Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v18n1p133

Comparing machine learning and conventional statistical approaches for injury prediction in young professional soccer players

By Berendes et al.

15 March 2025

Electronic Journal of Applied Statistical Analysis
Vol. 18, Issue 01, march 2025, 133-164
DOI: 10.1285/i20705948v18n1p133

Comparing machine learning and conventional statistical approaches for injury prediction in young professional soccer players

A case study and versatilely applicable framework

Ina-Marie Berendes*[a], Alexander Gerharz[a,b], Andreas Groll[a], and Mathias Kolodziej[b]

[a]*Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany*
[b]*Department of Sport Science, Borussia Dortmund, Adi-Preißler-Allee 9, 44309 Dortmund, Germany*

15 March 2025

Frequent injuries pose a problem in professional soccer that is being tackled with preventive measures. Consequently, injury prediction and prevention are also increasingly addressed from a statistical perspective. In a pilot study, several machine learning algorithms and conventional statistical approaches have been compared regarding their potential to predict time-loss non-contact lower-body injuries in professional youth soccer players, using data from a prospective cohort study with 56 players of which 22 were injured. The covariates considered here include basic soccer-related as well as neuromuscular and biomechanical features derived from physical testing. Lasso regularized logistic regression, naive Bayes, linear discriminant analysis, k-nearest neighbors, classification trees, random forests, XGBoost, and support vector machines are considered for binary classification and prediction of an injury occurrence. The prediction results from a cross-validated procedure are compared regarding multiple quality measures. Post Lasso logistic regression with a reduced penalty gives the best results with an accuracy of 0.625, a predictive likelihood of 0.593, and a Brier score of 0.228. The respective sensitivity and specificity are 0.773 and 0.529, with an AUC

*Corresponding author: ina-marie.berendes@tu-dortmund.de

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

of 0.672. Moreover, an XGBoost model slightly outperforms the Lasso model in terms of accuracy (0.661), while for the other performance measures it is dominated by the Lasso.

In addition to the specific results on the available injury data set, the proposed comparison procedure of several models for binary prediction provides a generally applicable analysis guideline. This roadmap can also be applied in other contexts where similarly structured small but rich data sets are available.

keywords: injury prediction, soccer, lasso regularization, machine learning, binary classification.

1. Introduction

Professional soccer is a sport with a high injury incidence. By considering different countries, Junge and Dvorak (2004) found that adult male professional players sustain approximately one performance affecting injury per year. Moreover, as a particular example, the incidence of hamstring injuries has increased severely over the span of two decades (Ekstrand et al., 2023). Young professional players are even more often affected by injuries than adult players (Pfirrmann et al., 2016).

First of all, injuries negatively affect the careers of the players and are naturally unpleasant for them. Moreover, they have the consequence that players will miss training sessions or matches. Finally, also their teammates can be affected by fear of an own injury or by low confidence in the team's competence due to replacements. Such negative emotions can also spread between players (social contagion; Hurley, 2016). Injuries in the team tangibly impact the performance (Hägglund et al., 2013). Not least, the absence of a professional player due to an injury costs the club a considerable amount of money (Ekstrand, 2013). All these issues have led to an increased focus on injury risk factors and causes, prevention, and prediction in sport science.

For these reasons, injury prevention in professional soccer has become an important aspect also in the field of statistical analysis. Machine learning models are promising instruments for accurate prediction. Beside conventional statistical approaches such as (regularized) logistic regression, linear discriminant analysis, k-nearest neighbors, and naive Bayes classifiers, possible machine learning methods include trees, random forests, extreme gradient boosting, and support vector machines. Most works attempt to model non-contact soft-tissue injuries, because they are seen as predictable and preventable (Rossi et al., 2022). Player-specific characteristics such as age, weight, height, or body mass index (BMI) are often considered in such models. Moreover, soccer players also complete several neuromuscular and biomechanical tests to assess their physical performance. Measurements from these tests can be valuable covariates in injury prediction models.

A study on elite male youth soccer players by Read et al. (2018) approached non-contact lower extremity injury prediction using multivariate binary logistic regression.

Rossi et al. (2018) used logistic regression, tree-based, and random forest models for injury forecasting. López-Valenciano et al. (2018) studied injuries of soccer players and handball players with the same injury definition as used here, employing tree models in combination with boosting and bagging. We will use similar covariates such as age, BMI, player position, kicking leg, as well as covariates describing core stability, dynamic postural control, and hip strength. Oliver et al. (2020) also used trees to predict injuries in elite male youth soccer players, while Rommers et al. (2020) employed the XGBoost method for the same purpose, both using covariate measurements collected before the season.

This work continues the research of Kolodziej et al. (2023), who used different Lasso-penalized logistic regression models (Friedman et al., 2010) to predict the injury status of young professional soccer players. An overview of publications on machine learning in team sports with the subtopic of injuries (Claudino et al., 2019) presents one study involving the Lasso method for injury risk modeling. A more extensive systematic review of injury prediction models in sport (Bullock et al., 2022) reveals the existence of three studies using Lasso-penalized models for injury prediction. However, they all focus on different study subjects other than young professional soccer players. In the present manuscript, we apply further conventional statistical as well as machine learning approaches to the data of Kolodziej et al. (2023) in an attempt to increase the prediction quality and provide a comprehensive analysis comparing state-of-the-art methods. Particularly, we examine if modern machine learning models are able to outperform the classical statistical approaches of Kolodziej et al. (2023). To the best of our knowledge, this work is the first to provide a thorough comparative study comparing conventional statistical methods with modern machine learning approaches in the context of young professional soccer. We incorporate the previous and the new models into a parallel cross-validation (CV) procedure to make results comparable and obtain fair prediction results on unseen test data. Regarded covariates are basic ones such as age, BMI, training age, player position, and kicking leg, as well as a range of covariates derived from physical testing. Moreover, we seek to identify the most important covariates for good injury prediction from the best-performing algorithms with the goal of assisting soccer coaches and healthcare providers in injury prevention.

Our available data set contains a rich selection of variables. However, due to large efforts needed for the data acquisition (and a restricted number of professional players available to be investigated), it consists of only 56 observations. Given the rather small sample size, our work mostly aims at providing a thorough guideline for the comparison of modeling approaches. The proposed analysis framework could be useful in other contexts where the same restrictions regarding data availability apply, which often is the case in professional sports.

The remainder of this manuscript is structured as follows. The data at hand are described in Section 2. In Section 3, the applied statistical modeling methods as well as our cross-validation framework for comparing the methods are ~~briefly~~ described. Section 4 provides both the previous results of Kolodziej et al. (2023) and those from our new analyses. Finally, Section 5 summarizes our main findings, addresses the limitations of these analyses, and gives prospects to potential future analyses.

2. Data

A professional athlete health care data set of a prospective cohort study of 56 young professional male soccer players, belonging to three different youth teams (under 16, under 17, under 19) from two German professional soccer clubs, is regarded in this work. The same data set has already been analyzed in Kolodziej et al. (2023), whereby we include one specific player that was injured twice only once into our data set. Exclusion criteria were the usage of prophylactic devices, a musculoskeletal injury reported no longer than six weeks ago, or any current physical indisposition with the risk of a performance impairment. Players outside of these criteria underwent testing before the season of 2018/19 at the start of the preparation phase. After completing a questionnaire guided by a research assistant, the practical screening took place in one sport science research center. Figure 1 shows the stages of the screening process. A 3D motion analysis device and force plates recorded the players while they performed a single-leg drop landing (SLDL) task and an unanticipated side-step cutting (USSC) task. Additionally, their postural control under three different conditions and the strength of their lower body was measured. The screening is described in detail in Kolodziej et al. (2021) and Kolodziej et al. (2022). Tables 5 to 8 in Appendix A list the physical variables derived from the measurements and transformations thereof used as covariates. For the 10 months of the season after the testing, it was documented whether the players suffered from any time-loss non-contact soft-tissue injury of the lower body. As specified in Rossi et al. (2022), a time-loss injury by definition leads to absence in physical training or matches for at least one day after the injury occurrence. Professional healthcare staff documented injuries using a standardized injury form without knowledge of the players' screening and testing results. In the following, the occurrence of such injuries is encoded as a binary (response) variable $y \in \{0, 1\}$, with $0 =$ *no injury* and $1 =$ *injury*. Incomplete observations missing more than half of the covariate values due to technical problems are excluded from the analysis, leading to our sample size of 56.

3. Methodology

In this section, we first briefly introduce all statistical and machine learning approaches considered in this work (Section 3.1). Next, in Section 3.2 we introduce several quality measures which will be used for comparison of the predictive performance of the approaches. Finally, we explain a cross-validation strategy and how hyperparameter tuning is done in Section 3.3.

3.1. Statistical and machine learning approaches

A classification problem consists of training observations with a categorical outcome y. An observation has features $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ to be used as predictors, and the category or class y it belongs to. Using the given data, the goal is to find a model, or learner, that can correctly predict the class of an observation from its feature values. New observations

