University of Essex

Research Repository

Using Machine Learning to Explain Recreational Running Injury: An Exploratory Classification Analysis

Bradley S. Neal, Queen Mary University of London/University of Essex

Stella Hadjiantoni, University of Essex

Allison H. Gruber, HH Morris Human Performance Laboratories, School of Public Health—Bloomington, Bloomington, IN, USA

Richard W. Willy, Ohio State University, Columbus, OH, USA

Bernard X.W. Liew, University of Essex

Accepted for publication in the Journal of Sport Rehabilitation

Research Repository link: https://repository.essex.ac.uk/44122/

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the published version if you wish to cite this paper.
https://doi.org/10.1123/jsr.2026-0073

www.essex.ac.uk

University of Essex

Abstract

Context: Running is good for overall health but has a poorly understood risk of injury. Wearable technology and machine learning (ML) offer solutions to this problem. Using ML, we aimed to determine a) the key-important factors that increase the probability of classifying a run as belonging to an injured group, b) the best performing ML model, C) the sample size required for future substantiative cohorts.

Design: Exploratory classification analysis

Methods: Heart rate/Inertial measurement unit/global positioning system data were extracted from the wristwatches of 88 recreational runners (65 injured, 23 uninjured) during a 12-week prospective cohort. Data were extracted from each run completed by participants and each run was considered an independent event. Variance inflation factor was used to check for collinearity, and the robust normaliser was used to scale numerical features. Seven ML models were used and validated using repeated stratified 10-fold cross validation. Accuracy, precision, and recall were used to evaluate ML model performance and SHapley Additive exPlanations to determine the relative importance of the individual factors.

Results: The final dataset had 29 numeric factors, with 4,758 run instances from 80 participants. Participants completed between 6 and 148 (mean 55.5) runs between 02/11/2022–05/02/2023. The four best performing ML models were CatBoost, GMB, Light GBM, and GXBoost (average model performance >0.90). CatBoost was the best performing model; 96% correct in classifying runs from the injured groups. Critical power was the most important factor across all four algorithms, with higher values increasing the probability of classifying a run as belonging to the injured group.

Conclusion: Future cohort studies using ML to explain RRI development should use boosting ML models, specifically CatBoost, and include critical power. We advise against interpreting these results as being a causal explanation to injury development and model performance may be overestimated due to our validation approach.

Key words: Epidemiology, prospective, feasibility, wearable technology

Key points

1. Machine learning models involving boosting demonstrate appropriate performance for classifying runs as belonging to an injured group
2. Factors reflecting running intensity consistently aided in correct classification

Introduction

There is an ongoing conflict between the reduced risk of premature mortality[1] and high risk of running-related injury (RRI)[2, 3] offered by recreational running. This paradox is partly driven by the scarcity of prospective studies investigating the aetiology of RRI,[4, 5] which limits the development of injury prevention strategies underpinned by robust evidence. Existing prospective cohort studies have sought to explore the association between anatomical and biomechanical factors,[4] psychological and behavioural factors,[6-8] and training load factors,[9, 10] and future RRI. The logistical challenges of recruiting an adequate number of participants for in-lab data collection are significant, with studies often limited by small samples (i.e., <100) and low numbers of events (injuries) and time points per variable.[11]

Wearable technology presents a potential solution to the logistical challenges of large prospective studies, as up to 88% of recreational runners use devices equipped with heart rate monitors (HRM), inertial measurement units (IMUs) and global positioning system (GPS) technology.[12, 13] A recently published feasibility study[14] demonstrated that wearable technology is a feasible method for the prospective collection of multiple key variables theoretically linked to RRI development. Beyond establishing feasibility, the study also identified a significant association was identified between acute load by calculated effort and subsequent RRI.[14] Other research groups have used wearable technology to classify how runners with and without a history of knee surgery differ in their training habits[15], and larger cohorts are planned to explore the relationship between training exposure and joint health[16].

