A Preventive Model For Hamstring Injuries in Professional Soccer: Learning Algorithms

Authors

Francisco Ayala[1], Alejandro López-Valenciano[1], Jose Antonio Gámez Martín[2], Mark De Ste Croix[3], Francisco J. Vera-Garcia[1], Maria del Pilar García-Vaquero[1], Iñaki Ruiz-Pérez[1], Gregory D. Myer[4, 5, 6]

Affiliations

1 Department of Sport Science, Sport Research Centre, Miguel Hernández University of Elche, Elche (Alicante), Spain.
2 Escuela Superior de Ingeniería Informática, Universidad de Castilla-La Mancha, Albacete, Spain
3 School of Sport and Exercise, University of Gloucestershire, Gloucester, United Kingdom of Great Britain and Northern Ireland
4 The SPORT Center, Division of Sports Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States
5 Department of Pediatrics and Orthopaedic Surgery, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States
6 The Micheli Center for Sports Injury Prevention, Waltham, MA, United States

Key words

injury prevention, injury risk, modeling, screening, decision-making

accepted 19.12.2018

Bibliography

DOI https://doi.org/10.1055/a-0826-1955
Published online: 2019
Int J Sports Med

ISSN 0172-4622

Correspondence

Dr. Francisco Ayala, PhD
Sports Research Centre
Miguel Hernandez University of Elche
Avda. de la Universidad s/n.
03202 Elche
Spain
Tel.: + 34/620/699 291, Fax: + 34/968/278 658
Franciscoayalarodriguez@gmail.com

ABSTRACT

Hamstring strain injury (HSI) is one of the most prevalent and severe injury in professional soccer. The purpose was to analyze 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 measurements. 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.

Introduction

Hamstring strain injury (HSI) is the most prevalent noncontact injury reported in professional male soccer (football) representing 12–14 % 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 16–60 % [24].

Prior to establishing injury prevention programs, 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 first identify soccer players at high risk of HSI and then apply tailored injury prevention programs, 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. First, 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 originally to define the cut-off values) [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), resulting in 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 measurements) 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 with HSI (e. g., based on odds ratios, certain values of the p statistic [mainly $p < 0.05$]). However, based on the general agreement that the etiology 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 utilizing 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 analyze 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.

Materials and Methods

Participants

A total of 96 male professional soccer players took part in the current study. Soccer players were recruited from 4 different soccer teams that were engaged in the 1st (one team, n = 25) and 2nd B (3 teams, n = 73) Spanish National Soccer League divisions.

The exclusion criteria were: a) presence of orthopedic 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 were 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 IJSM journal [26].

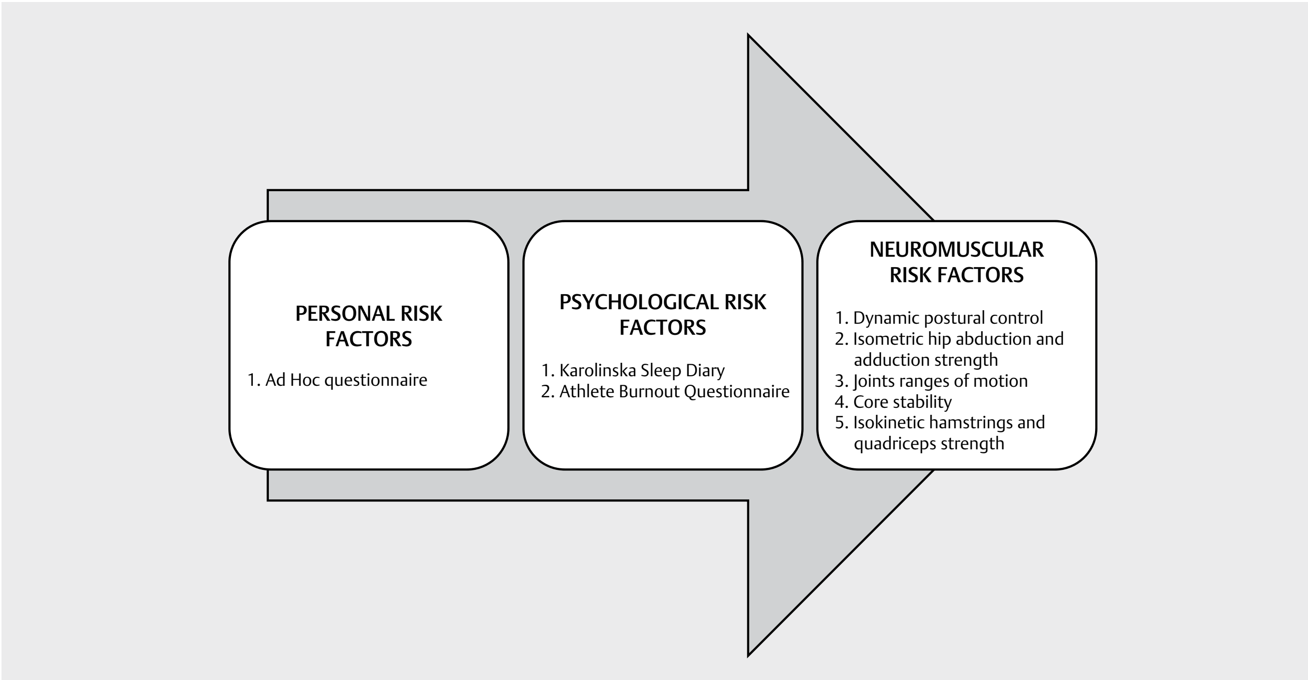
Study design

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

Players underwent a pre-season evaluation of a number of personal, psychological and neuromuscular measurements, 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 3 different parts (▶ **Fig. 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 measurements related to sleep quality and athlete burnout. Finally, the third part of the session was used to assess a number of neuromuscular measurements. A substantive number of individual, psychological and neuromuscular measurements coming from these 3 parts of the testing session were recorded (n = 229) with the aim of developing



▶ **Fig. 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.

a risk factor model that could reflect the suggested multifactorial nature of the HSI phenomenon.

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 (2 senior and 2 junior researchers, 2 technicians and 2 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 SDC1 displays a description of all the personal risk factors recorded.

Psychological risk factors

Sleep quality and athlete burnout variables were measured through 2 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 3 different dimensions that comprise athlete burnout: a) physical/emotional exhaustion; b) reduced sense of accomplishment; and c) sport devaluation. Supplementary file SDC2 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 measurements obtained from 5 different testing maneuvers: 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 maneuvers are not described below, and the reader is to refer to their original sources. Furthermore, Supplementary files SDC3–SDC7 display a description of the 5 testing maneuvers carried out and the neuromuscular risk factors recorded through each of the maneuvers.

The order of the tests was consistent for all participants (▶ **Fig. 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 maneuvers.

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-month 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 analyzing and comparing the behaviors 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 $M{:}F\rightarrow C$. Thus, the statistical analysis comprised 2 stages:

1. Data pre-processing. At this stage, the data set was prepared to apply the machine learning techniques. To optimize 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 SDC8.

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 measurements can be obtained from the SCV technique. A well-known approach to unifying these measurements and producing an evaluation criterion is to use the area under the ROC curve (AUC). In particular, the AUC corresponds to the probability of correctly identifying which one of the 2 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, 2 extra measurements from the confusion matrix were also used as evaluation criteria: a) true positive rate (TPrate): TPrate = (TP/TP + FN) also called sensitivity or recall, is the proportion of actual positives that are predicted to be positive; and b) true negative rate (TNrate): TNrate = (TN/TN + FP) or specificity, is the proportion of actual negatives that 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 hamstring 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 analyzed. 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 comprised 10 different cost sensitive ADTree classifiers (Supplementary files SDC9–SDC18). The cost matrix for cost-sensitive classifier was set to C [(0/1)/(11/0)] 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.

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 comprises 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 developed in the current study to identify athletes at high risk of HSI is higher than the model

▶ **Table 1** Average AUC, TPrate and TNrate results for all the decision tree methodologies in isolation and after applying 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			
J48	0.610	33.3	80.9
ADTree	**0.806**	**55.6**	**82.4**
Scart	0.734	72.2	75
CS-SBAG			
J48	0.584	33.3	75
ADTree	**0.846**	**66.7**	**82.4**
Scart	0.723	61.1	70.6

▶ **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

used in 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. 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 etiology 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 2 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 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 2 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 337 N 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 3 main categories of potential injury risk factors employed in the current study (psychological,

▶ **Table 4** Risk factor measurements included in the model for predicting HSI and the number of times that they appear in the classifiers; those that appear in 4 or more classifier are highlighted in bold.

