Robles-Palazón, Francisco Javier, Puerta-Callejón, José M, Gámez, José, De Ste Croix, Mark B ORCID: 0000-0001-9911-4355, Cejudo, Antonio, Santonja, Fernando, Sainz de Baranda, Pilar and Ayala, Francisco ORCID: 0000-0003-2210-7389 (2023) Predicting injury risk using machine learning in male youth soccer players. Chaos, Solitons and Fractals, 167. Art 113079. doi:10.1016/j.chaos.2022.113079

Supplementary file 1. TRIPOD checklist.

	Item	Recommendation	# Page[a]
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
Introduction			
Background/ objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	3-4
	3b	Specify the objectives, including whether the study describes the development or validation of the model, or both	5
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation datasets, if applicable	5-6
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up	6
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5
	5b	Describe eligibility criteria for participants.	5-6
	5c	Give details of treatments received, if relevant.	NA
Outcome	6	Clearly define the outcome that is predicted by the prediction model, including how and when assessed	13
	6b	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured	7-12
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors	NA
Sample size	8	Explain how the study size was arrived at.	NA
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method	15
Statistical analysis methods	10a	Describe how predictors were handled in the analyses	14-15
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation	14-15

	Item	Recommendation	# Page[a]
	10c	For validation, describe how the predictions were calculated	14
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	14-15
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	Provide details on how risk groups were created, if done.	13-15
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	14
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	16
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	5-6
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors, and outcome).	NA
Model development	14a	Specify the number of participants and outcome events in each analysis.	16
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	17-19
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	16-22
	15b	Explain how to use the prediction model.	17,20
Model performance	16	Report performance measures (with CIs) for the prediction model	19
Model updating	17	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	26
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.	22-24
	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	22-26

	Item	Recommendation	# Page[a]
Implications	20	Discuss the potential clinical use of the model and implications for future research.	26-27

[a] Page numbers specified are based on the authors' latest version accepted for publication (before the final version formatted and published by the journal).

Supplementary file 2. Description of the personal or individual injury risk factors recorded.

Name	**Labels**
Player position	Goalkeeper, Defender, Midfielder or Forward
Chronological age (y)	Numeric
Age group	U11-12, U13-14, U15-16 or U17-19
Dominant leg	Right, Left or Two-footed
12 months LE-ST time loss injury history	Yes or no
Years of playing football (y)	Numeric
Training frequency (days)	Numeric
Body mass (kg)	Numeric
Stature (cm)	Numeric
Body mass index (kg/m^2)	Numeric
Leg length (cm)	Numeric
Tibia length (cm)	Numeric
Maturity offset	Numeric
Age at peak height velocity	Numeric

Supplementary file 3. Description of the psychological injury risk factors recorded.

Name	**Labels**
Anxiety-Trait	Numeric
Profile of Mood States (POMS)	
Tension	Numeric
Depression	Numeric
Anger	Numeric
Vigour	Numeric
Fatigue	Numeric
Confusion	Numeric
Friendliness	Numeric
Psychological Characteristics related to the Sport Performance (CPRD)	
Stress control	Numeric
Performance evaluation	Numeric
Motivation	Numeric
Mental skills	Numeric
Team cohesion	Numeric
Global score	Numeric

Supplementary file 4. Measures obtained from the Jump tests.