Many prospective cohorts designed to explore the development of RRI are limited to analysing point-observational data (e.g., baseline training load) and a single incident (i.e., injury report). Whilst this approach is beneficial when strong linear relationships exist (e.g., smoking and lung cancer), it is likely that RRI result from the interactions of several complex factors.[17] IMU/GPS devices allow for on-going data collection each time a recreational runner completes a workout, providing data that can subsequently be analysed as a time series. Unlike traditional regression approaches, machine learning (ML) can model patterns

between multiple factors, even in the presence of complex non-linearity.[18] Sports scientists employ ML at an increasing rate, particularly when high volumes of athlete-specific data are available.[19]

We therefore investigated if there are key important factors that correctly classify a run as belonging to the injured versus the uninjured group by utilising ML. Identifying these factors should inform future hypothesis-driven studies designed to test causal associations with RRI development. We also aimed to provide a sample size calculation for future substantiative prospective cohort studies using ML. We did this by re-analysing an existing longitudinal dataset collected as part of a previously published feasibility study.[14]

Methods

Study design and ethics

We conducted a secondary exploratory analysis of previously published prospective observational feasibility data.[14] Ethical approval was obtained from the xxxx of xxxx Ethics Committee (ETH-2122-1352).

Participants

Recreational runners were eligible if they met a series of specific criteria, described in detail in the published feasibility study.[14] After self-declaring their eligibility, reporting their gender (Man; Woman; Prefer not to say; I am not represented), height (cm), and mass (kg), and providing written informed consent, participants followed a customised link to create an account with DashLX™ (Flagstaff, Arizona, USA), which linked their HRM/IMU/GPS wristwatch to a repository. DashLX™ employs a proprietary, manufacturer-agnostic analysis platform to integrate running data on environmental conditions, heart rate, training profiles, running speed and distance, and spatiotemporal metrics. The DashLX™ system is fully compliant with the General Data Protection Regulations and the California Consumer Privacy Act.

Participants were then instructed to continue their typical running routine and were sent a maximum of 12 consecutive weekly injury surveillance emails (Qualtrics, Seattle, USA), starting from day seven after study enrolment. Participants completed the published feasibility study[14] either by reporting an RRI and entering the injured group, or by responding to 75% (i.e., 9/12) emails without reporting a RRI and entering the uninjured group. For the purpose of this study, an RRI was operationally defined as an episode of pain stopping or limiting three consecutive runs or persisting for seven days (whichever was sooner), or having sought the advice of a medical professional.[20] Injury surveillance emails ceased if participants reported an RRI[14] or completed 12 consecutive weeks without reporting an RRI. This resulted in a sample of 88 participants (65 uninjured runners, 23 injured runners).

Demographic data are reported in Table 1. The most common injury sites were the ankle (n=5) and foot (n=5), followed by the knee (n=4), hip (n=4), low back (n=1), low back and leg (n=1), lateral thigh (n=1), medial tibia (n=1), and global lower limb (n=1).

[TABLE 1 HERE]

Data analysis

A ML approach was employed to investigate which factors increased the probability of correctly classifying a run as belonging to the injured group.

Data pre-processing

The dependent variable was injury status (uninjured versus injured). Every run from a given participant was labelled according to that participant's final status (i.e., uninjured or injured), as it was not possible to determine when individual participants became injured, but only the week in the original cohort when their injury was reported. Each time a participant completed a running workout during the maximum 12-week surveillance period, 123 unique factors were extracted from their HRM/IMU/GPS wristwatch and uploaded to the DashLX data repository (see supplementary file one).

Six of these factors were condensed into three by computing the change from a start and end value, and 26 factors were determined to be irrelevant to the research question (i.e., device vendor) were manually removed. 25 factors were subsequently removed due to a high percentage of missing values (e.g., average heart in zone one whilst walking), and a further 25 factors were removed by the low-pass filter threshold within the DashLX^{TM} algorithm entering zero values (e.g., total distance walking in heart rate zone one). Individual participants with a high percentage of missing values, defined as $\geq$60%, were also removed.

A simple within participant imputation technique was implemented for numerical factors in which missing values were <60%. Missing values were replaced with the mean for all the numerical factors that appeared to be affected less by outliers upon visual inspection (28 factors), and the median was used for factors where distributions appeared to be skewed (5

values). The variance inflation factor was calculated to detect highly collinear factors that need to be considered for building classification models. Fifteen factors were removed to tackle collinearity (see supplementary file one). The final dataset consisted of 29 numeric factors within the domains of heart rate, spatiotemporal, run duration, kinetic, and environmental variables. To enable different factors of different scales to have an equal opportunity of being selected in the model, the robust normaliser was used to scale the numerical factors ([x-median]/[inter quartile range]).

ML algorithms

XGBoost and RF are amongst the top performing ML models for injury risk in sport[19], with GBM, LightGBM, and CatBoost also previously reported in the literature.[21] Seven ML algorithms were implemented to align with the exploratory design of this study: Adaptive boosting (AdaBoost),[22] Gradient Boosting Machine (GBM),[23] Light Gradient Boosting Machine (LightGBM),[24] eXtreme Gradient Boosting (XGBoost),[25] CatBoost,[26] Random Forest (RF),[27] linear discriminant analysis,[28] and binomial logistic regression was tested as a baseline. Justification and strengths/weaknesses of the tested ML models are detailed in Table 2.

