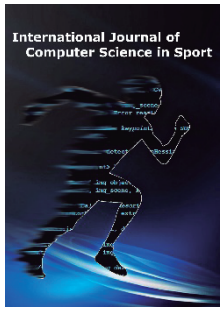
International Journal of Computer Science in Sport

Volume 20, Issue 2, 2021

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2021-0009



A Data Mining Approach to Predict Non-Contact Injuries in Young Soccer Players

Mandorino, M.[1], Figueiredo, A.J.[2], Cima, G.[3], Tessitore, A.[1]

[1]Department of Movement, Human and Health Sciences, University of Rome "Foro Italico", Rome, Italy

[2]University of Coimbra, Research Center for Sport and Physical Activity, Faculty of Sport Science and Physical Education, Coimbra, Portugal

[3]Computer, Control and Management Engineering Department, Sapienza University of Rome, Rome, Italy.

Abstract

Predicting and avoiding an injury is a challenging task. By exploiting data mining techniques, this paper aims to identify existing relationships between modifiable and non-modifiable risk factors, with the final goal of predicting non-contact injuries. Twenty-three young soccer players were monitored during an entire season, with a total of fifty-seven non-contact injuries identified. Anthropometric data were collected, and the maturity offset was calculated for each player. To quantify internal training/match load and recovery status of the players, we daily employed the session-RPE method and the total quality recovery (TQR) scale. Cumulative workloads and the acute: chronic workload ratio (ACWR) were calculated. To explore the relationship between the various risk factors and the onset of non-contact injuries, we performed a classification tree analysis. The classification tree model exhibited an acceptable discrimination (AUC=0.76), after receiver operating characteristic curve (ROC) analysis. A low state of recovery, a rapid increase in the training load, cumulative workload, and maturity offset were recognized by the data mining algorithm as the most important injury risk factors.

KEYWORDS: INJURY; YOUTH SOCCER; DATA MINING; PREDICTION; TRAINING LOAD

Introduction

Analyzing the injury risk factors to help prevent injuries in young soccer players has become a popular area of research in recent years. This increased interest arises from the need to prevent injuries, and therefore, avoid the side effects associated with them. In this regard, talent development stagnation (Richardson, Clarsen, Verhagen, & Stubbe, 2017), time-loss from sport participation and long-term sequelae (Timpka, Risto, & Björmsjö, 2008), as well as economic impact on the health-care system (Marshall, Lopatina, Lacny, & Emery, 2016), have been investigated. Notably, Polinder et al. (2016) estimated an annual cost of € 413 million in the Netherlands as a result of sport injuries, mainly caused by football/soccer due to the high incidence (30%).

Predicting and avoiding a potential injury is a challenging task. Indeed, an injury is a complex multifactorial process determined by the interaction of different modifiable (e.g. strength, flexibility) and non-modifiable (e.g. age, gender) factors (Bahr & Holme, 2003). By investigating the impact of non-modifiable factors on injury risk in young soccer players, Venturelli et al. (2011) identified an association between height and muscular strains. Kofotolis (2014) found an increase in ankle sprains with increasing age, while Johnson et al. (2020) observed a higher injury rate related to the peak height velocity (PHV) period. On the other hand, among the modifiable risk factors, different aspects have been investigated as the impact of neuromuscular control (Ko, Rosen, & Brown, 2018; Read, Oliver, De Ste Croix, Myer, & Lloyd, 2018) and strength level (De Ridder, Witvrouw, Dolphens, Roosen, & Van Ginckel, 2017). However, as observed by Meeuwisse et al. (2007), the interaction of all these factors is not sufficient to induce the onset of an injury. Indeed, according to the recent dynamic model proposed by Windt & Gabbett (2017) the workload represents the main vehicle that increases the athletes' susceptibility to injury.

To date, several studies have already analyzed the impact of workload on injury risk in soccer (Bacon & Mauger, 2017; Bowen, Gross, Gimpel, & Li, 2017; Brink et al., 2010), reporting different methods to monitor external (e.g. GPS) and internal (e.g. heart rate, perceived exertion) loads (Impellizzeri, Rampinini, Coutts, Sassi, & Marcora, 2004). According to the UEFA Elite Club Injury Study (McCall, Dupont, & Ekstrand, 2016), internal load markers were recognized to be relevant in identifying alarm bells related to the risk of injuries. By investigating these parameters in young soccer player, Brink et al. (2010) found an association between the session-rating of perceived exertion (S-RPE), monotony, and strain with traumatic injuries. Furthermore, in the same investigation, poor recovery values explained the raised injury predisposition. Similarly, Watson et al. (2017) reported that the injury risk increased with a high weekly and monthly training load. Besides, the acute: chronic workload ratio (ACWR), despite it has recently been questioned (Impellizzeri et al., 2021), was recognized in several studies as a valid tool in detecting dangerous conditions related to the onset of injuries (Andrade et al., 2020; Gabbett, 2016).