Risk Factor	Nº of Classifiers
Personal measurements	
Age	1
History of HSI last season	3
Maximal level of play achieved	1
Psychological measurements	
Sleep quality	5
Physical/emotional exhaustion	2
Reduced sense of accomplishment	4
Dynamic postural control measurements	
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 measurements	
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 measurements	
ROM-PHF$_{KE}$-Dominant Leg	4
ROM-ADF$_{KE}$-Non Dominant Leg	1
ROM-PHA-Dominant Leg	1
ROM-PHA-Non Dominant Leg	1
ROM-PHER-Dominant Leg	1
Core stability measurements	
CORE-USNF	1
Isokinetic knee flexion and extension strength measurements	
PT-Q$_{CON180}$-Dominant Leg	1
PT-Q$_{CON60}$-Dominant Leg	1
PT-Q$_{CON240}$-Non Dominant Leg	2
PT-Q$_{CON180}$-Dominant Leg	1
PT-H$_{CON300}$-Non Dominant Leg	1
PT-H$_{CON300}$-Dominant Leg	2
PT-H$_{CON240}$-Non Dominant Leg	1
PT-Q$_{ECC60}$-Non Dominant Leg	2
PT-H$_{ECC60}$-Non Dominant Leg	1
PT-H$_{ECC300}$-Dominant Leg	1
PT-H$_{ECC180}$-Non Dominant Leg	1
APT-H$_{CON}$- Dominant Leg	1
APT-H$_{ECC180}$-Dominant Leg	4
APT-H$_{ECC60}$-Dominant Leg	1
APT-H$_{ECC60}$-Non Dominant Leg	2
APT-Q$_{CON240}$-Non Dominant Leg	1
APT-Q$_{ECC30}$-Non Dominant Leg	3
APT-Q$_{ECC60}$-Dominant Leg	2
APT-Q$_{ECC60°/s}$-Non Dominant Leg	1
15-T-Q$_{ECC60}$-Dominant Leg	1
15-T-Q$_{ECC30}$-Non Dominant Leg	1
15-T-H$_{ECC60}$-Non Dominant Leg	1
15-T-H$_{ECC180}$-Non Dominant Leg	1
30-T-Q$_{ECC180}$-Non Dominant Leg	1
30-T-H$_{ECC30}$-Dominant Leg	2
45-T-Q$_{ECC180}$- Non Dominant Leg	1
45-T-H$_{ECC60}$-Dominant Leg	1
45-T-H$_{ECC180}$- Dominant Leg	1
UniRatio H/Q$_{CONV300}$-Non Dominant Leg	1
UniRatio H/Q$_{FUN60}$-Dominant Leg	1
15-UniRatio-H$_{30}$/Q$_{240}$-Non Dominant Leg	1
15-UniRatio H/Q$_{FUN180}$-Dominant Leg	3
15-UniRatio H/Q$_{CONv60}$-Dominant Leg	1
15-UniRatio H/Q$_{CONV240}$- Dominant Leg	1
15-UnilRatio H/Q$_{FUNC180}$-Non Dominant Leg	1
30-UniRatioH/Q$_{FUNC60}$-Dominant Leg	1
30-UniRatio-H/Q$_{CON180}$-Dominant Leg	1
45-UniRatioH/Q$_{FUNC60}$-Non Dominant Leg	1
45-UniRatio-H/Q$_{FUNC180}$-Non Dominant Leg	1
45-UniRatio-H/Q$_{CONV240}$-Dominant Leg	1
45-UniRatio-H/Q$_{CONV300}$-Non-Dominant Leg	2
45-UniRatio H/Q$_{CONV300}$-Dominant Leg	2
BilaRatio-Q$_{CON240}$	1
BilaRatio-H$_{CON180}$	1
BilaRatio-H$_{CON240}$	1

HSI: hamstring strain injury; Bila: bilateral; ISOM: Isometric; Add: adduction; Abd: abduction; ROM: range of motion; ADF: ankle dorsi-flexion; Q: quadriceps; H: hamstring; HF: hip flexion; HER: hip external rotation; Ant: anterior; Post: posterior; APT: angle of peak torque; ECC: eccentric; CON: concentric; PT: peak torque; T: torque; FUNC: functional; CONV: conventional; USNF: unstable sitting without feedback.

personal and neuromuscular) all have some representation in the final model selected and hence, this reinforces the idea that the etiology of HSI is multifactorial.

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 analyzed whether burnout and sleep quality measurements 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 3 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 4 out of 10 classifiers. This finding is in concordance with the results found by previous studies where hip flexion ROM (considered as an indirect measure of hamstring muscle 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 stabilize 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

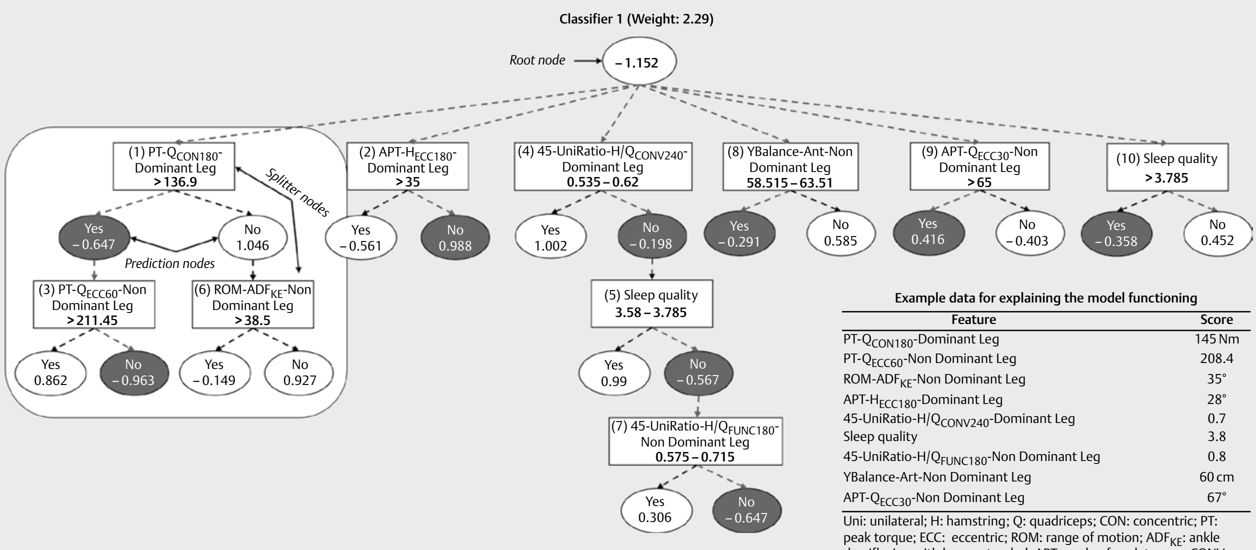
In term of practical applications, each classifier has a vote or decision (yes [high risk of HSI] or no [lower risk of HSI]), and the final decision regarding whether or not a player might suffer an injury is based on the combination of the votes of each individual classifier to each class (yes or no),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

Supplementary files SDC9–SDC18 show the weight of the vote of each classifier. For example, if a player gets 4 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 2 types of nodes: Elliptical prediction nodes and rectangular splitter nodes (▶ **Fig. 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).

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 ▶ **Fig. 2**, which correspond to a fictional soccer player. In addition, ▶ **Fig. 2** represents in grey the paths followed by the selected instance or example.



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-Art-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.

▶ **Fig. 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 comprising 6 main branches whose father nodes (first leaves) are the following: a) PT-QCON180-Dominant Leg, b) APTHECC180-Dominant Leg, c) 45-UniRatio-H/QCONV240-Dominant Leg, d) YBalance-Ant-Non-Dominant Leg, e) APT-QECC30-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 input 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-QCON180-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-QECC60-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-HECC180-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/QCONV240-Dominant Leg provides a total score of − 1.412 (− 0.198 + [− 0.567] + [− 0.647]), as the soccer player's values do not satisfy the condition presented in either father or child nodes. For its part, in the fourth branch, the soccer player does satisfy the condition of the father node, UniRatio-H/QCON60-Dominant Leg, that provides a score of − 0.291. Finally, and for both the fifth and sixth branches, the player again satisfies the condition presented in their respective father nodes (APT-QECC30-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 up the baseline score of the root node with the scores reported in each of the 6 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).

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 maximize 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.

Acknowledgements

There was no financial support for this project. Iñaki Ruiz-Pérez were supported by predoctoral grant given by Ministerio de Economía y Competitividad (FPI BES-2015-07200) from Spain. Francisco Ayala were supported by postdoctoral grant given by Seneca Foundation (postdoctoral fellowships funded by the regional sub program focuses on the postdoctoral development, 20366/PD/17) from Spain.

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; 15(1)

[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: 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 Trans Syst Man Cybern Part C Appl Rev 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 Medicine Journal 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: 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: 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

Supplemental Digital Content (SDC)

SDC 1: Description of the personal injury risk factors recorded.

Name	Labels
Player position	Goalkeeper, defender, midfielder or striker
Current level of play	1st division or 2nd B division
Dominant leg	Right, left or two-footed
Age	Sub21, sub23, senior [23–30 y] or veteran [>30 y]
Body mass (kg)	<71.75, 71.75–77.7 or >77.7
Stature (cm)	<1.775, 1.775–1.825 or >1.825
History of HSI last season	Yes or no
HSI: hamstring strain injury; y: years	

SDC 2: Description of the psychological risk factors recorded.

Name	Labels
Sleep quality	<3.58, 3.58–3.785 or >3.785
Athlete Burnout Questionnaire	
a) Physical/emotional exhaustion	<1.9, 1.9–2.155 or ≥2.155
b) Reduced sense of accomplishment	<2.67, 2.67–2.9 or >2.9
c) Sport devaluation	<1.1, 1.1–1.49 or >1.49

SDC 3: Description of the dynamic postural control testing maneuver and measurements obtained from it

Dynamic postural control

Dynamic postural control was evaluated using the Y-Balance device® and following the guidelines described by Shaffer et al. [1]. The distance reached in each direction (anterior, posteromedial and posterolateral) was normalized by dividing by the previously measured leg length to standardize the maximum reach distance ([excursion distance/leg length] × 100 = % maximum reach distance). The bilateral ratio (dominant / non-dominant score) of each direction was also calculated. Finally, to obtain a global measure of the balance test for each leg, data from each direction were averaged to calculate a composite score.