[TABLE 2 HERE]

Modelling validation

For model validation, a repeated (10 repeats) stratified 10-fold cross-validation was used given the imbalanced data set (i.e., a greater proportion of uninjured runners), to preserve the same proportion of uninjured and injured runners as the original dataset. As this study represents a secondary analysis, we decided *a priori* that the cohort of experienced runners were clearly targeting specific training elements when completing a run workout (e.g., base endurance versus threshold training), with high variability in distance, time, and effort metrics. We therefore felt it was most appropriate to consider each run to be an independent event as opposed to clustering runs within participant. For each of the 10 folds, the average and standard deviation of the extracted performance metrics were quantified.

To evaluate the performance of the seven ML models, five measures that assess the quality of fit of the produced models were used: accuracy (correct classifications/all classifiations), precision (the proportion of true positives against all classified [true and false] positive [i.e. injured] cases), recall (the proportion of true positives against all actual positive cases), F1-score (2 * Precision * Recall/Precision + Recall), and area under the receiver operator curve (AUC). Accuracy ranges from 0 (no correct classification) to 1 (perfect classification). Precision and recall both take values between 0 and 1, with 1 reflecting zero false positives and a 100% detection rate, respectively. The F1-score takes values between 0 and 1, with 1 indicating perfect classification performance. AUC ranges from 0 to 1, with 1 reflecting when the model can perfectly distinguish between all injured and uninjured cases correctly, 0.5 when the model cannot distinguish the classes, and 0 being when the model is perfectly incorrect in its discrimination. Practically, a model with a high precision is good at not making false positive classifications, whilst a model with a high recall is good at finding all positive cases. An average of all five evaluation measures was calculated to reflect overall model performance.

Factor importance

To determine the relative importance of factors in the performance of an ML model, the SHapley Additive exPlanations (SHAP) was used.[29] SHAP values (i.e., log-odds) quantify how each factor influences the model's probability that the observation (i.e., the run) belongs to an injured runner (i.e., the factor's relative importance to the model), and the magnitude (i.e., relative importance) and direction (i.e., increases or decreases the probability) of the contribution. SHAP values are not a predictor metric inferring causality to a factor being a risk factor for injury. The 20 most important factors for the four best-performing algorithms are represented on a SHAP summary plot, with the y-axis representing the factors and the x-axis representing the SHAP values (Figures 1a-d). Factors that increase the model's probability of classifying the run to an injured runner are shown on the right-hand side of zero on the x-axis. For each factor in the dataset, the raw value for every run is represented

by a coloured dot (red=high values; blue=low values), and the horizontal position represents the SHAP value for that specific factor.

To summarize the magnitude of the influence for a given factor's contribution to the model, violin plots showing the distribution of absolute SHAP values were produced for the 10 most important factors of the four best-performing models (Figures 2a-d). The violin plots were generated using a Kernel density estimator, and 95% confidence intervals for the mean absolute SHAP values are provided in a table within each subplot. For each model and factor, the distribution of the top-ranking factors (i.e., those with the largest mean absolute SHAP value) are located higher on the y-axis; factors with the largest mean absolute SHAP value appear farther from zero. The lower a factor's impact on the model, the closer its position to zero on the y-axis.

Sample size calculation

Using guidance from Riley et al.,[30] we used three individual variations of sample size estimate equations to calculate the required sample for a future substantiative cohort to develop a stable multivariable classification model using our final set of factors (n=29) and RRI prevalence (26%). The Cox-Snell R^2 statistic (R^2_{cs}) was estimated using the calculated AUC values from our ML models and shrinkage from 0.8 to 0.95 was considered. The largest required sample from the three equations has been presented as per this guidance.

Results

Seven participants (four uninjured and three injured) were removed from the final analysis for having >60% total data missing from their run workouts. One uninjured participant was removed as they completed only a single workout during the 12-week cohort period. The final dataset therefore contained data from 60 uninjured and 20 injured runners (n=80), with 4,758 run instances. The number of run workouts completed by each participant varied between 6 and 148 (mean 56) and spanned the period 02/11/2022 – 05/02/2023.

