORCID: https://orcid.org/0000-0001-9911-4355, Vera-García, Francisco, García-Vaquero, Maria, Ruiz-Pérez, Iñaki and Myer, Gregory (2019) A preventive model for hamstring injuries in professional soccer: Learning algorithms. International Journal of Sports Medicine, 40 (5). pp. 344-353. doi:10.1055/a-0826-1955

Official URL: https://doi.org/10.1055/a-0826-1955
DOI: http://dx.doi.org/10.1055/a-0826-1955
EPrint URI: https://eprints.glos.ac.uk/id/eprint/6383

Disclaimer

The University of Gloucestershire has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

The University of Gloucestershire makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

The University of Gloucestershire makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

The University of Gloucestershire accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

A preventive model for hamstring injuries in professional soccer: Learning algorithms

Professor Mark de Ste Croix

ABSTRACT

Hamstring strain injury (HSI) is one of the most prevalent and severe injury in professional soccer. The purpose was to analyse and compare the predictive ability of a range of machine learning techniques to select the best performing injury risk factor model to identify professional soccer players at high risk of HSIs. A total of 96 male professional soccer players underwent a pre-season screening evaluation that included a large number of individual, psychological and neuromuscular measures. Injury surveillance was prospectively employed to capture all the HSI occurring in the 2013/2014 season. There were 18 HSIs. Injury distribution was 55.6% dominant leg and 44.4% non-dominant leg. The model generated by the SmooteBoostM1 technique with a cost-sensitive ADTree as the base classifier reported the best evaluation criteria (area under the receiver operating characteristic curve score=0.837, true positive rate=77.8%, true negative rate=83.8%) and hence was considered the best for predicting HSI. The prediction model showed moderate to high accuracy for identifying professional soccer players at risk of HSI during pre-season screenings. Therefore, the model developed might help coaches, physical trainers and medical practitioners in the decision-making process for injury prevention.

Keywords: injury prevention, injury risk, modelling, screening, decision-making.

INTRODUCTION

Hamstring strain injury (HSI) is the most prevalent noncontact injury reported in professional male soccer (football) representing 12% to 14% of all injuries [16], accounting for 37% of all muscle injuries sustained [16,17,24] and resulting in a mean of 14 competition days lost per injury [range 1-128 days] [15]. Furthermore, the recurrence rate of HSIs remains substantial, ranging from 16% to 60% [24].

Prior to establishing injury prevention programmes, it may be of value to identify soccer players at high risk of HSI. Several prospective studies have identified a number of modifiable (e.g.: strength, joint ranges of motion [ROM], trunk stability) and non-modifiable (e.g.: age, sex, history of HSI) risk factors that have demonstrated a statistically significant relationship with HSI [3,9,12-14,20,25,37,38]. It should be noted that among all of these modifiable and non-modifiable risk factors, history of HSI is the only one that has been consistently identified as a primary risk factor for future injury [20,25]. However, the presence of a statistically significant association does not imply that there is a causal relationship between the factor and injury incidence and hence, this knowledge alone is likely insufficient to identify soccer players at high risk of HSI [6]. Accordingly, some studies have defined markers or cut-off scores for specific risk factors in an attempt to identify soccer players at high risk of HSI [12,13,20,37].

However, despite the substantive effort made in recent years by the scientific community and medical practitioners to firstly identify soccer players at high risk of HSI and then apply tailored injury prevention programmes, recent evidence has demonstrated that HSI incidence has not decreased, but has increased slightly over recent years [17].

Two different arguments appear to be behind the lack of generality of the proposed cut-off scores and this could explain why they cannot identify soccer players at high risk of HSI. Firstly, the generality of the cut-off scores proposed for certain injury risk factors (e.g.: strength imbalance, joints ROM) might be limited since their predictive abilities to identify new soccer players at high risk of HSIs has not been verified in a new population of players (e.g. a different group than that used to defining the cut-off values originally) [6,27]. This suggests that cut-off scores might be overfitted (i.e. their predictive ability is adjusted to the data set used in their learning process), with this yielding overly optimistic performance and hence, they may not be acceptable for screening purposes. This appears to be supported by the fact that the cut-off scores defined by some prospective studies (mainly those related to strength measures) have not been later ratified by others using similar designs and assessment methodologies but with different samples of soccer players [3,9,12-14,20,25,37,38]. For example, while Croisier et al. [12] and Dauty et al. [14] found that professional soccer players with reciprocal (functional) hamstring-to-quadriceps strength ratios (H/Q) lower than 0.8 were at higher risk of sustaining an HSI, van Dyk et al. [38] did not identify this strength ratio measure as a risk factor for HSI. The second issue with the current body of the literature is that most of the available studies have identified potential risk factors for HSI according to the presence of statistically significant relationships (based on odds ratios, certain values of the p statistic [mainly $p < 0.05$]) with HSI. However, based on the general agreement that the aetiology of HSI is multifactorial and that some relationships of conditional dependence might exist among factors, it is possible that the influence of a specific factor on the likelihood of suffering an HSI might not be statistically significant ($p < 0.05$) in itself, but relevant when it is used in conjunction with several other factors to develop a more robust predictive model. In other words, combining information from several modifiable and non-modifiable risk factors might lead to the development of a more robust model with an improved predictive ability.

The application of contemporary statistical approaches (e.g.: supervised learning algorithms) derived from Machine Learning and Data Mining environments, which have been specifically designed to deal with problems where a large number of factors are involved and the use of resampling techniques (i.e. cross-validation, bootstrap and leave-one-out), may overcome the limitations inherent to the current body of knowledge and it might shed new light to better identify athletes at high risk of HSI.

Lopez-Valenciano et al. [28] and Rossi et al. [31] have recently developed a muscle injury and a non-contact injury predictive model specifically for soccer players after having determined several modifiable and non-modifiable risk factors and by utilising supervised learning algorithms. The predictive power of these models is significantly higher than those reported in other models where traditional (lineal) approaches were applied [3,9,12-14,20,25,37,38].

Therefore, the main purpose of this study was to analyse and compare the predictive ability of a range of learning methods in order to select the best performing injury risk factor model to identify professional soccer players at high or low risk of HSI.

METHOD

Participants

A total of 96 male professional soccer players took part in the current study. Soccer players were recruited from four different soccer teams that were engaged in the 1st (one team, n = 25) and 2nd B (three teams, n = 73) Spanish National Soccer League divisions.

The exclusion criteria were: a) presence of orthopaedic problems that prevented the proper execution of one or more of the neuromuscular tests selected for this study; and b) players who were transferred to other clubs and did not finish the 9-month follow up period. Only primary injuries we used for any player sustaining multiple HSIs.