SDC 3 Measurements obtained from the dynamic postural control test.

Name	Labels	
	Dominant Leg	**Non-Dominant Leg**
YBalance-Anterior	<57.825, 57.825–63.035 or >63.035	<58.515, 58.515–63.51 or >63.51
YBalance-PosteroMedial	<101.215, 101.215–107.865 or >107.865	<102.42, 102.42–108.49 or >108.49
YBalance-PosteroLateral	<96.395, 96.395–104.93 or >104.93	<96.19, 96.19–103.71 or >103.71
BilaRatio-YBalance-Anterior	<0.965, 0.965–1.015 or >1.015	
BilaRatio-YBalance-PosteroMedial	<0.975, 0.975–1.005 or >1.005	
BilaRatio-YBalance-PosteroLateral	<0.985, 0.985–1.035 or >1.035	
YBalance-Composite	<85.44, 85.44–91.71 or >91.71	<86.73, 86.73–91.4 or >91.4-
Bila: bilateral		

Reference

1. 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.

SDC 4: Description of the isometric hip abduction and adduction strength testing maneuver and list of measures obtained from it

Isometric hip abduction and adduction strength test

Isometric hip abduction and adduction peak torques of the dominant and non-dominant limb were assessed with a portable hand-held dynamometer (Nicholas Manual Muscle Tester, Lafayette Indiana Instruments) in a supine lying position on a plinth with the participant's legs extended and following the methodology described by Thorborg et al. [1]. Briefly, participants performed 5 trials of 5-second isometric maximal voluntary contraction for each hip movement. The mean of the 3 most closely related trials were used for the subsequent statistical analyses. Unilateral hip abductor/adductor peak torque ratio defined as the hip adductor peak torque divided by hip abductor peak torque was calculated for each leg. Furthermore, the hip abduction and adduction bilateral ratios were also determined as the quotient of the dominant hip mean isometric peak value by the non-dominant hip mean isometric peak value. A side-to-side difference higher than 10 % was defined as bilateral asymmetry.

SDC 4 Measures obtained from the isometric hip abduction and adduction strength test.

Name	Labels	
	Dominant Leg	**Non-Dominant Leg**
PT_{ISOM}-HipAbd	< 190.64, 190.64–217.625 or > 217.625	< 194.025, 194.025–222 or > 222
PT_{ISOM}-HipAbd-Normalized	< 2.555, 2.555–2.91 or > 2.91	< 2.655, 2.655–2.92 or > 2.92
PT_{ISOM}-HipAdd	< 191.575, 191.575–219.625 or > 219.625	< 187.75, 187.75–215.5 or > 215.5
PT_{ISOM}-HipAdd-Normalized	< 2.635, 2.635–2.965 or > 2.965	< 2.555, 2.555–2.905 or > 2.905
UnRatio-ISOM-HipAbd/HipAdd	< 0.956, 0.956–1.095 or > 1.095	< 0.92, 0.92–1.015 or > 1.015
BilaRatio-PT_{ISOM}-HipAbd	No Asymmetry (< 10 %) or Asymmetry (≥ 10 %)	
BilaRatio-PT_{ISOM}-HipAdd	No Asymmetry (< 10 %) or Asymmetry (≥ 10 %)	

Bila: bilateral; Uni: unilateral; ISOM: isometric; PT: peak torque; Abd: abduction; Add: adduction.

Reference

1. 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

SDC5: Description of the lower extremity joints (hip, knee and ankle) range of motion assessment tests and measures obtained from them

Lower extremity joints range of motion assessment tests

The passive hip flexion with knee flexed and extended, extension, abduction, external and internal rotation; knee flexion; and ankle dorsiflexion with knee flexed and extended ROMs of the dominant and non-dominant legs were assessed following the methodology previously described [1]. Furthermore, for each joint ROM measure, side-to-side differences were also calculated. In this sense, when side-to-side difference > 6º was found, players were categorized as showing bilateral asymmetries whereas scores ≤ 6º were accepted as normal (non-bilateral asymmetries) [2].

SDC 5 Measures obtained from the lower extremity range of motion assessment tests.

Name	Labels	
	Dominant Leg	**Non-Dominant Leg**
ROM-PHF_{KF}	< 144.5, 144.5–151.5 or > 151.5	< 144.5, 144.5–152.5 or > 152.5
ROM-PHF_{KE}	< 77.5, 77.5–82.9 or > 82.9	< 78.5, 78.5–84.5 or > 84.5
ROM-PHE	< 7.5, 7.5–12.5 or > 12.5	< 9.25, 9.25–13.5 or > 13.5
ROM-PHABD	< 61.5, 61.5–68.5 or > 68.5	< 58.5, 58.5–66.5 or > 66.5
ROM-PHIR	< 44.5, 44.5–50.5 or > 50.5	< 42.5, 42.5–48.5 or > 48.5
ROM-PHER	< 47.5, 47.5–52.5 or > 52.5	< 46.5, 46.5–55.5 or > 55.5
ROM-PKF	< 121.5, 121.5–132 or > 132	< 120.5, 120.5–130.5 or > 130.5
ROM-$PAKDF_{KE}$	< 34.25, 34.25–39.5 or > 39.5	< 35.25, 35.25–38.5 or > 38.5
ROM-$PAKDF_{KF}$	< 35.5, 35.5–40.5 or > 40.5	< 36.75, 36.75–39.75 or > 39.75
BilaRatio- ROM-PHF_{KF}	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	
BilaRatio- ROM-PHF_{KE}	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	
BilaRatio- ROM-PHE	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	
BilaRatio- ROM-ABD	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	
BilaRatio- ROM-PHIR	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	
BilaRatio- ROM-PHER	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	
BilaRatio- ROM-PKF	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	
BilaRatio- ROM-$AKDF_{KE}$	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	
BilaRatio- ROM-$AKDF_{KF}$	No Asymmetry (≤ 6º) or Asymmetry (> 6º)	

PROM: passive range of motion; HF_{KF}: hip flexion with the knee flexed; HF_{KE}: hip flexion with the knee extended; HE: Hip extension; HABD: hip abduction at 90º of hip flexion; HIR: hip internal rotation; HER: hip external rotation; KF: knee flexion; $AKDF_{KE}$: ankle dorsi-flexion with the knee extended; $AKDF_{KF}$: ankle dorsi-flexion with the knee flexed; Bila: bilateral.

References

1. 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

2. 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

SDC 6: Description of the trunk stability testing maneuver and measurements obtained from it (names and labels)

Trunk stability

The unstable sitting protocol described by Barbado et al. [1] was used to assess participant's ability to control trunk posture and motion while sitting. Briefly, after a familiarization / practice period (2 min), participants performed different static and dynamic tasks while sitting on an unstable seat:

- One static stability task without visual feedback (test 1) and another with visual feedback (test 2). In test 1 participants were asked to sit still in their preferred seated position on the unstable seat, while in test 2 participants were requested to adjust their center of pressure position to a target point located in the center of a screen placed in front of them.
- Three dynamic stability tasks with visual feedback, in which participants were asked to track the target point, which moved along 3 possible trajectories (anterior-posterior, medial-lateral and circular).

All tasks were performed twice. The duration of each trial was 70 s and the rest period between trials was 1 min. Participants performed each trial with arms crossed over the chest. All participants were able to maintain the sitting position without grasping a support rail.

The mean radial error was used as a global measure to quantify the trunk performance during the trials. This variable was calculated as the mean of vector distance magnitude of the center of pressure from the target point trials (trials with visual feedback) or from the participant's own mean center of pressure position (trials without visual feedback).

► **SDC 6** Measurements obtained from the Trunk stability test.

Name	Labels
USNF	<5.125, 5.125–6.46 or>6.46
USWF	<4.74, 4.74–5.72 or>5.72
USML	7.345, 7.345–8.925 or>8.925
USAP	<7.445, 7.445–8.87 or>8.87
USCD	<9.47, 9.47–11.185 or>11.185
GLOBAL	<6.88, 6.88–8.24 or>8.24

USNF: unstable sitting without feedback; USWF: unstable sitting with feedback; USML: unstable sitting while performing medial-lateral displacements with feedback; USAP: unstable sitting while performing anterior-posterior displacements with feedback; USCD: unstable sitting while performing circular displacements with feedback.

References

1. 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

SDC 7: Description of the Isokinetic hamstring and quadriceps strength testing manoeuvre and measures obtained from it (names and labels)

Isokinetic hamstring and quadriceps strength assessment

A Biodex System-4 isokinetic dynamometer (Biodex Corp., Shirley, NY, USA) and its respective manufacture software were used to determine isokinetic concentric and eccentric torques during knee extension and flexion actions in both limbs following the methodology described by Ayala et al. [1,2].

The dynamometer was calibrated according to the manufacturer's instructions before the start of each test session. In each testing session only the dominant leg, determined through interview and defined as the leg preferred when kicking a ball, was tested.

