Journal of Sports Sciences

ISSN: 0264-0414 (Print) 1466-447X (Online) Journal homepage: www.tandfonline.com/journals/rjsp20

Development and internal validation of machine learning prognostic models of sports injuries using self-reported data in athletics (track and field): The influence of quantity and quality of features

Spyridon Iatropoulos, Pierre-Eddy Dandrieux, Pascal Edouard & Laurent Navarro

To cite this article: Spyridon Iatropoulos, Pierre-Eddy Dandrieux, Pascal Edouard & Laurent Navarro (13 Jun 2025): Development and internal validation of machine learning prognostic models of sports injuries using self-reported data in athletics (track and field): The influence of quantity and quality of features, Journal of Sports Sciences, DOI: 10.1080/02640414.2025.2517971

To link to this article: https://doi.org/10.1080/02640414.2025.2517971

JOURNAL OF SPORTS SCIENCES
https://doi.org/10.1080/02640414.2025.2517971

Routledge
Taylor & Francis Group

SOCIAL AND BEHAVIORAL SCIENCES

Development and internal validation of machine learning prognostic models of sports injuries using self-reported data in athletics (track and field): The influence of quantity and quality of features

Spyridon Iatropoulos[a], Pierre-Eddy Dandrieux[a,b], Pascal Edouard[a,c] and Laurent Navarro[b]

[a]Laboratoire Interuniversitaire de Biologie de la Motricité, Université Jean Monnet Saint-Etienne, Lyon 1, Université Savoie Mont-Blanc, Saint-Etienne, France; [b]Centre CIS, F-42023, Mines Saint-Etienne, University Lyon, University Jean Monnet, Laboratoire Interuniversitaire de Biologie de la Motricité, Saint-Etienne, France; [c]Department of Clinical and Exercise Physiology, Sports Medicine Unit, Faculty of Medicine, University Hospital of Saint-Etienne, Saint-Etienne, France

ABSTRACT

To compare the performance of sports injury prognostic machine learning models when trained on (i) baseline data (i.e. collected once) vs. monitoring data (i.e. collected frequently over a period), (ii) raw monitoring data vs. time-integrating engineered features of the same data, and (iii) different numbers of features. Self-reported data collected during a previous randomised controlled trial in athletics athletes over 39 weeks constituted the dataset for model development. Baseline features, monitoring features, and two time-integrating feature engineering strategies were employed. Seven machine learning algorithms were trained with different groups and numbers of features and validated internally with bootstrapping. The models' discrimination was statistically compared using t-tests or Mann–Whitney tests ($\alpha = 0.00026$). A dataset of 4537 cases including 149 injuries was derived from 165 athletes. Monitoring features outperformed baseline features in 5 out of 7 algorithms ($p < 0.00026$). The two feature engineering strategies showed marginal differences (1–8%) in 4 out of 7 algorithms ($p < 0.00026$). Larger numbers of features showed consistent improvements of performance for 6 out of 7 algorithms. Developing injury prediction ML models based on self-reported data in the sport of athletics seems promising but highly influenced by the quality and quantity of features.

ARTICLE HISTORY
Received 01 September 2024
Accepted 03 June 2025

KEYWORDS
Artificial intelligence; injury prediction; injury forecasting; bootstrapping validation; feature engineering

Introduction

Athletes participating in athletics (track and field) run the risk of injuries (Edouard et al., 2024). Given the impact of athletics injuries on health and sport participation, there is a great interest to develop and promote injury risk reduction approaches (Edouard et al., 2024). One approach could be the detection of athletes at risk in order to individualise and optimise their injury risk reduction strategies. In this context, development of predictive algorithms using machine learning (ML) techniques has been attempted in many research studies in the past decade (Bullock et al., 2022; Claudino et al., 2019; Eetvelde et al., 2021). Although plenty of injury prediction models have been published in a range of individual and team sports (Bullock et al., 2022; Claudino et al., 2019; Eetvelde et al., 2021), to our knowledge, only one study involving athletes of all athletics disciplines exists (Dandrieux et al., 2025), and one focusing on middle- and long-distance runners (Lövdal et al., 2021). Given the vast differences in the data characteristics among sports, model generalisability and transferability to other sports might be limited. Thus, there is a need to continue developing athletics-specific models.

Developing prognostic models of sports injuries can be achieved using self-reported data of potentially injury-influencing factors, such as training load (Briand et al., 2022; Dandrieux et al., 2025; Goggins et al., 2022; Hecksteden et al., 2023; Huang et al., 2022), sleep quality (Dandrieux et al., 2025; Lyubovsky et al., 2022; Vallance et al., 2020), and perceived states of physical and/or psychological well-being (Briand et al., 2022; Dandrieux et al., 2025; Hecksteden et al., 2023; Huang et al., 2022; Lövdal et al., 2021; Lyubovsky et al., 2022; Vallance et al., 2020). Alternative data sources, such as global positioning systems (GPS) (Carey et al., 2018; Lövdal et al., 2021; Lyubovsky et al., 2022; Vallance et al., 2020), heart-rate monitoring (Lövdal et al., 2021), and/or variability (HRV) trackers (Briand et al., 2022; Lyubovsky et al., 2022), physical tests (Huang et al., 2022; Mandorino et al.,

CONTACT Spyridon Iatropoulos spyros.iatrop@gmail.com Laboratoire Interuniversitaire de Biologie de la Motricité, Université Jean Monnet saint-Etienne, Lyon 1, Université Savoie Mont-Blanc Institution, Saint-Etienne 42270, France
Supplemental data for this article can be accessed online at https://doi.org/10.1080/02640414.2025.2517971

2022), and biochemical analysis of biological fluids (e.g., urine) (Huang et al., 2022) may present as seemingly objective and analytical, but they have not proved their added value compared to the questionnaire-based ones with respect to ML model performance (Lyubovsky et al., 2022; Vallance et al., 2020). Besides, the cost and limited availability in non-elite sport settings of such technologies may prevent their general adoption. Hence, researchers should focus on harnessing the maximal potential of self-reported data which is cost-effective, minimally invasive and scalable to all performance levels.

Also notable in the literature of sports injury prognostic ML models is the large variability in the methodology and the resulting predictive performance of the various models (Bullock et al., 2022; Claudino et al., 2019; Eetvelde et al., 2021). This raises important questions not only about which methodology can accomplish higher performance scores, but more importantly, which methodology could be better integrated in a given sport field context. One such methodological aspect concerns the frequency of data collection. Many studies have based their models on variables that were measured only once (i.e., screening), typically at the start of the season, such as personal information (age, sex, height, weight, past injury history) (López-Valenciano et al., 2018; Rommers et al., 2020), neuromuscular tests (López-Valenciano et al., 2018; Oliver et al., 2020; Rommers et al., 2020; Ruddy et al., 2018; Ruiz-Pérez et al., 2021), physiological tests (Rommers et al., 2020), biomechanical tests (Jauhiainen et al., 2022), psychological tests (López-Valenciano et al., 2018; Oliver et al., 2020; Ruiz-Pérez et al., 2021), or physical examination tests (Jauhiainen et al., 2022; Oliver et al., 2020), which might play a role to the athlete's risk profile (Jauhiainen et al., 2022; López-Valenciano et al., 2018; Oliver et al., 2020; Rommers et al., 2020; Ruddy et al., 2018; Ruiz-Pérez et al., 2021). However, it is uncertain whether these variables remain more or less unaltered during the injury surveillance period of interest. Moreover, injury risk is expected to fluctuate over time under the influence of other time-dependent factors such as the training/competition load (internal and external), the physiological and psychological state of the athlete, and arising injuries and illnesses (Briand et al., 2022; Hecksteden et al., 2023; Meeuwisse et al., 2007). Thus, repeated measurement of such time-varying variables (i.e., monitoring) has been growingly implemented during the development of injury prediction models (Briand et al., 2022; Carey et al., 2018; Cohan et al., 2021; Huang et al., 2022; Lövdal et al., 2021; Rossi et al., 2018; Vallance et al., 2020). Lately, researchers are trying to integrate both screening and monitoring data to capture both injury susceptibility and dynamic risk exposure during model development (Goggins et al., 2022; Hecksteden et al., 2023; Lyubovsky et al., 2022; Mandorino et al., 2022).