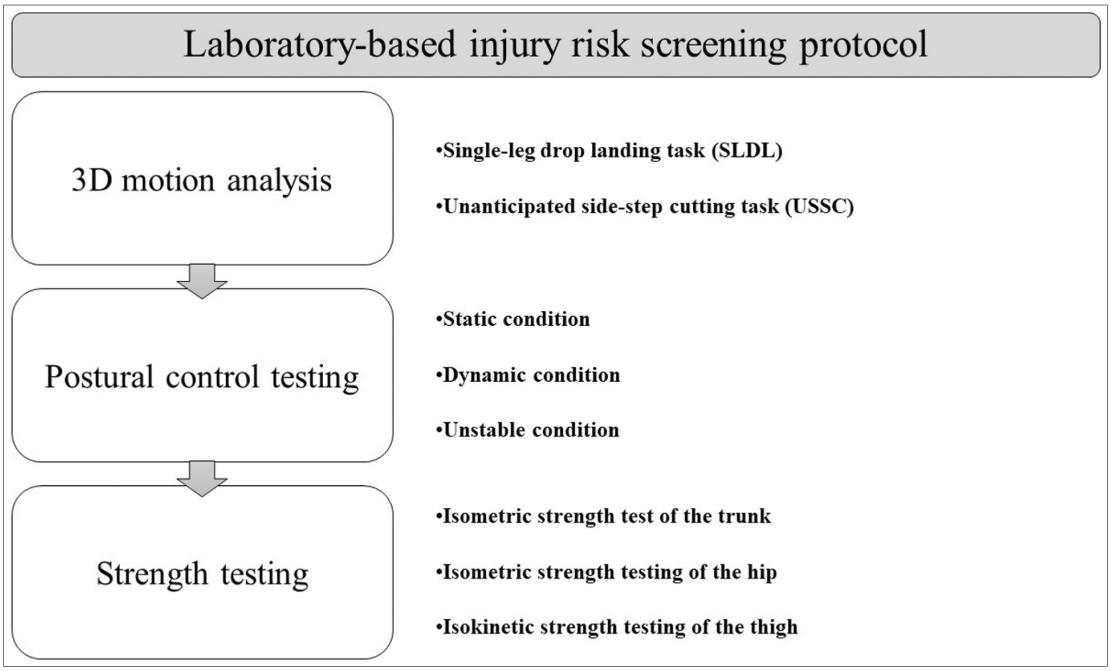
Laboratory-based injury risk screening protocol
3D motion analysis
•Single-leg drop landing task (SLDL)
•Unanticipated side-step cutting task (USSC)
Postural control testing
•Static condition
•Dynamic condition
•Unstable condition
Strength testing
•Isometric strength test of the trunk
•Isometric strength testing of the hip
•Isokinetic strength testing of the thigh

Figure 1: Injury risk screening protocol (Kolodziej et al., 2023)

with a given feature vector and unknown class can then be processed by the learner to predict their class or their posterior class probabilities (see, e.g., Hastie et al., 2009).

Logistic regression

In a logistic regression model (see, e.g., Tutz, 2012), the probability $\pi(\boldsymbol{x})$ of a binary event y conditioned on covariates $\boldsymbol{x} = (x_1, \dots, x_p)^\top$ is modeled with unknown but fixed coefficients $\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}$. Then, $\pi(\boldsymbol{x})$ is transformed via the logit function, i.e.,

$$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \boldsymbol{x}^\top\boldsymbol{\beta} \iff \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = \exp\left(\boldsymbol{x}^\top\boldsymbol{\beta}\right) \iff \pi(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{x}^\top\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}^\top\boldsymbol{\beta}\right)}.$$

The coefficient vector $\boldsymbol{\beta}$ is estimated via maximum likelihood (ML) techniques (see, e.g., McCullagh and Nelder, 1989, for further details).

Following Tutz (2012), the estimated probability for $y_i = 1$ for an observation with features $\boldsymbol{x}_i$ is calculated via

$$\hat{\pi}(\boldsymbol{x}_i) = \frac{\exp\left(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}\right)}.$$

Lasso regression

If there are many possible predictors available, variable selection can be useful to obtain a sparse model with only the most relevant features included and, thus, better interpretability. The *Least Absolute Shrinkage and Selection Operator* (Lasso; Tibshirani, 1996) regression is a suitable regularization method that results in variable selection. Via penalization of the regression coefficients, it shrinks all coefficients towards 0, and only retains the most important covariates in the model. It adds the penalty $-\frac{\lambda}{2}\sum_{j=1}^{p}|\beta_j|$ to the conventional log-likelihood. The penalty parameter λ controls the strength of the penalization and amount of shrinkage of the coefficients, and hence, the number of covariates in the model. It needs to be tuned. As λ increases from 0, the regression coefficients are shrunk, and the number of covariates retained in the model descends from p down to 0 (James et al., 2021). This property makes Lasso particularly useful in high-dimensional data situations, especially with more covariates than observations ($p > n$). The Lasso estimate is given by the $\boldsymbol{\beta}$ that maximizes the penalized log-likelihood (Friedman et al., 2010).

The *relaxed Lasso* (Meinshausen, 2007) is a two-staged modification of the usual Lasso procedure. It maximizes the penalized log-likelihood

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \phi\frac{\lambda}{2}\sum_{j=1}^{p}|\beta_j|$$

with respect to $\boldsymbol{\beta}$. As before, $\lambda \in [0,\infty)$ controls the penalty strength and, hence, how many of the most important covariates are retained in the model. The second tuning parameter ϕ determines how much the regression coefficients of the selected variables are finally shrunk, while $\phi = 1$ leads to the classic Lasso estimator. For $\phi \in (0,1)$, there is less shrinkage on the coefficients the smaller ϕ is. In the special case $\phi = 0$, the feature selection is controlled by λ in a first step. Then, a model containing only the selected features is estimated in the classical way, meaning only the corresponding regression coefficients are (re-)estimated in a second, unregularized estimation step. This case is also known as the so-called *post Lasso*. For simplicity and to avoid time-consuming multidimensional tuning, we restrict ourselves to the two extreme cases, $\phi = 1$ (regular Lasso) and $\phi = 0$ (post Lasso).

Naive Bayes

The naive Bayes method classifies observations directly (Cleve and Lämmel, 2020), calculating posterior class probabilities from their features $\boldsymbol{x}$, the prior class probabilities π_k, and the class-specific densities $f_k(\boldsymbol{x})$, $k = 1,\dots,K$. The posterior probability for class k of an observation with feature vector $\boldsymbol{x}$, i.e., $\mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x})$, can then be derived via the Bayes' theorem. The simplifying assumption that the p features are independent in each class allows for factorizations of the density functions into a product of the single variable densities $f_{kj}(x_j)$, $j = 1,\dots,p$, see, e.g., Hastie et al. (2009) and James et al. (2021) for details. The method is quite simple, however, it has to be noted that the probability predictions can suffer from a large amount of included features (Friedman, 1997).

Linear discriminant analysis

In linear discriminant analysis (LDA; Hastie et al., 2009), the feature vector in class k is assumed to be distributed according to $\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with class-specific means $\boldsymbol{\mu}_k$ and a common covariance matrix $\boldsymbol{\Sigma}$. Bayes' theorem is then used to calculate posterior class probabilities using multivariate normal densities $f_k(\boldsymbol{x})$ and prior class probabilities π_k. A large number of features $p > n$ can lead to problems due to singular matrices (Yu and Yang, 2001).

k-nearest neighbors

The k-nearest neighbors (KNN) method classifies an observation via considering its k closest observations in the feature space based on a suitable distance measure. Majority vote among the classes of its k nearest neighbors determines the class it is assigned to. Moreover, the relative frequencies of classes of the closest neighbors serve as posterior probability estimates. Note that the number of considered neighbors k is a tuning parameter (see, e.g., Hastie et al., 2009). Again, the curse of dimensionality can lead to a poor performance of KNN when including many features (Friedman, 1997).

Classification trees

The Classification and Regression Tree (CART) method (Breiman et al., 1984) is a stepwise, top-down, greedy procedure to split the feature space and thus a data sample into classes. In each iteration, a single covariate x_j and a cut point s are chosen to split the feature space into two regions $R_1(j, s) = \{x|x_j < s\}$ and $R_2(j, s) = \{x|x_j \geq s\}$, which also splits the observations into two sets accordingly. The choice of the split point and splitting variable is made such that the total impurity of the two new sets with regard to the class variable, measured, e.g., by the Gini index, is reduced as much as possible in comparison to the whole set. This binary splitting procedure is repeated until a stopping criterion such as a minimal set size (James et al., 2021) or impurity improvement (Breiman et al., 1984) is reached. To prevent overfitting, the tree can be pruned. The predicted class for an observation is the majority class in the final leaf the observation ends up in, and posterior probabilities are the relative frequencies of the classes in that leaf (James et al., 2021). As tuning parameters, we use both the minimal number of observations in a set, and the complexity parameter `cp` that stops the tree growing if the impurity is not decreased by at least the factor `cp`. If the response is continuous, so-called regression trees can be constructed in a similar fashion. Since any tree split only involves a single best feature separately, trees can handle a large input number of features.

Random forests

The random forest (Breiman, 2001) extends the tree concept and contains a multitude of trees that are grown in parallel. Several bootstrap samples, each of size n, from the original data are drawn (with replacement), and a fully sized tree is grown on each sample. For every split, only a subset of a pre-specified size `mtry` $\leq p$ of randomly selected covariates may be used, where `mtry` is the main tuning parameter of the method. An observation is classified based on the majority vote over all trees in the forest. As the procedure is based on individual trees, it inherits the property to be able to easily handle a large input number of features (James et al., 2021).

XGBoost

Boosting, another approach to combine many learners, works sequentially instead of parallelly (James et al., 2021). Boosting of trees can be conducted via a gradient descent procedure (Friedman, 2001). Extreme Gradient Boosting (XGBoost; Chen and Guestrin, 2016) is a further development of this method with additional regularization to prevent overfitting. After the first tree is fitted, every subsequent tree is trained on the residuals based on the true outcome and a prediction based on all previous trees (James et al., 2021). The prediction is updated by adding the new prediction contribution scaled by a small learning rate $\nu \in (0, 1]$, making the learner "weak" (Friedman, 2002). The number of learners M, the learning rate ν, and the maximal tree depth are the main tuning parameters we focus on. Again, as also this procedure is based on individual trees, it inherits the property to be able to easily handle a large input number of features.

Support vector machine

Support vector machines (SVMs; Cortes and Vapnik, 1995) construct a hyperplane in the feature space to separate the data with regard to the outcome. This is done such that a margin on both sides of the hyperplane preferably does not contain many sample observations. Observations within the margin or even on the wrong side of the hyperplane are penalized, with the cost of such violations being a tuning parameter. Moreover, the total allowed violation is limited. Nonlinear boundaries between the classes are possible by introducing kernels. We use the radial kernel $K(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{i_1 j} - x_{i_2 j})^2\right)$, with $\gamma > 0$ representing a second tuning parameter (see, e.g., James et al., 2021).

Shapley Additive Explanation values

Lundberg and Lee (2017) propose a game theoretical approach to assign so-called SHapley Additive exPlanation (SHAP) values to features in a model as a measure of their importance in the prediction for one observation. The mean absolute value over all observations then represents the global importance of a feature in the model.