Participants were secured in a supine position on the dynamometer with the hip passively flexed at 10–20º and the body head was maintained at 0º of flexion. The axis of rotation of the dynamometer lever arm was aligned with the lateral epicondyle of the knee. The force pad was placed approximately 3 cm superior to the medial malleolus with the foot in a relaxed position. Adjustable strapping across the pelvic, posterior thigh proximal to the knee and foot localised the action of the musculature involved. The range of movement was set from 0° (0º was determined as maximal voluntary knee extension for each participant) to 90° knee flexion. During the isokinetic testing procedure, the cushion setting on the control panel for the ends of the range of motion was set to its lowest (hardest) setting in order to reduce the effect of limb deceleration on the reciprocal motion.

The isokinetic examination was separated into two parts. The first part of the examination was the assessment of the hamstrings and quadriceps muscles during concentric/concentric (CON/CON) cycles with quadriceps undertaken first. After a 5 min rest period the eccentric/eccentric (ECC/ECC) testing cycle was performed. In both testing methods, two cycles of knee flexions and extensions were performed at 4 pre-set constant angular velocities in the following order: 60, 180, 240 and 300º/s (slow to fast). The passive eccentric mode was chosen so that the full range of movement would be completed for every action, which is important for the calculation of H/Q ratios using joint angle-specific torque values. Furthermore, this study employed continuous CON/CON and ECC/ECC cycles because they may have made the movement easier to understand and perform compared to CON/ECC cycles. The two testing parts (CON/CON and ECC/ECC) were separated by a 5 min rest interval and a rest of 30 s was allowed between action cycles.

For both CON/CON and ECC/ECC cycles, participants were encouraged to push/resist as hard and as fast as possible and to complete the full range of motion. Participants were told to abort the test if they felt any discomfort or pain. During the test, all participants were given visual feedback from the system monitor. They were also verbally encouraged by the investigator to give their maximal effort, and the instructions were standardized by using key words such as "resist" and "hard and fast as possible".

Four different torque values (peak torque [PT] and 3 joint angle-specific torque values (15º, 30º and 45º) and the joint angle of peak torque (APT) were extracted for each movement (flexion and extension), muscle action (concentric, eccentric) and velocity (60, 180, 240 and 300º/s for concentric actions and 30, 60 and 180º/s for eccentric actions). In each of the 3 trials at each velocity, the PT and APT were reported as the single highest torque output and corresponding joint angle. For each isokinetic variable, the average of the 3 sets at each velocity was used for subsequent statistical analysis. When a variation > 5 % was found in the PT, angle-specific toque and APT values between the 3 trials, the mean of the two most closely related torque values were used for the subsequent statistical analyses.

Reciprocal (conventional and functional) hamstrings to quadriceps ratios as well as bilateral hamstrings and quadriceps ratios were also calculated using peak torque and joint angle-specific torque values extracted for each velocity.

Thus, the conventional hamstrings to quadriceps ratios were calculated as the ratio between the torque values produced concentrically by hamstrings and quadriceps muscles during the isokinetic tests. Functional hamstrings to quadriceps ratios were calculated as the ratio between the toque values produced eccentrically by hamstrings muscles and concentrically by the quadriceps muscles. Bilateral hamstrings and cuadriceps ratios were calculated dividing the PT value of the dominant limb by the PT value of the non-dominant leg.

Finally, the functional knee flexion to knee extension ratio proposed by Croisier et al. [3] was also calculated as the ratio between the torques (peak and angle-specific values) values produced eccentrically by the hamstrings at 30º/s and concentrically by the quadriceps muscles at 240º/s.

SDC 7 Description of the measures obtained from the isokinetic hamstring and quadriceps strength assessment.

Measure	Labels	
	Dominant Leg	Non-Dominant Leg
Concentric Muscle Actions		
PT-Q_{60}	< 172.6, 172.6–198.25 or > 198.25	< 161.1, 161.1–188.65 or > 188.65
PT-H_{60}	< 78.5, 78.5–98.2 or > 98.2	< 72.9, 72.9–89.1 or > 89.1
PT-Q_{180}	< 115.75, 115.75–136.9 or > 136.9	< 115.7, 115.7–136.3 or > 136.3
PT-H_{180}	< 62.8, 62.8–79 or > 79	< 62.75, 62.75–76.25 or > 76.25
PT-Q_{240}	< 102.8, 102.8–125.85 or > 125.85	< 100.4, 100.4–121.45 or > 121.45
PT-H_{240}	< 60.2, 60.2–74.8 or > 74.8	< 59.2, 59.2–71.85 or > 71.85
PT-Q_{300}	< 96.35, 96.35–113.2 or > 113.2	< 89.8, 89.8–109.85 or > 109.85
PT-H_{300}	< 57.05, 57.05–71.35 or > 71.35	< 52.85, 52.85–63.65 or > 63.65
APT-Q	< 45, 45–60 or > 60	
APT-H	< 25, 25–35 or > 35	
Eccentric Muscle Actions		
PT-H_{30}	< 77, 77–101.25 or > 101.25	< 72.15, 72.15–86.9 or > 86.9
PT-Q_{30}	< 171.95, 171.95–221.6 or > 221.6	< 160, 160–207.75 or > 207.75
15-T-H_{30}	< 59.75, 59.75–89.4 or > 89.4	< 55.05, 55.05–77.15 or > 77.15
15-T-Q_{30}	< 28.15, 28.15–48 or > 48	< 28.7, 28.7–46.15 or > 46.15
30-T-H_{30}	< 65.8, 65.8–82.8 or > 82.8	< 59.9, 59.9–76.2 or > 76.2
30-T-Q_{30}	< 82.35, 82.35–110.15 or > 110.15	< 73.8, 73.8–100.15 or > 100.15
45-T-H_{30}	< 61.35, 61.35–80 or > 80	< 56.2, 56.2–69.85 or > 69.85
45-T-Q_{30}	< 127.3, 127.3–159.5 or > 159.5	< 114.05, 114.05–149.05 or > 149.05
PT-H_{60}	< 78.65, 78.65–101.9 or > 101.9	< 69.3, 69.3–88.7 or > 88.7
PT-Q_{60}	< 180.45, 180.45–230.35 or > 230.35	< 164.4, 164.4–211.45 or > 211.45
15-T-H_{60}	< 66.85, 66.85–85.5 or > 85.5	< 56.3, 56.3–79.65 or > 79.65
15-T-Q_{60}	< 32.9, 32.9–44.4 or > 44.4	< 28.95, 28.95–44.5 or > 44.5
30-T-H_{60}	< 67.95, 67.95–87.75 or > 87.75	< 60.25, 60.25–78.15 or > 78.15
30-T-Q_{60}	< 76.8, 76.8–100 or > 100	< 74.2, 74.2–102.48 or > 102.48
45-T-H_{60}	< 63.95, 63.95–80.25 or > 80.25	< 59.45, 59.45–74.05 or > 74.05
45-T-Q_{60}	< 120.65, 120.65–159.05 or > 159.05	< 119.65, 119.65–148.8 or > 148.8
PT-H_{180}	< 76.25, 76.25–98.7 or > 98.7	< 71.6, 71.6–90 or > 90
PT-Q_{180}	< 163.3, 163.3–201.35 or > 201.35	< 163.15, 163.15–194.3 or > 194.3
15-T-H_{180}	< 47.35, 47.35–72.9 or > 72.9	< 51.75, 51.75–75.9 or > 75.9
15-T-Q_{180}	< 41.5, 41.5–53.2 or > 53.2	< 38.65, 38.65–53.95 or > 53.95
30-T-H_{180}	< 68.15, 68.15–85.35 or > 85.35	< 60.05, 60.05–83.45 or > 83.45
30-T-Q_{180}	< 97.1, 97.1–117.8 or > 117.8	< 82.4, 82.4–114.4 or > 114.4
45-T-H_{180}	< 74.05, 74.05–89.1 or > 89.1	< 66.85, 66.85–81.3 or > 81.3
45-T-Q_{180}	< 144.3, 144.3–168 or > 168	< 131.5, 131.5–167.35 or > 167.35