Regardless the data sources and the collection frequency, there is typically limited capacity in most ML algorithms to discern the data patterns that lead to injury. Especially for time-dependent data, the representation of the timely order of events (i.e., 'X preceding Y' is different from 'X following Y') in an algorithmically interpretable way remains a challenge. One solution could be the engineering of features, a process of manipulating and/or combining raw features to develop new ones, which is routinely used in ML and mainly informed by domain knowledge which can be found in human expertise (Zheng, 2018). So far, for injury prediction tasks, no (Lövdal et al., 2021; Vallance et al., 2020), or limited feature engineering was implemented (Briand et al., 2022; Carey et al., 2018; Goggins et al., 2022; Hecksteden et al., 2023; Huang et al., 2022; Lyubovsky et al., 2022; Mandorino et al., 2022; Rossi et al., 2018). In the latter case, researchers capitalised mainly on sophisticated estimations of the total training load for a given day such as the session's rate of perceived exertion (session-RPE = Duration * RPE) (Carey et al., 2018; Goggins et al., 2022; Hecksteden et al., 2023; Huang et al., 2022; Lyubovsky et al., 2022; Mandorino et al., 2022), as well as for a period of time, including the linearly (LWMA) (Goggins et al., 2022; Hecksteden et al., 2023), or exponentially weighted moving averages (EWMA) (Briand et al., 2022; Carey et al., 2018; Goggins et al., 2022; Huang et al., 2022; Rossi et al., 2018), variations of the acute-to-chronic workload ratio concept (ACWR) (Briand et al., 2022; Rossi et al., 2018), and load monotony (Huang et al., 2022; Rossi et al., 2018). However, such computation methods may resemble more a one-dimensional projection of the past training load to the present, rather than a temporal sequence of data, which could restrict the ML to identify patterns associated with injury beyond those highlighted by the engineered features. Equally important, the incorporation of time through feature engineering should not be confined to the training load variables but it should be expanded to all time-varying variables which may contribute to the injury risk, such as sleep and physiological status, but this has not been yet explored. Identifying whether feature engineering and which technique can improve performance of sports injury prognostic models could help not only in exploiting the potential of collected data to the maximum, but also in enhancing our understanding of how the fluctuation of some variables can influence the injury risk.

In this context, we developed ML models to predict injuries based solely on athletes' self-reported data in

the sport of athletics. The primary aim was to compare the models' performance when trained on (i) screening (baseline) data *vs*. monitoring data, (ii) raw monitoring data *vs*. time-integrating engineered features of the same data, and (iii) different numbers of features. Two secondary aims were (iv) to assess the calibration of the developed models, and (v) to explore which individual features could be of importance in prognostic modelling. Our aims did not include identifying the best methodology to maximise the predictive performance of these models, but to explore how different features may influence such performance.

Methods

Study design

We used self-reported data collected in a cluster-randomised controlled trial called 'PREVATHLE', including competitive athletics athletes followed during 39 weeks of the 2017–2018 athletics season (from October 2017 to July 2018) (Edouard et al., 2021). Athletes in the intervention group were asked to perform at least twice a week the Athletics Injury Prevention Programme (AIPP) developed by Edouard et al. (2020), and the control group was asked to follow their regular training plan. The original 'PREVATHLE' study protocol was reviewed and approved by the Committee for the protection of persons (CPP Ouest II – Angers, number: 2017-A01980-53) and registered on the ClinicalTrials.gov (Identifier: NCT03307434) (Edouard et al., 2021). The use of the data for the present study was reviewed and approved by the Saint-Etienne University Hospital Ethical Committee (Institutional Review Board: IORG0007394; IRBN IRBN292023/CHUSTE). All participants included in the present study were informed about the present study's aim and procedure, that their data were used for this new analysis, and their rights to refuse that their data be used for research. The Ethical Committee required no new signed informed consent. There was no pre-registered protocol for the present study, but it was based on the data collected in the 'PREVATHLE' cluster randomised controlled trial registered on the ClinicalTrials.gov (Identifier: NCT03307434). We reported the methodology and results according to the TRIPOD +AI guidelines (Collins, Moons, et al., 2024).

Patient and public involvement

There was no patient or public involvement.

Equity, diversity and inclusion statement

All athletes licensed at the French Federation of Athletics (FFA) for competition in a club of at least 15 athletes (i.e., included cluster), without any contraindications for competitive athletics activity attested by the license at the FFA, aged between 15 and 40 years, and having access to Internet, without any restriction based on sex, race/ethnicity/culture, socioeconomic level, or representation from marginalised groups were eligible to be included in this study.

The research team included two junior researchers and two senior researchers, from a variety of disciplines (sports medicine, physical medicine and rehabilitation, sports science, and data science), and two different countries in Europe (France, Greece).

Population

Athletes were invited to participate in the PREVATHLE study at an individual level at the start of the 2017–18 athletics season by an e-mail invitation sent by the FFA on the 23 October 2017. For more details on athletes' recruitment please see Edouard et al. (2021) Inclusion criteria were: athletes licensed at the FFA in a club of at least 15 athletes, without any contraindications for competitive athletics activity attested by the license at the FFA, aged between 15 and 40 years, and having access to Internet. We excluded athletes if they refused to participate in the study or if they were unable to express agreement or signing the informed consent. The athletes had to provide written informed consent for participation, as well as their parents for those under 18 years of age. For this study, we only included athletes with 100% of weekly response proportion (calculated by dividing the number of completed weekly questionnaires by the maximum number of questionnaires expected to be completed) and full completion of the baseline/screening questionnaire, so no missing data was expected.

Sample size justification

An a-priori sample size estimation was not deemed necessary, i) because the aims of our study did not include developing or validating a model with the intention to be directly used in clinical and sport practice and ii) because the data originated from an already completed study. All available data were attempted to be used and the feasibility of our study was indicated by the

comparable amount of data used in previous studies of sports injury prognostic modelling (Bullock et al., 2022).

Data collection

We collected data through online questionnaires: i) at the start of the season using a survey developed in Google Forms (Google®) sent to all included athletes and ii) during the season and follow-up using a secured website called 'Prevathle' (Windows Server 2013 R2 64 bits – SP2; IBM DOMINO 9.01 fix pack 8). Questionnaires were sent at the start for baseline/screening data and on a weekly basis throughout the season (automatically sent every Monday) for monitoring data. Baseline/screening data was collected once at the start of the season: sex, age, height, body mass, discipline, typical athletics weekly training volume, typical other sport weekly training volume, injury history during the preceding season, and group allocation. Monitoring data was collected weekly throughout the season about the preceding week: number of hours of training and of competition, level of intensity of training and of competition, number of completed AIPP sessions, mean number of sleep hours per night, level of fatigue, illness and injury complaints. The primary outcome was injury complaints related to athletics practice that leads to restrictions in athletics participation, called 'injury complaint with participation restriction' (ICPR) (Edouard et al., 2021). For more details on data collection please see Edouard et al., (2021).

Data pre-processing

We encoded the sex (Male: 0, Female: 1) and grouped the athletics discipline (Explosive – sprints, hurdles, jumps, throws, combined events: 0; Endurance – middle/long distances, road running, race walking: 1) variables into binary. We also created a binary variable for weekly ICPR (no ICPR: 0; ICPR: 1) (Supplemental data A).

Looking for a clinically relevant task, we chose as prediction outcome the binary ICPR variable of the following week. Thus, each case of our dataset (i.e., one row of data) would have the data collected over a period of 3 weeks for one athlete as predictor variables (i.e., input), and the binary ICPR variable of the following 4th week as the prediction outcome (Figure 1). If an ICPR was reported during any of the 3 weeks of input data, the case was excluded. This was