Prior to study participation, experimental procedures and potential risks were fully explained to the participants in verbal and written form, and written informed consent was obtained from them. The Institutional Research Ethics committee of Miguel Hernandez University of Elche approved the study protocol prior to data collection (DPS.FAR.02.14) and followed the ethical standards of the journal IJSM [26].

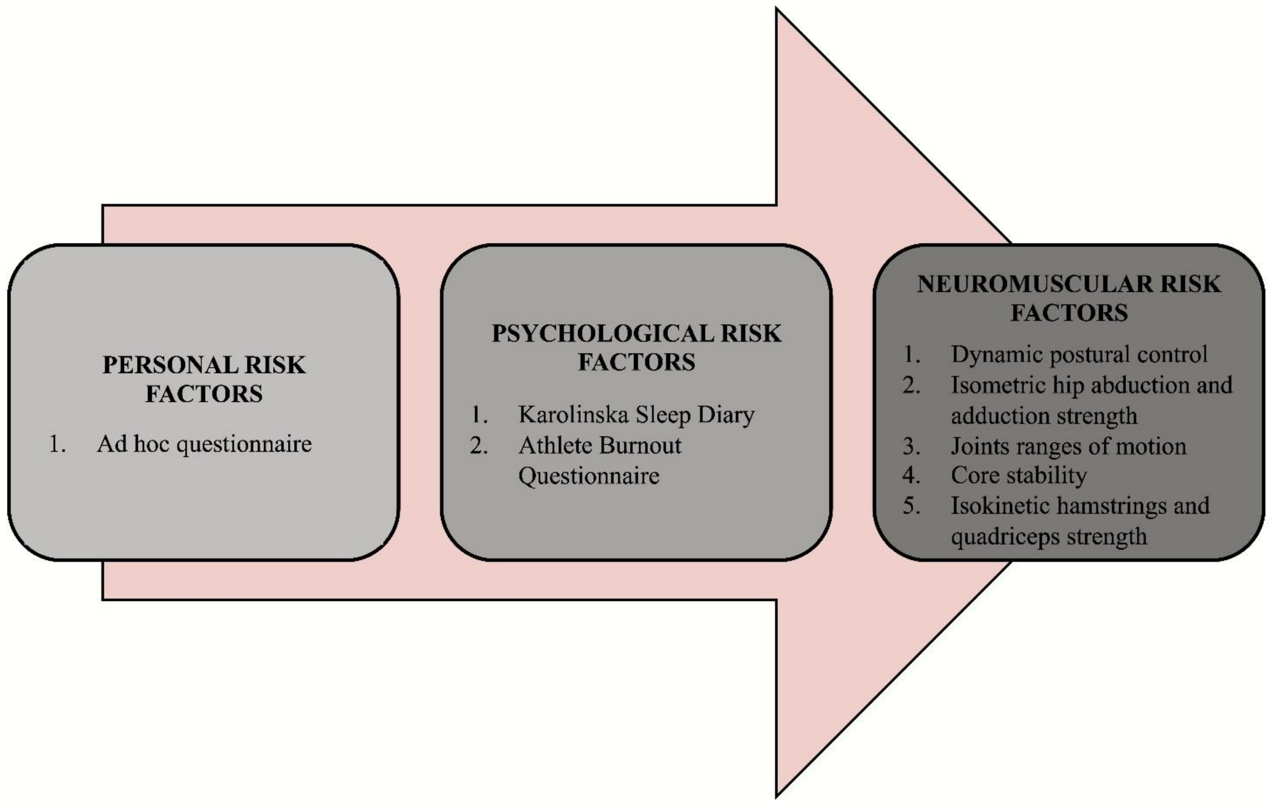
Study design

A prospective cohort design was used to address the purposes of this study. In particular, all the HSIs accounted for within the 9 months (2013/2014 season, from the second week of August to the second week of May) following the initial testing session were prospectively collected for all players.

Players underwent a pre-season evaluation of a number of personal, psychological and neuromuscular measures, most of them considered potential sport-related injury risk factors. In each soccer team, the testing session was conducted at the middle-end of the pre-season phase of the year (end of July or beginning of August).

Testing procedure

The testing session was divided into three different parts (Figure 1). The first part of the test session was used to obtain information related to the participants' personal or individual characteristics. The second part was designed to assess psychological measures related to sleep quality and athlete burnout. Finally, the third part of the session was used to assess a number of neuromuscular measures. A substantive number of individual, psychological and neuromuscular measures coming from these three parts of the testing session were recorded (n = 229) with the aim of developing a risk factor model that could reflect the suggested multifactorial nature of the HSI phenomenon.


PERSONAL RISK FACTORS
1. Ad hoc questionnaire
PSYCHOLOGICAL RISK FACTORS
1. Karolinska Sleep Diary
2. Athlete Burnout Questionnaire
NEUROMUSCULAR RISK FACTORS
1. Dynamic postural control
2. Isometric hip abduction and adduction strength
3. Joints ranges of motion
4. Core stability
5. Isokinetic hamstrings and quadriceps strength

(Figure 1.)

Each of the 8 testers who took part in this study conducted the same tests throughout all the testing sessions and they were blinded to the purposes of this study. All testers were members (two senior and two junior researchers, two technicians and two PhD students) of the same research team and had more than 4 years of experience in neuromuscular assessment.

Personal or individual risk factors

The ad hoc questionnaire designed by Olmedilla et al. [29] was used to record personal or individual features that have been defined as potential non-modifiable risk factors for sport injuries. Through this questionnaire sport-related background (player position, current level of play, dominant leg [defined as the participant´s kicking leg]) and demographic (age, body mass and stature) features were recorded. In addition, the presence within the last season (yes or no) of HSIs with a total time taken to resume full training and competition > 8 days was also recorded (self-reported). Supplementary file 1 displays a description of all the personal risk factors recorded.

Psychological risk factors

Sleep quality and athlete burnout variables were measured through two validated and worldwide used Likert scales. The Spanish version of the Karolinska Sleep Diary [1] was used to measure the sleep quality of the soccer players. The Spanish version of the Athlete Burnout Questionnaire [2] was used to assess the three different dimensions that comprise athlete burnout: a) physical/emotional exhaustion; b) reduced sense of accomplishment; and c) sport devaluation. Supplementary file 2 displays a description of all the psychological risk factors recorded.

Neuromuscular risk factors

Prior to the neuromuscular risk factor assessment, all participants performed the dynamic warm-up designed by Taylor et al. [35]. The overall duration of the entire warm-up was approximately 15-20 min. The assessment of the neuromuscular risk factors was carried out 3-5 min after the dynamic warm-up.