SDC 7 Continued

Measure	Labels	
	Dominant Leg	Non-Dominant Leg
APT-H	< 25, 25–35 or > 35	
APT-Q	< 50, 50–65 or > 65	
Unilateral Conventional Ratios		
H/Q_{CONV60}	< 0.47, 0.47–0.60 or > 0.60	
$H/Q_{CONV180}$	≤ 0.60 or > 0.60	
$H/Q_{CONV240}$	≤ 0.60 or > 0.60	
$H/Q_{CONV300}$	< 0.6, 0.6–0.8 or > 0.8	
Angle-Specific Unilateral Conventional Ratios		
15-H/Q_{CONV60}	< 0.93, 0.93–1.165 or > 1.165	< 0.915, 0.915–1.17 or > 1.17
15-$H/Q_{CONV180}$	< 1.06, 1.06–1.425 or > 1.425	< 1.075, 1.075–1.505 or > 1.505
15-$H/Q_{CONV240}$	< 0.8, 0.8–1.175 or > 1.175	< 0.75, 0.75–1.065 or > 1.065
15-$H/Q_{CONV300}$	< 0.54, 0.54–0.885 or > 0.885	< 0.565, 0.565–0.885 or > 0.885
30-H/Q_{CONV60}	< 0.645, 0.645–0.76 or > 0.76	< 0.625, 0.625–0.735 or > 0.735
30-$H/Q_{CONV180}$	< 0.695, 0.695–0.835 or > 0.835	< 0.66, 0.66–0.82 or > 0.82
30-$H/Q_{CONV240}$	< 0.665, 0.665–0.785 or > 0.785	< 0.645, 0.645–0.755 or > 0.755
30-$H/Q_{CONV300}$	< 0.835, 0.835–1.085 or > 1.085	< 0.87, 0.87–1.075 or > 1.075
45-H/Q_{CONV60}	< 0.435, 0.435–0.515 or > 0.515	< 0.425, 0.425–0.515 or > 0.515
45-$H/Q_{CONV180}$	< 0.505, 0.505–0.595 or > 0.595	< 0.495, 0.495–0.585 or > 0.585
45-$H/Q_{CONV240}$	< 0.535, 0.535–0.62 or > 0.62	< 0.515, 0.515–0.615 or > 0.615
45-$H/Q_{CONV300}$	< 0.545, 0.545–0.645 or > 0.645	< 0.515, 0.515–0.61 or > 0.61
Unilateral Functional Ratios		
H/Q_{FUNC60}	< 0.6, 0.6–0.7 or > 0.7	
$H/Q_{FUNC180}$	≤ 0.80 or > 0.80	
H_{30}/Q_{240}	< 0.8, 0.8–1.0 or > 1.0	
Angle-Specific Unilateral Functional Ratios		
15-H/Q_{FUNC60}	< 0.915, 0.915–1.175 or > 1.175	< 0.875, 0.875–1.12 or > 1.12
15-$H/Q_{FUNC180}$	< 0.8, 0.8–1.315 or > 1.315	< 0.985, 0.985–1.32 or > 1.32
15-H_{30}/Q_{240}	< 1.42, 1.42–1.785 or > 1.785	< 1.18, 1.18–1.63 or > 1.63
30-H/Q_{FUNC60}	< 0.605, 0.605–0.735 or > 0.735	< 0.545, 0.545–0.695 or > 0.695
30-$H/Q_{FUNC180}$	< 0.755, 0.755–0.945 or > 0.945	< 0.715, 0.715–0.865 or > 0.865
30-H_{30}/Q_{240}	< 0.875, 0.875–1.05 or > 1.05	< 0.765, 0.765–0.965 or > 0.965
45-H/Q_{FUNC60}	< 0.435 ,0.435–0.525 or > 0.525	< 0.415, 0.415–0.5 or > 0.5
45-$H/Q_{FUNC180}$	< 0.665, 0.665–0.76 or > 0.76	< 0.575, 0.575–0.715 or > 0.715
45-H_{30}/Q_{240}	< 0.635, 0.635–0.775 or > 0.775	< 0.585, 0.585–0.71 or > 0.71
Bilateral Ratios		
H/H_{CON60}	No Asymmetry or Asymmetry	
H/H_{CON180}	No Asymmetry or Asymmetry	
H/H_{CON240}	No Asymmetry or Asymmetry	
Q/Q_{CON60}	No Asymmetry or Asymmetry	
Q/Q_{CON180}	No Asymmetry or Asymmetry	
Q/Q_{CON240}	No Asymmetry or Asymmetry	
H/H_{ECC60}	No Asymmetry or Asymmetry	
H/H_{ECC180}	No Asymmetry or Asymmetry	

PT: peak torque; H: hamstring; Q: quadriceps; CON: concentric; ECC: eccentric; APT: angle of peak torque.

References

1. 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.

2. 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.

3. 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.

SDC 8: Description of the statistical analysis carried 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.

Data pre-processing.

To optimize the performance of the different learning algorithms used in the data processing stage, standard pre-processing methods such as data cleaning and data discretization were applied.

First, those players who did not complete all the neuromuscular tests for any reason (6 players) were removed. This exclusion criterion was based on the fact that if a player had not completed a neuromuscular test a large number of features would be absent and this might have a negative impact on the performance of the models generated. Furthermore, 4 players were also removed because they left their respective teams before the follow-up procedure was completed. Second, an investigation regarding the presence of outliers was carried out using boxplots and the detected outliers were removed. The third step consisted of looking for missing data. To address this issue, frequency tables and diagrams were built. Thus, missing data were replaced by the mean value of the corresponding feature of the specific level of play (1st or 2nd B divisions) of the players. For example, if a 1st division player did not report his height for any reason, then the average value of his counterpart 1st division players was inputted. It should be noted that none of the features reported a percentage of missing data and/or outliers higher than 5 %. The SPSS Statistical software (V21.0) was used to carry out these data cleaning processes.

After having applied the above-mentioned data cleaning methods, an imbalance (showing an imbalance ratio of 0.26) and high dimensional data set comprised of 86 soccer players (instances) and 229 potential risk factors (features) was created.

The final step comprised the discretization of the continuous features as this has been shown to be an effective measure to improve the performance of several classifiers [4]. Thus, continuous features were discretized applying the unsupervised discretization algorithm available in the well-known Weka (Waikato Environment for Knowledge Analysis) Data Mining software and using the equal frequency binning approach (3 intervals). We selected 3 intervals in order to reflect taxonomy of low, moderate and high scores that might make the final models more comprehensible. In those features where the graphical representation of the data allowed the authors to suggest alternative cut-off values, a comparative analysis was run in order to identify the discretization approaches (algorithm vs. authors visual inspection) that displayed the best predictive ability. The approach reporting the better predictive results was used for the discretization of each feature. Consequently, lower extremity ROM and isokinetic angle of peak torque (APT) features, as well as both the reciprocal knee flexion to knee extension ratios and bilateral knee flexion and extension ratios were discretized using the graphical representation of the data as a guide; whereas the remaining features were discretized using the Weka unsupervised discretization algorithm (Supplementary files SDC1–SDC7).

Data processing

Part of the taxonomies for external (oversampling) and internal (ensembles) methods for learning with imbalanced data sets proposed by Elkarami et al. [5] and Galar et al. [7] were used to build models for predicting HSI in professional soccer players. Thereby, the algorithms of each of the above mentioned families (oversampling and ensembles) that showed the best goodness scores in the latter mentioned studies were used to train models. The model with the highest validity metrics was considered the best for predicting HSI based on the current data set.

To achieve founded conclusions, 3 decision tree algorithms were selected to be used in the oversampling and ensemble methodologies as base classifiers: J48, which is an algorithm for generating a pruned or unpruned C4.5 decision tree [8]; ADTree, which is an alternating decision tree [6]; and SimpleCart, which implements minimal cost-complexity pruning. Hence a decision tree is a set of conditions organized in a hierarchical structure [1]. An instance is classified by following the path of satisfied conditions from the root of the tree until a leaf is reached, which will correspond with a class label.

All the decision trees selected were made cost sensitive to minimize the cost of misclassification of the minority class by using the filter cost sensitive classifier algorithm available in Weka workbench. Thus, the training data were reweighted according to the costs assigned to each class. The set up of the definitive cox matrix was based on the best performance reported after testing all the possibilities. For the sake of brevity and the lack of space, the codes of the algorithms used in this study are not presented. Instead, only the names of the algorithms have been specified and the reader is referred to the original sources. Furthermore, all the classification algorithms used are available in the Weka Data Mining software.

Although there are several data oversampling methods, we used one of the most popular methodologies that is the classic synthetic minority oversampling technique (SMOTE) [2]. The main concept behind SMOTE is to create new minority class examples by interpolating several minority class instances that lie together for oversampling the training set. With this technique, the positive class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Three different levels of balance in the training data were analyzed (25:75; 40:60; 50:50) and the best in term of predictive ability was reported. Additionally, the interpolations that are computed to generate new synthetic data were made considering the 5 nearest neighbors of minority class instances using the Euclidean distance.

Regarding ensemble learning algorithms, the algorithm families designed to deal with skewed class distributions in data sets were included: Boosting-based and Bagging-based. The Boosting-based ensembles that were considered in the current study were SMOTEBoostM1 [3] and RUSBoost [9]. With respect to Bagging-based ensembles, it was included from the OverBagging group, OverBagging (which uses random oversampling) and SMOTEBagging [10].

Finally, the behavior of some specific combination of class-balanced ensembles with cost-sensitive base classifiers was also studied. The final cox matrix set up was based on the best performance reported after testing all the possibilities.

The following table summarizes the list of algorithms grouped by families and also shows the abbreviations that have been used along the experimental framework and a short description of them.

▶ **SDC 8** Algorithms used in the data processing phase.