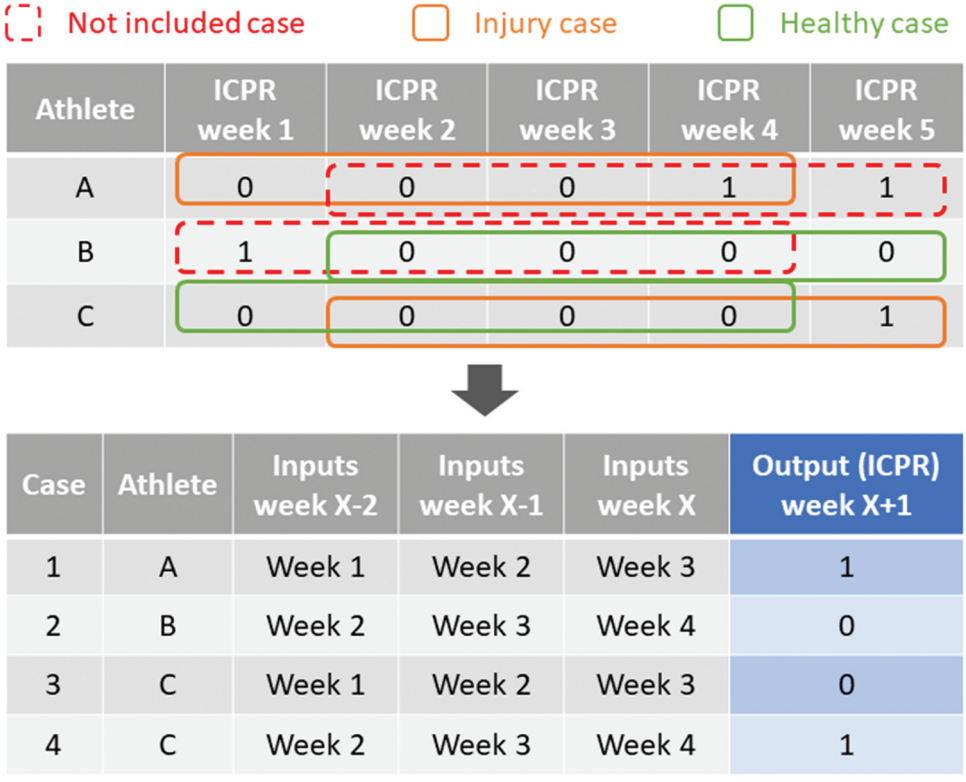

Figure 1. Examples of case selection for creating the training dataset of the experiments using imaginary data. Above, three athletes (A, B, C) reported weekly for 5 consecutive weeks whether they sustained an injury complaint leading to participation restriction (i.e., ICPR = 1) or not (i.e., ICPR = 0). Any case that consisted of 3 consecutive weeks without ICPR was included below, either as an injury case if there was an ICPR reported the subsequent week, or as a healthy case if there was no ICPR reported the subsequent week. Please note that an athlete can contribute more than one cases in the training dataset (e.g., athlete C).

based on a previous study (Lövdal et al., 2021) and reflected that predicting the injury status of an already injured athlete was considered clinically irrelevant and could potentially overestimate the models' performance. Besides, a study in professional football showed that the risk of injury was above baseline for the first 25 days after return-to-sport, indicating that it would not be very relevant to 'predict' an injury during a period when it is already likely (Zhang et al., 2024). Thus, the final dataset would consist only of the cases of each athlete where he/she was ICPR-free for 3 consecutive weeks and the task would be to predict the presence or absence of ICPR on the following week (Figure 1).

Feature selection

Initially, each instance of an athlete's week consisted of 9 baseline features (sex, age, height, body mass, discipline, typical athletics weekly training volume, typical other sport weekly training volume, and group allocation) and 9 monitoring features for each of the current and previous 2 weeks (sleep, status, training volume and intensity, AIPP sessions per week, competition volume and intensity, other sports volume and intensity) leading to a total of 27 monitoring features. For each of the 9 monitoring variables, we developed 2 classes of engineered features: the rolling and the differential features. The rolling features consisted of the 3-week moving averages ($n = 9$) and standard deviations ($n = 9$) of each feature (total rolling features = 18). The differential features consisted of a simplified first ($n = 9$) and second derivative ($n = 9$) of each feature's 3-week time-series data. Specifically, given a feature X, we calculated the change of value from the previous to the current week ($X_0 - X_{-1}$) and the change of this change in the last 3 weeks [$(X_0 - X_{-1}) - (X_{-1} - X_{-2})$], similarly to calculating velocity and acceleration from displacement data (total differential features = 18).

So, each case of the dataset consisted of 9 baseline, 27 monitoring, and 36 engineered features for a total of 72 features (Supplemental data B).

We performed 8 different experiments, corresponding to the primary aim (*i* and *ii*), using the same model development process while changing only the selected features during training. Each experiment included either i) the 9 baseline (*BASE*), ii) the 9 current-week monitoring (*MON1*), iii) the combination of 9 baseline and 9 current-week monitoring features (*BASE-MON1*), iv) the 18 2-week monitoring (*MON2*), v) the 27 3-week monitoring (*MON3*), vi) the 18 rolling (*ROLL*), vii) the 18 differential (*DIFF*), or viii) all 72 available features (*ALL*) (Appendix B). Data preparation was performed using the 'Pandas' (v.2.1.2) library of the Python programming language (The Pandas Development Team, 2020).

Model development

For all the experiments, we used a minimum-maximum scaler to normalise the values of each feature before training (i.e., $X_{scaled} = (X_{raw} - X_{min})/(X_{max} - X_{min})$). Seven different binary ML classifiers were trained on the same data separately, 4 single-learner models (logistic regression (*LOG*), linear support vector machine (*LIN-SVM*), support vector machine with radial basis function kernel (*RBF-SVM*), decision tree (*DT*)), and 3 multiple-learner 'ensemble' models (random forest (*RF*), extreme gradient boosting (*XGB*), adaptive boosting (*ADA*)). These algorithms were selected as the most frequently reported and used in the literature for injury prognostic modelling (Bullock et al., 2022; Leckey et al., 2024). Hyperparameter tuning was done by training each algorithm multiple times, separately, using a different combination of hyperparameters from the following prespecified values: parameter C (0.7, 1, 1.3); positive class weight (1, 8, 64, *'balanced'*); maximum tree depth (5, 9, 15); ensemble number of estimators (50, 75, 100); *XGB* learning rate (0.03, 0.1, 0.3). Since finding the optimal hyperparameter tuning was not the focus of this study, the above limited selection of hyperparameter values was based on the default values provided by the 'Scikit-learn' (v.1.3.2) library (Pedregosa et al., 2011) and empirical knowledge of the influence of each hyperparameter on performance. Specifically, for the class weight, we selected options that could probably handle the class imbalance of our dataset. For all the other hyperparameters, we chose 3 values that were realistic, different enough so that a difference in performance could be identifiable, and which would not be computationally intensive. The selection might not have been optimal but it allowed for the results not to be influenced by the hyperparameter tuning. Thus, a total of 12 *LOG*, 12 *LIN-SVM*, 12 *RBF-SVM*, 12 *DT*, 36 *RF*, 9 *XGB*, and 3 *ADA* models were trained at each experiment. Model development and evaluation was performed using the 'Scikit-learn' (v.1.3.2) library of the Python programming language (Pedregosa et al., 2011). Both probability and classification predictions could be derived using the built-in methods of the 'Scikit-learn' library, assuming a classification threshold of 0.5.

Internal validation

Non-parametric bootstrapping was used to evaluate the models' performance since the sample size was not large

enough to use the train-test split or the k-fold cross-validation methods (Collins, Dhiman, et al., 2024). Thus, performance metrics were corrected for optimism (i.e., an estimation of overfitting) using a 200-sample bootstrapping technique, as described by Collins et al. (Collins, Dhiman, et al., 2024) Optimism was the difference between the bootstrap model's apparent performance (i.e., the performance on the same data that it was trained on) and its performance on the original dataset (Collins, Dhiman, et al., 2024). For calibration, where scores were centered around a perfect score of 1, optimism was calculated as a ratio and correction was done by dividing the apparent performance by the optimism ratio. For each experiment, the combination of hyperparameters of each algorithm that resulted in the highest average optimism-corrected area under the receiver operator curve (AUROC) was selected for statistical analysis.

Number of features

To evaluate the influence of the number of features independently of their quality we created 10 random sequences of the 72 available features. For each random sequence, an iterative process of model development was performed using a 10-sample bootstrap and the same hyperparameters (parameter C = 1, positive class weight = 64, maximum tree depth = 9, ensemble number of estimators = 50). Specifically, given the random order of features, the first model was trained only on the first feature of the sequence, the second model was trained on the first two features of the sequence and so on. At the n^{th} step of the process, a model was trained on the *n* first features of the sequence, so that 72 models for each of the seven ML algorithms were built using between 1 and 72 features. This iterative process was repeated for each of the 10 random sequences of the 72 features. At the end, there were 10 different models trained on 1 feature, 10 models trained on 2 features, and so on, up to 72 features. For a given number of features, the 10 corresponding models were trained on randomly selected features so that the average performance of the 10 models would be influenced only by the number of features and not by their quality.