In the experimental session, participants were assessed from a number of neuromuscular performance measures obtained from 5 different testing manoeuvres: 1) dynamic postural control [33], 2) isometric hip abduction and adduction strength [36], 3) lower extremity joint ROMs [10], 4) trunk stability [7] and 5) isokinetic hamstrings and quadriceps strength [5]. For a matter of space, the testing manoeuvres are not described below, and the reader is to refer to their original sources. Furthermore, Supplementary files 3-7 display a description of the five testing manoeuvres carried out and the neuromuscular risk factors recorded through each of the manoeuvres.

The order of the tests was consistent for all participants (Figure 1) and was established with the intention of minimizing any possible negative influence among variables. A 5-min rest interval was given between consecutive testing manoeuvres.

Injury Surveillance

Following the recommendations made by the International Injury Consensus Group [22], a HSI was defined as an acute pain in the hamstrings location that occurred during training or competition and resulted in the immediate termination of play and inability to participate in the next training session or match. HSIs were confirmed through a clinical examination (identifying pain on palpation, pain with isometric contraction, and pain with muscle) by team doctors. Players were considered injured until the club medical staff (medical doctor or physiotherapist) allowed full participation in training and availability for match selection.

The club medical staff of each club recorded HSIs on an injury form that was sent to the study group each month. For all HSIs, team medical staff provided the following details to investigators: leg injured (dominant/non-dominant), injury severity based on lay off time from soccer (slight/minimal [0-3 days], minor [4-7 days], moderate [8-28 days], and severe [>28 days]), date of injury, moment (training or match), whether it was a recurrence (defined as an HSIs that occurred in the same leg and during the same season as the initial injury), and total time taken to resume full training and competition. At the conclusion of the 9 months follow up period, all data from the individual clubs were collated into a central database, and discrepancies were identified and followed up at the different clubs to be resolved. Some discrepancies among medical staff teams were found to diagnose minimal HSIs and to record their total time lost. To resolve these inconsistencies in the injury surveillance process (risk of misclassification of the players), only HSIs showing a time lost > 4 day (minor to severe) were selected for the subsequent statistical analysis.

Statistical analysis

The statistical analysis framework carried out in this study for analysing and comparing the behaviours of several machine learning techniques with the aim of finding the best model for predicting HSIs in professional soccer players was based on a supervised learning perspective. From a statistical standpoint, the problem can be stated as follows: given a set of features F (in our case risk factors) and a target (discrete) variable (in our case HSI [yes or no]), named class, C, we want to estimate/learn a mapping function M:F→C. Thus, the statistical analysis comprised two stages:

1. Data pre-processing. At this stage, the data set was prepared to apply the machine learning techniques. To optimise this aspect, pre-processing methods such as data cleaning and data discretization were applied.
2. Data processing. At this stage, the most powerful techniques reported by Elkarami et al. [18] and Galar et al. [23] to address learning with imbalanced data sets were applied in order to build models for predicting HSIs. In particular, a study on the performance of some proposals for pre-processing, cost-sensitive learning and ensemble-based methods was carried out. Three classic decision tree algorithms were used as base classifiers in each method: J48 [30], ADTree [21] and SimpleCart [8].

A complete description of the statistical techniques carried out in both stages, data pre-processing and data processing, has been written in the Supplementary file 8.

In order to evaluate the performance of the decision tree algorithms, the 3-fold stratified cross validation (SCV) technique was used. That is, we split the dataset into 3 folds, each one containing 33,3% of the patterns of the dataset. For each fold, the algorithm was trained with the examples contained in the remaining folds and then tested with the current fold. A wide range of classification performance measures can be obtained from the SCV technique. A well-known approach to unify these measures and to produce an evaluation criterion is to use the area under the ROM curve (AUC). In particular, the AUC corresponds to the probability of correctly identifying which one of the two stimuli is noise and which one is signal plus noise [23]. Thus, the AUC was used as a single measure of a classifier's performance for evaluating which model is better on average. Furthermore, two extra measures from the confusion matrix were also used as evaluation criteria: a) true positive rate (TPrate): TPrate = $\frac{TP}{TP+FN}$ also called sensitivity or recall, is the proportion of actual positives which are predicted to be positive; and b) true negative rate (TNrate): TNrate = $\frac{TN}{TN+FP}$ or specificity, is the proportion of actual negatives which are predicted to be negative.

RESULTS

Hamstrings muscle strain injuries epidemiology

There were 18 HSIs over the follow up period and all of them were used to train the models. Injury distribution between the legs was 55.6% dominant leg and 44.4% non-dominant leg. In term of severity, most of injures were categorized as moderate (n = 15) while only 3 cases were considered minor and no severe injuries were recorded.

Predictive model for lower extremity muscle injuries

Table 1 shows the average AUC, TPrate and TNrate results for all oversampling and ensemble learning methods separately for each decision tree base classifier. Highlighted in bold is the method that obtained the best performing result within each method. Furthermore, highlighted in grey is the model considered as the best for predicting HSI.

The ADTree base classifier reported the best performance in most of the methods analysed. In fact, the final model was built using the SmoteBoostM1 ensemble method with the ADTree as the base classifier using a reweighted training instance (cost-sensitive) approach.

Therefore, the final model selected to predict HSI in professional soccer players was comprised by 10 different cost sensitive ADTree classifiers (Supplementary files 9-18). The cost matrix for cost-sensitive classifier was set to $C\left\{\frac{0}{1}\middle|\frac{11}{0}\right\}$ where a false negative had a cost of 11 and a false positive had a cost of 1. This cost matrix was selected because it reported the best predictive performance in this particular scenario after having tested all the possible combinations.

The confusion matrix and the main cross validation results of the final model are shown in tables 2 and 3 respectively.

Table 1: Average AUC, TPrate and TNrate results for all the decision tree methodologies in isolation and after having been applied in them the oversampling and ensemble techniques selected

Technique	AUC	TPrate	TNrate
Cost-sensitive base classifiers			
J48	0.474	16.7	77.9
ADTree	0.675	33.3	80.9
Scart	**0.756**	**77.8**	**69.1**
Oversampling techniques			
CS-SMT			
J48	0.547	33.3	76.5
ADTree	**0.759**	**50**	**79.4**
Scart	0.603	50	69.1
Boosting-based Ensembles			
CS-SBOM1			
J48	0.669	33.3	89.7
ADTree	**0.837**	**77.8**	**83.8**
Scart	0.661	50	79.4
CS-RUSB			
J48	0.723	66.7	66.2
ADTree	**0.750**	**77.8**	**63.2**
Scart	0.695	77.8	57.4
Bagging-based Ensembles			
CS-OB			

Table 2: Confusion Matrix

A	B	Classified as
14	4	A = Injured
11	57	B = Non Injured

Table 3: Cross validation results for the final prediction model

Correctly classified instances	71 (82.6%)
Incorrectly classified instances	15 (17.4%)
Kappa statistic	0.539
Mean absolute error	0.199
AUC	0.837

DISCUSSION

The current study is the first (to the best of our knowledge) that has built a model to predict HSI by applying a novel multifactorial approach and whose predictive ability has been determined through the exigent resampling technique called cross-validation. In this study the HSI risk model is comprised of 10 classifiers with a tree-shape structure and was developed thanks to the application of learning algorithms (on the training subsets) widely used in the Data Mining setting. Thus, the model reports an AUC score of 0.837 with true positive and negative rates of 77.8% and 83.8% respectively.