A detailed analysis of studies investigating modifiable and non-modifiable injury risk factors shows two main drawbacks: (1) these studies often adopt linear models. This is in contrast with the nature of injuries, which are the result of the complex non-linear interaction between many different conditions (Bittencourt et al., 2016; Bourdon et al., 2017); (2) these investigations are often limited in identifying an association between workload, modifiable and non-modifiable factors, and the onset of injuries. However, as pointed out by Fanchini et al. (2018), association is a concept that must not be confused with the ability of prediction, for which it is not possible to rely on the conventional linear statistical models (e.g., multivariate linear regression). Predictive analytics (i.e., the ability to forecast future outcomes based on historical data) requires more recent data mining technologies and techniques. The reason behind the use of data mining

techniques is linked to the nature of sport injuries. In fact, an injury is a multifactorial phenomenon and data mining allows to identify non-trivial, non-linear and unsuspected relations in the data (Montella, de Oña, Mauriello, Riccardi, & Silvestro, 2020). Moreover, differently from traditional statistical models which use probability theory to make inferences about population parameters of interest (Johnson, Borkowf, & Albert, 2007), the main goal of data mining is to build predictive models (Bhardwaj & Pal, 2012).

To the best of our knowledge, only few recent studies have adopted predictive analytics for injuries in soccer. Rossi et al. (2018) adopted machine learning techniques to predict non-contact injuries in twenty-six professional soccer players. Ayala et al. (2019), relying on pre-season psychological and neuromuscular measures, developed learning models to predict hamstring injuries in professional soccer. Similarly, Oliver et al. (2020) and Rommers et al. (2020), collected pre-season neuromuscular measures to identify young soccer players at risk of injury.

Thus, there is still a lack of knowledge regarding the impact of the training load on injury risk in young soccer players, who may be exposed to a different predisposition to injury compared to adult players due to growth processes (Vänttinen, Blomqvist, Nyman, & Häkkinen, 2011) and to different intensity and volume of training. For this reason, the results found in an adult population cannot be generalized to young athletes.

Therefore, the main purposes of the current study were to exploit data mining techniques in order to predict non-contact injuries in young soccer players, to identify the complex interactions between training load markers, modifiable and non-modifiable factors, and to shift from a unidimensional to a multidimensional approach.

Methods

Participants

Twenty-three U14 male soccer players (mean ± SD: age 13.5 ± 0.26 years, body mass 51.3 ± 8.5 kg, height 164 ± 7,3 cm) belonging to the same team were monitored during an entire season (2018/2019). Participants trained 3 days per week and competed once a week in an U14 sub-elite championship. A total of 119 trainings and 30 matches were monitored. The approval for data collection was obtained from the club as player's data were routinely collected over the course of the season (Winter & Maughan, 2009). The study was conducted in accordance with the Declaration of Helsinki (2013) and approved by the local research ethics committee of the University of Rome "Foro Italico" (code CARD 64/2020).

Injuries data collection

The team's physical therapist and strength and conditioning coach, supervised by the team's physician, provided an injury standardized data collection. To avoid variations in injuries definitions and methodologies, which can determine significant differences in results, the data collection followed the indications of the "consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries" (Fuller et al., 2006). According to this model, an injury was registered if the player was unable to take full part in future soccer training or match play. For the purpose of this study, only non-contact injuries were included in the data mining model, excluding contact injuries, which by their nature are not predictable. A descriptive analysis of injuries in relation to type and severity has been provided in Figure 1. Regarding their severity, injuries were classified as follows: slight (0 day); minimal (1-3 days); mild (4-7 days), moderate (8-28 days), severe (>28 days) and career-ending injuries.