Outcomes

The main performance outcome which was statistically analysed was the optimism-corrected AUROC. Secondary performance outcomes were the optimism-corrected recall, precision, f1-score, f2-score, accuracy and Brier scores. To assess the influence of the number of features on performance, we calculated the average optimism-corrected AUROC scores of each ML algorithm at each different number of features across the 10 random sequences initially created. Calibration outcomes were the optimism-corrected mean calibration (calibration-in-the-large), calculated as the ratio between the average model probability and the proportion of the positive class in the sample and the optimism-corrected calibration slope, calculated as the slope of a linear regression model fitted on the relationship between the predicted probabilities and the actual outcomes of each case (Van Calster et al., 2019). For features' importance, the median and range of each feature's rank of importance in each ML algorithm was derived using the 'SHAP' distribution (v0.45.0) (Lundberg & Lee, 2017).

Statistical analysis

The performance differences of different feature groups (aims *i* and *ii*) were analysed with two-tailed independent-sample t-tests or their non-parametric equivalent Mann–Whitney tests whenever the assumption of normality was not met, using the 'statsmodels' (v.0.14.1) library of the Python programming language (Seabold & Perktold, 2010). In order to account for multiple testing (196 tests) the level of significance was set at $\alpha = 0.00026$ (= 0.05/196). The influence of the number of features on performance (aim *iii*) was assessed visually and descriptively, without any hypothesis testing. The secondary aims (*iv* and *v*) were examined descriptively. The code of the data pre-processing, model development, evaluation, and statistical analysis can be found in the following repository: https://github.com/spyrosiatrop/ML_Injury_Prediction_Prevathle_RCT.

Results

Population and case selection

Among the 840 athletes included in the PREVATHLE randomised controlled trial, 168 athletes had 100% response rate of weekly questionnaires, while 3 of them had missing data on their baseline questionnaire. So, 165 athletes were included in the present study for the prediction model development. From the 6435 available cases (165 athletes over 39 weeks of follow-up), 1898 were excluded because they did not consist of three consecutive weeks without ICPR before the prediction week. Thus, a total of 4537 cases including 149 injuries (3.3%) were considered during model development. Each athlete contributed a median of 30 cases in the final dataset (interquartile range: 25–33; absolute range: 4–33).

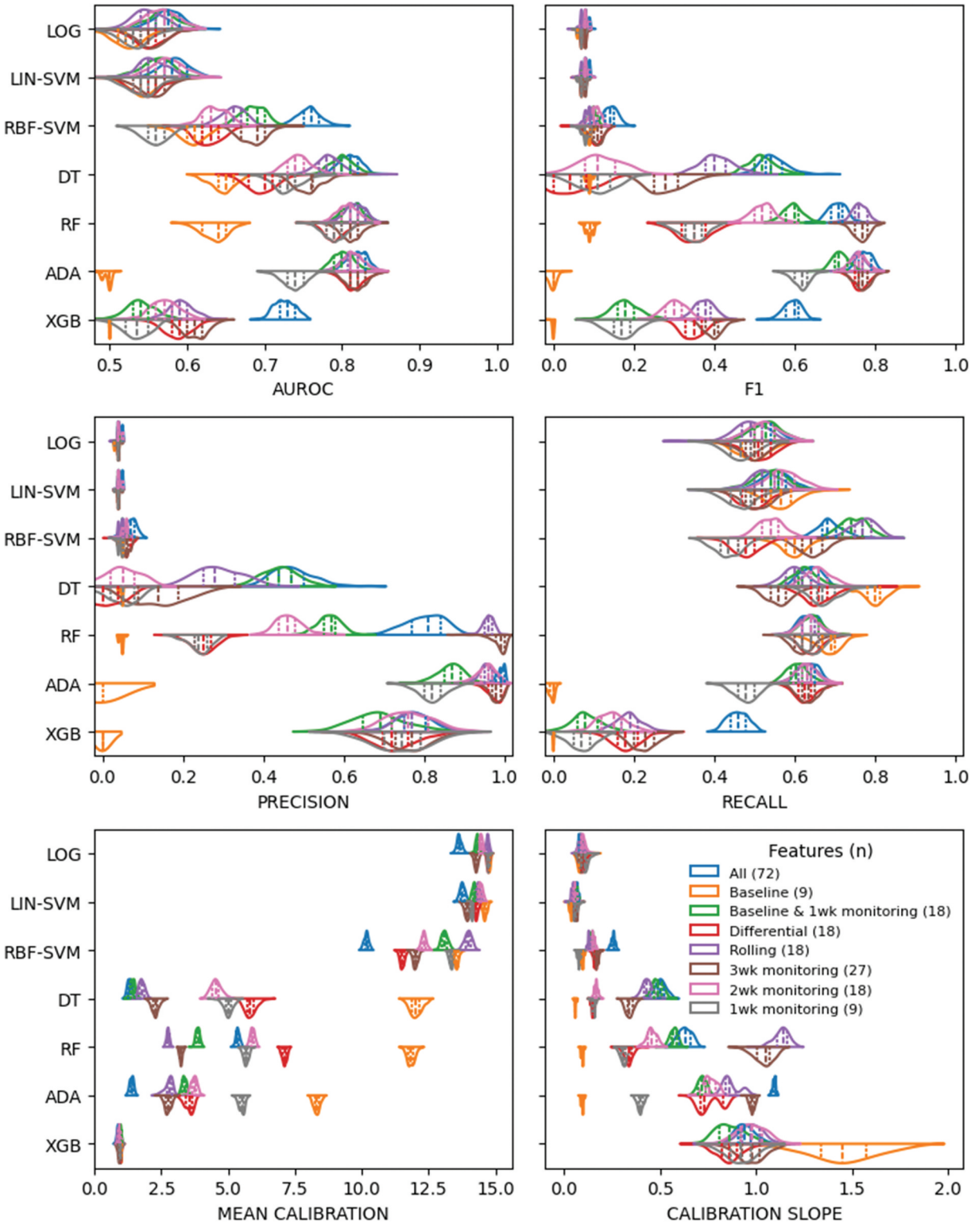

Figure 2. Performance and calibration scores of developed models categorised by group of features and ML algorithm. The optimism-corrected bootstrap distributions are presented along with vertical lines signifying the mean and the interquartile range. Please note that the AUROC scores are presented between 0.5 and 1.

Table 1. Comparisons of optimism-corrected AUROC among different groups of features across all ML algorithms. The sum of the models that show each particular difference (i.e., < 'smaller than', ≈ 'no difference', > 'larger than') when comparing the group mentioned on the left with the group on top is presented. The differences were estimated by Mann–Whitney U tests, with α = 0.00026.

	1-week monitoring (9)			2-week monitoring (18)			3-week monitoring (27)			All (72)			Baseline & 1-week monitoring (18)			Baseline (9)			Differential (18)			Rolling (18)		
	<	≈	>	<	≈	>	<	≈	>	<	≈	>	<	≈	>	<	≈	>	<	≈	>	<	≈	>
1-week monitoring (9)				7	0	0	7	0	0	7	0	0	6	1	0	1	1	5	4	2	1	7	0	0
2-week monitoring (18)	0	0	7				3	2	2	6	1	0	3	1	3	0	0	7	1	1	5	4	1	2
3-week monitoring (27)	0	0	7	2	2	3				5	2	0	2	3	2	0	0	7	0	2	5	2	3	2
All (72)	0	0	7	0	1	6	0	2	5				0	1	6	0	0	7	0	1	6	0	1	6
Baseline & 1-week monitoring (18)	0	1	6	3	1	3	2	3	2	6	1	0				0	0	7	2	0	5	2	2	3
Baseline (9)	5	1	1	7	0	0	7	0	0	7	0	0	7	0	0				6	1	0	7	0	0
Differential (18)	1	2	4	5	1	1	5	2	0	6	1	0	5	0	2	0	1	6				4	3	0
Rolling (18)	0	0	7	2	1	4	2	3	2	6	1	0	3	2	2	0	0	7	0	3	4			

Feature selection

The optimism-corrected performance scores of each ML algorithm for each group of features are presented in Figure 2 and in Supplemental data C as tabular data. The results of the statistical analysis are summarised in Table 1. Comparing *MON1* to *BASE*, optimism-corrected AUROC was better in five algorithms (*LOG, DT, RF, XGB & ADA*; $p < 0.00026$) with a median improvement of 0.08

(range: 0.01–0.24), not different in *LIN-SVM* ($p = 0.055$), and worse in *RBF-SVM* ($p < 0.0026$). The combination of baseline and monitoring features (*BASE-MON1*) scored higher than *MON1* in all models ($p < 0.00026$) except *XGB* ($p = 0.008$). However, when matching for the number of features (*MON2*) the results were balanced with *RBF-SVM, DT & RF* scoring higher, while *LIN-SVM, XGB & ADA* scoring lower than *MON2* ($p < 0.00026$).