Cost-sensitive base classifiers		
Abbr.	*Method*	*Short Description*
J48	J48	Algorithm for generating a pruned or unpruned C4.5 decision tree
SCart	SimpleCart	Algorithm for implementing minimal cost-complexity pruning
ADTree	ADTree	Alternating decision tree
Resampling techniques		
Abbr.	*Method*	*Short Description*
CS-SMT	SMOTE	Each cost-sensitive decision tree applied on data set previously pre-processed with Smote
Boosting-based ensembles with a cost-sensitive base classifier		
Abbr.	*Method*	*Short Description*
CS-SBOM1	SmoteBoost	AdaBoost.M1 with Smote in each iteration and with an asymmetric classification cost matrix in the base classifier
CS-RUS	RusBoost	AdaBoost.M2 with random undersampling in each iteration and with an asymmetric classification cost matrix in the base classifier
Bagging-based ensembles with a cost-sensitive base classifier		
Abbr.	*Method*	*Short Description*
CS-OBAG	OverBagging	Bagging with oversampling of the minority class and with an asymmetric classification cost matrix in the base classifier
CS-SBAG	SmoteBagging	Bagging where each bag's Smote quantity varies and with an asymmetric classification cost matrix in the base classifier

References

1. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Wadsworth & Brooks. Monterey, CA 1984

2. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artificial Intelligence Res* 2002;16:321–357

3. Chawla N, Lazarevic A, Hall L, Bowyer K. SMOTEBoost: Improving prediction of the minority class in boosting. *Paper presented at the European Conference on Principles of Data Mining and Knowledge Discovery* 2003:107–119

4. Ekbal A. Improvement of prediction accuracy using discretization and voting classifier. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on IEEE* 2006;2:695–698

5. 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

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

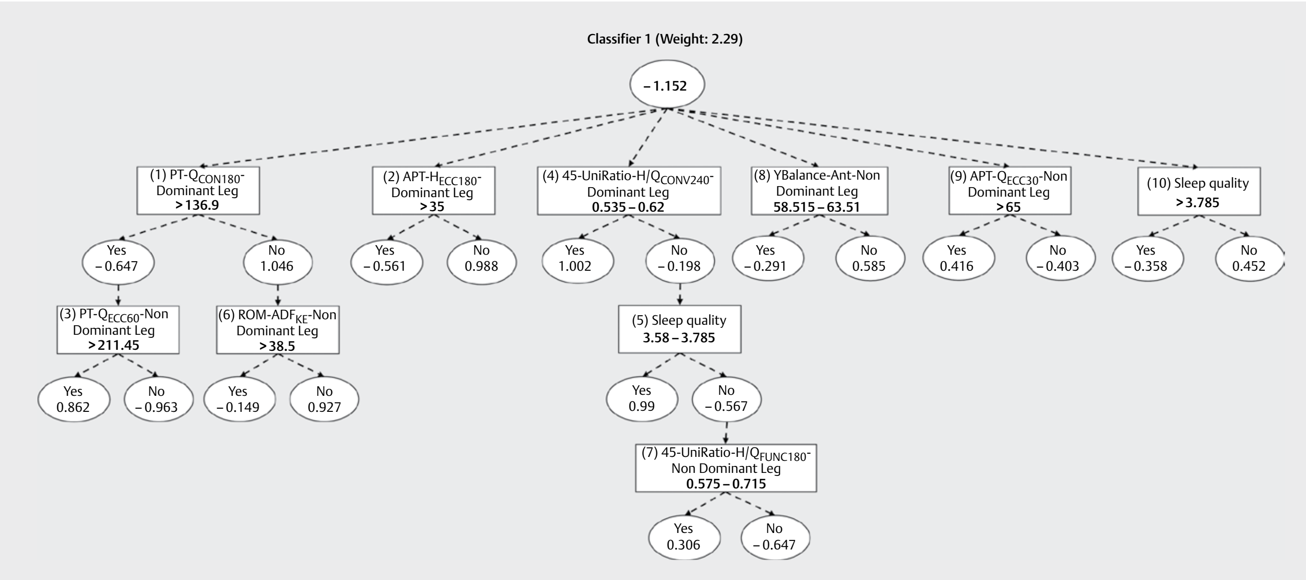
7. Galar M, Fernandez A, Barrenechea E, Bustince H, & Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 2012;42:463–484

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

9. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 2010;40:185–197

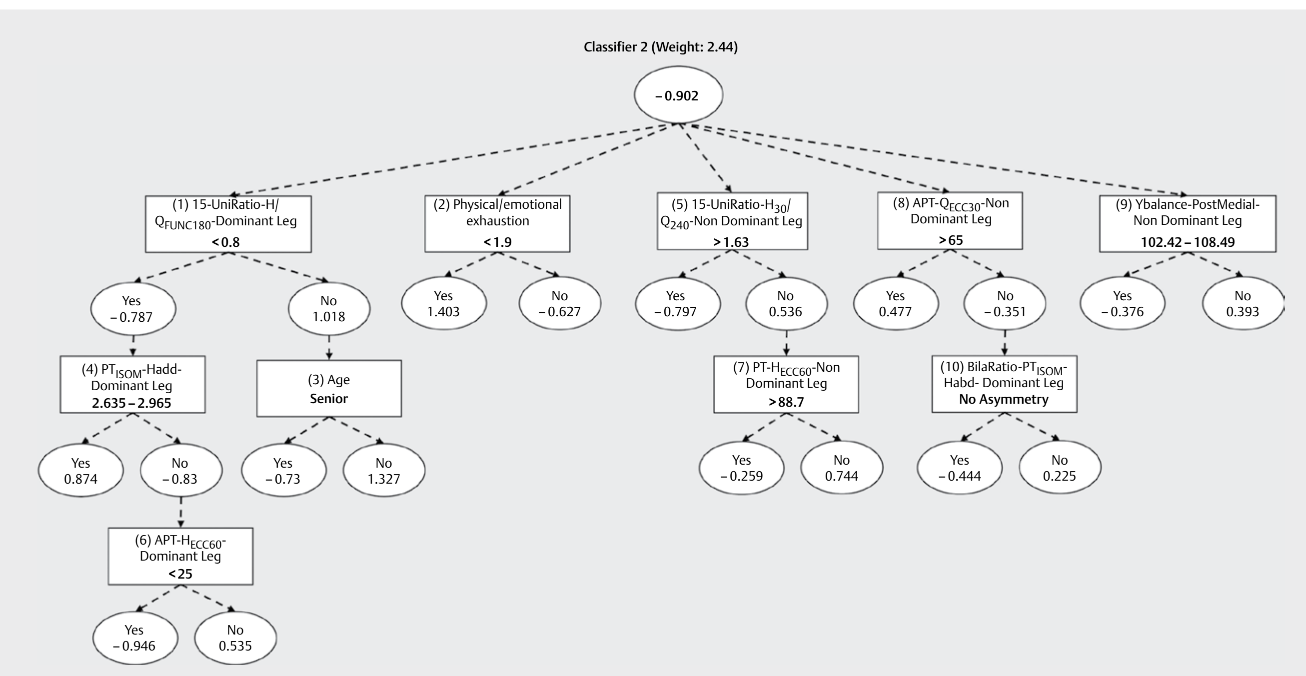
10. Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models. *Paper presented at the Computational Intelligence and Data Mining 2009. CIDM'09* IEEE Symposium on; 2009:324–331

SDC 9: First classifier



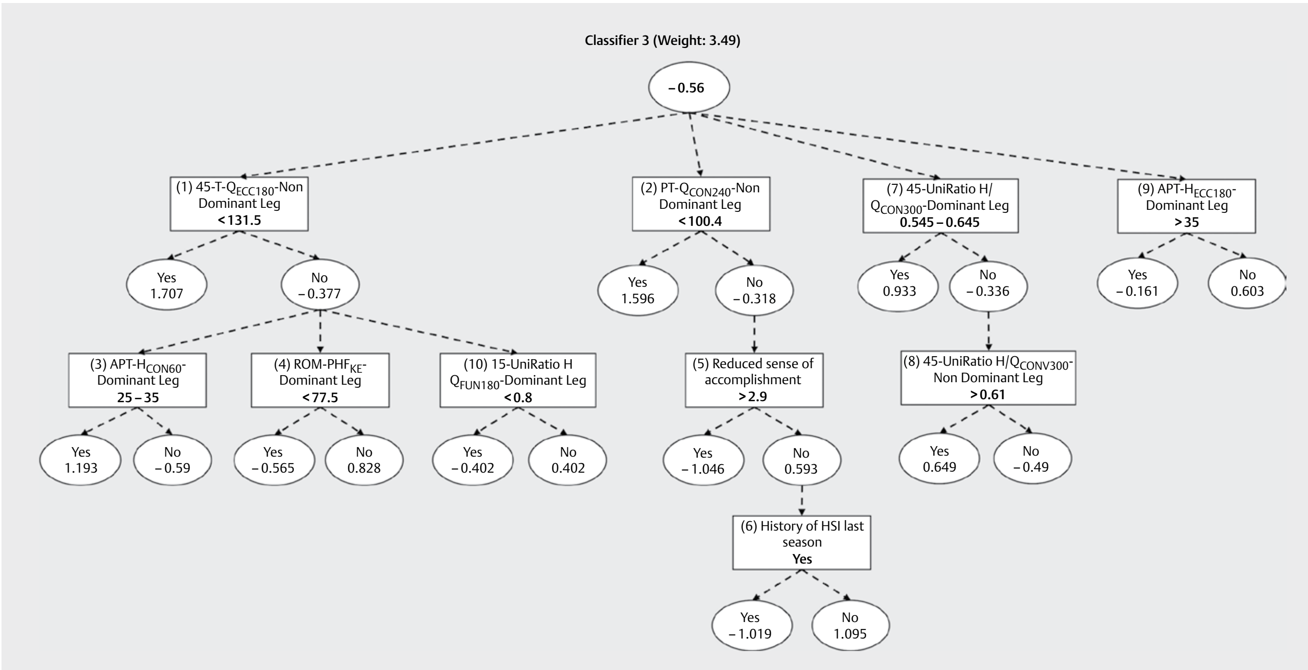
Graphical representation of the first classifier of the predictive model for muscle injuries.