3.2. Metric definition

We are searching for a model that can not only identify players at risk of an injury, but that has an overall good prediction performance from a global statistical perspective. Therefore, we jointly regard all following measures to assess the accordance between the models' predictions and the true outcomes:

- the global *accuracy*, which gives the relative frequency of observations for which the prediction matches the true class; a common and overall measure of model accuracy (Yale et al., 2018);
- the *sensitivity*, which is the relative frequency of observations with true class 1 that are predicted as class 1; important for injury prediction;
- the *specificity*, which is the percentage of true class 0 observations that are predicted as class 0; important for non-injury prediction;
- *Youden's index* (Youden, 1950), given by sensitivity + specificity − 1, representing a summary of the two above measures; important if correct prediction of both classes is equally relevant;
- the *AUC*, which is the area under the ROC (receiver operating characteristic) curve; relative predictive power of a model (Yale et al., 2018), model's ability to discriminate between the classes (Van Calster et al., 2019);
- the (average) *predictive Bernoulli likelihood*, given by $\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i^{y_i}\left(1-\hat{\pi}_i\right)^{1-y_i}$, representing the average probability of a correct class prediction regardless of class;
- the *Brier score*, given by $\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\pi}_i\right)^2$, representing the average quadratic deviation between true and predicted class; does not only take into account the class predictions, but also the predicted probabilities.

All above measures take values in $[0;1]$. A large value in this interval indicates a good performance for all measures but the Brier score, which is an error measure where a small value is desirable.

3.3. Leave-one-out cross-validation implementation and hyperparameter tuning

The comparison of the approaches from Section 3.1 is done via a nested cross-validation procedure. We perform a leave-one-out (LOO) CV on an outer level. Cycling through all observations, each one is once taken out, leaving 55 observations for training. These are randomly split into 15 CV folds for an inner CV that is used for hyperparameter tuning. Using 15-fold CV leads to a good compromise of having an appropriately large number of folds, which are not too small with three or four observations each. The same folds are used for all methods for better comparability. The best tuning parameters for the logistic Lasso regression approaches are chosen consulting the deviance, which is the suggested measure by the `glmnet` authors (Tay et al., 2023). For all other approaches, the Brier score is used to select the tuning parameters with the best predictions. All methods are then fitted once again on all 55 observations with these respective best hyperparameters. Predictions for the respective left out observation are saved and the measures from Section 3.2 are used to assess them[1]. The simple constant model is exempt from this procedure and uses the proportion of injured players in the whole data set as a predicted probability for every observation, and in the following analyses serves as a benchmark.

The hyperparameters mentioned in Section 3.1 are tuned using an even grid of nine possible values each, which are given in Table 1. For Lasso logistic regression, however, a grid of 100 λ values is created automatically via the `cv.glmnet` function from the `glmnet` package. For post Lasso, ϕ is set to 0. Finally, note that there are no natural tuning parameters for naive Bayes and LDA.

Let θ be the threshold value for converting a predicted probability for class 1 into a prediction of class 1, i.e., a learner predicts class 1 for an observation if its predicted probability is greater than or equal to θ. We conduct the analysis for $\theta = 0.5$, $\theta = 0.4$, and for the respective optimal θ for each method according to Youden's index (Youden, 1950).

Table 1: Grids of hyperparameters used in tuning (using 9 different values for each grid)

Method	Parameter	Smallest Value	Largest Value	Log Scale
k-nearest Neighbors	k	7	15	no
Classification Tree	`cp`	0.00001	0.1	yes
	min. # obs.	2	10	no
Random Forest	`mtry`	2	18	no
XGBoost	max. depth	1	9	no
	ν	0.00001	0.1	yes
	M	10	100	no
Support Vector Machine	cost	0.0001	10000	yes
	γ	0.0001	10000	yes

3.4. Construction of bootstrap confidence intervals

Principally, a single value for every measure from Section 3.2 and every method from Section 3.1 results from the analysis of our data set using the procedure from Section 3.3. To quantify the uncertainty of these values, said procedure is repeated for a number of

[1] An important technical detail is that this is done on the whole data set level, since this is necessary to calculate e.g. the sensitivity, specificity, and AUC. Casalicchio and Burk (2024) refer to this as micro-averaging.

B bootstrap samples from the original data set. The samples are drawn in a stratified manner such that the proportion of injured players in each sample corresponds to the true proportion (Fox, 2015). Moreover, it is ensured that in every bootstrap sample at least 35 % of the observations (corresponding to at least 20 players) are drawn exactly once. Only these unique observations from every sample undergo the CV-type procedure of being left out of the data set described above. Thus, predicted probabilities for these observations are obtained from separate training data, analogous to the process in the original data set, avoiding information leakage and over-optimism. This bootstrap procedure generates B values for every combination of quality measure and modeling approach. In order to be able to assume normality in every set of B values, we suggest to choose $B \geq 30$. Then, Gaussian confidence intervals can be constructed using the estimated means and standard deviations.

All analyses were carried out with R (R Core Team, 2022) in version 4.3.0. The specific packages used for the different modeling approaches are `glmnet` (Tay et al., 2023) for all Lasso models, `e1071` (Meyer et al., 2023) for naive Bayes and the SVM, `MASS` (Venables and Ripley, 2002) for the LDA, `kknn` (Schliep and Hechenbichler, 2016) for KNN, `rpart` (Therneau and Atkinson, 2023) for trees, `ranger` (Wright and Ziegler, 2017) for random forests, and `xgboost` (Chen et al., 2023) for XGBoost.

To make the tuning process of all employed approaches as much comparable as possible, all models except for the Lasso-regularized logistic regressions, which were tuned via `cv.glmnet` from `glmnet` (Tay et al., 2023), were tuned with the package `mlr3verse` (Lang and Schratz, 2023).

The `pROC` package (Robin et al., 2011) was used to calculate AUC values and the `iml` package (Molnar et al., 2018) for SHAP values. The package `rpart.plot` (Milborrow, 2022) was used for classification tree plotting.

4. Application on injury data

Descriptive characteristics of the injury data set and the previous results of Kolodziej et al. (2023) are briefly presented in Section 4.1. In Section 4.2, we apply all classification methods introduced in Section 3.1 to the data using CV, and illustrate the main findings on the prediction quality. Important covariates for several learning methods are compared in Section 4.4.

4.1. Descriptive analysis and previous results

Our injury data set contains 56 observations of 86 covariates and the outcome variable. Overall, 22 of 56 players (39.3 %) suffered an injury in the observation time of 10 months. Tables 5 to 8 in Appendix A contain the arithmetic means and standard deviations of all physical continuous covariates on the whole data set as well as separately for injured and non-injured players. In Figures 3 to 7 in Appendix B, the distributions of the remaining variables are illustrated.

Kolodziej et al. (2023) analyzed the injury data using Lasso logistic regression models, using main effects of all available covariates to model the injury probability. The penalty strength λ was set to either its optimal value, λ_{opt}, according to a 15-fold CV, or to $\lambda_{1\text{se}}$, a slightly smaller value[2]. The regression coefficients were either shrunk by the respective λ, or not shrunk using post Lasso. These two modeling options in combination with the two λ candidates led to four possible Lasso models.

All models competed in a LOO CV prediction competition similar to the one described in Section 3.3. Altogether, the performance of the post Lasso model with the smaller penalty $\lambda_{1\text{se}}$ turned out to be best. Of the four models, it reached the largest predictive likelihood of 0.58, and also the largest AUC value of 0.63. Moreover, it clearly achieved the largest sensitivity of 0.35. Its specificity of 0.79 was, albeit smaller, in the range of the other models.

4.2. Results of the leave-one-out cross-validation competition on predictive power

The LOO CV strategy from Section 3.3 is used to compare the methods from Section 3.1 fairly and assess their predictive performance on unseen data. We include all available covariates from Kolodziej et al. (2023) in the modeling approaches in linear form only (but note that the tree-based approaches of course can account for non-linear and interaction effects nevertheless). For comparison and as a benchmark, we also regard a deterministic featureless learner that simply uses the relative frequency of injured players in the data as the predicted probability for class 1, which is called *constant* in the following.

Table 2 presents the predictive performance results of the LOO CV for the threshold $\theta = 0.5$. The best accuracy of 0.661 is reached by XGBoost, corresponding to a correct prediction for 37 out of 56 players. The post Lasso with λ_{opt} has one and the post Lasso with $\lambda_{1\text{se}}$ and the SVM have two matches less. The latter Lasso approach, Naive Bayes, the tree, and XGBoost all achieve the largest sensitivity of 0.409. Of these, XGBoost has the largest specificity of 0.824. It also shows the largest value of 0.233 of Youden's index over all methods. Finally, the post Lasso with $\lambda_{1\text{se}}$ reaches the largest AUC of 0.672, the largest predictive likelihood of 0.593, and the smallest Brier score of 0.228 over all other methods. With $\theta = 0.5$, the post Lasso with $\lambda_{1\text{se}}$ and XGBoost show good values for several quality measures.

We repeat the analysis for $\theta = 0.4$, which is very close to the proportion of observations of class 1 in the whole data set. The results are shown in Table 9 in Appendix C. The KNN algorithm reaches the largest accuracy of 0.625, predicting the classes of 35 out of 56 players correctly. It also shows the largest specificity of 0.853, while its sensitivity is 0.227. Conversely, XGBoost achieves the largest sensitivity of 0.727, and a specificity of 0.294. The largest value of Youden's index of 0.131, however, is reached by the post Lasso with $\lambda_{1\text{se}}$. The modeling approach also still shows the largest AUC, the largest predictive likelihood, and the smallest Brier score, all of which do not depend on the chosen threshold θ.

[2] This is the smallest λ within one CV standard error of λ_{opt}. It leads to a weaker regularization than λ_{opt}, so it typically results in a slightly larger model.