Feature engineering

Comparing different feature engineering strategies, *ROLL* scored higher than *DIFF* in *LIN-SVM, RBF-SVM, DT & RF* ($p < 0.00026$) with a median improvement of 0.03 (range: 0.01–0.08), while there was no difference in *LOG, XGB & ADA* ($p > 0.003$). Compared to *MON3* (from where the engineered features were derived), *ROLL* scored lower in *RBF-SVM, XGB & ADA* ($p < 0.00026$), higher in *DT & RF* ($p < 0.00026$) and not differently in *LOG & LIN-SVM* models ($p > 0.007$), while *DIFF* scored lower in all models except *LOG & ADA* ($p > 0.028$). Compared to *MON2* (equal number of features), *ROLL* scored higher in four models (*RBF-SVM, DT, RF & XGB* ($p < 0.00026$), while *DIFF* scored higher only in *XGB* ($p < 0.00026$) (Figure 2, Table 1).

Number of features

The number of features seemed to be an important contributor to increased discriminatory performance for all models except the *LIN-SVM* (Figure 3). The *ADA, RF & DT* models showed a rapid improvement within the first 20 features and only marginal improvements thereafter. In contrast, *LOG, XGB & RBF-SVM* showed a more linear improvement, with additional features improving performance even at larger numbers of features.

Calibration

The estimated optimism-corrected mean calibration (calibration-in-the-large) was between 0.89 and 0.99 only for the *XBG* models, while all other models had a mean calibration between 1.4 and 15. Similarly, optimism corrected calibration slopes were between 0.86 and 1.45 for the *XGB* models, while most of the other models had smaller slope values between 0.04 and 1.09 (Figure 2).

Features' importance

The top 30 features according to their median ranking across all algorithms are presented in Table 2. The past season injury history was the most important feature in 4 out of 7 algorithms (median rank: 1, rank range: 1–16). The average range of a feature's ranking across different algorithms was ~ 43 positions, which is 60% of the maximum possible range (71 positions).

Discussion

The main findings of this study were that both the quality and the quantity of features can influence the discriminatory performance of sports injury prognostic ML models, but this was highly dependent on the ML algorithm used.

Feature selection

Compared to baseline features (*BASE*), monitoring features (*MON1*) improved discriminatory performance by

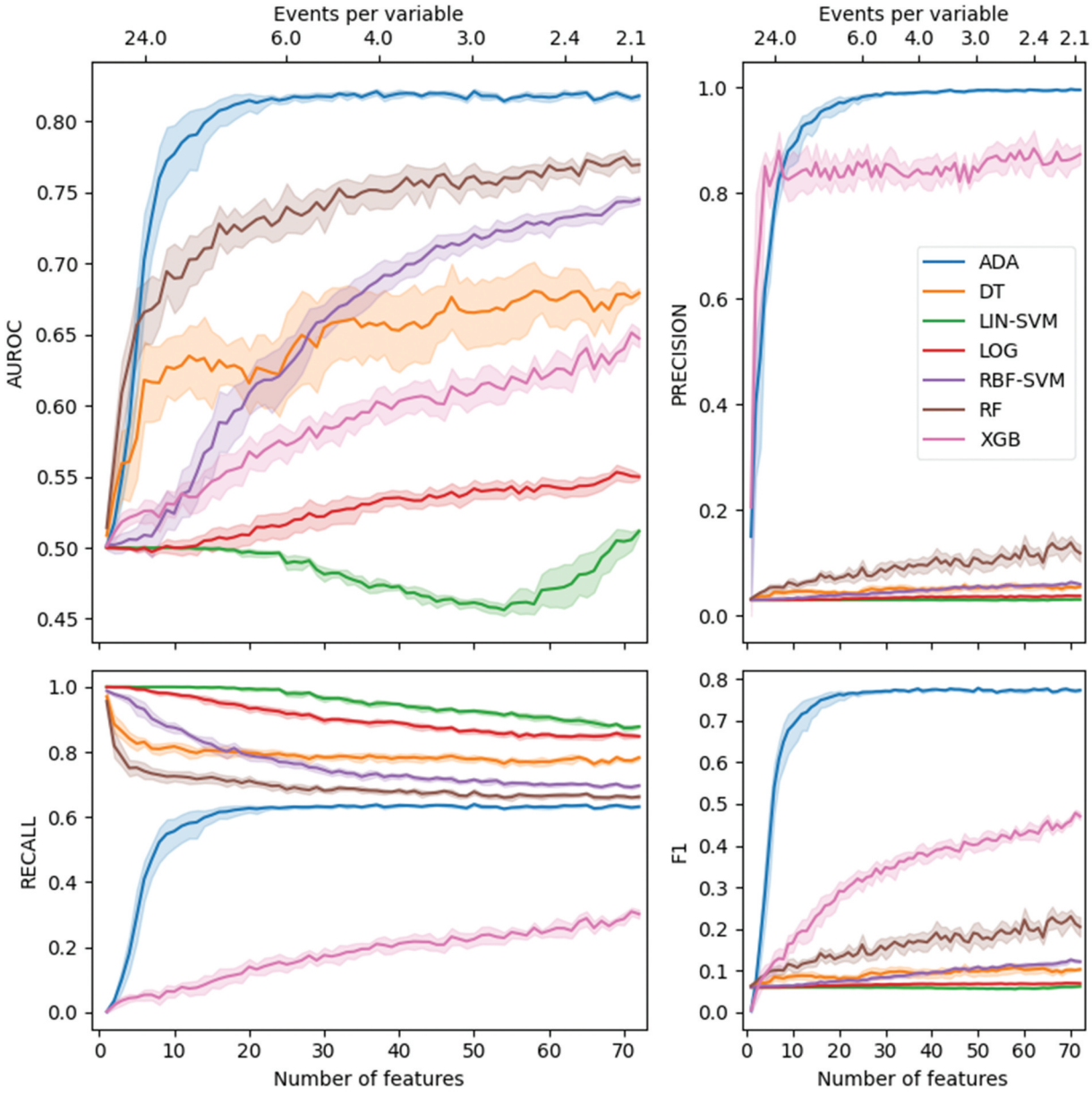

Figure 3. Performance scores of each ML algorithm while trained on different numbers of randomly selected features. The optimism-corrected scores are presented as means and standard errors.

~15% in all but the *SVM* algorithms. Their combination (*BASE-MON1*) seemed to outperform both *MON1* and *BASE* separately, but this could be attributed to the increased number of features because, compared to equal number of only monitoring features (*MON2*), *BASE-MON1* scored higher in only half of the ML algorithms. This reinforces suggestions by previous studies that combination of monitoring and baseline features may be beneficial for injury prediction (Hecksteden et al., 2023; Mandorino et al., 2022). This also highlights that such benefits may stem from the increased number of features rather than the integration of any specific features.

Feature engineering

Despite feature engineering has been used to some extent in previous research (Carey et al., 2018; Rossi et al., 2018), the additional value of using different feature engineering strategies for sports injury prognostic modelling has not been yet explored. In this study, two different feature engineering strategies did not show to improve the models' performance. Particularly, both the time-ordered (*DIFF*) and time-unordered (*ROLL*) engineering strategies scored lower than the 3-week monitoring features (*MON3*) from which they were derived in most of the algorithms, showing that dimension

Table 2. The 30 most important features based on their median ranking across seven ML algorithms. Algorithms were trained on all 72 available features, and the rank of each feature was derived from its mean absolute SHAP value.