SDC 10: Second classifier



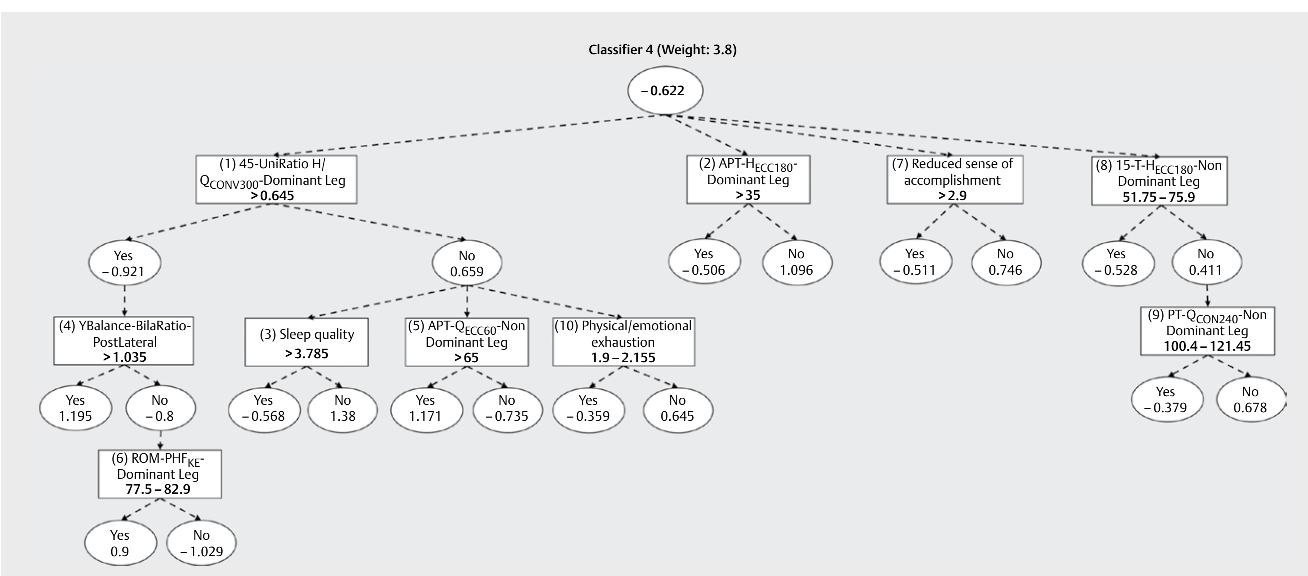
Graphical representation of the second classifier of the predictive model for muscle injuries.

SDC 11: Third classifier



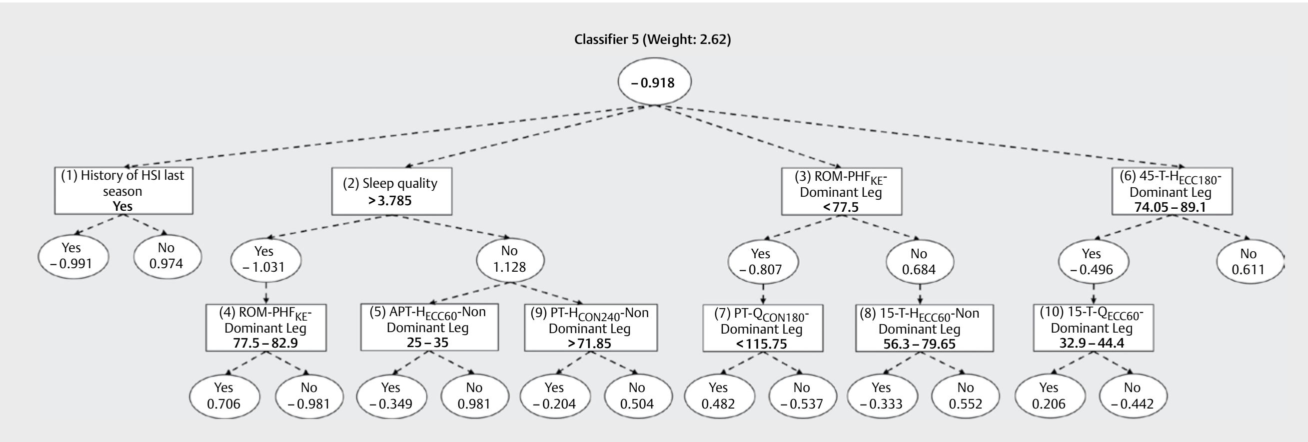
Graphical representation of the third classifier of the predictive model for muscle injuries.

SDC 12: Fourth classifier



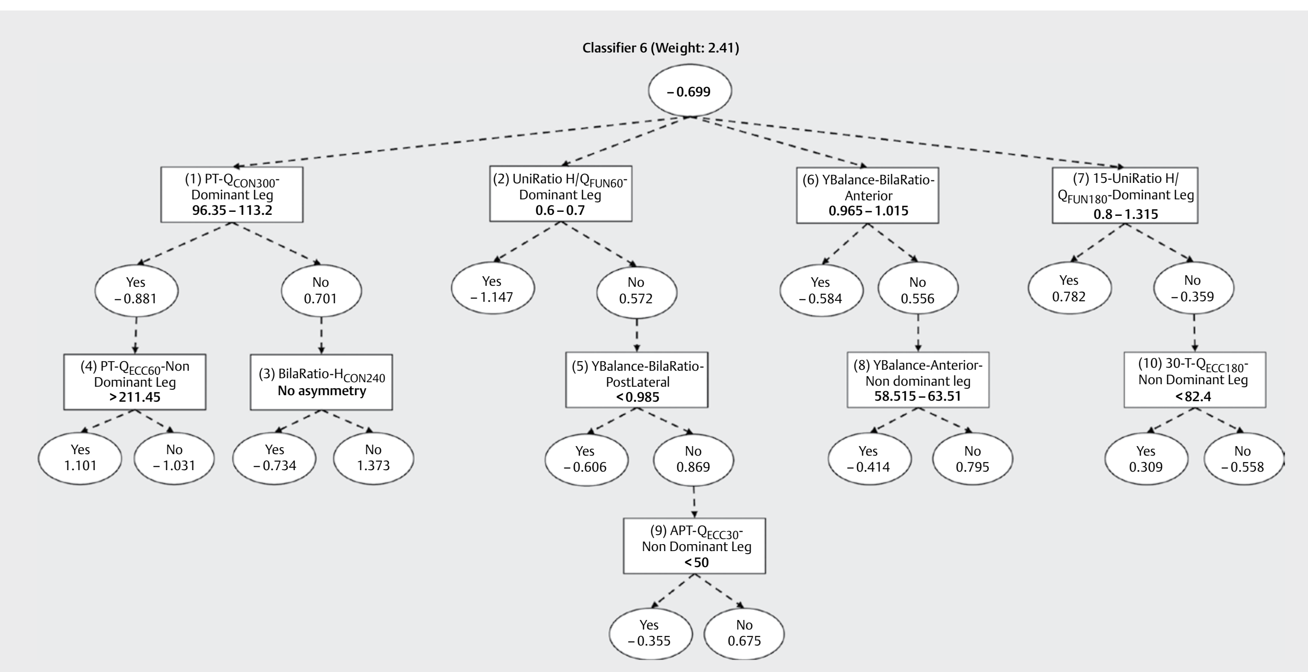
Graphical representation of the fourth classifier of the predictive model for muscle injuries.

SDC 13: Fifth classifier



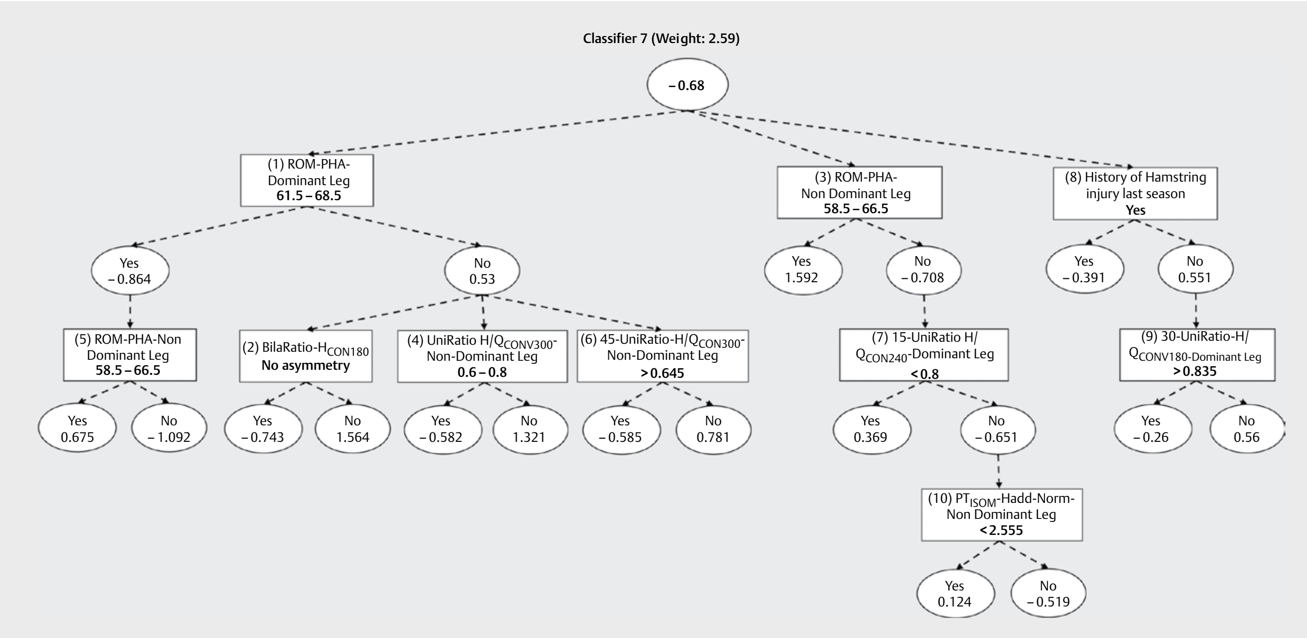
Graphical representation of the fifth classifier of the predictive model for muscle injuries.