Name	Labels	
	Dominant leg	Non-dominant leg
Tuck Jump Assessment (TJA)		
FPPA	≤0 (none), 1–9 (minor), 10–20 (moderate), >20 (severe)	≤0 (none), 1–9 (minor), 10–20 (moderate), >20 (severe)
BIL-FPPA	No Asymmetry or Asymmetry	
HF_IC (°)	Numeric	
KF_IC (°)	Numeric	
AF_IC (°)	Numeric	
HF_PF (°)	Numeric	
KF_PF (°)	Numeric	
AF_PF (°)	Numeric	
HF_ROM (°)	Numeric	
KF_ROM (°)	Numeric	
AF_ROM (°)	Numeric	
Drop Vertical Jump (DVJ)		
H (cm)	Numeric	
CT (ms)	Numeric	
RSI (mm/ms)	Numeric	
FPPA	≤0 (none), 1–9 (minor), 10–20 (moderate), >20 (severe)	≤0 (none), 1–9 (minor), 10–20 (moderate), >20 (severe)
BIL-FPPA	No Asymmetry or Asymmetry	
KMD	≤0 (none), 0.1-3.0 (minor), 3.1-6.0 (moderate), >6.0 (severe)	≤0 (none), 0.1-3.0 (minor), 3.1-6.0 (moderate), >6.0 (severe)
BIL-KMD	No Asymmetry or Asymmetry	
KASR	Varus or Valgus	
KSD (cm)	Numeric	
HF_IC (°)	Numeric	
KF_IC (°)	Numeric	
AF_IC (°)	Numeric	
HF_PF (°)	Numeric	
KF_PF (°)	Numeric	
AF_PF (°)	Numeric	
HF_ROM (°)	Numeric	
KF_ROM (°)	Numeric	
AF_ROM (°)	Numeric	
Countermovement Jump (CMJ)		
H (cm)	Numeric	

Name	Labels	
	Dominant leg	Non-dominant leg
Single-leg countermovement jump (SLCMJ)		
H (cm)	Numeric	Numeric
BIL-H	No Asymmetry or Asymmetry	
Take-off pVGRF ($N\cdot kg^{-1}$)	Numeric	Numeric
Landing-pVGRF ($N\cdot kg^{-1}$)	Numeric	Numeric
pLFT (ms)	Numeric	Numeric
Take-off BIL-pVGRF	No Asymmetry or Asymmetry	
Landing BIL-pVGRF	No Asymmetry or Asymmetry	
BIL-pLFT	No Asymmetry or Asymmetry	
Horizontal Jump tests		
SLJ (cm)	Numeric	
SHD (% leg length)	Numeric	Numeric
SHD-BIL	No Asymmetry or Asymmetry	

SLJ: standing long jump; SHD: single hop for distance; H: height; CT: contact time; RSI: reactive strength index; FPPA: frontal plane projection angle; HF: hip flexio; KF: knee flexion; AF: ankle flexion; IC: initial contact; PF: peak flexion; ROM: range of motion; KSD: knee separation distance; KASR: knee-to-ankle separation ratio; KMD: knee medial displacement; pVGRF: peak vertical ground reaction force; pLFT: peak landing force timing; BIL: bilateral ratio.

Supplementary file 5. Measures obtained from the Sprint.

Name	**Labels**
10m-Sprint (s)	Numeric
20m-Sprint (s)	Numeric
10to20m-Sprint (s)	Numeric
Vmax ($m \cdot s^{-1}$)	Numeric
M_F0 ($N \cdot kg^{-1}$)	Numeric
V(0) ($m \cdot s^{-1}$)	Numeric
Pmax ($W \cdot kg^{-1}$)	Numeric
DRF (%)	Numeric
FV ($N \cdot s \cdot m^{-1} \cdot kg^{-1}$)	Numeric
RF-10m ($N \cdot kg^{-1}$)	Numeric
RFPeak (%)	Numeric

Vmax: maximal velocity; M_F0: theoretical maximal force; V(0): theoretical maximal velocity; Pmax: maximal power; DRF: decrease in the ratio of horizontal-to-resultant force; FV: slope of the force-velocity relationship; RF: ratio of the net horizontal-to-resultant force; RFPeak: maximal ratio of horizontal-to-resultant force.

Supplementary file 6. Measures obtained from the ROM-Sport battery.