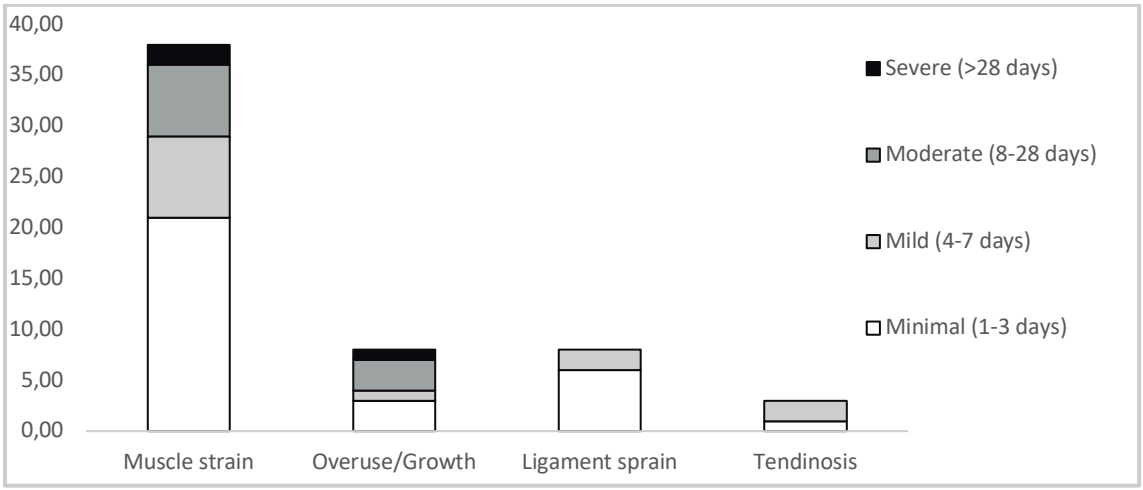


Figure 1. Injury type and severity of injuries. Severity of injuries; Slight (0 day), Minimal (1-3 days), Mild (4-7 days), Moderate (8-28 days), Severe (>28 days).

Anthropometric data

Players' standing and sitting height were measured through a fixed stadiometer (SECA 213, measuring range 20–205 cm, SECA Germany), while body mass was measured through a portable balance (SECA 762, Germany).

The Mirwald et al. (2002) algorithm was adopted to predict years from the peak height velocity (PHV), defined as maturity offset (R=0.94, and SEE=0.59). The present study employed the following male specific equation (Mirwald et al., 2002):

$$-9.236 + (0.0002708 * (Leg\ Length * Sitting\ Height)) + (-0.001663 * (Age * Leg\ Length)) + (0.007216 * (Age * Sitting\ Height)) + (0.02292 * (Weight/Height * 100)) \quad (1)$$

Collection of internal load markers and state of recovery

The session-RPE method (S-RPE) (Foster et al., 2001) was adopted to quantify players' training and match loads. S-RPE scores were calculated by multiplying the duration of each training or match for every single player by the rate of perceived exertion (RPE) value quantified through the CR-10 Borg's scale modified by Foster et al. (2001). The weekly load (WL) was obtained by adding the loads of all training sessions and matches over the course of a week, while cumulative loads were calculated for 2, 3, and 4 weeks (WL2, WL3, WL4). Moreover, monotony (the mean daily load divided by the standard deviation of the load over one week) and strain (weekly load multiplied by monotony) were estimated (Foster, 1998; Foster et al., 2001). The acute: chronic workload ratio (ACWR) was determined by dividing the weekly workload (acute load) by the average weekly workload over the previous 4 weeks (chronic load) (Gabbett, 2016; Malone et al., 2017). Evidence supporting an association between ACWR and injury is inconsistent and controversial (Impellizzeri et al., 2021), however, the debate on the usefulness of this parameter is still open (Seshadri et al., 2021; Zouhal, Boullosa, Ramirez-Campillo, Ali, & Granacher, 2021). Therefore, the ACWR was considered in the current study pending further studies that may definitively clarify the role of this parameter.

Players' perceived recovery status was estimated adopting the modified 10-point total quality recovery (TQR) scale (Gjaka, Tschan, Francioni, Tishkuaj, & Tessitore, 2016; Sansone, Tschan, Foster, & Tessitore, 2020) before each training and match. Athletes quantified the recovery status considering their psychophysical cues (i.e. mood states and muscle soreness), as suggested by the authors of the scale (Kenttä & Hassmén, 1998).

Statistical Analyses

The injury incidence was calculated as the number of injuries per 1000h of play exposure. Furthermore, training injury incidence and match injury incidence were calculated separately.

Data mining model setting

The classification and regression tree model (CART) was built to predict whether a player would get injured during the next training session based on training loads data (RPE, S-RPE, WL, WL2, WL3, WL4, Monotony, Strain ACWR), recovery status (TQR) and his biological characteristics (maturity offset, height and body mass). Therefore, all these features were inserted in the model as predictors and modelled on the binomial target variables, NO-INJURY (NI) and NON-CONTACT INJURY (NCI).

IBM SPSS Modeler 18.1 software was employed to develop the CART and to evaluate the performance of the model.

Data pre-processing

Standard pre-processing techniques were adopted to optimize the performance of the learning model. As first step, a data cleaning process was performed. Tuples reporting anomalies or errors were deleted and missing training loads data (0.6%) were replaced by the mean value of the player's corresponding parameter. All training load features were re-scaled adopting z-score transformation. Differently, a discretization process was employed for maturity offset, height, and weight parameters in order to find meaningful intervals (low, medium, high). Particularly, a fixed-width binning method was adopted, and three different bins (b1, b2, b3) were created. The following thresholds were identified:

- Maturity offset: (b1: $-1.3 \leq$ maturity offset < -0.55 years, b2: $-0.55 \leq$ maturity offset < 0.19 years, b3: $0.19 \leq$ maturity offset ≤ 0.95 years).
- Height: (b1: $146 \leq$ height < 156 cm, b2: $156 \leq$ height < 166 cm, b3: $166 \leq$ height ≤ 176 cm).
- Body mass: (b1: $34 \leq$ body mass < 46 kg, b2: $46 \leq$ body mass < 58 kg, b3: $58 \leq$ body mass ≤ 70 kg).