Table 3 shows the best-performing algorithms using the Accuracy, Precision, Recall, F1, and AUC scores. The four best overall performing algorithms were CatBoost, GMB, Light GBM, and XGBoost, each with an average model performance of >0.90. The best performing algorithm was CatBoost, which was 96% correct in its classification of injury development. The two worst overall-performing algorithms were LDA and LR, with an average performance across all metrics of ~0.4. CatBoost had the best accuracy (0.96 ±0.01), recall (0.87 ±0.02), F1-score (0.92 ±0.03), and AUC (0.93 ±0.02); whilst AdaBoost had the best precision (0.98 ±0.03).

Figures 1a-d illustrate how the 20 most important factors for the four best-performing ML models impact the output using SHAP values (log-odds). Critical power was the most important factor across all four algorithms, indicating that it contributed more than other factors to the models' ability to classify a run as belonging to the injured group. Higher (red) raw critical power values were generally associated with positive SHAP values reflecting an **increase** in the predicted log-odds of classifying a run as belonging to the injured group. Consistent with its greatest level of importance, critical power produced the largest changes in log-odds with mean absolute SHAP values ranging from 0.84 (95%CI 0.84, 0.85) in the XGBoost model to 1.35 (95%CI 1.34, 1.37) in the CatBoost model (Figure 2).

[FIGURE 1 HERE]

Average vertical stiffness (kvert) was the second most important factor across all four algorithms, with higher (red) raw values being generally associated with negative SHAP

values reflecting a **reduction** in the predicted log-odds of classifying a run as belonging to the injured group. Mean absolute SHAP values ranged from 0.52 (95%CI 0.52, 0.53) in the XGBoost model, to 0.88 (95%CI 0.87, 0.89) in the CatBoost model (Figure 2).

Average stride time was the third most important factor across all four algorithms, with lower (blue) raw stride time being generally associated with positive SHAP values reflecting an **increase** in the predicted log-odds of classifying a run as belonging to the injured group. Mean absolute SHAP values ranged from 0.28 (95%CI 0.27, 0.28) in the XGBoost model to 0.60 (95%CI 0.60, 0.61) in the CatBoost model (Figure 2).

Figures 2a-d illustrate the distribution of the absolute SHAP values of the top 10 most important factors from the four best-performing algorithms. Factors with larger mean absolute SHAP values (e.g., critical power) are located further from zero on the y-axis, indicating that they produce larger chances in the model's predicted log-odds of classifying runs as belonging to the injured group.

[FIGURE 2 HERE]

Sample size requirement

The equation $n = P/((S-1)*ln(1-(R^2_{cs}/S))$ consistently resulted in the largest required sample, where P is the number of predictors, S is the shrinkage, $\ln$ denotes the natural logarithm, and R^2_{cs} is the Cox-Snell R^2 statistic. Dependent on shrinkage and AUC (to estimate R^2_{cs}) values from our four top performing ML models, between 629 (S=0.8; AUC=0.93) and 3,145 (S=0.95; AUC=0.76) participants would be required for a future substantiative cohort. Figure 3 demonstrates the required number of participants relative to a range of shrinkage and AUC values.

[FIGURE 3 HERE]

Discussion

We employed ML to a) identify key important factors associated with correctly classifying a run as belonging to a member of the injured group b) the best performing ML model and c) provide a sample size calculation for developing a future classification model. We identified four ML algorithms with acceptable performance, from which the best performing model (CatBoost) was overall 93% correct in its classifications. We decided *a priori* to analyse each run as an independent event and this, along with the low sample size, may explain the overall high performance for the tested boosting models. Between 629 and 3,145 participants would be required for a future substantiative cohort.

The four top performing ML algorithms were all boosting-based methods, which are well suited for multifactorial data with non-linear relationships and interactions. In contrast, the poorest performance occurred in the two linear models, which primarily assume linear associations between exposures and outcomes. The boosting algorithms' ability to combine information from previously trained weak classifiers, learning and correcting from previous models errors, can improve classification accuracy even with imbalanced datasets. The success of the boosting algorithms aligns with the earlier work of Lövdal et al.,[18] who used the XGBoost model to predict RRI in elite Dutch track runners, reporting AUC values of between 0.68-0.72. This model type is important for future ML studies designed to predict RRI onset to consider, with variable injury incidence rates[31] likely to lead to moderately imbalanced datasets (i.e., 75% uninjured versus 25% injured runners). Such a substantiative cohort would need to recruit 941 to 1,152 participants with the recommended shrinkage of >0.80[30] and an AUC of 0.90-0.93.