Table 2: Quality measure results for the threshold $\theta = 0.5$ (best value in **bold font**)

	Acc	Sens	Spec	Youd	AUC	Pred. L.	Brier
Lasso λ_{opt}	0.607	0.091	0.941	0.032	0.586	0.532	0.238
Lasso λ_{1se}	0.589	0.136	0.882	0.019	0.508	0.527	0.257
Post Lasso λ_{opt}	0.643	**0.409**	0.794	0.203	0.638	0.577	0.239
Post Lasso λ_{1se}	0.625	0.364	0.794	0.158	**0.672**	**0.593**	**0.228**
Naive Bayes	0.518	**0.409**	0.588	-0.003	0.515	0.536	0.420
LDA	0.482	0.364	0.559	-0.078	0.370	0.440	0.510
KNN	0.589	0.045	0.941	-0.013	0.549	0.546	0.252
Tree	0.518	**0.409**	0.588	-0.003	0.473	0.508	0.424
Random Forest	0.536	0.136	0.794	-0.070	0.449	0.509	0.258
XGBoost	**0.661**	**0.409**	0.824	**0.233**	0.549	0.509	0.253
SVM	0.625	0.091	**0.971**	0.061	0.434	0.542	0.232
Constant	0.607	——	——	——	0.500	0.523	0.239

Abbreviations: Acc: Accuracy, Sens: Sensitivity, Spec: Specificity, Youd: Youden's index, AUC: Area under the ROC curve, Pred. L.: Predictive Bernoulli likelihood, Brier: Brier score

Lastly, we conduct the analysis again, using the respective best threshold θ that leads to the maximal value of Youden's index for each method. Table 3 displays these best threshold values and the resulting quality measures, along with bootstrap confidence intervals described in Section 4.3. For the constant learner, two thresholds θ lead to a maximal Youden's index, namely $\theta = 0$ and $\theta = 1$. Consequently, we have two best constant models, which predict either 1 or 0 for all observations, respectively, and their accuracies sum up to 1. The XGBoost method has the largest accuracy of 0.661 again, and its best θ is very close to 0.5. Even if its sensitivity is only at 0.5, the other measures are often better than most approaches, at least. The post Lasso with λ_{opt} and the SVM have large sensitivity values, the LDA has a large specificity. The best performance of both measures combined in the form of Youden's index is reached by the post Lasso approach with λ_{1se}. Again, it also has the best AUC value, predictive likelihood, and Brier score, since these measures do not depend on the threshold θ. While the post Lasso with λ_{opt} shows good values for many measures, it is still often slightly outperformed by the one with λ_{1se}.

Altogether, the post Lasso with λ_{1se} shows the best results for four out of seven quality measures considered here. The XGBoost has the largest accuracy of all methods, and often one of the best results among the newly employed methods.

Table 3: Best thresholds θ according to Youden's index and quality measure results, together with 95 % bootstrap confidence intervals for the respective best threshold (best value in **bold font**). Colors indicate if the interval covers the true value (yes / no).

	Threshold θ	Accuracy	Sensitivity	Specificity	Youden's index	AUC	Predictive likelikood	Brier score
Lasso λ_{opt}	0.38079	0.625 [0.603; 0.955]	0.727 [0.289; 1]	0.559 [0.488; 1]	0.286 [0.182; 0.897]	0.586 [0.463; 0.976]	0.532 [0.491; 0.761]	0.238 [0.099; 0.302]
Lasso λ_{1se}	0.26484	0.482 [0.62; 0.968]	0.864 [0.432; 1]	0.235 [0.534; 1]	0.099 [0.247; 0.919]	0.508 [0.508; 0.976]	0.527 [0.512; 0.861]	0.257 [0.069; 0.312]
Post Lasso λ_{opt}	0.13758	0.554 [0.501; 0.879]	**0.955** [0.253; 1]	0.294 [0.209; 1]	0.249 [0.030; 0.729]	0.638 [0.300; 0.904]	0.577 [0.444; 0.724]	0.239 [0.138; 0.361]
Post Lasso λ_{1se}	0.26999	0.625 [0.572; 0.864]	0.773 [0.266; 1]	0.529 [0.300; 1]	**0.302** [0.152; 0.729]	**0.672** [0.400; 0.903]	**0.593** [0.476; 0.754]	**0.228** [0.140; 0.351]
Naive Bayes	0.98490	0.643 [0.663; 0.966]	0.273 [0.415; 1]	0.882 [0.572; 1]	0.155 [0.306; 0.912]	0.515 [0.600; 0.987]	0.536 [0.593; 0.864]	0.420 [0.109; 0.380]
LDA	0.99995	0.607 [0.639; 0.935]	0.136 [0.376; 1]	**0.912** [0.467; 1]	0.048 [0.217; 0.894]	0.370 [0.468; 0.993]	0.440 [0.526; 0.872]	0.510 [0.071; 0.393]
KNN	0.37413	0.643 [0.471; 0.926]	0.364 [0.249; 1]	0.824 [0.222; 1]	0.187 [0.052; 0.724]	0.549 [0.382; 0.909]	0.546 [0.482; 0.674]	0.252 [0.155; 0.319]
Tree	0.95833	0.625 [0.643; 0.974]	0.364 [0.458; 1]	0.794 [0.609; 1]	0.158 [0.235; 0.965]	0.473 [0.541; 1]	0.508 [0.571; 0.953]	0.424 [0.036; 0.391]
Random Forest	0.41372	0.571 [0.745; 0.995]	0.455 [0.569; 1]	0.647 [0.707; 1]	0.102 [0.526; 0.961]	0.449 [0.785; 1]	0.509 [0.624; 0.786]	0.258 [0.066; 0.184]
XGBoost	0.49077	**0.661** [0.704; 0.975]	0.500 [0.431; 1]	0.765 [0.669; 1]	0.265 [0.351; 0.946]	0.549 [0.585; 1]	0.509 [0.578; 0.831]	0.253 [0.055; 0.258]
SVM	0.34893	0.518 [0.705; 0.994]	**0.955** [0.538; 1]	0.235 [0.550; 1]	0.190 [0.482; 0.944]	0.434 [0.735; 1]	0.542 [0.673; 0.869]	0.232 [0.039; 0.178]
Constant 1	0	0.393	——	——	——	0.500	0.523	0.239
Constant 0	1	0.607	——	——	——	0.500	0.523	0.239

4.3. Uncertainty quantification

In order to obtain uncertainty intervals for the quality measures from Section 4.2, we employ the bootstrap procedure described in Section 3.4. A set of B values for every quality measure and method except for the constant models is obtained from the bootstrapped LOO CV. Due to the large computational burden of this procedure, we decided to choose $B = 30$. In order to justify that an assumption of normality is acceptable for these sets, we carefully investigated corresponding density plots. Exemplary ones for the accuracy, the predictive likelihood and the Brier Score for the post Lasso approach with λ_{1se}, the random forest an KNN are provided in Figure 8 in Appendix D. For all other measures and methods, these look quite similar. Therefore, we proceed with the normality assumption and construct a respective 95 % confidence interval (CI) from every value set, belonging to one measure and modeling approach. The intervals are given in Table 3, where a color scheme indicates whether the value obtained from the original data lies within the respective CI.

All employed measures can take values between 0 and 1, but the interval widths for the different measures show differences. Some CIs for the sensitivity or specificity cover almost the whole $[0, 1]$-interval, while for the predictive likelihood and Brier score, much smaller CIs are obtained. In most cases, the models' performances on the bootstrap data sets are overall slightly better than on the original data. For many modeling approaches, the original performance measure value is not even included in the bootstrap CI. However, for KNN, the classical Lasso and both post Lasso approaches, the intervals cover the true value for all seven performance measures. Moreover, the true values often lie approximately in the middle of their intervals for these approaches. Therefore, we conclude that such bootstrap CIs give a reasonable range for the performance measures of these methods in situations comparable to ours.

4.4. Important covariates

In the following, we compare the proposed most important covariates of three models. Firstly, the post Lasso approach with a smaller penalty λ_{1se} emerges as the overall best one from our comparisons. Secondly, XGBoost is a modern, popular and powerful machine learning modeling approach and shows the largest accuracy. Thirdly, the classification tree, which despite not showing a convincing performance, is highly interpretable. After re-tuning the hyperparameters in a 15-fold CV using all observations, we fit the three approaches on the whole data set with the obtained best hyperparameters and calculate variable importances based on SHAP values. The Lasso model selects three covariates (see also Kolodziej et al., 2023). The resulting classification tree is shown in Figure 2 and uses four different splitting variables. Here it is to be noted that due to the small sample size and the nature of the tree algorithm, a slightly differing data set might profoundly change the structure of the resulting tree (see Hastie et al., 2009). The 10 largest SHAP values for XGBoost are given in Table 4.

The covariate with the largest absolute coefficient ($\hat{\beta} = -0.987$) and the second largest SHAP value of 0.101 in the Lasso model is the concentric knee extension torque. In

XGBoost, the feature also has the largest variable importance of 0.044. Moreover, it is the third tree splitting variable and shows the largest SHAP value of 0.183 in this model as well. Both tree and Lasso model link a higher value of the covariate to a lower probability to be injured.

The second covariate chosen by the Lasso model, the hip external and internal rotation moment in the SLDL task, has a regression coefficient of $\hat{\beta} = -0.963$ and is thus associated with a smaller injury probability. Moreover, in the Lasso model, it has the third largest SHAP value of 0.092. The other two models don't use this feature primarily.

The center of pressure (COP) sway of a force plate the athletes stood on to test their postural control under static conditions is the third covariate chosen by the Lasso model, with a regression coefficient of $\hat{\beta} = 0.731$. Here, it has the largest SHAP value of 0.144. The COP sway has the second largest SHAP value of 0.036 in XGBoost, and it is the first splitting variable in the classification tree with the second largest SHAP value of 0.151. In both the Lasso and the tree model, a larger COP sway is associated with a larger injury probability.

The second splitting variable in the tree with the third largest SHAP value of 0.110 is the hip adduction and abduction moment in the USSC task. A higher value of the covariate is related to a higher injury risk. The covariate is the third most important variable in the XGBoost model, however, the SHAP value of 0.017 is only half as large as the second largest value.

The plantar and dorsal moment of the ankle in the USSC task appears as the fourth splitting variable in the classification tree, shows a SHAP value of 0.096, and predicts a higher injury probability with a larger value. XGBoost contains this feature as the seventh most important one.