These methods ensure that each feature equally contributes to the learning process (D. Singh & Singh, 2020). As previously reported, CART decision tree algorithm was employed in the current study. Decision tree-based models generally do not require a re-scaling process. However, it might help with data manipulation and when it is needed to compare the performance with other algorithms. Therefore, re-scaling was adopted in the current study although not necessary to increase CART performance.

At this stage, to evaluate data mining algorithm performance, the dataset was split in training set and test set. A total of 2501 observations formed the dataset, 70% of them made up the training set and 30% the test set. However, the dataset was highly imbalanced since the class NI was the most represented condition (98%). Class imbalance refers to the condition when one class is less represented than the other (Kuhn & Johnson, 2013). An imbalanced dataset, as in our case, can

impair the predictive ability of the data mining model (Kuhn & Johnson, 2013). As suggested in previous studies (Carey et al., 2018; Ruddy et al., 2018), to cope class imbalance, synthetic over-sampling techniques (SMOTE) were applied to training set. The SMOTE technique generates randomly new examples of the minority class allowing to rebalance the dataset (Chawla, 2005). The SMOTE technique was applied only on the training set, while the test set was separated to preserve the original samples.

Finally, to improve prediction performance of the model and to have a better understanding of data, a feature selection was performed. Particularly, a filter method based on the use of Pearson's correlation coefficient was employed in the current study (Chandrashekar & Sahin, 2014).

Algorithm selection

A wide range of algorithms were developed in data mining field. Support Vector Machine (SVM), Neural Networks (NN) and ensemble learning methods such as Random Forests (RF) became popular in binary classification tasks due to their ability to identify non-linear patterns (Cortez & Embrechts, 2013). Despite their accurate predictions, these algorithms are generally considered black-box models due to their inability to be easily understood by humans (Bourdon et al., 2017; Cortez & Embrechts, 2013). Considering the importance of providing practical implications in sports science field (Bourdon et al., 2017), CART decision tree algorithm was selected for the specific purpose of the current study. Indeed, CART decision tree presents several advantages: (1) it can be used for binomial and multinomial classification (2) it can handle both numerical and categorical data, and (3) as discussed in previous studies (S. Singh & Gupta, 2014), CART provides an easy interpretability of the outcome model from a human perspective.

All data pre-processing steps performed before the CART algorithm setting were presented in Figure 2.