It is important to distinguish that SHAP values indicate the most influential factors that increase the probability of correctly classifying a run as belonging to a member of the injured group. SHAP values are not a predictor metric inferring causality to a factor being a risk factor for injury. Rather, factors that consistently improved the model's classification probability, including critical power, vertical stiffness, and stride time, should be interpreted

as possible targets for hypothesis-driven studies designed to test causal relationships to injury, rather than evidence of injury mechanisms at this stage.

High critical power was the most important factor linked to correctly classifying runs as belonging to the injured group in all four top performing ML models, with larger critical power values increasing the model's probability of correctly classifying the run. This result builds on the initial analysis of this dataset, where a significant association between acute load by calculated effort (of which critical power is a component) and subsequent RRI was identified[14]. Critical power reflects an athlete moving from aerobic to anaerobic metabolism,[32] and a longitudinal increase therefore reflects a runner gaining fitness. The increased training intensity required to gain fitness may contribute to injury development, and this result again aligns well with the findings of Lövdal et al.,[18] who reported that the number of kilometres completed in heart rate zone five (i.e., with high critical power) was the most important factor for RRI classification in their model. It could also be that critical power simply differs systematically between our injured and uninjured groups. These findings do not confirm that critical power is causally associated with injury, but rather that it is a promising variable for future hypothesis-driven, causal investigations.

High vertical stiffness was the second most important factor that helped the model distinguish between runs from the injured versus uninjured groups, with higher values associated with a lower probability of the model classifying the run as belonging to the injured group. Vertical stiffness is defined as the ratio of the peak vertical ground reaction force and the vertical displacement of the centre of mass between foot strike and its minima during stance.[33] Malisoux et al.,[34] have previously reported no prospective relationship between vertical stiffness and RRI development in a large cohort of 848 runners. This conflict may be explained by Malisoux et al., measuring vertical stiffness using an instrumented treadmill as opposed to being calculated from IMU data as with our study. It could also be that high vertical stiffness is not associated with RRI development. Our results

suggest future cohort studies with appropriate sample sizes and follow-up length should continue to explore if vertical stiffness calculated from IMU data predicts RRI onset.

Low average stride time was the third most important factor across models, with lower stride time values generally associated with a higher predicted probability of classifying runs as belonging to the injured group. A low average stride time is consistent with high critical power being the most important factor in our models, as higher critical power is likely to equate to higher running speeds, and there is an inverse relationship between speed and stride time (i.e., as speed increases, stride time decreases).[35] This phenomena is also more likely to occur in more experienced runners,[35] and the cohort included in this study are experienced given the nature of the original inclusion criteria. Average stride time should be a factor of interest in future cohort studies but may be highly colinear with running speed and intensity.

Strengths & limitations

Whilst similar methods have been employed in other sports, we believe that this is the first attempt to use ML to classify RRI status using a HRM/IMU/GPS dataset. Whilst we initially included a high number of factors given our exploratory design, we took a systematic approach to reducing them based on plausibility, percentage of missing values, in-built low-pass filters, and collinearity. Whilst the eventual 29 factors clearly breaches the 10 participants per factor 'rule of thumb',[36] retaining this number of factors has allowed a detailed analysis of importance to guide future research, while recognising that this choice may have implications for model robustness.

This is an exploratory analysis of previously collected data, and we advocate caution in making robust inferences on causality, which SHAP values cannot provide. Whilst this dataset includes many run instances, the overall sample size is small, and so we encourage the reader to remember that our aim was exploratory rather than confirmatory. We excluded seven participants from this exploratory re-analysis due to data loss. Whilst we employed imputation for elements of missing data, these participants had >60% data loss across the

feature set often related to heartrate data. This may be explained by these participants simply disabling the heartrate feature, but three participants who had >60% total data loss all used the Apple Watch, which means this device may not be suitable for future cohort studies using HRM/GPS/IMU-derived data. The variables used in this study were derived from commercial wrist-worn devices, which may differ in accuracy from measurements obtained using foot- or shank-mounted sensors or laboratory-based motion capture[37]. Model classification performance and feature contributions should be interpreted with the limitations of commercial wrist-worn device measurements.

Cross-validation was performed at the run level rather than the participant level and we acknowledge this may inflate model classification performance due to within-participant similarity. We decided *a priori* to perform cross-validation in this way given our exploratory aim of identifying candidate factors that distinguish runs between injured and uninjured groups. It is important to emphasise that the results of this study describe how the tested ML models distinguished runs between groups in the context of run-level cross-validation and do not imply a causal or biological relationship between any factor and RRI. Higher values of a factor increasing the probability of a run being classified as belonging to the injured group does not imply that lower values increase the probability of a run being classified as belonging to the uninjured group. Our findings should be interpreted as descriptive of model classification behaviour and used to guide future hypothesis-driven studies, rather than to draw conclusions about causal or protective effects.