The predictive ability of the model built in the current study to identify athletes at high risk of HSI is higher than the only study published to date that has used supervised learning algorithms with the aim of predicting the incidence of HSI in Australian footballers [32]. Ruddy et al. [32] investigated the ability of some individual (age, history of HSI last season, stature, mass and primary playing position) and strength (eccentric hamstring strength) risk factors to identify Australian footballers at high risk of HSI through the use of some supervised learning algorithms (Naive Bayes, Logistic regression, Random forest, Support vector machine, Neural network) reporting AUC scores lower than 0.6. Perhaps the limited number of risk factors determined by Ruddy et al. [32] to build the models may explain the discrepancy found with the predictive scores reported in the current study. Based on the general agreement that the aetiology of HSI is multifactorial and that no powerful individual predictors have been found, the combination of information from several modifiable and non-modifiable risk factors might lead to the development of a more robust model with an improved predictive ability. On the other hand, the predictive ability of the model built in the current study was similar to those reported by the two predictive models available in the existing literature that were built using a large number of risk factors and thank to the application of a supervised learning algorithm (decision tress), with the aim of identifying professional soccer players at high risk of muscle injury [28] and non-contact injury [31]. Lopez-Valenciano et al. [28] built an injury risk factor-based model to identify professional soccer and handball players at high risk of lower extremity muscle injuries, which comprised of 10 classifiers with a tree-shape structure (SmooteBoost technique with a cost-sensitive ADTree as base classifier). Fifty-two features reported an AUC score of 0.747 with true positive and negative rates of 65.9% and 79.1% respectively. Unlike Lopez-Valenciano et al. [28] who prospectively recorded lower extremity muscle injuries (hamstrings, quadriceps, adductors and triceps surae), the current study only focused on HSIs. Perhaps, the fact that the current study built an injury-specific predictive model might explain the slightly better predictive performance results obtained in comparison with the non-specific injury risk model developed by Lopez-Valenciano et al. [28]. Likewise, Rossi et al. [31], included 16 weeks of training workload data, collected via GPS, built a non-contact injury model that reports a true positive and negative rate of 76% and 100%, respectively. In contrast to the model developed by Rossi et al. [31] our model was conceived to be used as a single session pre-participation screening tool for the prevention of muscle injuries rather than needing to determining training load over a number of weeks using GPS technology and hence, it is less time consuming and more injury-specific.

On the other hand, the predictive ability of the current model to identify soccer players at high risk of HSI is much higher than those reported in models from previous studies in which less exigent validation processes were applied [3,9,12-14,20,25,37,38]. For example, van Dyk et al. [38] found that two independent predictors were associated with the risk of HSI (hamstring eccentric strength and quadriceps concentric strength) from regression analysis, but the ROC analysis demonstrated an AUC lower than 0.6. Likewise, Timmins et al. [37] stated that those soccer players showing eccentric knee flexion strength scores lower than 337N had 4.4 times greater risk of a subsequent HSI in comparison with stronger players. However, the reported value of the ROC for this cut-off score was only 0.65.

In the current study the learning process of the model started with 229 features, however the final model only considered 66 of them relevant (Table 4). This finding indicates that the range of variables required to identify high and low risk players is manageable in real world settings and would considerably reduce the time required in the pre-season screening processes aimed at identifying athletes at high risk of HSIs. The three main categories of potential injury risk factors employed in the current study (psychological, personal and neuromuscular) all have some representation in the final model selected and hence, this reinforces the idea that the aetiology of HSI is multifactorial.

Table 4: Risk factor measures included in the model for predicting HSI and the number of times that they appear in the classifiers, In bold are highlighted those that appear in four or more classifiers

Risk Factor	Nº of Classifiers
Personal measures	
Age	1
History of HSI last season	3
Maximal level of play achieved	1
Psychological measures	
Sleep quality	5
Physical/emotional exhaustion	2
Reduced sense of accomplishment	4
Dynamic postural control measures	
YBalance-Ant-Non Dominant Leg	2
Ybalance-PostMedial-Non Dominant Leg	1
YBalance-PostLateral-Non Dominant Leg	1
YBalance-BilaRatio-Anterior	1
YBalance-BilaRatio-PostLateral	2
Isometric hip abduction and adduction strength measures	
PT$_{ISOM}$-Hadd-Dominant Leg	1
PT$_{ISOM}$-Hadd-Norm-Non Dominant Leg	2
PT$_{ISOM}$-Hadd-Norm-Dominant Leg	1
BilaRatio-PT$_{ISOM}$-Habd- Dominan Leg	1
Lower extremity joints range of motion measures	
ROM-PHF$_{KE}$-Dominant Leg	4

The main features related to the psychological category of burnout (physical/emotional exhaustion and reduced sense of accomplishment) were important, but specifically sleep quality was an important risk factor as it was the most consistent variable present in the classifiers (5 out of 10 classifiers). This is the first study that has analysed whether burnout and sleep quality measures are predictive of HSI, alongside other known variables, and therefore direct comparisons are not possible. However, this finding is in concordance with the results found by Cresswell and Eklund [11] who reported statistically significant correlations between sport-injuries and feelings of sport devaluation in a cohort of professional rugby players. Perhaps, the feeling of frustration experienced by players with a short-term history of HSIs might lead them to lose concentration and this can impair the neuromuscular readiness to perform high-intensity intermittent actions during both training and match play, and thus might increase the risk of HSI.

Furthermore, previous HSI, identified by the variable “history of HSI last season” also reported a high presence among the classifiers of the model, evident in three out of 10. This finding is in agreement with the findings of several previous studies [20,25], although not all [3], in which previous HSI has been identified as an independent predictor for HSI in professional soccer players. Remaining deficits in physical conditioning or proprioception or altered movement patterns after a previous injury may provide a plausible link to an anatomically unrelated injury in a following season [25].