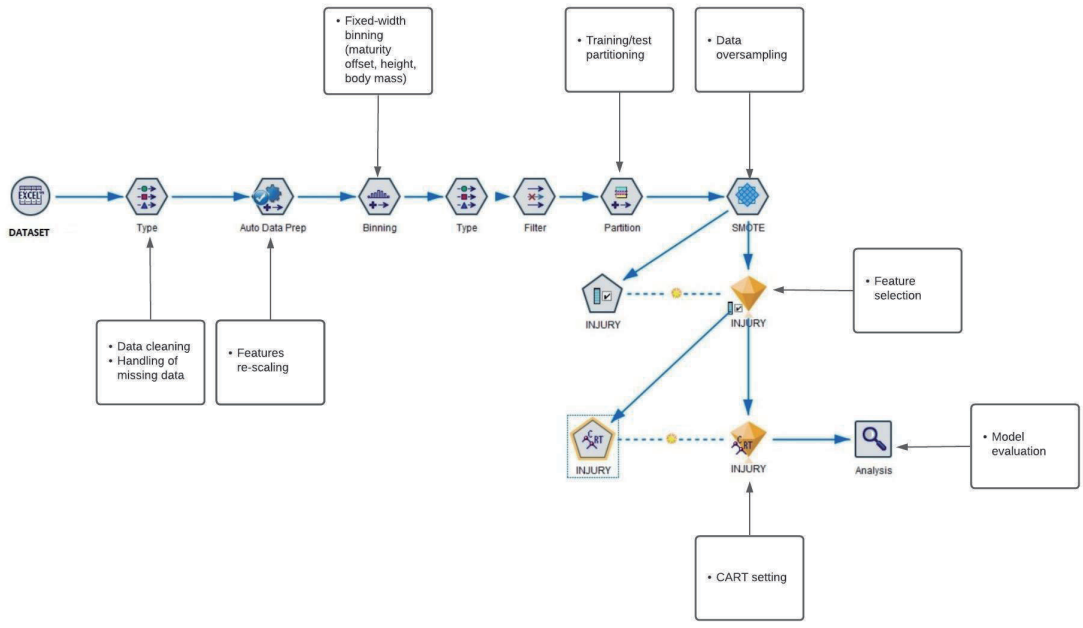


Figure 2. Data pre-processing steps

Classification trees

A tree is a directed acyclic graph whose undirected version is still acyclic. A tree is called a binary tree if each node has at most two outgoing edges (an edge is a link from a node, called parent node, to another node, called child node). In a binary tree, the children of a parent are referred to as the left children and the right children. The root node is the only node in the tree that has no incoming edges. A node with no children is called a leaf node.

Classification tree is a nonlinear and nonparametric supervised learning technique for predictive analytics used in statistics, data mining, and machine learning. Our study particularly focused on the framework of the CART algorithm (Breiman, Friedman, Stone, & Olshen, 1984), using the Gini index as node impurity measure.

In our study, we adopted two criteria to stop the growth of the tree: (1) minimum decrease in the impurity equal to 0.0001; and (2) maximum number of levels of the tree equal to eight. Moreover, pruning technique was employed to avoid overfitting. Considering the imbalanced dataset, a cost-sensitive leaning approach was employed. The cost matrix for cost-sensitive classifier was set to C $\begin{bmatrix} 0 & 1 \\ 1.5 & 0 \end{bmatrix}$ where a false positive had a cost of 1.5 and a false negative had a cost of 1. This setting was selected to minimize false positives.

The hyperparameters used within the CART model were summarized in table 1.

Table 1. Hyperparameters tuning

MODEL	HYPERPARAMETERS
CART	1. Maximum Tree Depth: 8 2. Overfitting prevention: pruning technique 3. Misclassification costs: C $\begin{bmatrix} 0 & 1 \\ 1.5 & 0 \end{bmatrix}$ 4. Impurity Measure: Gini 5. Minimum change in impurity: 0.0001

Model evaluation

To measure the performance of the data mining model, sensitivity and specificity were calculated. Sensitivity (true positive rate) measures the ability of the model to correctly detect the positive condition (NCI). It was calculated as follows: Sensitivity= [true positive / (true positive + false negative)] x 100. Instead, Specificity (true negative rate) measures the ability of the model to correctly detect the negative condition (NI). It was calculated as follows: Specificity = [true negative / (true negative + false positive)] x 100.

Moreover, once the classification tree model was provided, a ROC analysis was adopted to evaluate its predictive accuracy. Through this method, a graphical plot was produced, where the sensitivity on the vertical axis against the specificity on the horizontal axis were represented. When ROC curve was created, the area under the curve (AUC) was calculated to estimate the classification accuracy of the model. According to Hosmer Jr et al. (2013), an AUC of 0.5 suggests no discrimination, between 0.51 and 0.69 is considered poor discrimination, 0.70-0.79 is acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 outstanding.

Results

A total of 57 non-contact injuries were recorded during the entire soccer season. An overall injury incidence of 14.5 per 1000 h was revealed, with a training injury incidence of 12.3 and a match injury incidence of 29.5 per 1000 playing hours. More than half of injuries (67%) were classified as muscle strains. Further information regarding injury type and severity are presented in Figure 1.

A classification tree analysis was adopted to predict the onset of injuries. After the feature selection process, the following independent injury risk factors were included in our model: height, body mass, maturity offset, RPE, strain, WL2, WL3, WL4, ACWR and TQR.

The model's sensitivity was 70% and its specificity was 79%. The outcomes produced by the CART algorithm tested on the test set were presented in the confusion matrix (Table 2). After the ROC analysis, the model exhibited an AUC of 0.76, showing an acceptable discrimination.

Table 2. Confusion Matrix of INJURY classification

		PREDICTED CLASS	
		INJURY	NO-INJURY
ACTUAL CLASS	INJURY	14 (TP)	6 (FN)
	NO-INJURY	154 (FP)	587 (TN)

TP=true positive
FP=false positive
FN=false negative
TN=true negative

The CART classification tree produced 23 nodes, of which 12 terminal nodes. To facilitate the understanding of the decision tree model, the values, which were previously re-scaled adopting z-score transformation or clustered through the binning procedure, were reconverted into their original value. The CART tree, modelled on the training set data, was presented in Figure 3. Red color was used to display the NI condition, while blue color the NCI condition. The first split based on the recovery status (TQR). A TQR higher than 8 AU (node 2) produced a greater proportion of the NI condition (NI=80.2%). The node 2 further split in relation to WL3. WL3 lower than 2532 AU, associated with a strain lower than 909.68 AU (node 9), produced only NI conditions.