Name	Labels	
	Dominant Leg	**Non-Dominant Leg**
ROM-PHF_{KF} (º)	Numeric	Numeric
ROM-PHF_{KE} (º)	Numeric	Numeric
ROM-PHE (º)	Numeric	Numeric
ROM-PHABD (º)	Numeric	Numeric
ROM-$PHABD_{HF}$ (º)	Numeric	Numeric
ROM-PHADD (º)	Numeric	Numeric
ROM-PHIR (º)	Numeric	Numeric
ROM-HER (º)	Numeric	Numeric
ROM-PKF (º)	Numeric	Numeric
ROM-ADF_{KE} (º)	Numeric	Numeric
ROM-ADF_{KF} (º)	Numeric	Numeric
ROM-BIL-PHF_{KF}	No Asymmetry or Asymmetry	
ROM-BIL-PHF_{KE}	No Asymmetry or Asymmetry	
ROM-BIL-PHE	No Asymmetry or Asymmetry	
ROM-BIL-PHABD	No Asymmetry or Asymmetry	
ROM-BIL-$PHABD_{HF}$	No Asymmetry or Asymmetry	
ROM-BIL-PHADD	No Asymmetry or Asymmetry	
ROM-BIL-PHIR	No Asymmetry or Asymmetry	
ROM-BIL-PHER	No Asymmetry or Asymmetry	
ROM-BIL-PKF	No Asymmetry or Asymmetry	
ROM-BIL-ADF_{KE}	No Asymmetry or Asymmetry	
ROM-BIL-ADF_{KF}	No Asymmetry or Asymmetry	

ROM: range of motion; PHF_{KF}: passive hip flexion with the knee flexed; PHF_{KE}: passive hip flexion with the knee extended; PHE: passive hip extension; PHABD: passive hip abduction; $PHABD_{HF}$: passive hip abduction at 90º of hip flexion; PHADD: passive hip adduction; PHIR: passive hip internal rotation; PHER: passive hip external rotation; PKF: passive knee flexion; ADF_{KE}: passive ankle dorsiflexion with the knee extended; ADF_{KF}: passive ankle dorsiflexion with the knee flexed; BIL: bilateral ratio.

Supplementary file 7. Measures obtained from the Y-Balance test.

Name	Labels	
	Dominant Leg	Non-Dominant Leg
YBalance-Anterior (%leg length)	Numeric	Numeric
YBalance-PosteroMedial (%leg length)	Numeric	Numeric
YBalance-PosteroLateral (%leg length)	Numeric	Numeric
BIL-YBalance-Anterior	No Asymmetry or Asymmetry	
BIL-YBalance-PosteroMedial	No Asymmetry or Asymmetry	
BIL-YBalance-PosteroLateral	No Asymmetry or Asymmetry	
YBalance-Composite (%leg length)	Numeric	Numeric

BIL: bilateral ratio.

Supplementary file 8. Brief description of the statistical techniques used.

Four classifiers based on different paradigms, namely decision trees with C4.5 and ADTree, Support Vector Machines with SMO and the well-known k-Nearest Neighbor (KNN) as an Instance-Based Learning approach were selected to be used in the resampling, ensemble and cost-sensitive learning methodologies as base classifiers. The configuration of each base classifier was optimised through the use of the metaclassifier MultiSearch (it performs a search of an arbitrary number of parameters of a classifier and chooses the best pair found for the actual filtering and training) with the F-score as evaluation criterion for evaluate classifier performance (C4.5: confidence factor [from 0.05 to 0.75], ADTree: number of interactions [from 5 to 50], SMO: complexity [from 1 to 10] and ridge [from -10 to 5], KNN: number of neighbours [from 1 to 5]).

With regard to the resampling techniques, four (two oversampling and two undersampling algorithms) of the most popular methodologies were selected, which are the synthetic minority oversampling technique (SMOTE), random oversampling (ROS), random undersampling (RUS) and Wilson's edited nearest neighbour rule (ENN). In the four resampling techniques selected, a level of balance in the training data near the 40/60 was attempted. In addition, the interpolations that are computed to generate new synthetic data are made considering the k-3-nearest neighbours of minority class instances using the Euclidean distance.