Another feature that consistently appears in the predictive model is hip flexion ROM with the knee passively extended (ROM-PHF$_{KE}$), which is presented in four out of 10 classifiers. This finding is in concordance with the results found by previous studies where hip flexion ROM (consider as an indirect measure of hamstrings muscles flexibility) has been identified as a primary risk factor for HSI [39]. A possible explanation for this might be attributed to the fact that players with limited ROM-PHF$_{KE}$ may have hamstring muscles that are not sufficiently prepared to store and release the high amount of elastic energy generated during repeated high intensity movements that are intrinsic to soccer play (i.e. sudden acceleration and deceleration, rapid changes of directions, jumping and landing tasks), and this might predispose such players to HSI [40].

The findings of the current study also highlight that poor reciprocal hamstring-to-quadriceps ratios, calculated using angle specific torque values close to full extension, are present in the identification of players at high risk of HSI in comparison with their homologous ratios calculated by using peak toque values. Likewise, hamstring and quadriceps eccentric torque values obtained close to knee extension (15º, 30º and 45º) also seem to adopt a critical role in the predictive model. A possible explanation for this could be attributed to the higher ecological validity of the angle-specific reciprocal H/Q ratios to describe the function of the knee [4]. Biomechanical studies have indicated that HSIs are more prone to occur during the latter part of the swing phase of sprinting (closer to full knee extension) when the hamstrings are working eccentrically (energy absorption) to decelerate the knee extension movement (generated among others by the concentric action of the quadriceps muscles) before foot contact, that is, as the muscle develops maximal tension while lengthening to stabilise the knee joint [34]. However, peak concentric and eccentric torque production is likely to occur in the mid-late range of the movement (around 40°–80° of knee flexion [0º = full knee extension]) [19]. Therefore, this joint angle discrepancy inherent between any peak torque H/Q ratio and where the HSI is likely to occur may reduce its validity to assess the muscular balance of the knee. This aspect could justify the reason why the angle-specific H/Q ratios play a more significant role in the likelihood of sustaining an HSI, as they may be more relevant to describe the muscular control of the knee.

Therefore, our model suggests that the angle of peak torque measured during eccentric (hamstrings) knee extension movements is important for predicting in-season HSI, as this variable is present in some classifiers. This finding supports the hypothesis of Brockett, Morgan and Proske [9] who suggest that in order to prevent HSI where players are able to achieve the peak torque throughout the given ROM is more relevant than the net peak torque value.

The model built also provides a main role to the isokinetic strength features to predict future HSIs, with 45 features out of 66. These results are not in agreement with some previous findings [38,41] who suggest that isokinetic testing cannot predict the risk of hamstring injury in subsequent professional competition. Based on our findings regarding angle specific torque data it may be that insufficient ecological validity of the isokinetic methodologies used in the above studies could explain this discrepancy. Additionally, van Dyk et al. [38] and Zvijac et al. [41] examined the relationship between torque and the likelihood of sustaining a hamstring employing isokinetic protocols with the participants adopting a seated position (80°–110° hip flexion). This seated position is not representative of the hip position during sporting tasks (i.e. sprinting, cutting) and does not replicate hamstrings and quadriceps muscle length–tension relationships that occur in the late phase of sprinting, were hamstring injury is likely to occur [34]. In contrast to these studies, we adopted a prone position (10–20° hip flexion), which has been suggested as being more functionally relevant in term of simulating the injury mechanism [5,34].

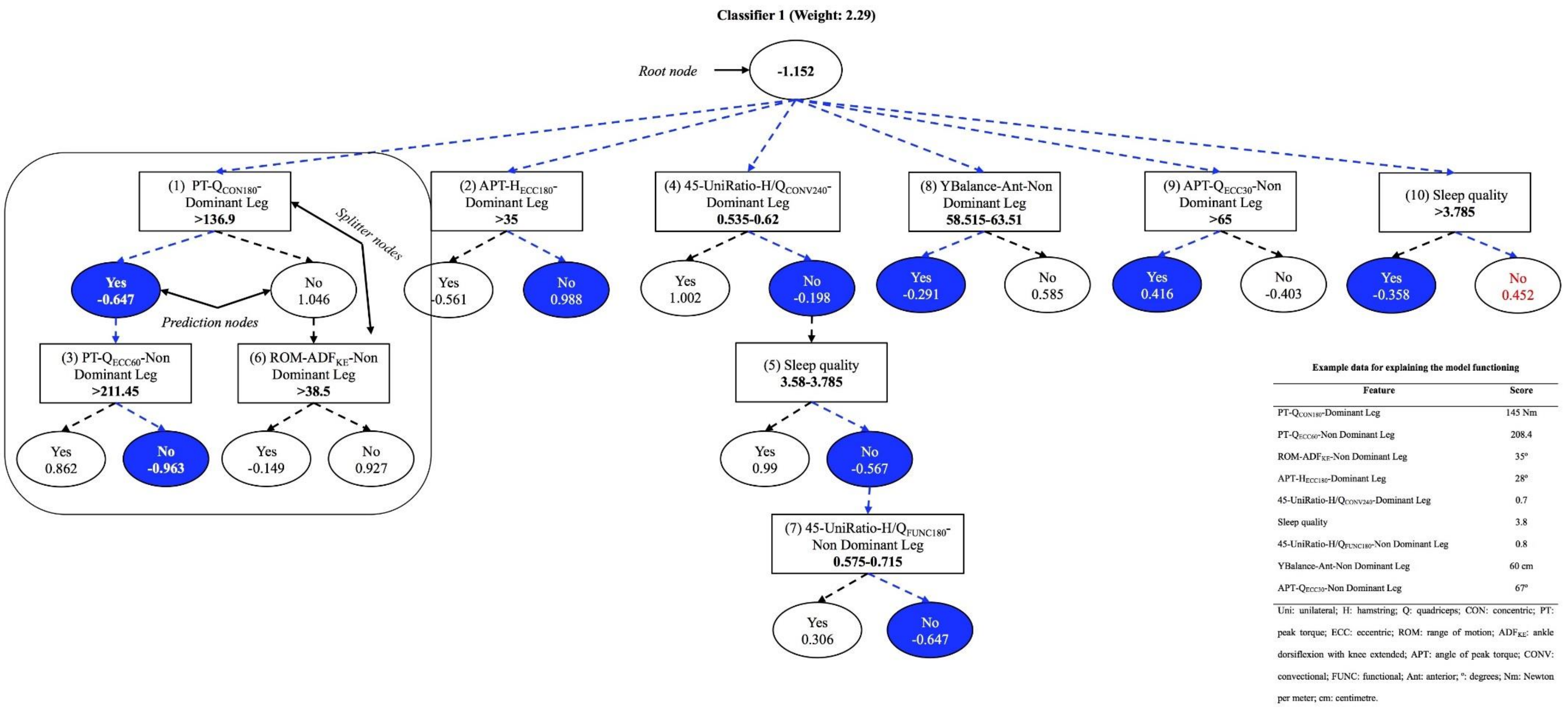
Clinical implications

In term of practical applications, each classifier has a vote or decision (yes [high risk of HSI] or no [lower risk of HSI]), and the final decision regarding whether or not a player might suffer an injury is based on the combination of the votes of each individual classifier to each class (yes or no), where the weight of each classifier's vote is a function of its accuracy.