Conclusion

Four ML models involving boosting all demonstrated acceptable performance in correctly classifying a run as belonging to the injured group using HRM/IMU/GPS data. The CatBoost model demonstrated the best overall performance. Consistent key factors reflecting increased running intensity increased the model's probability of correctly classifying runs. Future substantiative cohorts should consider using ML models to analyse complex, non-linear data and ensure that metrics of running intensity are collected. The results from this study can inform future hypothesis-driven studies designed to test causal associations with RRI development.

Acknowledgements

The authors have no acknowledgements to make.

Conflict of interest

The authors confirm that they have no conflicts of interests to declare.

References

1. Lee DC, Brellenthin AG, Thompson PD, Sui X, Lee IM, Lavie CJ. Running as a Key Lifestyle Medicine for Longevity. *Prog Cardiovasc Dis.* 2017;60(1):45-55.

2. Fokkema T, Hartgens F, Kluitenberg B, et al. Reasons and predictors of discontinuation of running after a running program for novice runners. *J Sci Med Sport.* 2019;22(1):106-111.

3. Desai P, Jungmalm J, Borjesson M, Karlsson J, Grau S. Recreational Runners With a History of Injury Are Twice as Likely to Sustain a Running-Related Injury as Runners With No History of Injury: A 1-Year Prospective Cohort Study. *J Orthop Sports Phys Ther.* 2021;51(3):144-150.

4. Ceyssens L, Vanelderen R, Barton C, Malliaras P, Dingenen B. Biomechanical Risk Factors Associated with Running-Related Injuries: A Systematic Review. *Sports Med.* 2019;49(7):1095-1115.

5. Saragiotto BT, Yamato TP, Hespanhol Junior LC, Rainbow MJ, Davis IS, Lopes AD. What are the main risk factors for running-related injuries? *Sports Med.* 2014;44(8):1153-1163.

6. Johnston R, Cahalan R, Bonnett L, et al. General health complaints and sleep associated with new injury within an endurance sporting population: A prospective study. *J Sci Med Sport.* 2020;23(3):252-257.

7. Leon-Guereno P, Tapia-Serrano MA, Sanchez-Miguel PA. The relationship of recreational runners' motivation and resilience levels to the incidence of injury: A mediation model. *PLoS One.* 2020;15(5):e0231628.

8. Mousavi SH, Hijmans JM, Minoonejad H, Rajabi R, Zwerver J. Factors Associated With Lower Limb Injuries in Recreational Runners: A Cross-Sectional Survey Including Mental Aspects and Sleep Quality. *J Sports Sci Med.* 2021;20(2):204-215.

9. Maupin D, Schram B, Canetti E, Orr R. The Relationship Between Acute: Chronic Workload Ratios and Injury Risk in Sports: A Systematic Review. *Open Access J Sports Med.* 2020;11:51-75.

10. Nakaoka G, Barboza SD, Verhagen E, van Mechelen W, Hespanhol L. The Association Between the Acute:Chronic Workload Ratio and Running-Related Injuries in Dutch Runners: A Prospective Cohort Study. *Sports Med.* 2021;51(11):2437-2447.

11. Peduzzi P, Concato J, Kemper E, Holford T, Feinstein A. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology.* 1996;49(12):1373-1379.

12. Cloosterman KLA, Fokkema T, de Vos RJ, van Oeveren B, Bierma-Zeinstra SMA, van Middelkoop M. Feasibility and usability of GPS data in exploring associations between training load and running-related knee injuries in recreational runners. *BMC Sports Sci Med Rehabil.* 2022;14(1):78.

13. Janssen M, Walravens R, Thibaut E, Scheerder J, Brombacher A, Vos S. Understanding Different Types of Recreational Runners and How They Use Running-Related Technology. *International Journal of Environmental Research and Public Health.* 2020;17(7).

14. Neal BS, Bramah C, McCarthy-Ryan MF, et al. Using wearable technology data to explain recreational running injury: A prospective longitudinal feasibility study. *Phys Ther Sport.* 2024;65:130-136.

15. Carey DL, Girdwood M, De Oliveira Silva D, et al. See how they run: Characteristics of 114,324 runs from 227 runners with and without a history of knee surgery. *J Sports Sci.* 2026;44(7):868-879.