Instead, node 1 split according to ACWR. An ACWR lower than 0.76 reduced the probability of the NCI condition. Node 4 also split in relation to WL3. A WL3 lower than 5455 AU together with a WL4 higher than 3740 AU increased the prevalence of the NCI condition (node 12, NCI=74.73%). Instead, node 16 further split according to maturity offset. Of the 115 observations present inside the node 17 (maturity offset < -0.55), 56.5% were represented by the condition NI. Differently, of the 1186 observations inside the node 18, 79.59% were represented by the condition NCI. Both nodes (17, 18) produced terminal nodes based on the TQR values.

Variable importance analysis (Figure 4) identified TQR as the most important variable.

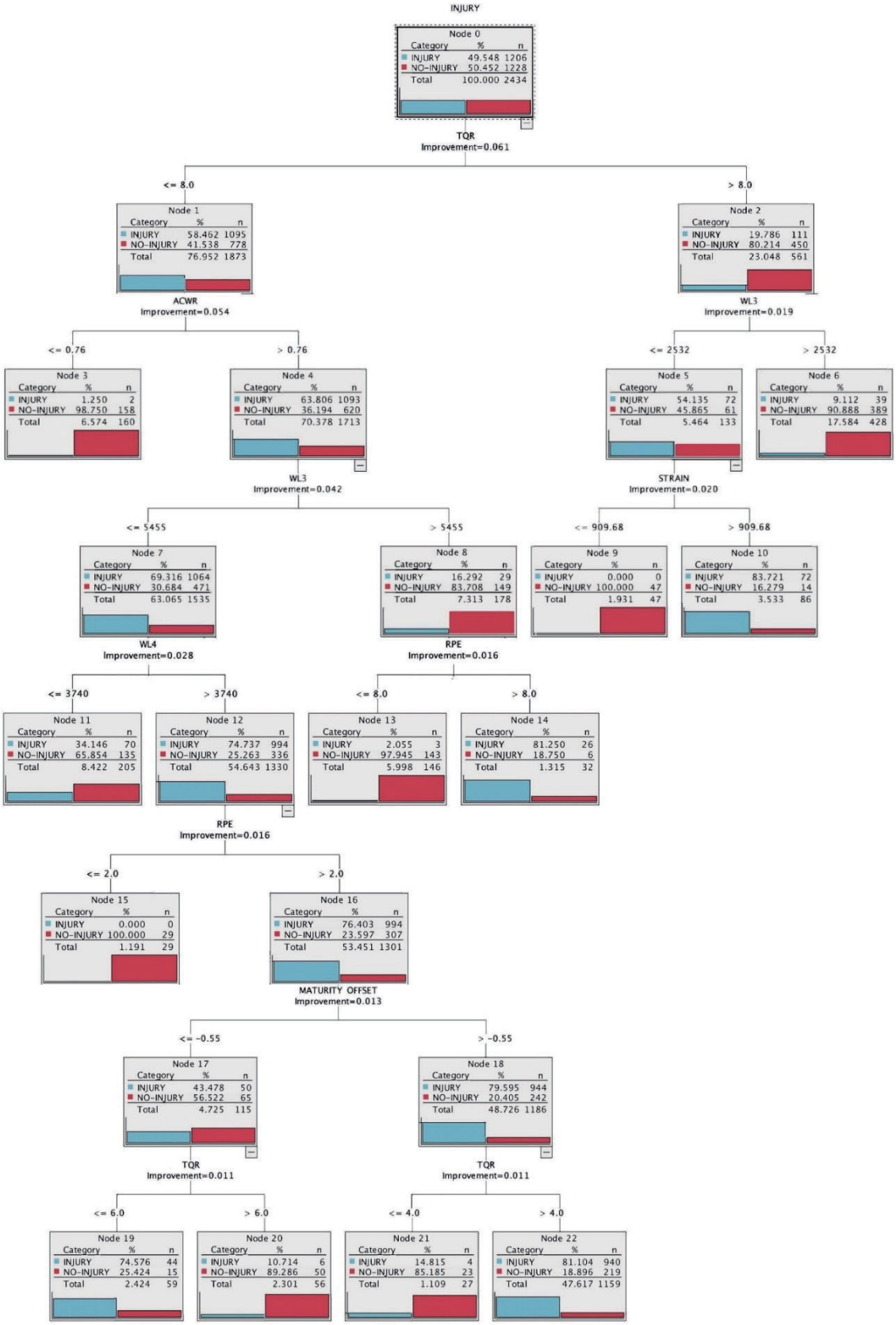


Figure 3. CART classification tree model. TQR=total quality recovery; ACWR= acute: chronic workload ratio; WL=weekly workload; WL3= cumulative workload of the previous 3 weeks; WL4= cumulative workload of the previous 4 weeks