Supplementary files 9-18 show the weight of the vote of each classifier. For example, if a player gets four Yes answers or votes in the classifiers (numbers 1, 4, 7 and 9); while the remaining answers to the other classifiers are No, then the final decision will be calculated as follow:

- Yes´ weight = 2.29 (classifier 1) + 3.8 (classifier 4) + 2.59 (classifier 7) + 2.56 (classifier 9) = 11.24
- No´s weight = 2.44 (classifier 2) + 3.49 (classifier 3) + 2.62 (classifier 5) + 2.41 (classifier 6) + 2.76 (classifier 8) + 2.65 (classifier 10) = 16.37
- Final decision = No weight > Yes weight ⇒ No (low risk of HSI)

Unlike traditional tree models the classification of instances by ADTree is not determined by a single path traversed in the tree, but rather by the additive score of a collection of paths. The ADTree is graphically represented with two types of nodes: Elliptical *prediction nodes* and rectangular *splitter nodes* (Figure 2). Each splitter node is associated with a value indicating the rule condition: If the feature represented by the node satisfied the condition for a given instance, the prediction path will go through the left child node, otherwise the path will go through the right child node. The final classification score produced by the tree is found by summing the values from all the prediction nodes reached by the instance, with the root node being the precondition of the classifier. If the summed score is greater than zero, the instance is classified as true (low risk of HSI).


Classifier 1 (Weight: 2.29)
Root node
-1.152
Splitter nodes
Prediction nodes
(1) PT-Q_{CON180}-Dominant Leg >136.9
Yes -0.647
No 1.046
(3) PT-Q_{ECC60}-Non Dominant Leg >211.45
Yes 0.862
No -0.963
(6) ROM-ADF_{KE}-Non Dominant Leg >38.5
Yes -0.149
No 0.927
(2) APT-H_{ECC180}-Dominant Leg >35
Yes -0.561
No 0.988
(4) 45-UniRatio-H/$Q_{CONV240}$-Dominant Leg 0.535-0.62
Yes 1.002
No -0.198
(5) Sleep quality 3.58-3.785
Yes 0.99
No -0.567
(7) 45-UniRatio-H/$Q_{FUNC180}$-Non Dominant Leg 0.575-0.715
Yes 0.306
No -0.647
(8) YBalance-Ant-Non Dominant Leg 58.515-63.51
Yes -0.291
No 0.585
(9) APT-Q_{ECC30}-Non Dominant Leg >65
Yes 0.416
No -0.403
(10) Sleep quality >3.785
Yes -0.358
No 0.452

Example data for explaining the model functioning

Feature	Score
PT-Q_{CON180}-Dominant Leg	145 Nm
PT-Q_{ECC60}-Non Dominant Leg	208.4
ROM-ADF_{KE}-Non Dominant Leg	35°
APT-H_{ECC180}-Dominant Leg	28°
45-UniRatio-H/$Q_{CONV240}$-Dominant Leg	0.7
Sleep quality	3.8
45-UniRatio-H/$Q_{FUNC180}$-Non Dominant Leg	0.8
YBalance-Ant-Non Dominant Leg	60 cm
APT-Q_{ECC30}-Non Dominant Leg	67°

Uni: unilateral; H: hamstring; Q: quadriceps; CON: concentric; PT: peak torque; ECC: eccentric; ROM: range of motion; ADF_{KE}: ankle dorsiflexion with knee extended; APT: angle of peak torque; CONV: convectional; FUNC: functional; Ant: anterior; °: degrees; Nm: Newton per meter; cm: centimetre.

(Figure 2.)

To better explain how coaches and sport practitioners should use the model to predict HSI, we have explained the classifier number 1 or ADTree-1 using the data displayed in figure 2, which correspond to a fictional soccer player. In addition, figure 2 represents in blue the paths followed by the selected instance or example.

Limitations

The model developed in the present study was built with the goal of allowing sport medicine practitioners to accurately identify professional soccer players at high risk of HSI during pre-season screenings. To address this issue, we used several predictors (risk factors) as well as external (oversampling) and internal (ensembles) methods and a decision tree (ADTree) as base classifier in order to build a model with moderate to good predictive accuracy. This set up allowed us to build a powerful model (AUC = 0.837; TPrate = 77.8%; TNrate = 83.8%), which was also very complex in nature. Therefore, although the model fulfils the goal for which it was built (making predictions); its complexity (10 different classifiers and 66 predictors) does not afford the opportunity to answer the question concerning why HSI happens.

Another potential limitation of the current study is the population used. The sport background of participants was professional soccer and the generalizability to other sport modalities and level of play cannot be ascertained. Likewise, the number of HSIs recorded over the follow up period may be considered a priori as small for a prospective cohort study aimed at developing a model to predict a specific type of injury. However, the large number of features recorded during the pre-season evaluation, the 18 HSIs sustained by the soccer players over the follow up period and the machine learning statistical approach applied allowed us to build a robust predictive model to identify professional male soccer players at risk of HSIs.

Finally, it should also be noted that the model is dependent on the predictors used in the training process and hence, practitioners must follow the same assessment methodologies used in the current study in order to replicate the current results to maximise the applicability to their populations.

CONCLUSIONS

To the best of our knowledge this is the first study to use a cross-validation process using data mining techniques to concurrently explore a wide range of HSI risk factors to be able to identify high risk soccer players. This technique appears to permit the identification of high risk soccer players with an AUC value of 0.837, significantly higher than previously reported. The current study reinforces that HSI is multifactorial due to the number and range of variables identified in the classifiers. This provides additional challenges for practitioners wanting to screen athletes and identify them as high or low risk due to the time restraints in real world settings**.**

Conflict of Interests

The authors declare no conflict of interest.