Table 4: Top 10 SHAP values for XGBoost fitted on the whole data set

	XGBoost	
Rank	Variable	SHAP value
1	Concentric Knee Extension	0.044
2	COP Sway	0.036
3	Hip Adduction/Abduction Moment USSC	0.017
4	Knee Adduction/Abduction PEAK SLDL	0.017
5	BMI	0.010
6	Knee External/Internal Rotation IC SLDL	0.009
7	Ankle Plantarflexion/Dorsalflexion Moment USSC	0.007
8	Hip External/Internal Rotation IC SLDL	0.005
9	Eccentric Knee Extension Ratio	0.005
10	Knee Flexion/Extension PEAK USSC	0.003

Following Van Calster et al. (2019), the calibration of risk prediction models is an

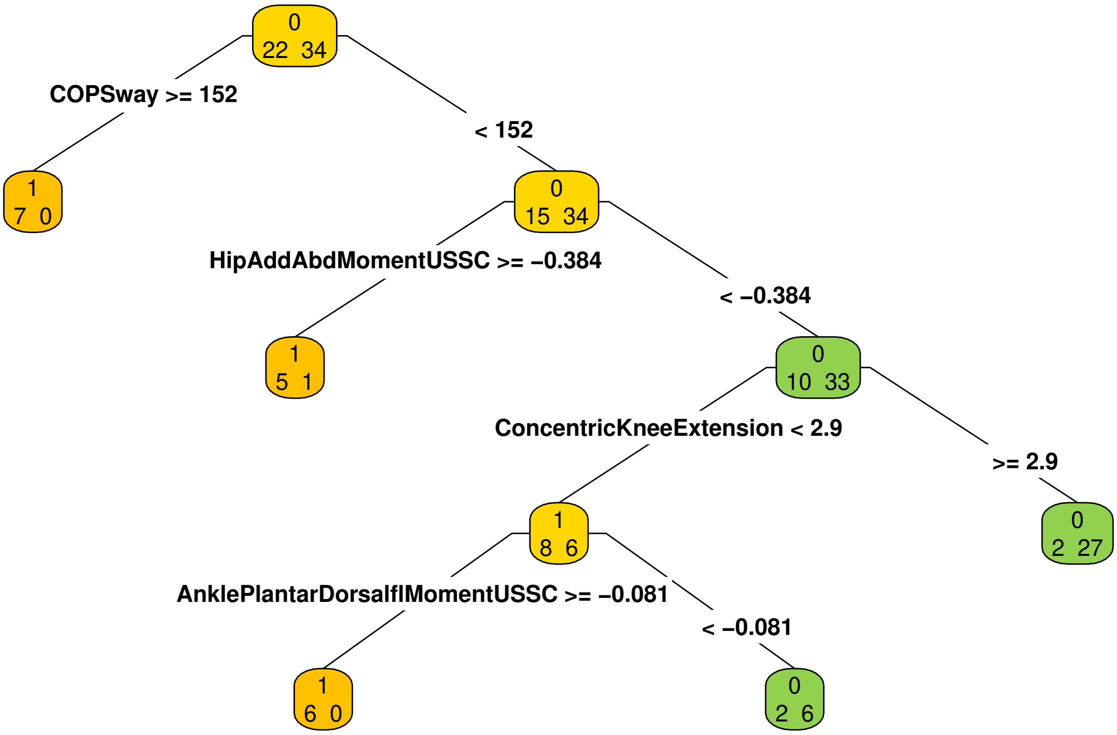
0
22 34
COPSway >= 152
< 152
1
7 0
0
15 34
HipAddAbdMomentUSSC >= −0.384
< −0.384
1
5 1
0
10 33
ConcentricKneeExtension < 2.9
>= 2.9
1
8 6
0
2 27
AnklePlantarDorsalflMomentUSSC >= −0.081
< −0.081
1
6 0
0
2 6

Figure 2: Classification tree obtained on the whole data set. In each node, the upper 0 or 1 gives the prediction of the respective node. In the second line of each node, on the left side is the number of injured players within the node and on the right side the number of uninjured players. The splitting criteria for sorting the data is added on the links between the nodes.

important aspect. Mean calibration is assessed via comparison of the true relative event frequency and the average predicted risk from the regarded models. In our data, we observe a proportion of 0.393 of injured players. This is also the average predicted risk by the post Lasso model and the classification tree. XGBoost predicts an injury probability of 0.456 on average, thus in general overestimating the risk. We also calculated the so-called calibration intercepts and slopes as suggested by Van Calster et al. (2019) for weak calibration assessment. By construction, the post Lasso model numerically meets the target values of 0 for the calibration intercept and 1 for the slope. For the classification tree, the deviation from these target values is less than 1%[3]. For XGBoost, the calibration intercept is -0.271. However, since the predicted class 1 probabilities are even perfectly linearly separable with regard to the true classes, calculation of the calibration slope is not possible.

5. Conclusion, limitations, and discussion

On a small scale, our work approaches the problem of predicting the binary injury status in young professional soccer players. Player-specific as well as biomechanical and neuromuscular features are considered for prediction. The latter are collected in a practical screening, consisting of tests of postural control, strength, and motion. Preliminary work on the data by Kolodziej et al. (2023) investigates the prediction capacity of several Lasso-regularized logistic regression models, yielding a model with a reduced penalty and a post coefficient re-estimation step as the best one. In our work, we compare it to the prediction power of other conventional and machine learning approaches on the same task. We also provide bootstrap-based confidence intervals to quantify the uncertainty of the obtained prediction performance measures.

All in all, no new method substantially outperforms the best post Lasso model with λ_{1se}, which reaches the largest value of Youden's index of 0.302, the largest AUC of 0.672, the largest predictive likelihood of 0.593, and the smallest Brier score of 0.228, using its best threshold. Its sensitivity and specificity values are 0.773 and 0.529, respectively. As the most important predictors for the injury status, it contains the concentric extensor peak torque of the knee, the transversal plane moment of the hip in a single-leg drop landing task, and the sway of the COP of a force plate in a postural control test (see also Kolodziej et al., 2023). The best XGBoost model and classification tree place the largest importance for injury prediction on the COP sway, and use the concentric knee extensor peak torque as the second or third most important variable. The corresponding high relevance of these two variables in the three algorithms generally confirms the results from Kolodziej et al. (2023) and is a strong argument for their importance in injury prediction. However, a drawback of our classical modeling approach is the inability to draw any conclusions on causality of these variables on the injury risk.

When investigating another population of players with similar inclusion criteria, our

[3] Note that for 13 observations, we changed the predicted class 1 probability from the respective tree node from 1 to 0.999. This was necessary for transforming the predicted probabilities for the logistic model fitting for the calibration evaluation.

findings of important predictors could possibly be generalized. However, caution is to be applied when it comes to identifying the best performing models. In our study with the small sample size of $n = 56$, it is entirely possible/probable that modern machine learning methods can not realize their full potential. Therefore, the Lasso model might be outperformed by other approaches with the inclusion of more observations, and we warn against generalizing the goodness of the Lasso model here to other settings. In case of availability of data where $n > p$, conventional statistical methods that struggle with high-dimensional data might also improve.

From a larger perspective beyond the specific problem of injuries in young professional soccer, we hope to provide a thorough and extensive guideline here on how several modeling approaches in a binary classification and prediction setting can be compared. We propose a roadmap for the analysis of small data sets with (typically expensive) high-quality variables, seeking to extract as much information as possible from data collected with significant efforts.

In our analysis, after choosing best thresholds for every modeling approach, we regard all prediction quality measures in parallel, regardless of whether they depend on a threshold. A legitimate alternative might be a two-stage process where the best modeling approaches are first identified based on threshold-independent measures like AUC, predictive likelihood, and Brier score. In a second step, the optimal threshold could be determined for the chosen approaches. In our case, the post Lasso model with λ_{1se} shows the best performance of all approaches for all three mentioned global measures.

The reliable prediction of soccer injuries in young professional male players remains a challenge. While the best model found here can identify a large proportion of players that are not going to be injured, it is less capable to detect players that are going to suffer an injury. Hence, the search for prediction models and larger data sets allowing for better overall predictive performance continues.

Optimization might be possible by not only conducting physical tests before the season, but also regularly within the season. However, the expenditure of time necessary for more tests would be a new issue. Moreover, while larger data sets of many teams would be desirable, dividing the work of collecting the data might affect their quality when it is not done in a standardized way (Petrie and Falkstein, 1998).

In the future, we intend to analyze the data again with time-to-event approaches, regarding the time between the screening and a potential injury as a more informative outcome compared to a binary one. Our hope is that the utilization of this additional information in the data may lead to an increase in prediction quality.

Acknowledgement

The authors would like to thank the participating youth academies, coaches, and players for their collaboration and the test team for their contribution to the data collection.

Appendix A Physical covariates

Table 5: Means ± standard deviations of the postural control covariates and strength covariates for all players, and separately for injured and non-injured players

	Covariate	all	injured	not injured
Postural Control				
Static	COP Sway	119.193 ± 24.318	129.730 ± 26.363	112.375 ± 20.529
Dynamic	DPSI	4.598 ± 0.937	4.520 ± 1.054	4.649 ± 0.866
Unstable	Path of Platf.	0.392 ± 0.189	0.443 ± 0.226	0.360 ± 0.155
Strength				
Trunk	Flexion	2.337 ± 0.482	2.209 ± 0.466	2.421 ± 0.481
	Extension	4.936 ± 1.054	4.801 ± 0.935	5.023 ± 1.129
	Flex. + Ext.	3.637 ± 0.670	3.505 ± 0.606	3.722 ± 0.704
	Flex./Ext.	0.488 ± 0.117	0.473 ± 0.120	0.497 ± 0.116
	Lat. Flex.	2.411 ± 0.498	2.309 ± 0.456	2.477 ± 0.520
	Lat. Flex. l/r	0.974 ± 0.176	0.979 ± 0.119	0.971 ± 0.206
	Tr. Rot.	1.978 ± 0.340	1.914 ± 0.334	2.020 ± 0.342
	Tr. Rot. l/r	1.016 ± 0.160	1.049 ± 0.189	0.995 ± 0.138
	Core Score	16.053 ± 2.561	15.456 ± 2.410	16.438 ± 2.615
Hip	Abduction	1.917 ± 0.324	1.896 ± 0.321	1.931 ± 0.330
	Adduction	2.074 ± 0.531	1.985 ± 0.463	2.131 ± 0.570
	Abd./Add.	0.959 ± 0.190	0.981 ± 0.166	0.945 ± 0.206
Knee	Conc. Ext.	3.084 ± 0.475	2.855 ± 0.472	3.233 ± 0.421
	Conc. Ext. l/r	1.004 ± 0.143	1.041 ± 0.178	0.981 ± 0.111
	Conc. Flex.	1.667 ± 0.255	1.565 ± 0.201	1.733 ± 0.268
	Conc. Flex. l/r	0.971 ± 0.110	0.988 ± 0.103	0.960 ± 0.115
	Ecc. Ext.	3.652 ± 0.706	3.382 ± 0.701	3.827 ± 0.661
	Ecc. Ext. l/r	1.016 ± 0.185	1.051 ± 0.193	0.993 ± 0.179
	Ecc. Flex.	2.124 ± 0.417	1.980 ± 0.322	2.216 ± 0.448
	Ecc. Flex. l/r	1.012 ± 0.137	1.020 ± 0.142	1.006 ± 0.135
	Conv. K. Rat.	0.545 ± 0.070	0.555 ± 0.070	0.539 ± 0.070
	Func. K. Rat.	0.693 ± 0.118	0.702 ± 0.116	0.688 ± 0.121