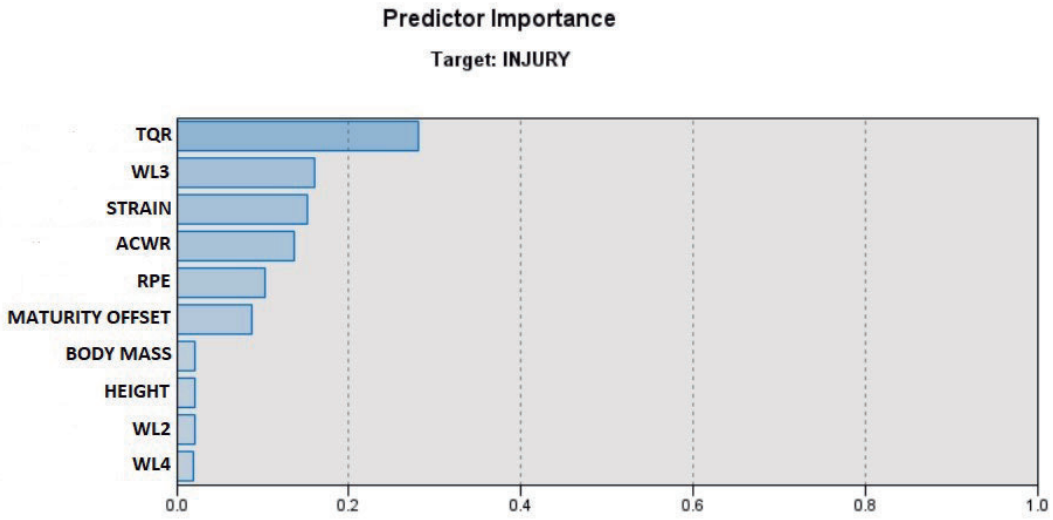


Figure 4. Variable importance for the dependent variable INJURY. TQR=total quality recovery; ACWR= acute: chronic workload ratio; WL2= cumulative workload of the previous 2 weeks; WL3= cumulative workload of the previous 3 weeks; WL4= cumulative workload of the previous 4 weeks.

Discussion

The data mining techniques applied in this study have proven to be effective in predicting with acceptable accuracy the onset of non-contact injuries in young soccer players. Particularly, our model revealed how recovery status, internal load markers, modifiable (height, body mass) and non-modifiable factors (maturity status), interacting each other, can modify the predisposition to risk of injuries. Among the wide range of data mining algorithms, the CART was selected. Although other algorithms such as k-nearest neighbors (KNN) or extreme gradient boosting (XGBoost) showed a good predictive ability in previous studies (Rommers et al., 2020; Vallance, Sutton-Charani, Imoussaten, Montmain, & Perrey, 2020), they are often considered as black-box models. The reason of this fact is linked to their difficulties of interpretation and their inability to provide information regarding the interaction between the different features (Cortez & Embrechts, 2013).

The main purpose of this study was not only to predict non-contact injuries in young soccer players, but also to identify the complex interactions between training load markers, modifiable and non-modifiable factors adopting a data mining approach. Therefore, the CART algorithm was implemented to support coaches and physical trainers in understanding the mechanism underlying the risk of injury. Being able to predict injuries in young athletes would allow to limit the side effects associated with them (e.g., talent development stagnation, health care costs, long-term sequelae). Furthermore, identifying the mechanism underlying them would help sport practitioners in properly managing the weekly training load and in promoting adequate prevention strategies.

The model implemented in our study exhibited an acceptable discrimination (AUC=0.76) according to Hosmer Jr et al. (2013) classification. The AUC value is in line with previous studies (Rossi et al., 2018). Moreover, the sensitivity and specificity found in the current study (70% and 79%) were similar to the results (77% and 84%) reported by Ayala et al. (2019). However, it is necessary to specify that these authors conducted their studies on different populations, using different parameters and adopting different algorithms.

The classification tree model built in the current study produced two branches starting from the recovery status (Figure 3), that was recognized as the most relevant feature after variable importance analysis (Figure 4). When the TQR value was higher than 8 AU, a higher proportion of the NI condition (80.2%) was identified by the model. However, when the high recovery status was accompanied by a lower WL3 (<2532 AU) and higher strain values (>909.68 AU), the risk of injury increased, as witnessed by the higher proportion of the NCI condition (NCI=83.72%). In accordance with previous studies, a higher cumulative workload may be protective against injury (Malone et al., 2017). Indeed, a higher workload over a chronic period may induce positive physical adaptation, reducing the influence of fatigue and consequently the risk of injury (Hulin et al., 2014). In line with these findings, also in the current study, when a low cumulative workload was associated to a higher strain, the probability of injury increased. The players' low 'fitness' level (chronic load) exhibited in this study (node 5) might not be sufficient to cope with a high strain (Delecroix, Mccall, Dawson, Berthoin, & Dupont, 2019), determining higher predisposition to injury (node 10).