REFERENCES

1. Åkerstedt T, Hume, K, Minors D, Waterhouse, JIM. The subjective meaning of good sleep, an intraindividual approach using the Karolinska Sleep Diary. *Percept Mot Skills* 1994;*79*:287-296.
2. Arce C, De Francisco C, Andrade E, Seoane G, Raedeke T. Adaptation of the Athlete Burnout Questionnaire in a Spanish sample of athletes. *Span J Psychol* 2012;15:1529-1536.
3. Arnason A, Sigurdsson SB, Gudmundsson A, Holme I, Engebretsen L, Bahr R. Risk factors for injuries in football. *Am J Sports Med* 2004;32:5S-16S.
4. Ayala F, De Ste Croix M, Sainz de Baranda P, Santonja F. Absolute reliability of hamstring to quadriceps strength imbalance ratios calculated using peak torque, joint angle-specific torque and joint ROM-specific torque values. *Int J Sports Med* 2012;33:909-916.
5. Ayala F, Puerta-Callejón JM, Flores-Gallego MJ, García-Vaquero MP, Ruiz-Pérez I, Caldearon-López A, Parra-Sánchez S, López-Plaza D, López-Valenciano A. A bayesian analysis of the main risk factors for hamstring injuries. *Kronos* 2016;1-15.
6. Bahr R. Why screening tests to predict injury do not work - and probably never will…: a critical review. *Br J Sports Med* 2016;50:776-780.
7. Barbado D, Lopez-Valenciano A, Juan-Recio C, Montero-Carretero C, van Dieën JH, Vera-Garcia FJ. Trunk stability, trunk strength and sport performance level in judo. *PloS one* 2016;11:e0156267.
8. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Wadsworth & Brooks. Monterey, CA 1984.
9. Brockett CL, Morgan DL, Proske U. Predicting hamstring strain injury in elite athletes. *Med Sci Sports Exerc* 2004;36:379-387.
10. Cejudo A, Sainz de Baranda P, Ayala F, Santonja F. Normative data of lower-limb muscle flexibility in futsal players. *Rev Int Med Cienc Act Fis Deporte* 2014;14:509-525.
11. Cresswell SL, Eklund RC. The nature of player burnout in rugby: Key characteristics and attributions. *J Appl Sport Psychol* 2006;18:219-239.

12. Croisier JL, Ganteaume S, Binet J, Genty M, Ferret JM. Strength imbalances and prevention of hamstring injury in professional soccer players a prospective study. *Am J Sports Med* 2008;36:1469-1475.

13. Croisier JL, Forthomme B, Namurois MH, Vanderthommen M, Crielaard JM. Hamstring muscle strain recurrence and strength performance disorders. *Am J Sports Med* 2002;30:199-203.

14. Dauty M, Menu P, Fouasson-Chailloux A, Ferréol S, Dubois C. Prediction of hamstring injury in professional soccer players by isokinetic measurements. *Muscles Ligaments Tendons J* 2016;6:116-123.

15. Ekstrand J, Hägglund M, Waldén M. Epidemiology of muscle injuries in professional football (soccer). *Am J Sports Med* 2011;39:1226-1232.

16. Ekstrand J, Hagglund M, Waldén M. Injury incidence and injury patterns in professional football: the UEFA injury study. *Br J Sports Med* 2011;45:553-558.

17. Ekstrand J, Waldén M, Hägglund M. Hamstring injuries have increased by 4% annually in men's professional football, since 2001: a 13-year longitudinal analysis of the UEFA Elite Club injury study. *Br J Sports Med* 2016;50:731-737.

18. Elkarami B, Alkhateeb A, Rueda L. Cost-sensitive classification on class-balanced ensembles for imbalanced non-coding RNA data. *In: Proceedings of the Student Conference (ISC), 2016 IEEE EMBS International* 2016:1-4.

19. Forbes H, Bullers A. Lovell A, McNaughton LR, Polman RC, Siegler JC. Relative torque profiles of elite male youth footballers: effects of age and pubertal development. *Int J Sports Med* 2009;30(08):592-597.

20. Fousekis K, Tsepis E, Poulmedis P, Athanasopoulos S, Vagenas G. Intrinsic risk factors of non-contact quadriceps and hamstring strains in soccer: a prospective study of 100 professional players. *Br J Sports Med* 2011;45:709-714.

21. Freund Y, Mason L. The alternating decision tree learning algorithm. *In: Proceedings of the icml* 1999;99:124-133.

22. Fuller CW, Ekstrand J, Junge A, Andersen TE, Bahr R, Dvorak J, Hägglund M, McCrory P, Meeuwisse WH. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Scand J Med Sci Sports* 2006;16:83-92.

23. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 2012;42:463-484.

24. Hagglund M, Walden M, Ekstrand J. Injury incidence and distribution in elite football: a prospective study of the Danish and the Swedish top divisions. *Scand J Med Sci Sports* 2005;15:21-28.

25. Hägglund M, Waldén M, Ekstrand J. Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons. *Br J Sports Med* 2006;40:767-772.

26. Harriss DJ, Macsween A, Atkinson G. Standards for ethics in sport and exercise science research: 2018 Update. *Int J Sports Med* 2017; 38: 1126–1131.

27. Jovanovic M. Uncertainty, heuristics and injury prediction. *Aspetar Sports Med J* 2017;6:18-24.

28. López-Valenciano A, Ayala F, Puerta JM, De Ste Croix M, Vera-Garcia FJ, Hernández-Sánchez S, Ruiz-Perez I, Myer GD. A Preventive Model for Muscle Injuries: A Novel Approach based on Learning Algorithms. *Med Sci Sports Exerc* 2018; 50(5), 915-927.

29. Olmedilla A, Laguna M, Redondo AB. Injury and psychological characteristics in handball players. *Rev Andal Med Deporte* 2011;4:6-12.

30. Quinlan JR. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)* 1996;28:71-72.

31. Rossi A, Pappalardo L, Cintia P, Iaia FM, Fernàndez J, Medina D. Effective injury forecasting in soccer with GPS training data and machine learning. *PloS one* 2018; 13(7):e0201264.

32. Ruddy JD, Shield AJ, Maniar N, Williams MD, Duhig S, Timmins RG, Hickey J, Bourne MN, Opar DA. Predictive Modeling of Hamstring Strain Injuries in Elite Australian Footballers. *Med Sci Sports Exerc* 2018;50:906-914.

33. Shaffer SW, Teyhen DS, Lorenson CL, Warren RL, Koreerat, CM, Straseske CA, Childs JD. Y-balance test: a reliability study involving multiple raters. *Mil Med* 2013;178:1264-1270.

34. Sun Y, Wei S, Zhong Y, Fu W, Li L, Liu Y. How joint torques affect hamstring injury risk in sprinting swing–stance transition. *Med Sci Sports Exerc* 2015;47:373-380.

35. Taylor KL, Sheppard JM, Lee H, Plummer N. Negative effect of static stretching restored when combined with a sport specific warm-up component. *J Sci Med Sport* 2009;12:657-661.

36. Thorborg K, Petersen J, Magnusson SP, Hölmich P. Clinical assessment of hip strength using a hand-held dynamometer is reliable. *Scand J Med Sci Sports* 2010;20:493-501.