16. De Oliveira Silva D, Johnston RTR, Mentiplay BF, et al. Trajectory of knee health in runners with and without heightened osteoarthritis risk: the TRAIL prospective cohort study protocol. *BMJ Open.* 2023;13(2):e068040.

17. Bittencourt NFN, Meeuwisse WH, Mendonca LD, Nettel-Aguirre A, Ocarino JM, Fonseca ST. Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition-narrative review and new concept. *Br J Sports Med.* 2016;50(21):1309-1314.

18. Lovdal SS, Den Hartigh RJR, Azzopardi G. Injury Prediction in Competitive Runners With Machine Learning. *Int J Sports Physiol Perform.* 2021;16(10):1522-1531.

19. Leckey C, van Dyk N, Doherty C, Lawlor A, Delahunt E. Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis. *Br J Sports Med.* 2024.

20. Yamato TP, Saragiotto BT, Lopes AD. A consensus definition of running-related injury in recreational runners: a modified Delphi approach. *J Orthop Sports Phys Ther.* 2015;45(5):375-380.

21. Hecksteden A, Schmartz GP, Egyptien Y, Aus der Funten K, Keller A, Meyer T. Forecasting football injuries by combining screening, monitoring and machine learning. *Sci Med Footb.* 2023;7(3):214-228.

22. Fruend Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 1997;55:119-139.

23. Friedman JH. Greedy function approximation: a gradient boosting machine *The Annals of Statistics* 2001;29(5):1189-1232.

24. Ke G, Meng Q, Finley T, et al. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.2017.

25. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*2016:785-794.

26. Prokhorenkova L, Gusec G, Vorobev A, Dorogush A, Gullin A. CatBoost: unbiased boosting with categorical features. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.*2018.

27. Breiman L. Random forests. *Machine Learning.* 2001;45:5-32.

28. Hawkins DM, McLachlan GJ. High-Breakdown Linear Discriminant Analysis. *Journal of the American Statistical Association.* 1997;92(437):136-143.

29. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput Methods Programs Biomed.* 2022;214:106584.

30. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ.* 2020;368:m441.

31. Videbaek S, Bueno AM, Nielsen RO, Rasmussen S. Incidence of Running-Related Injuries Per 1000 h of running in Different Types of Runners: A Systematic Review and Meta-Analysis. *Sports Med.* 2015;45(7):1017-1026.

32. Kolbe T, Dennis S, Selley E, Noakes T, Lambert M. The relationship between critical power and running performance. *Journal of Sports Sciences.* 1995;13(3):265-269.

33. McMahon TA, Cheng GC. The mechanics of running: how does stiffness couple with speed? *Journal of Biomechanics* 1990;23:65-78.

34. Malisoux L, Gette P, Delattre N, Urhausen A, Theisen D. Spatiotemporal and Ground-Reaction Force Characteristics as Risk Factors for Running-Related Injury: A Secondary Analysis of a Randomized Trial Including 800+ Recreational Runners. *Am J Sports Med.* 2022;50(2):537-544.

35. van Oeveren BT, de Ruiter CJ, Beek PJ, van Dieen JH. Optimal stride frequencies in running at different speeds. *PLoS One.* 2017;12(10):e0184273.

36. Seabrook JA. Powering Nutrition Research: Practical Strategies for Sample Size in Multiple Regression. *Nutrients.* 2025;17(16).

37. Mach MSE, Hayley; Lally, E; Earl-Boehm, Jennifer E. Foot-Placed Inertial Measurement Units Are Valid Against Shank-Placed Units When Measuring Temporospatial Running Variables. *Journal of Sport Rehabilitation.* 2025;34(6):14.

38. Bentejac C, Csorgo A, Martinez-Munoz G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review.* 2021;54(3):1937-1967.

Table 1. Baseline characteristics of eligible participants. Data are presented as mean (standard deviation).