	Models							Statistics		
Features	ADA	DT	LIN SVM	LOG	RBF SVM	RF	XGB	Median	Min	Max
base_injury	16	1	1	1	2	1	14	1	1	16
base_bodyweight	10	11	5	2	3	4	8	5	2	11
wk_train_hr_roll3_std	3	4	48	53	57	6	4	6	3	57
wk_train_hr_roll3_mean	8	2	25	24	32	3	7	8	2	32
wk_train_hr_acc02	1	6	54	60	67	10	10	10	1	67
base_age	6	9	56	33	11	2	13	11	2	56
wk_status	13	56	2	3	4	33	29	13	2	56
base_height	5	7	16	34	21	12	20	16	5	34
wk_train_int_roll3_std	37	21	13	16	9	17	56	17	9	56
base_train_hr	11	18	9	13	28	20	25	18	9	28
wk_train_int	32	57	10	7	6	19	23	19	6	57
wk_train_hr_-2	2	15	64	43	26	5	21	21	2	64
wk_sport_int_roll3_std	39	35	12	8	13	21	54	21	8	54
wk_sleep_roll3_std	22	57	20	31	17	16	63	22	16	63
wk_status_acc02	21	12	46	48	50	18	22	22	12	50
wk_status_-2	35	25	18	21	18	23	46	23	18	46
wk_train_int_roll3_mean	24	3	22	23	23	9	33	23	3	33
wk_sleep_acc02	27	14	23	25	47	8	18	23	8	47
wk_sport_hr_roll3_std	20	10	24	28	56	45	11	24	10	56
wk_sleep	15	46	19	15	24	46	50	24	15	50
wk_status_d01	17	48	26	27	25	24	27	26	17	48
wk_sport_hr_acc02	25	26	47	56	62	27	12	27	12	62
wk_status_roll3_std	28	51	29	17	20	15	31	28	15	51
wk_sport_hr	52	23	28	18	29	30	43	29	18	52
base_event_end	65	29	6	6	15	60	41	29	6	65
wk_status_roll3_mean	26	8	41	37	44	13	30	30	8	44
wk_train_int_-2	48	57	14	30	16	26	69	30	14	69
wk_sleep_d01	30	34	27	26	30	34	42	30	26	42
wk_sleep_roll3_mean	7	30	43	40	45	28	15	30	7	45
wk_sport_int_-2	50	32	11	10	14	52	48	32	10	52

reduction was not successful in preserving or improving performance. Again, larger number of features seemed to influence this outcome because performance of engineered features was more comparable to an equal number of original monitoring features (*MON2*). Contrary to our hypothesis, the time-unordered strategy seemed superior to the time-ordered in most ML algorithms but by only a small margin (~5%). It is possible that our choice of estimating the rate of change of features was not optimal to discriminate between healthy and injury cases and potentially other time-ordered strategies could do better. Nonetheless, since the purpose of feature engineering is to reduce the volume of data by preserving as much of the information as possible and not add any new information, it may not be at all necessary if the increased number of features is not a limiting factor of the model's performance, as our results indicated.

Number of features

Regarding the influence of the features' quantity on model performance, we found that except for the *LIN-SVM*, all models improved in discrimination performance (i.e., AUROC) the more features they used. Remarkably, given our dataset consisted of 149 injuries and higher number of features resulted in lower events-per-variable ratios (down to ~2), the performance did not deteriorate (Figure 3). This is in accordance with a simulation study warning against the use of the events-per-variable criterion (e.g., EPV > 10) to calculate a-priori sample sizes and number of predictors for ML models (van Smeden et al., 2018). Such consistently high performance with additional features in our case may be partially attributed to the default regularisation factor (L1 or L2) of each algorithm in 'Sci-kit learn' (Pedregosa et al., 2011), which could have minimised the relative importance of some features (Jung, 2022). However, this impact is even smaller with larger datasets according to a simulation study (van Smeden et al., 2018). Besides, regularisation is an indispensable element of most ML model development in injury prognostic modelling tasks, and exploring different regularisation methodologies was not the aim of our study.

Calibration

Although we aimed to compare the predictive performance of the developed models, calibration and reliability (i.e., stability) assessment are equally important to evaluate the usefulness of ML models (Riley et al., 2023;

Van Calster et al., 2019). The estimated optimism-corrected mean calibration (calibration-in-the-large) was close to 1 (0.89 to 0.99) only for the *XBG* models, while all the other models provided a mean probability of injury 1.4 to 15 times larger than the actual prevalence of injury events (3.28%) in the dataset, signifying poorer calibration (Figure 2 and Supplemental data D). Similarly, optimism corrected calibration slopes were close to 1 (0.86 to 1.45) mainly for the *XGB* models, while most of the other models had smaller slope values, highlighting a systematic overestimation of the probability of injury (Figure 2 and Supplemental data D). By assessing calibration visually through instability plots, it was evident that, first, models mostly overestimated the probability of injury, except for the *XGB* algorithm, and second, there was great instability within each model, which is a limiting factor against its generalised use (Figure 4 and Supplemental data E).

Features' importance

Finally, it should be underlined that despite the relatively good predictive performance of some of the developed models (largest optimism-corrected AUROC mean = 0.82), we could not infer any causative relationships between the available features and injury risk. Despite past season injury history and body mass were highly important among all algorithms, in general, the importance of each feature was dependent on the algorithm used. Indeed, we found an extreme variability of the importance ranking of most features, with an average range of 43 positions, out of a maximum possible range of 71 positions (Table 2). For example, the moving average of the weekly training volume of the past 3 weeks (*wk_train_hr_roll_mean*) was the 2nd most important feature in the *DT* algorithm but only the 32nd in the *RBF-SVM* algorithm. In other words, each algorithm learned to predict injuries by relying on different features, even when all features were available. These results indicate that features' importance derived from ML models should probably not be used as a method for identifying risk factors of injuries (Markus et al., 2022).

Predictive performance

Although not part of our aims, we can notice a high variability among the predictive performances of all the developed models (Figure 2). For example, focusing on the AUROC scores, *LOG*, *LIN-SVM*, and most of *XGB* algorithms did not perform much better than mere guessing (AUROC $\leq$ 0.6) while *ADA*, *RF*, and some *DT* models scored the highest (AUROC$\approx$ 0.8). This can imply a general limitation in our dataset to successfully discriminate the injury cases based on the available self-reported data and that perhaps more objective and diverse data sources could improve performance (Lövdal et al., 2021; Lyubovsky et al., 2022). However, our results were similar to most of other studies in this field (Bullock et al., 2022; Leckey et al., 2024).

Strengths & weaknesses

The examination of the research questions with seven different ML algorithms and the use of bootstrapping as internal validation technique give a firm basis for our conclusions. Furthermore, this is to our knowledge the first study on sports injury prediction modelling that is being reported with the new TRIPOD+AI guidelines (Collins, Moons, et al., 2024).

One limitation is the relatively few injury cases in our dataset (149) which resulted in great class imbalance (3.28%). Although this could be interpreted as an extra layer of difficulty for our models, it is also promising of higher performance provided more injury cases. Besides, we did not choose to synthetically improve this class imbalance, as this could harm the performance and the calibration of our models (van den Goorbergh et al., 2022). An inherent limitation of our dataset was the weekly frequency of data collection, which may have biased the reporting of monitoring variables and also may have mitigated the time-dependency of our data (e.g., the exact day of injury and the days lost were not accurately represented). Also, our dataset did not differentiate among different types of injuries, which may have added noise during model development, given the potentially different causality of different injuries. Lastly, we could not know to what extent the present findings can be generalised to other sports injury prognostic models involving different populations, sports, and/or datasets.

Future perspectives

Given the aims of this study, we cannot recommend the usability of these developed models by athletes and their entourage. However, we showed that self-reported data has potential for developing prognostic models of injury. Future studies could explore the usability of such models, and especially in the challenging cases where there is frequent missing data. Also, the possible benefits of daily reporting frequency compared to weekly as well as ways to improve the calibration and stability of such models could be investigated. Another direction could be to understand how such models could be better integrated in actual sport contexts and facilitate daily practice of athletes and their entourage while limiting