Abbreviations: COP: center of pressure, DPSI: Dynamic Postural Stability Index, Platf.: Platform, Flex.: Flexion, Ext.: Extension, Lat.: Lateral, l/r: left/right, Tr.: Transversal, Rot.: Rotation, Abd.: Abduction, Add.: Adduction, Conc.: Concentric, Ecc.: Eccentric, Conv.: Conventional, Func.: Functional, K. Rat.: Knee Ratio

Table 6: Means ± standard deviations of the joint kinematics covariates in the SLDL in degrees for all players, and separately for injured and non-injured players

	Covariate	all	injured	not injured
Joint Kinematics in the SLDL				
Ankle	Ev./Inv. IC	8.342 ± 3.724	8.464 ± 4.081	8.263 ± 3.536
	Ev./Inv. PEAK	9.468 ± 3.835	9.832 ± 4.087	9.232 ± 3.706
	Pl.fl./Do.fl. IC	24.339 ± 5.185	24.661 ± 4.822	24.132 ± 5.469
	Pl.fl./Do.fl. PEAK	26.315 ± 6.036	27.377 ± 6.028	25.628 ± 6.030
	Ext./Int. Rot. IC	1.730 ± 3.952	1.353 ± 4.029	1.975 ± 3.942
	Ext./Int. Rot. PEAK	3.154 ± 4.484	2.783 ± 4.653	3.394 ± 4.425
Hip	Add./Abd. IC	-9.642 ± 2.859	-10.059 ± 2.713	-9.372 ± 2.958
	Add./Abd. PEAK	0.360 ± 3.299	0.564 ± 3.519	0.229 ± 3.196
	Flex./Ext. IC	-30.004 ± 6.974	-30.057 ± 7.649	-29.970 ± 6.620
	Flex./Ext. PEAK	-44.372 ± 8.698	-44.399 ± 9.986	-44.354 ± 7.915
	Ext./Int. Rot. IC	-4.764 ± 4.573	-4.694 ± 5.678	-4.810 ± 3.785
	Ext./Int. Rot. PEAK	0.627 ± 4.230	1.353 ± 4.953	0.157 ± 3.693
Knee	Flex./Ext. IC	14.926 ± 4.927	14.651 ± 4.598	15.103 ± 5.188
	Flex./Ext. PEAK	52.218 ± 4.618	52.873 ± 4.396	51.794 ± 4.772
	Ext./Int. Rot. IC	-2.056 ± 3.966	-1.596 ± 4.196	-2.354 ± 3.844
	Ext./Int. Rot. PEAK	10.060 ± 4.433	9.553 ± 4.675	10.387 ± 4.309
	Add./Abd. IC	1.966 ± 2.935	1.639 ± 3.027	2.178 ± 2.899
	Add./Abd. PEAK	-0.833 ± 4.024	-1.934 ± 4.658	-0.120 ± 3.441

Abbreviations: SLDL: single-leg drop landing task, Ev.: Eversion, Inv.: Inversion, IC: initial contact (first moment of contact with the force plate), PEAK: peak value (maximal value within 100 ms after IC), Pl.fl.: Plantarflexion, Do.fl.: Dorsalflexion, Ext.: External, Int.: Internal, Rot.: Rotation, Add.: Adduction, Abd.: Abduction, Flex.: Flexion, Ext.: Extension

Table 7: Means ± standard deviations of the joint kinematics covariates in the USSC in degrees for all players, and separately for injured and non-injured players

	Covariate	all	injured	not injured
Joint Kinematics in the USSC				
Ankle	Ev./Inv. IC	12.858 ± 5.801	13.156 ± 6.115	12.665 ± 5.674
	Ev./Inv. PEAK	22.108 ± 4.928	22.400 ± 5.336	21.919 ± 4.719
	Pl.fl./Do.fl. IC	-2.652 ± 9.044	-0.931 ± 10.749	-3.766 ± 7.716
	Pl.fl./Do.fl. PEAK	3.448 ± 6.490	5.075 ± 6.863	2.396 ± 6.108
	Ext./Int. Rot. IC	-3.207 ± 5.028	-3.425 ± 5.820	-3.066 ± 4.531
	Ext./Int. Rot. PEAK	0.150 ± 5.173	-1.145 ± 5.297	0.988 ± 4.990
Hip	Add./Abd. IC	-7.095 ± 4.380	-6.951 ± 4.525	-7.188 ± 4.349
	Add./Abd. PEAK	-3.499 ± 4.535	-3.731 ± 4.635	-3.348 ± 4.533
	Flex./Ext. IC	-66.918 ± 8.794	-64.476 ± 10.712	-68.498 ± 7.022
	Flex./Ext. PEAK	-68.429 ± 8.495	-66.038 ± 10.058	-69.976 ± 7.045
	Ext./Int. Rot. IC	2.711 ± 4.684	2.667 ± 5.954	2.739 ± 3.742
	Ext./Int. Rot. PEAK	6.241 ± 5.402	6.634 ± 7.087	5.986 ± 4.063
Knee	Flex./Ext. IC	43.161 ± 11.834	39.393 ± 12.017	45.598 ± 11.221
	Flex./Ext. PEAK	56.807 ± 5.251	55.873 ± 5.855	57.411 ± 4.815
	Ext./Int. Rot. IC	7.247 ± 5.737	6.451 ± 6.415	7.763 ± 5.289
	Ext./Int. Rot. PEAK	12.656 ± 5.402	11.534 ± 5.968	13.382 ± 4.958
	Add./Abd. IC	-1.106 ± 4.343	-1.149 ± 4.054	-1.079 ± 4.579
	Add./Abd. PEAK	-6.094 ± 4.730	-6.370 ± 4.714	-5.915 ± 4.802

Abbreviations: USSC: unanticipated side-step cutting task, Ev.: Eversion, Inv.: Inversion, IC: initial contact (first moment of contact with the force plate), PEAK: peak value (maximal value within 100 ms after IC), Pl.fl.: Plantarflexion, Do.fl.: Dorsalflexion, Ext.: External, Int.: Internal, Rot.: Rotation, Add.: Adduction, Abd.: Abduction, Flex.: Flexion, Ext.: Extension

Table 8: Means ± standard deviations of the joint moment covariates in $\frac{\text{Nm}}{\text{kg}}$ and the vertical ground reaction forces in $\frac{\text{N}}{\text{kg}}$ for all players, and separately for injured and non-injured players

	Covariate	all	injured	not injured
PEAK	**Joint Moments**			
Ankle	Ev./Inv. USSC	-0.352 ± 0.097	-0.356 ± 0.084	-0.349 ± 0.106
	Ev./Inv. SLDL	-0.417 ± 0.116	-0.427 ± 0.125	-0.411 ± 0.112
	Pl.fl./Do.fl. USSC	-0.131 ± 0.178	-0.138 ± 0.192	-0.127 ± 0.172
	Pl.fl./Do.fl. SLDL	0.056 ± 0.021	0.054 ± 0.023	0.057 ± 0.020
	Ext./Int. Rot. USSC	-0.065 ± 0.061	-0.058 ± 0.062	-0.070 ± 0.060
	Ext./Int. Rot. SLDL	-0.363 ± 0.101	-0.363 ± 0.092	-0.363 ± 0.107
Hip	Add./Abd. USSC	-0.626 ± 0.268	-0.596 ± 0.278	-0.646 ± 0.264
	Add./Abd. SLDL	-0.990 ± 0.209	-1.005 ± 0.200	-0.980 ± 0.217
	Flex./Ext. USSC	3.752 ± 0.683	3.631 ± 0.694	3.829 ± 0.675
	Flex./Ext. SLDL	3.656 ± 0.543	3.765 ± 0.533	3.586 ± 0.545
	Ext./Int. Rot. USSC	-0.504 ± 0.222	-0.515 ± 0.187	-0.497 ± 0.244
	Ext./Int. Rot. SLDL	-0.651 ± 0.220	-0.724 ± 0.295	-0.604 ± 0.141
Knee	Flex./Ext. USSC	-2.238 ± 0.502	-2.258 ± 0.515	-2.224 ± 0.501
	Flex./Ext. SLDL	-2.503 ± 0.392	-2.499 ± 0.445	-2.505 ± 0.361
	Ext./Int. Rot. USSC	-0.320 ± 0.242	-0.334 ± 0.275	-0.311 ± 0.222
	Ext./Int. Rot. SLDL	-0.095 ± 0.073	-0.115 ± 0.106	-0.082 ± 0.037
	Add./Abd. USSC	0.741 ± 0.271	0.742 ± 0.257	0.740 ± 0.284
	Add./Abd. SLDL	0.875 ± 0.202	0.841 ± 0.186	0.897 ± 0.212
PEAK	**Vertical Ground Reaction Force**			
	vGRF USSC	20.877 ± 2.498	20.385 ± 2.491	21.195 ± 2.487
	vGRF SLDL	38.238 ± 3.857	37.996 ± 4.602	38.395 ± 3.355

Abbreviations: PEAK: peak value (maximal value within 100 ms after initial contact), Ev.: Eversion, Inv.: Inversion, USSC: unanticipated side-step cutting task, SLDL: single-leg drop landing task,Pl.fl.: Plantarflexion, Do.fl.: Dorsalflexion, Ext.: External, Int.: Internal, Rot.: Rotation, Add.: Adduction, Abd.: Abduction, Flex.: Flexion, Ext.: Extension, vGRF: vertical ground reaction force