	All (n=65)	Men (n=43)	Women (n=22)	All (n=23)	Men (n=13)	Women (n=10)
Variable	Uninjured participants			Injured participants		
Age (years)	37.5 (6.8)	37.4 (6.7)	37.7 (7.2)	34.6 (6.9)	33.6 (5.8)	35.8 (8.3)
Height (m)	1.7 (0.1)	1.8 (0.1)	1.7 (0.1)	1.7 (0.1)	1.8 (0.1)	1.7 (0.1)
Mass (kg)	67.9 (9.0)	71.6 (6.7)	60.5 (8.5)	67.9 (12.6)	75.2 (9.8)	58.4 (8.9)
BMI (kg/m^2)	22.4 (2.1)	22.5 (1.9)	22.2 (2.5)	22.1 (2.5)	23.0 (2.1)	21.0 (2.7)
Average run days per week (n)	4 (2)	4 (2)	4 (1)	4 (2)	4 (3)	4 (2)
Average weekly run distance (km)	43.7 (31.3)	46.4 (36.3)	38.6 (18.8)	45.2 (33.7)	52.6 (40.7)	34.9 (18.5)
Total runs reported during 12-week cohort	3698	2648	1050	1285	861	424

Notes: BMI = Body Mass Index; N/A = not applicable.

Table 2: Summary of the strengths and weaknesses of the implemented ML algorithms[38]

Algorithm	**Strengths**	**Weaknesses**
AdaBoost	Simple, interpretable, robust for small and imbalanced datasets	Sensitive to noisy data and outliers; less effective on complex nonlinear patterns
GBM	High accuracy, handles complex relationships well	Computationally expensive; prone to overfitting without tuning
LightGBM	Very fast, memory-efficient, great for large datasets	Can overfit on small datasets; less interpretable
XGBoost	Excellent accuracy, robust, widely used in competitions	More complex tuning; slower than LightGBM on very large data sets
CatBoost	Handles categorical data natively, reduces overfitting, strong accuracy	Slower than LightGBM; less interpretable than linear models
Random Forest	Robust to overfitting and imbalanced dayasets, easy to use, interpretable feature importance	Less accurate than boosting methods; slower for very large datasets

Table 3. Performance of the eight machine learning models. Values reflect the mean and standard deviation across the 10-folds.

	Accuracy	Precision	Recall	F1-score	AUC	Average
CatBoost	**0.96** (0.01)	0.97 (0.01)	0.87 (0.02)	**0.92** (0.03)	**0.93** (0.02)	**0.93**
LightGBM	0.95 (0.01)	0.96 (0.02)	0.85 (0.03)	0.90 (0.02)	0.92 (0.02)	0.92
XGBoost	0.94 (0.01)	0.89 (0.03)	**0.88** (0.03)	0.88 (0.02)	0.92 (0.02)	0.90
GBM	0.94 (0.01)	0.94 (0.02)	0.82 (0.03)	0.88 (0.02)	0.90 (0.02)	0.90
Random Forest	0.87 (0.01)	0.78 (0.04)	0.67 (0.04)	0.72 (0.03)	0.80 (0.02)	0.77
AdaBoost	0.78 (0.01)	**0.98** (0.03)	0.16 (0.03)	0.28 (0.05)	0.58 (0.02)	0.56
LDA	0.75 (0.01)	0.53 (0.10)	0.09 (0.02)	0.16 (0.03)	0.53 (0.01)	0.41
LR	0.74 (0.01)	0.51 (0.11)	0.08 (0.02)	0.14 (0.03)	0.53 (0.01)	0.40

Key: AUC – area under the receiver operator curve.

Figures 1a-d: SHAP summary plots for (a) CatBoost, (b) LightGBM, (c) GBM, and (d) XGBoost. Factors are ordered from largest (top of vertical axis) to smallest (bottom of vertical axis) overall importance within each model; factor order is different for each model. Each dot represents the SHAP value (i.e., log-odds) for a single run for that factor. SHAP values are plotted along the horizontal axis and represent the contribution of that factor to the model's predicted probability of classifying the run as belonging to an injured runner (positive SHAP values increase this predicted probability; negative SHAP values reduce this predicted probability. Point colour represents the raw factor value for that run (red dots represent high raw factor values of that factor; blue dots represent low raw factor values of that factor). Definitions of each factor are provided in supplementary file one.

Figures 2a-d: Violin plots of a) CatBoost, b) LightGBM, c) GBM and d) XGBoost. Each violin shows the distribution of absolute SHAP values for each of the top 10 most influential factors in that model, estimated using a Kernel density estimator. Factors with larger or more variable contributions have distributions farther from zero (vertical axis). The 95% confidence intervals corresponding to the mean absolute SHAP are shown in the top right of each plot. Factor definitions are provided in supplementary file one.

Figure 3: Minimum required sample size to develop a classification model RRI under various assumptions of target shrinkage (S) and anticipated AUC values (horizontal axis), based on the Riley et al.,[30] framework and assuming a 26% outcome prevalence and 29 candidate factors.