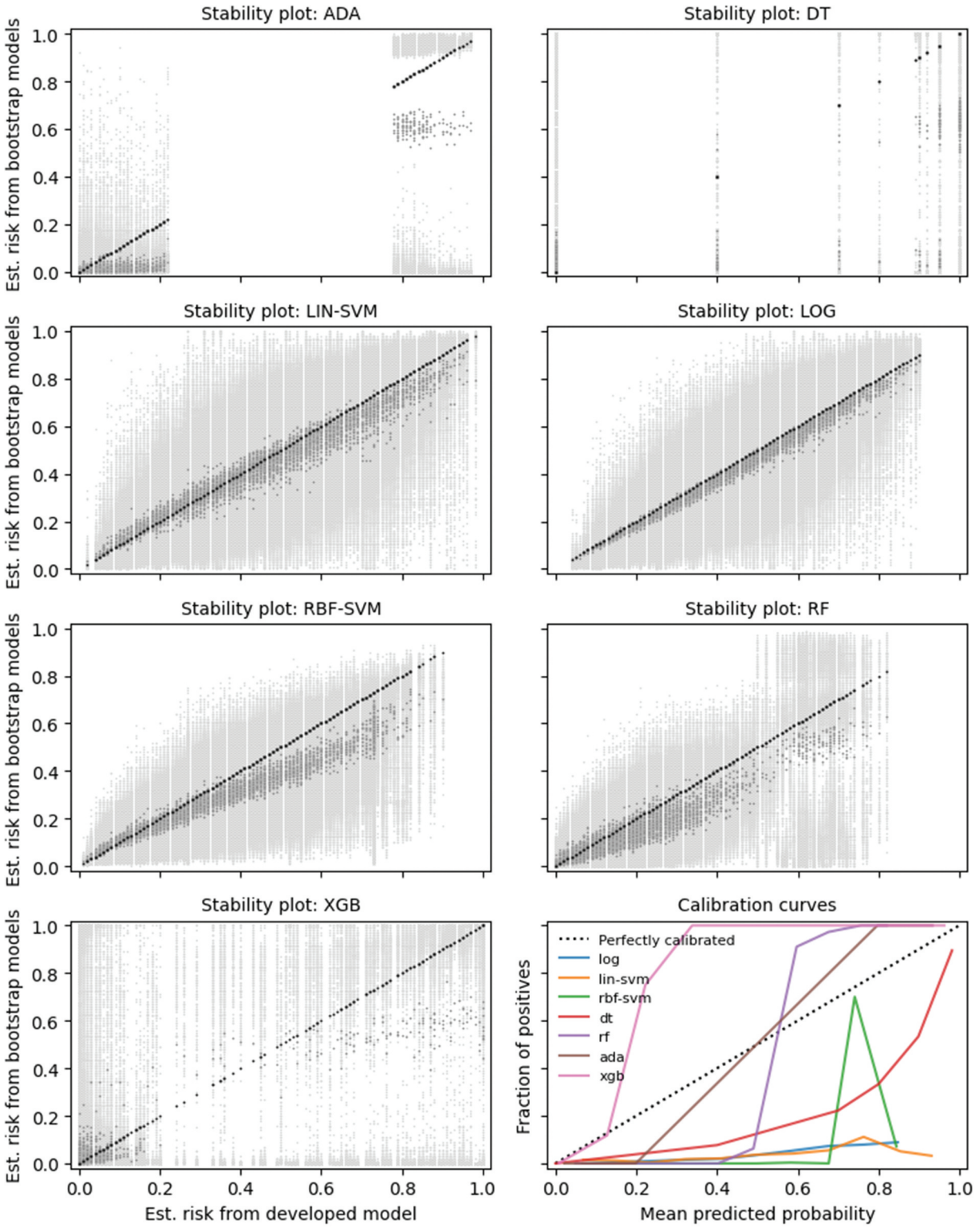

Figure 4. Instability plot and calibration curve of each ML algorithm using all 72 features. The light-grey points signify the estimated probability of each of the 200 bootstrap-generated models against the estimated probability of the original model (black dots) for the same case. Darker gray colour signifies the average probability of all bootstrap samples for a given case.

exposure to possible hazards associated with their use (Dandrieux et al., 2024).

Practical recommendations

Based on our results, it seems important that developing ML models for sports injury prognostic modelling should employ different categories of features, such as both monitoring and screening features. Certain tree-based and ensemble algorithms tend to have higher discrimination performance than other methods, but we advise that different algorithms should be tried, since which is the best one can be feature-dependent and data-specific. Additionally, feature engineering could be incorporated mainly to highlight certain associations between features and not so much to reduce the dimensionality (i.e., number of predictors) of the dataset. For this, it seems that common regularisation techniques can handle large number of features without a deterioration in discrimination performance. Researchers and sport scientists working on injury prognostic modelling should be aware that the usefulness of a particular feature/predictor is model-dependent and that knowledge on which features work better in one setting does not generalise to other settings. Thus, a trial-and-error approach should be adopted at all stages of model development.

Conclusion

Developing injury prediction ML models based on self-reported data in the sport of athletics showed promising results. The predictive performance was influenced by both the quality and the quantity of features, and researchers should take both of these into account when developing sports injury prognostic models. Monitoring features seemed to outperform baseline features, while feature engineering strategies in this dataset did not show a clear advantage, although their benefits may be more dependent on the algorithm used. Larger numbers of features showed to improve performance and not be limited by the number of injury cases.

Acknowledgments

The authors highly appreciate the cooperation of the athletes who participated in this study. They would like to thank Pedro Branco, Florian Celli, Yann Celli, Joris Chapon, Emilien Chave, Emmanuelle Chazal, Emmanuelle Cugy, Benoit Delattre, Romain Delattre, Frédéric Depiesse, Florent Derail, Romain Dolin, Anthony Leclair, Nicolas Morel, Damien Oliveras, Agathe Salque, Romain Slotala, Romain Vernede, for their help in developing the Athletics Injury Prevention Programme, Marie Peurrière and Laurie Sahuc for their help in writing the document for the Committee for the protection of persons, Pierre Gardet for developing the data collection system, Jean-Michel Serra ex-team physician of the French Athletics Federation for his support in the project, and the French Athletics Federation for sharing the questionnaire through their email lists.

Disclosure statement

PE is an Associate Editor for the British Journal of Sports Medicine, the BMJ Open Sport & Exercise Medicine, and the Scandinavian Journal of Medicine & Science in Sports. LN is an associate editor for the Journal of Sports Analytics.

Funding

No funding was received for this study.

ORCID

Spyridon Iatropoulos http://orcid.org/0000-0003-3381-2403

Data availability statement

Data are available upon reasonable request, similar to the publication of the results of the cluster-randomised controlled trial. Requests for data sharing from appropriate researchers and entities will be considered case-by-case. Interested parties should contact Pascal Edouard (pascal.edouard@univ-st-etienne.fr).

Author's contributions

PE conceived the initial study, contributed to the design of the trial, and in the data collection; All co-authors designed the present study and discussed the analyses; SI performed the analyses; SI performed the first draft of manuscript, and all co-authors contributed to the critical revision for important intellectual content and approval of the final manuscript. SI is the guarantor of the manuscript.

Ethics

This randomised controlled trial was reviewed and approved by the Committee for the protection of persons (CPP Ouest II – Angers, number: 2017-A01980-53) and was registered in the ClinicalTrials.gov (ClinicalTrials.gov Identifier: NCT03307434). This present study and analysis were reviewed and approved by the Saint-Etienne University Hospital Ethical Committee (Institutional Review Board: IORG0007394; IRBN292023/CHUSTE).

References

Briand, J., Deguire, S., Gaudet, S., & Bieuzen, F. (2022, July). Monitoring variables influence on random Forest models to forecast injuries in short-track speed skating. *Frontiers in Sports and Active Living*, *4*. https://doi.org/10.3389/fspor.2022.896828

Bullock, G. S., Mylott, J., Hughes, T., Nicholson, K. F., Riley, R. D., & Collins, G. S. (2022). Just how confident can we be in predicting sports injuries? A systematic review of the methodological conduct and performance of existing musculoskeletal injury prediction models in sport. *Sports Medicine*, *52*(10), 2469–2482. https://doi.org/10.1007/s40279-022-01698-9

Carey, D. L., Ong, K., Whiteley, R., Crossley, K. M., Crow, J., & Morris, M. E. (2018). Predictive modelling of training loads and injury in Australian football. *International Journal of Computer Science in Sport*, *17*(1), 49–66. https://doi.org/10.2478/ijcss-2018-0002

Claudino, J. G., de Oliveira Capanema, D., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., & Nassis, G. P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. *Sports Medicine - Open*, *5*(1). https://doi.org/10.1186/s40798-019-0202-3

Cohan, A., Schuster, J., & Fernandez, J. (2021). A deep learning approach to injury forecasting in NBA basketball. *Journal of Sports Analytics*, *7*(4), 277–289. https://doi.org/10.3233/jsa-200529

Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G. M., van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): From development to external validation. *BMJ*, *384*, e074819. https://doi.org/10.1136/bmj-2023-074819

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., & Wynants, L. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, e078378. https://doi.org/10.1136/bmj-2023-078378

Dandrieux, P.-E., Navarro, L., Blanco, D., Ruffault, A., Ley, C., Bruneau, A., Iatropoulos, S., Chapon, J., Hollander, K., & Edouard, P. (2025). Association between the use of daily injury risk estimation feedback (I-REF) based on machine learning techniques and injuries in athletics (track and field): Results of a prospective cohort study over an athletics season. *BMJ Open Sport & Exercise Medicine*, *11*(1), e002331. https://doi.org/10.1136/bmjsem-2024-002331