Appendix B Covariate distributions

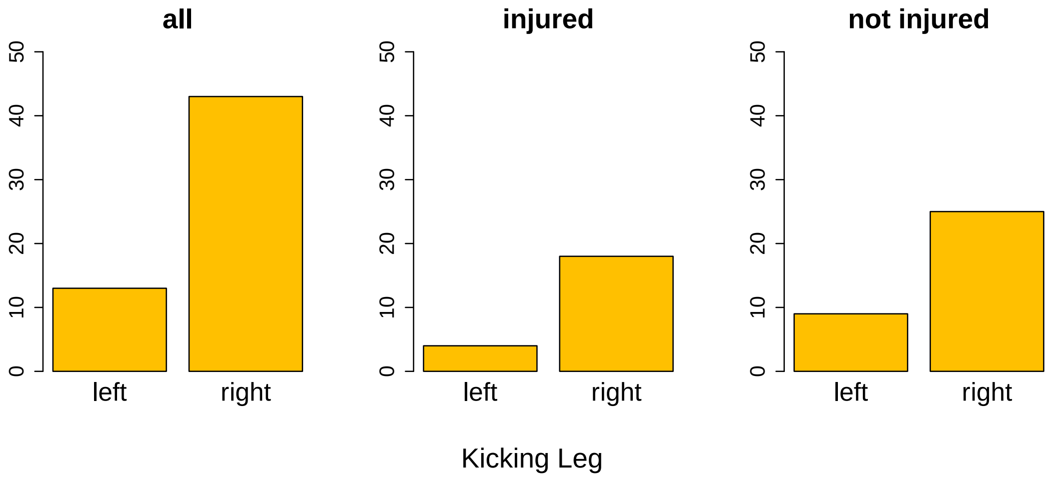

Figure 3: Distribution of the kicking leg for all players ($n = 56$), the injured players ($n_{\text{injured}} = 22$), and the non-injured players ($n_{\text{not injured}} = 34$).

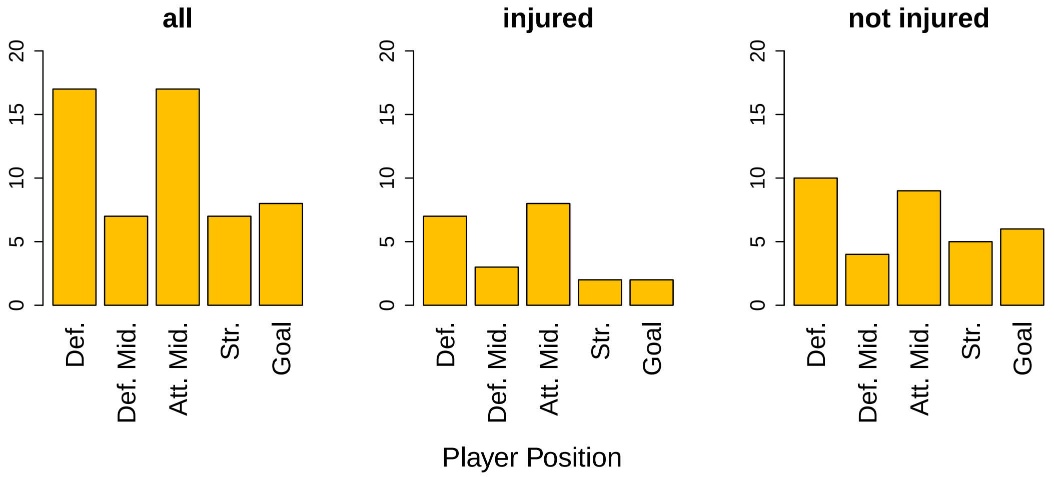

Figure 4: Distribution of the player position for all players ($n = 56$), the injured players ($n_{\text{injured}} = 22$), and the non-injured players ($n_{\text{not injured}} = 34$).
Abbreviations: Def.: Defense, Def. Mid.: Defensive Midfield, Att. Mid:, Attacking Midfield, Str.: Striker

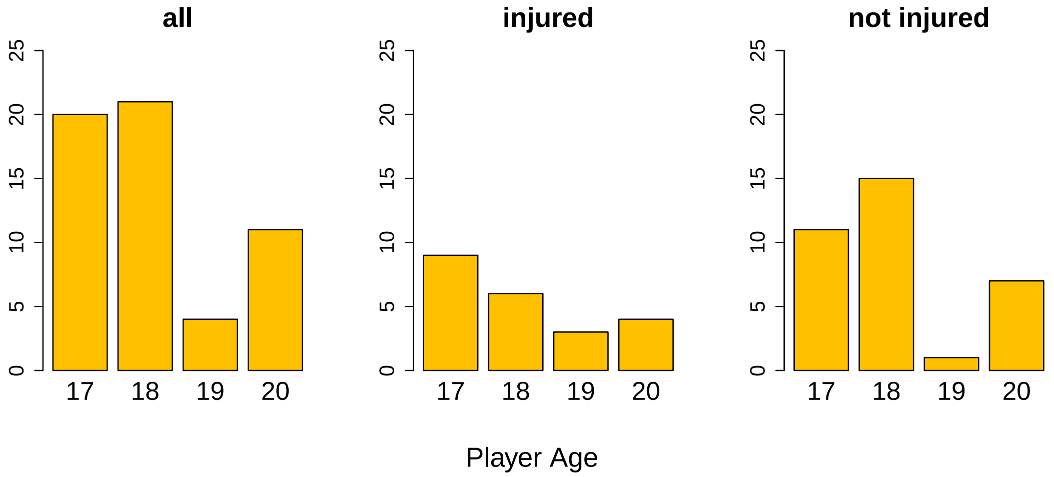

Figure 5: Distribution of the age for all players ($n = 56$), the injured players ($n_{\text{injured}} = 22$), and the non-injured players ($n_{\text{not injured}} = 34$).

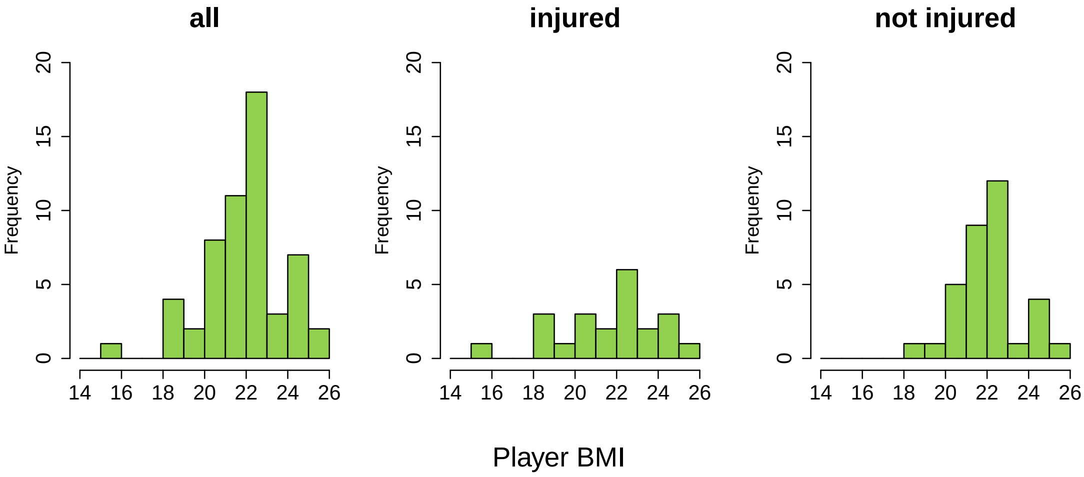

Figure 6: Distribution of the BMI for all players ($n = 56$), the injured players ($n_{\text{injured}} = 22$), and the non-injured players ($n_{\text{not injured}} = 34$).

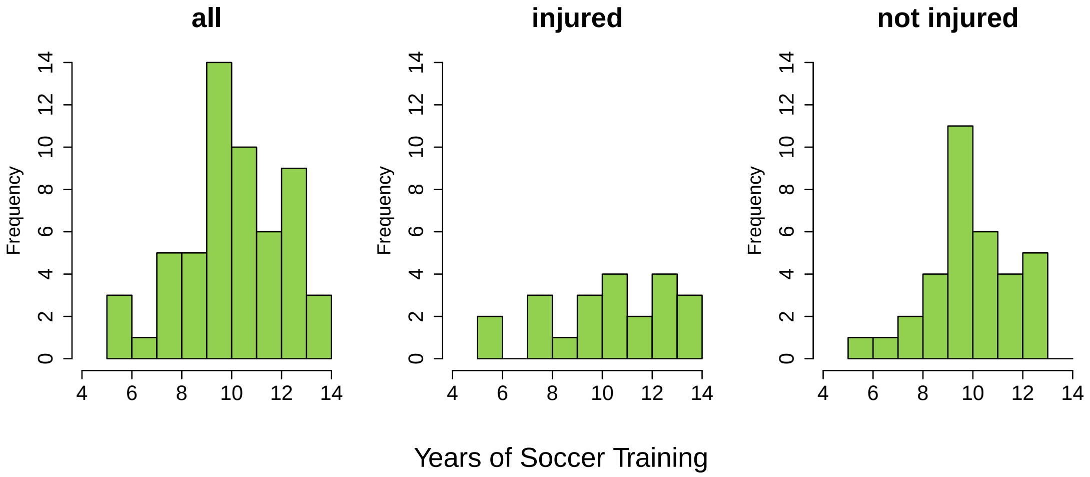

Figure 7: Distribution of the years of soccer training for all players ($n = 56$), the injured players ($n_{\text{injured}} = 22$), and the non-injured players ($n_{\text{not injured}} = 34$).