Regarding ensemble learning algorithms, classic ensembles such as Bagging, AdaBoost and AdaBoot.M1 were included in this study. Furthermore, the algorithm families designed to deal with skewed class distributions in data sets were also included: Boosting-based and Bagging-based. The Boosting based ensembles that were considered in the current study were SMOTEBoost and RUSBoost. Concerning Bagging based ensembles, it was included from the OverBagging group, OverBagging (which uses ROS), UnderBagging (which uses RUS) and SMOTEBagging. The number of internal classifiers used within each ensemble learning algorithm was set 100 (always the same) base classifiers (C4.5, ADTree, SVM and KNN) by default.

Concerning the cost-sensitive learning algorithms, two different algorithms were used, namely MetaCost and cost-sensitive classifier. Cost-sensitive learning solutions incorporating both the data (external) and algorithmic level (internal) approaches assume higher misclassification costs for samples in the minority class and seek to minimise the high cost errors. For the both cost-sensitive algorithms selected, the cox matrix set-up was to:

c = $\begin{Bmatrix} 0 & 2 \\ 1 & 0 \end{Bmatrix}$ where a false negative has a cost of 2 and a false positive had a cost of 1.

The behaviour of some specific combinations of class-balanced ensembles with cost-sensitive base classifiers was also studied. The algorithm Random Forest in isolation and in combination with the resampling techniques was also explored due to its good results showed in previous studies (1). Finally, to allow comparison of the constructed models to a baseline model, a ZeroR classifier was also used.

For the sake of brevity and the lack of space, the codes of the algorithms used in this study are not presented here. Instead, we have only specified the names and refer the reader to similar previously published studies in elite soccer (2,3) and futsal (4) using machine learning techniques. Furthermore, all the classification algorithms used are available in Weka Data Mining software (version 3.8.3).

References

1. Bergeron MF, Landset S, Maugans TA, Williams VB, Collins CL, Wasserman EB, et al. Machine learning in modeling high school sport concussion symptom resolve. Med Sci Sport Exerc. 2019;51(7):1362–71.

2. Ayala F, López-Valenciano A, Gámez Martín JA, De Ste Croix M, Vera-Garcia FJ, García-Vaquero MP, et al. A preventive model for hamstring injuries in professional soccer: learning algorithms. Int J Sports Med. 2019;40(5):344–53.

3. López-Valenciano A, Ayala F, Puerta JM, De Ste Croix M, Vera-García F, Hernández-Sánchez S, et al. A preventive model for muscle injuries: a novel approach based on learning algorithms. Med Sci Sports Exerc. 2018;50(5):915–27.

4. Ruiz-Pérez I, López-Valenciano A, Hernández-Sánchez S, Puerta-Callejón JM, De Ste Croix M, Sainz de Baranda P, et al. A field-based approach to determine soft tissue injury risk in elite futsal using novel machine learning techniques. Front Psychol. 2021;12:1–15.

Supplementary file 9. Scheme of the algorithms selected in data set.

Lower extremity non-contact soft tissue injuries
UBAG [SMO]
(1) meta.FilteredClassifier '-F \"unsupervised.attribute.ReplaceMissingValues \" -S 1 -W meta.AttributeSelectedClassifier -- -E \"CfsSubsetEval -P 1 -E 1\" -S \"GreedyStepwise -T -1.7976931348623157E308 -N -1 -num-slots 1\" -W meta.MultiSearch -- -E FM -search \"weka.core.setupgenerator.MathParameter -property classifier.classifier.calibrator.ridge -min -10.0 -max 5.0 -step 1.0 -base 10.0 -expression pow(BASE,I)\" -class-label 1 -algorithm \"meta.multisearch.DefaultSearch -sample-size 100.0 -initial-folds 2 -subsequent-folds 10 -initial-test-set . -subsequent-test-set . -num-slots 1\" -log-file /Applications/weka-3-8-3 -S 1 -W meta.Bagging -- -P 100 -S 1 -num-slots 1 -I 100 -W meta.FilteredClassifier -- -F \"supervised.instance.SpreadSubsample -M 1.5 -X 0.0 -S 1\" -S 1 -W functions.SMO -- -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 250007\" -calibrator \"functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4\"' -4523450618538717400