37. Timmins RG, Bourne MN, Shield AJ, Williams MD, Lorenzen C, Opar DA. Short biceps femoris fascicles and eccentric knee flexor weakness increase the risk of hamstring injury in elite football (soccer): a prospective cohort study *Br J Sports Med*. 2016;50:1524-1535.

38. van Dyk N, Bahr R, Whiteley R, Tol JL, Kumar BD, Hamilton B, Farooq A, Witvrouw E. Hamstring and quadriceps isokinetic strength deficits are weak risk factors for hamstring strain injuries: A 4-year cohort study. *Am J Sports Med* 2016;44:1789-1795.

39. Witvrouw E, Danneels L, Asselman P, D'Have T, Cambier D. Muscle flexibility as a risk factor for developing muscle injuries in male professional soccer players. *Am J Sports Med* 2003;31:41-46.

40. Witvrouw E, Mahieu N, Danneels L, McNair P. Stretching and injury prevention. *Sports Med* 2004;34:443-449.

41. Zvijac JE, Toriscelli TA, Merrick S, Kiebzak GM. Isokinetic concentric quadriceps and hamstring strength variables from the NFL Scouting Combine are not predictive of hamstring injury in first-year professional football players. *Am J Sports Med* 2013;41:1511-1518.

FIGURES LEGEND

Figure 1: Graphical representation of testing procedure. The order of the different tests used to record the personal or individual, psychological and neuromuscular risk factors in the testing session is shown.

Figure 2: Graphical representation of the first classifier. Prediction nodes are represented by ellipses and splitter nodes by rectangles. Each splitter node is associated with a real valued number indicating the rule condition, meaning: If the feature represented by the node satisfies the condition value the prediction path will go through the left child node, otherwise the path will go through the right child node. The numbers before the feature names in the prediction nodes indicate the order in which the different base rules were discovered. This ordering can to some extent indicate the relative importance of the base rules.

This classifier number 1 reports an initial score of -1.152 in its root node. Furthermore, this classifier shows a tree-shape structure comprised by six main branches whose father nodes (first leaves) are the following: a) PT-Q_{CON180}-Dominant Leg, b) APT-H_{ECC180}-Dominant Leg,

c) 45-UniRatio-H/$Q_{CONV240}$-Dominant Leg, d) YBalance-Ant-Non-Dominant Leg, e) APT-Q_{ECC30}-Non-Dominant Leg and f) Sleep quality. All the classifier´s main branches must be addressed, and the scores obtained in each branch (resulting from the data inputted in the father and child [if necessary] nodes) must be summed to the score initially reported by the root node in order to get the final vote of the classifier (yes = negative score [high risk of injury] or no = positive score [low risk of injury]) for the player.

Thus, and if we start by addressing the branch whose father node is the feature PT-Q_{CON180}-Dominant Leg, it is shown that the score reported by the soccer player (145 Nm) satisfies the condition present in the node (>136.9 Nm) and hence, he obtains the score of -0.647 from the prediction node Yes. This circumstance drives to the child node represented by the feature PT-Q_{ECC60}-Non-Dominant Leg. In this case, the player does not satisfy the condition presented in the just-mentioned feature, in other words, the value reported (208.4 Nm) is not higher than 211.45 Nm. Therefore, here the player achieves a score of -0.963 coming from the predictive node 'No'. As a consequence, the final result of this branch is the sum of -0.647 plus -0.963, ergo -1.61 points.

The pathway to follow in the branch whose father node is the feature titled APT-H_{ECC180}-Dominant Leg is shorter than the one previously described, and here the player demonstrated a score of 28º, which does not satisfy the established condition (>35º). Consequently, in this second branch, the player obtains a score of 0.988 from the predictive node "No".

The third branch, composed by the father node titled 45-UniRatio-H/$Q_{CONV240}$-Dominant Leg provides a total score of -1.412 (-0.198 +[- 0.567] + [-0.647]), as the soccer player´s values does not satisfy the condition presented in neither father nor child nodes.

For its part, in the fourth branch, the soccer player does satisfy the condition of the father node, UniRatio-H/Q_{CON60}-Dominant Leg, that provides a score of -0.291.

Finally, and for both the fifth and sixth branches, the player again does satisfy the condition presented in their respective father nodes (APT-Q_{ECC30}-Non-Dominant Leg and Sleep quality respectively) and hence, the scores obtained were 0.416 and -0.358 respectively.

All in all, and after summing the baseline score of the root node with the scores reported in each of the six branches of the classifier, a total score of -3.419 was achieved. This final score is a negative value, and this supposes a "Yes" vote with a weight of 2.29. The final classification will be based on the combination of the votes of each individual classifier to each class (yes or no).

SUPPLEMENTAL DIGITAL CONTENT (SDC)

- SDC 1**:** Description of the personal injury risk factors recorded (names and labels).
- SDC 2: Description of the psychological risk factors recorded (names and labels).
- SDC 3: Description of the dynamic postural control testing manoeuvre and measures obtained from it (names and labels).
- SDC 4: Description of the isometric hip abduction and adduction strength testing manoeuvre and list of measures obtained from it (names and labels).
- SDC 5: Description of the lower extremity joints (hip, knee and ankle) range of motion assessment tests and measures obtained from them (names and labels).
- SDC 6: Description of the trunk stability testing manoeuvre and measures obtained from it (names and labels).
- SDC 7: Description of the Isokinetic hamstring and quadriceps strength testing manoeuvre and measures obtained from it (names and labels).
- SDC 8**:** Description of the statistical analysis carries out.

  A list of algorithms (n = 68) grouped by families, the abbreviations that have been used along the experimental framework and a short description of them are displayed.
- SDC 9: First classifier.

  Graphical representation of the first classifier of the predictive model for muscle injuries.

- SDC 10: Second classifier.

  Graphical representation of the second classifier of the predictive model for muscle injuries.

- SDC 11: Third classifier.

  Graphical representation of the third classifier of the predictive model for muscle injuries.

- SDC 12: Fourth classifier.

  Graphical representation of the fourth classifier of the predictive model for muscle injuries.

- SDC 13: Fifth classifier.

  Graphical representation of the fifth classifier of the predictive model for muscle injuries.

- SDC 14: Sixth classifier.

  Graphical representation of the sixth classifier of the predictive model for muscle injuries.

- SDC 15: Seventh classifier.

  Graphical representation of the seventh classifier of the predictive model for muscle injuries.

- SDC 16: Eighth classifier.

  Graphical representation of the eighth classifier of the predictive model for muscle injuries.

- SDC 17: Ninth classifier.

  Graphical representation of the ninth classifier of the predictive model for muscle injuries.

- SDC 18: Tenth classifier.

  Graphical representation of the tenth classifier of the predictive model for muscle injuries.