Appendix C Quality measure results

Table 9: Quality measure results for the threshold $\theta = 0.4$ (best value in **bold font**)

	Acc	Sens	Spec	Youd	AUC	Pred. L.	Brier
Lasso λ_{opt}	0.571	0.409	0.676	0.086	0.586	0.532	0.238
Lasso λ_{1se}	0.518	0.364	0.618	-0.019	0.508	0.527	0.257
Post Lasso λ_{opt}	0.554	0.455	0.618	0.072	0.638	0.577	0.239
Post Lasso λ_{1se}	0.589	0.455	0.676	**0.131**	**0.672**	**0.593**	**0.228**
Naive Bayes	0.536	0.455	0.588	0.043	0.515	0.536	0.420
LDA	0.482	0.364	0.559	-0.078	0.370	0.440	0.510
KNN	**0.625**	0.273	**0.853**	0.126	0.549	0.546	0.252
Tree	0.518	0.409	0.588	-0.003	0.473	0.508	0.424
Random Forest	0.518	0.500	0.529	0.029	0.449	0.509	0.258
XGBoost	0.464	**0.727**	0.294	0.021	0.549	0.509	0.253
SVM	0.554	0.136	0.824	-0.040	0.434	0.542	0.232
Constant	0.607	——	——	——	0.500	0.523	0.239

Abbreviations: Acc: Accuracy, Sens: Sensitivity, Spec: Specificity, Youd: Youden's index, AUC: Area under the ROC curve, Pred. L.: Predictive Bernoulli likelihood, Brier: Brier score

Appendix D Bootstrap results

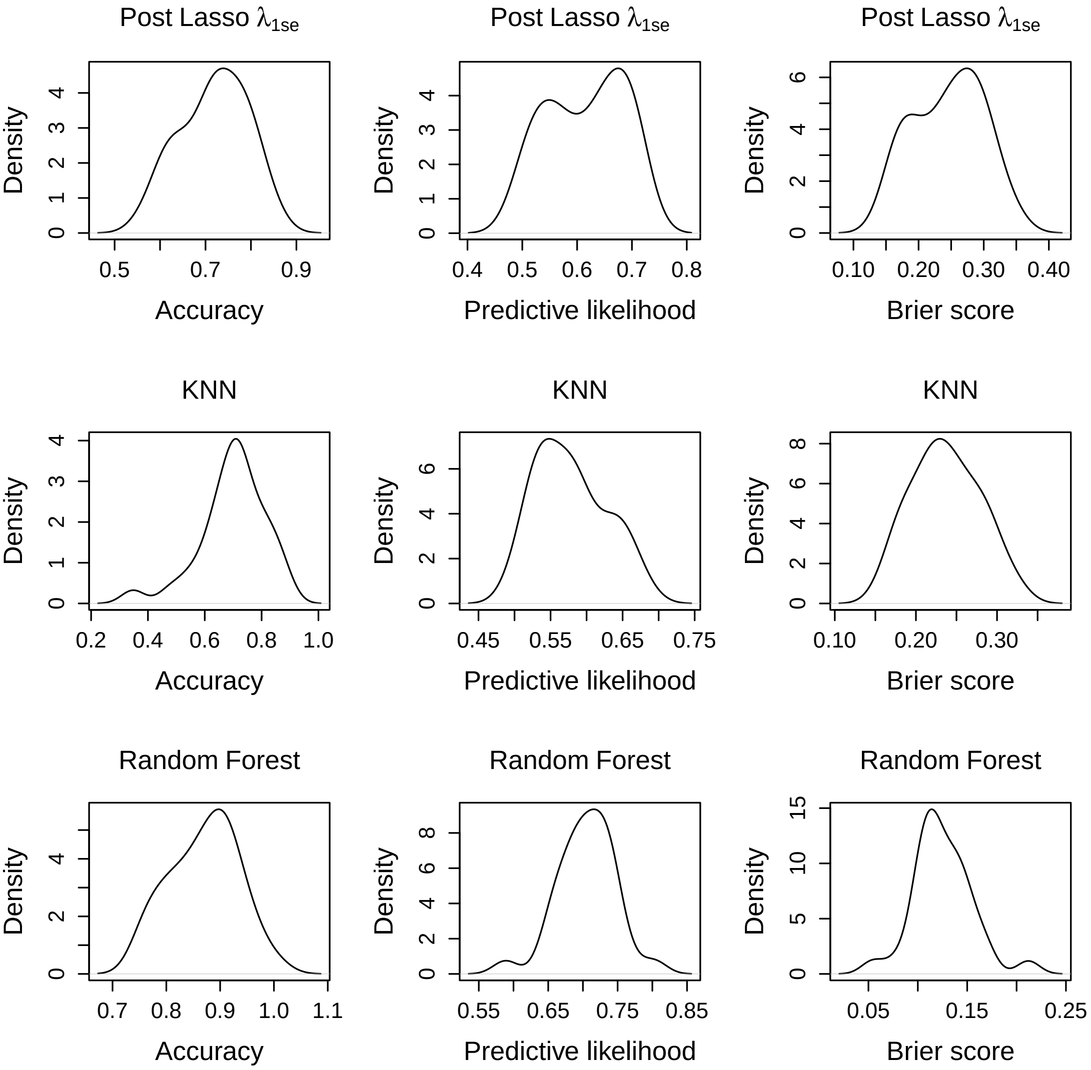

Figure 8: Density plots of bootstrapped values for three selected performance measures and modeling approaches, respectively

References

Breiman, L. (2001). Random forests. *Mach Learn*, 45:5–32.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and regression trees*. Routledge, New York.

Bullock, G. S., Mylott, J., Hughes, T., Nicholson, K. F., Riley, R. D., and 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.

Casalicchio, G. and Burk, L. (2024). Evaluation and benchmarking. In Bischl, B., Sonabend, R., Kotthoff, L., and Lang, M., editors, *Applied Machine Learning Using mlr3 in R*. CRC Press.

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York. Association for Computing Machinery.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and Yuan, J. (2023). *xgboost: Extreme Gradient Boosting*. R package version 1.7.5.1.

Claudino, J. G., Capanema, D. d. O., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., and 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*, 5:1–12.

Cleve, J. and Lämmel, U. (2020). *Data mining*. De Gruyter Oldenbourg, Berlin, Boston, third edition.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Ekstrand, J. (2013). Keeping your top players on the pitch: the key to football medicine at a professional level. *Br J Sports Med*, 47(12):723–724.

Ekstrand, J., Bengtsson, H., Waldén, M., Davison, M., Khan, K. M., and Hägglund, M. (2023). Hamstring injury rates have increased during recent seasons and now constitute 24% of all injuries in men's professional football: the UEFA Elite Club Injury Study from 2001/02 to 2021/22. *British Journal of Sports Medicine*, 57(5):292–298.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data*

Analysis, 38(4):367–378.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Hägglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H., and Ekstrand, J. (2013). Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, 47(12):738–742.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Series in Statistics. Springer, New York.

Hurley, O. A. (2016). Impact of player injuries on teams' mental states, and subsequent performances, at the Rugby World Cup 2015. *Frontiers in Psychology*, 7:807.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An introduction to statistical learning: with applications in R.* Springer Texts in Statistics. Springer, New York.

Junge, A. and Dvorak, J. (2004). Soccer injuries: a review on incidence and prevention. *Sports Medicine*, 34:929–938.

Kolodziej, M., Groll, A., Nolte, K., Willwacher, S., Alt, T., Schmidt, M., and Jaitner, T. (2023). Predictive modeling of lower extremity injury risk in male elite youth soccer players using least absolute shrinkage and selection operator regression. *Scandinavian Journal of Medicine & Science in Sports*, 33(6):1021–1033.

Kolodziej, M., Nolte, K., Schmidt, M., Alt, T., and Jaitner, T. (2021). Identification of neuromuscular performance parameters as risk factors of non-contact injuries in male elite youth soccer players: a preliminary study on 62 players with 25 non-contact injuries. *Frontiers in Sports and Active Living*, 3:615330.

Kolodziej, M., Willwacher, S., Nolte, K., Schmidt, M., and Jaitner, T. (2022). Biomechanical risk factors of injury-related single-leg movements in male elite youth soccer players. *Biomechanics*, 2(2):281–300.

Lang, M. and Schratz, P. (2023). *mlr3verse: easily install and load the 'mlr3' package family.* R package version 0.2.8.

López-Valenciano, A., Ayala, F., Puerta, J. M., Croix, M. D. S., Vera-García, F., Hernández-Sánchez, S., Ruiz-Pérez, I., and Myer, G. (2018). A preventive model for muscle injuries: a novel approach based on learning algorithms. *Medicine & Science in Sports & Exercise*, 50(5):915–927.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models.* Routledge, New York, second edition.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2023). *e1071: misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* R package version 1.7-14.

Milborrow, S. (2022). *rpart.plot: plot 'rpart' Models: an enhanced version of 'plot.rpart'.* R package version 3.1.1.

Molnar, C., Bischl, B., and Casalicchio, G. (2018). iml: An r package for interpretable machine learning. *JOSS*, 3(26):786.

Oliver, J. L., Ayala, F., De Ste Croix, M. B., Lloyd, R. S., Myer, G. D., and 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.

Petrie, T. A. and Falkstein, D. L. (1998). Methodological, measurement, and statistical issues in research on sport injury prediction. *Journal of Applied Sport Psychology*, 10(1):26–45.

Pfirrmann, D., Herbst, M., Ingelfinger, P., Simon, P., and Tug, S. (2016). Analysis of injury incidences in male professional adult and elite youth soccer players: a systematic review. *Journal of Athletic Training*, 51(5):410–424.

R Core Team (2022). *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Read, P. J., Oliver, J. L., De Ste Croix, M. B., Myer, G. D., and Lloyd, R. S. (2018). A prospective investigation to evaluate risk factors for lower extremity injury risk in male youth soccer players. *Scandinavian Journal of Medicine & Science in Sports*, 28(3):1244–1251.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.

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

Rossi, A., Pappalardo, L., and Cintia, P. (2022). A narrative review for a machine learning application in sports: an example based on injury forecasting in soccer. *Sports (Basel)*, 10(1):5.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., and Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS ONE*, 13(7):e0201264.

Schliep, K. and Hechenbichler, K. (2016). *kknn: weighted k-nearest neighbors.* R package version 1.3.1.

Tay, J. K., Narasimhan, B., and Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1).

Therneau, T. and Atkinson, B. (2023). *rpart: recursive partitioning and regression trees.* R package version 4.1.23.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tutz, G. (2012). *Regression for categorical data.* Cambridge University Press.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230.

Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S.* Springer, New York, fourth edition.

Wright, M. N. and Ziegler, A. (2017). ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Yale, K., Nisbet, R., and Miner, G. D. (2018). *Handbook of Statistical Analysis and Data Mining Applications.* Elsevier, London, San Diego, Cambridge, Oxford, second edition.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Yu, H. and Yang, J. (2001). A direct lda algorithm for high-dimensional data - with application to face recognition. *Pattern Recognition*, 34(10):2067–2070.