SDC 14: Sixth classifier



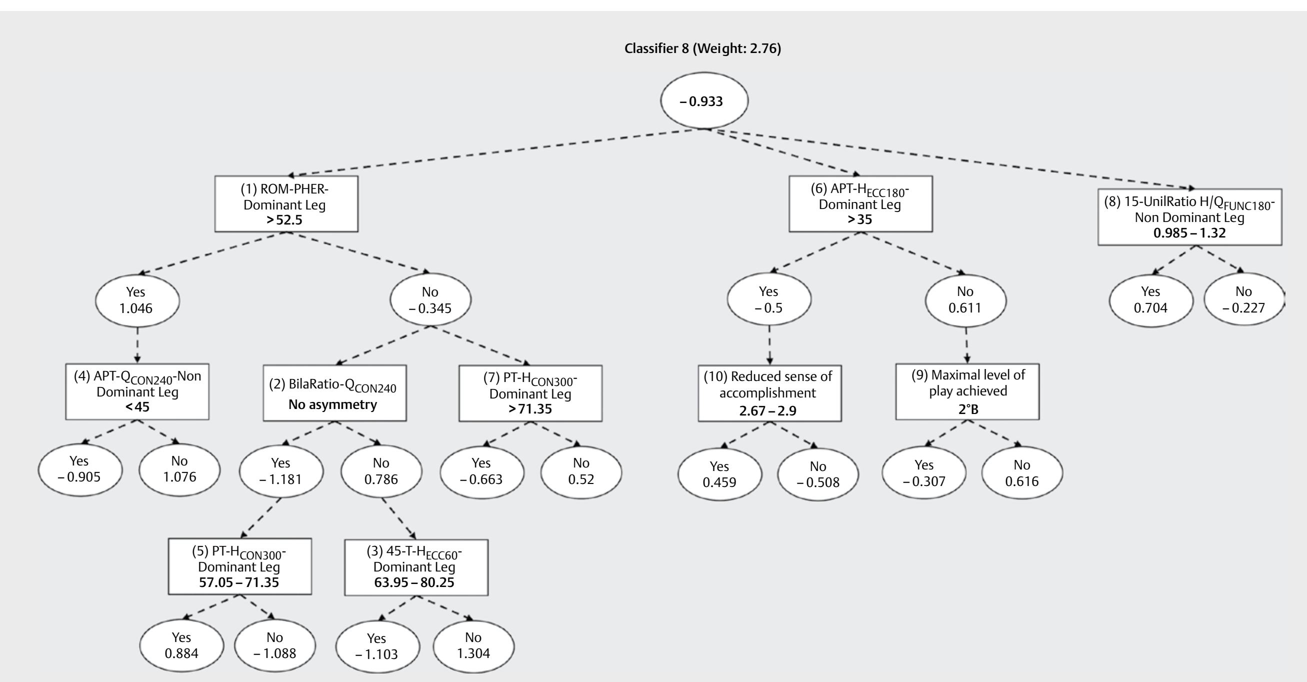
Graphical representation of the sixth classifier of the predictive model for muscle injuries.

SDC 15: Seventh classifier



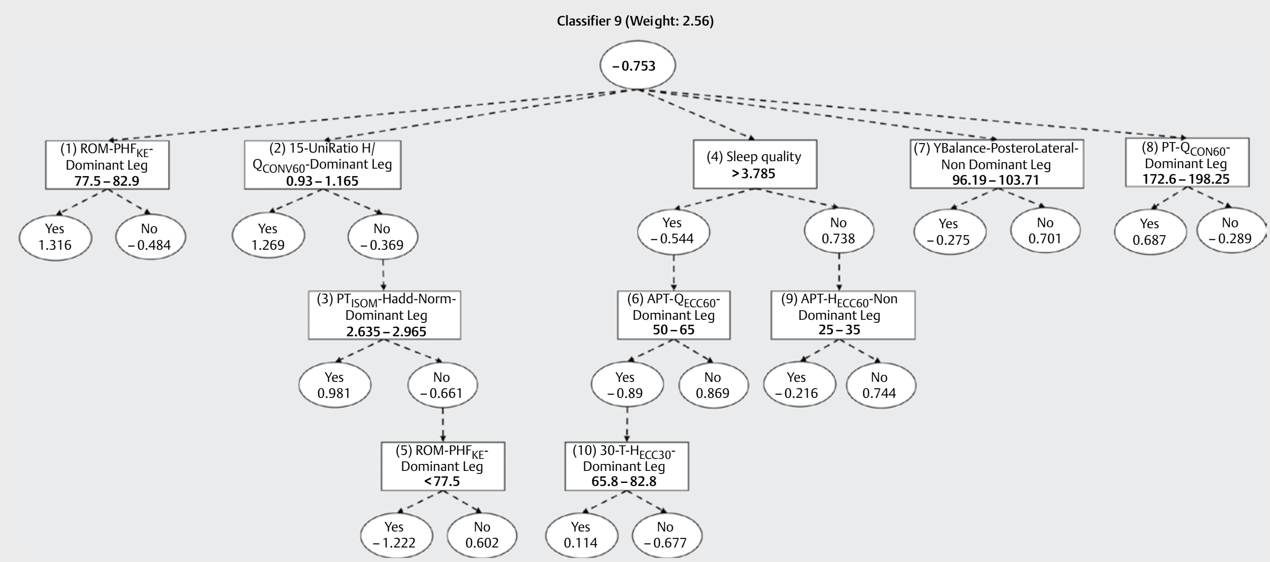
Graphical representation of the seventh classifier of the predictive model for muscle injuries.

SDC 16: Eighth classifier



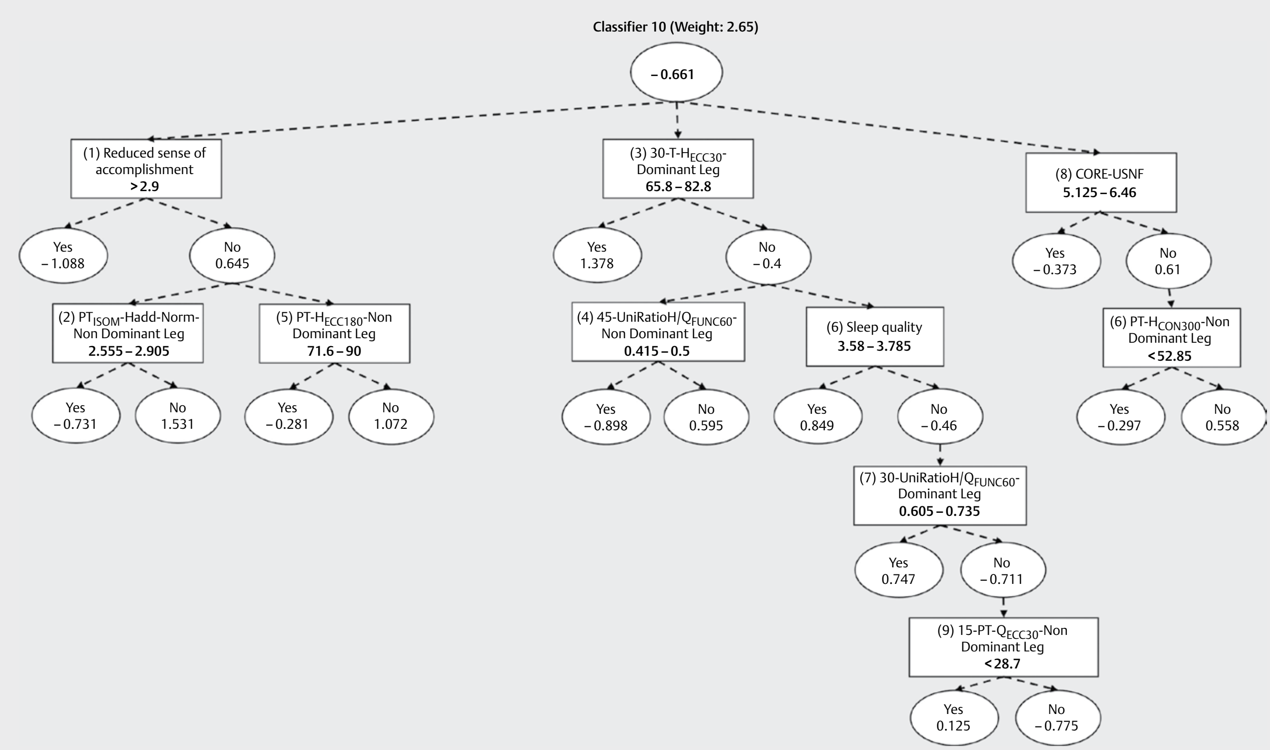
Graphical representation of the eigth classifier of the predictive model for muscle injuries.

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.