Dandrieux, P.-E., Navarro, L., Chapon, J., Tondut, J., Zyskowski, M., Hollander, K., & Edouard, P. (2024). Perceptions and beliefs on sports injury prediction as an injury risk reduction strategy: An online survey on elite athletics (track and field) athletes, coaches, and health professionals. *Physical Therapy in Sport*, *66*, 31–36. https://doi.org/10.1016/j.ptsp.2024.01.007

Edouard, P., Cugy, E., Dolin, R., Morel, N., Serra, J.-M., Depiesse, F., Branco, P., & Steffen, K. (2020). The athletics injury Prevention Programme can help to reduce the occurrence at short term of participation restriction injury complaints in athletics: A prospective cohort study. *Sports*, *8*(6), 84. https://doi.org/10.3390/sports8060084

Edouard, P., Dandrieux, P.-E., Iatropoulos, S., Blanco, D., Branco, P., Chapon, J., Mulenga, D., Guex, K., Guilhem, G., Jacobsson, J., Mann, R., McCallion, C., Mosser, C., Morin, J.-B., Prince, C., Ruffault, A., Timpka, T., Alonso, J.-M., & Navarro, L. (2024, June). Injuries in athletics (track and field): A narrative review presenting the current problem of injuries. *Deutsche Zeitschrift für Sportmedizin*, *75*(4), 132–141. http://dx.doi.org/10.5960/dzsm.2024.601

Edouard, P., Steffen, K., Peuriere, M., Gardet, P., Navarro, L., & Blanco, D. (2021). Effect of an unsupervised exercises-based athletics injury Prevention Programme on injury complaints leading to participation restriction in athletics: A cluster-randomised controlled trial. *International Journal of Environmental Research and Public Health*, *18*(21), 11334. https://doi.org/10.3390/ijerph182111334

Eetvelde, H. V., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. *Journal of Experimental Orthopaedics*, *8*(1). https://doi.org/10.1186/s40634-021-00346-x

Goggins, L., Warren, A., & Osguthorpe, D., (2022). Detecting injury risk factors with algorithmic models in elite women's pathway cricket 1. *International Journal of Sports Medicine*, *43*(4), 344–349. https://doi.org/10.1055/a-1502-6824

Hecksteden, A., Schmartz, G. P., Egyptien, Y., Aus der Fünten, K., Keller, A., & Meyer, T. (2023). Forecasting football injuries by combining screening, monitoring and machine learning. *Science and Medicine in Football*, *7*(3), 214–228. https://doi.org/10.1080/24733938.2022.2095006

Huang, Y., Huang, S., Wang, Y., Li, Y., Gui, Y., & Huang, C. (2022). A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning. *Frontiers in Physiology*, *13*. https://doi.org/10.3389/fphys.2022.937546

Jauhiainen, S., Kauppi, J.-P., Krosshaug, T., Bahr, R., Bartsch, J., & Äyrämö, S. (2022). Predicting ACL injury using machine learning on data from an extensive screening test battery of 880 female elite athletes. *The American Journal of Sports Medicine*, *50*(11), 2917–2924. https://doi.org/10.1177/03635465221112095

Jung, A. (2022). Regularization. In *Machine learning* (pp. 135–151). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-8193-6_7

Leckey, C., van Dyk, N., Doherty, C., Lawlor, A., & Delahunt, E. (2024). Machine learning approaches to injury risk prediction in sport: A scoping review with evidence synthesis. *British Journal of Sports Medicine*, *59*(7), 491–500. https://doi.org/10.1136/bjsports-2024-108576

López-Valenciano, A., Ayala, F., PUerta, J. M., De Ste Croix, M. B. A., Vera-Garcia, F. J., Hernández-Sánchez, S., Ruiz-Pérez, I., & Myer, G. D. (2018). A preventive model for muscle injuries. *Medicine and Science in Sports and Exercise*, *50*(5), 915–927. https://doi.org/10.1249/MSS.0000000000001535

Lövdal, S. S., Hartigh, R. J. R. D., & Azzopardi, G. (2021). "Injury prediction in competitive runners with machine learning". in. *International Journal of Sports Physiology & Performance*, *16*(10), 1522–1531. https://doi.org/10.1123/IJSPP.2020-0518

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, & S. Bengio (Eds.), Advances in neural information processing systems 30 (pp. 4765–4774). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Lyubovsky, A., Liu, Z., Watson, A., Kuehn, S., Korem, E., & Zhou, G. (2022). A pain free nociceptor: Predicting football injuries with machine learning. *Smart Health*, *24*, 100262. https://doi.org/10.1016/j.smhl.2021.100262

Mandorino, M., Figueiredo, A. J., Cima, G., & Tessitore, A. (2022). Predictive analytic techniques to identify hidden relationships between training load, fatigue and muscle strains in young soccer players. *Sports*, *10*(1), 3. https://doi.org/10.3390/sports10010003

Markus, A. F., Rijnbeek, P. R., & Reps, J. M. (2022). Why predicting risk can't identify 'risk factors': Empirical assessment of model stability in machine learning across observational health databases. In Z. Lipton, R. Ranganath, & M. Sendak (Eds.), *Proceedings of the 7th machine learning for healthcare conference proceedings of machine learning research* (Vol. 182, pp. 828–852). PMLR. https://proceedings.mlr.press/v182/markus22a.html

Meeuwisse, W. H., Tyreman, H., Hagel, B., & Emery, C. (2007). A dynamic model of etiology in sport injury: The recursive nature of risk and causation. *Clinical Journal of Sport Medicine*, *17*(3), 215–219. https://doi.org/10.1097/JSM.0b013e3180592a48

Oliver, J. L., Ayala, F., De Ste Croix, M. B. A., Lloyd, R. S., Myer, G. D., & Read, P. J. (2020). "Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *Journal of Science and Medicine in Sport*, *23*(11), 1044–1048. https://doi.org/10.1016/j.jsams.2020.04.021

The Pandas Development Team. (2020). pandas-dev/pandas: Pandas. *Zenodo*. https://doi.org/10.5281/zenodo.3509134

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12* (85), 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Riley, R. D., Pate, A., Dhiman, P., Archer, L., Martin, G. P., & Collins, G. S. (2023). Clinical prediction models and the multiverse of madness. *BMC Medicine*, *21*(1), 502. https://doi.org/10.1186/s12916-023-03212-y

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., Lenoir, M., D'HONDT, E., & Witvrouw, E. (2020). A machine learning approach to assess injury risk in elite youth football players. *Medicine and Science in Sports and Exercise*, *52*(8), 1745–1751. https://doi.org/10.1249/MSS.0000000000002305

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLOS ONE*, *13*(7), e0201264. https://doi.org/10.1371/journal.pone.0201264

Ruddy, J. D., Shield, A. J., Maniar, N., Williams, M. D., Duhig, S., Timmins, R. G., Hickey, J., Bourne, M. N., & Opar, D. A. (2018). Predictive modeling of hamstring Strain injuries in elite Australian footballers. *Medicine and Science in Sports and Exercise*, *50*(5), 906–914. https://doi.org/10.1249/MSS.0000000000001527

Ruiz-Pérez, I., López-Valenciano, A., Hernández-Sánchez, S., Puerta-Callejón, J. M., De Ste Croix, M., Sainz de Baranda, P., & Ayala, F. (2021). A field-based approach to determine soft tissue injury risk in elite futsal using novel machine learning techniques. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.610210

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. IN *9th Python in Science Conference*, Austin, Texas, June 28–July 3.

Vallance, E., Sutton-Charani, N., Imoussaten, A., Montmain, J., & Perrey, S. (2020). Combining internal- and external-training-loads to predict non-contact injuries in soccer. *Applied Sciences (Switzerland)*, *10*(15), 5261. https://doi.org/10.3390/APP10155261

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The achilles heel of predictive analytics. *BMC Medicine*, *17*(1). https://doi.org/10.1186/s12916-019-1466-7

van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, *29*(9), 1525–1534. https://doi.org/10.1093/jamia/ocac093

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2018). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, *28* (8), 2455–2474. https://doi.org/10.1177/0962280218784726

Zhang, G., Brink, M., Aus der Fünten, K., Tröß, T., Willeit, P., Meyer, T., Lemmink, K., & Hecksteden, A. (2024). The time course of injury risk after return-to-play in professional football (soccer). *Sports Medicine*, *55*(1), 193–201. https://doi.org/10.1007/s40279-024-02103-3

Zheng, A. (2018). *Feature engineering for machine learning*. O'Reilly Media.