Moving to the other branch, a TQR score lower than 8.0 AU produced a higher proportion of the NCI condition (NCI=58.46%). Indeed, an imbalance between recovery status and training load may represent a dangerous condition in terms of injury risk, as previously highlighted by Kenttä & Hassmén (1998). However, according to Wang et al. (2020), the probability of non-contact injury drastically reduced with an ACWR lower than 0.76 (node 3). Differently, an ACWR higher than 0.76 (node 4) produced a greater proportion of the NCI condition (63.80%). Nevertheless, the threshold value found in the current study by the classification tree model (0.76) is lower compared to the 'danger zone' (>1.5) identified by Gabbett (2016). This may be explained by the individual characteristics of the athletes (e.g., chronological age, biological age, years of training) which may lead to a different individual training loads tolerance. Although the usefulness of the ACWR was recently questioned (Impellizzeri et al., 2021), several studies identified an association between this parameter and the risk of non-contact injuries (Fanchini et al., 2018; McCall, Dupont, & Ekstrand, 2018). However, despite the association, the poor ACWR predictive ability revealed by these studies may be related to the fact that this parameter was studied within unidimensional or linear approaches. As previously stated, an injury is a complex non-linear multifactorial phenomenon, therefore, as other training load markers, ACWR could provide relevant information if inserted inside a multidimensional approach as that of data mining. The role of this parameter in a complex multidimensional approach is outside the scope of this paper, and we leave it as a future work.

Continuing the interpretation of the classification tree model, node 4 further split according to WL3. Also in this case, a higher WL3 (>5455 AU) reduced the probability of non-contact injury (NCI=83.70%). Nevertheless, a higher proportion of the NCI condition (NCI=81.25%) was identified if a higher WL3 was associated with extremely intense training sessions (node 14, RPE>8 AU). Similarly, a RPE higher than 2 AU (node 16) increased the probability of NCI if associated with a WL4 greater than 3740 AU (NCI=76.40%). As previously reported, high cumulative workload may prevent the risk of non-contact injuries. However, these results also highlight that injury risk may increase when high loads are maintained for a long period (Jaspers et al., 2018). We could speculate that high cumulative loads raise the state of fatigue in young players, making them more prone to injury. Consequently, fatigued athletes, subjected to strenuous exercise (high RPE values), may exhibit higher predisposition to injury. Therefore, a daily, weekly, and monthly training load monitoring is essential to promptly identify alarm bells related to the risk of non-contact injury.

In addition, node 16 further split in relation to maturity offset, showing a higher proportion of the NCI condition for players characterized by values greater than -0.55 (node 18). The predicted maturity offset, defined as the years before or after PHV (Mirwald et al., 2002), is a useful non-

invasive somatic indicator that predicts the time during which the athletes will experience their adolescent growth spurt (Malina, Bouchard, & Bar-Or, 2004). Particularly, the 6 months before and after PHV (maturity offset ranging from -0.5 to +0.5) have been identified as a critical period for the onset of injuries in young soccer players (Bult, Barendrecht, & Tak, 2018; van der Sluis et al., 2014). During this time, known also as the period of 'adolescent awkwardness' (Philippaerts et al., 2006), young athletes experience a decline in performance and motor control. The alteration in motor coordination combined with a rapid growth in muscle-, tendon-, ligament- and bone-structures may increase the risk of injuries. The classification tree model developed in the current study confirms these results: the players entering in the PHV period (maturity offset >-0.55) exhibited a greater probability of non-contact injuries compared to less mature players. Continuing the tree model discussion, it is worthy to note how both node 17 and node 18 split according to the TQR score. Interestingly, low TQR values determined a dangerous condition for less mature players (node 19), while TQR scores higher than 4 AU increased probability of NCI in more mature players. These results allow us to further emphasize the concept that players, characterized by different biological status, may exhibit contrasting physiological responses to training and show a different predisposition to injury (Towlson et al., 2020). The other features (e.g., height, body mass, WL2, WL4), as displayed in Figure 4, were not considered relevant within the prediction model.

In summary, our classification tree model confirmed that: (1) a poor recovery status (node 1) may increase the risk of injuries, (2) a 'spike' in the training load (node 4), as well as an inadequate training stimulus (low chronic training load), may increase the susceptibility to injury (node 5 and node 7), (3) maturity status may influence the predisposition to non-contact injuries, particularly in more mature players (node 18), (4) strenuous exercise may increase susceptibility to injury (node 14). Therefore, the classification tree model allowed to overcome the concept of association, to increase the ability to predict injuries (Fanchini et al., 2018), and to move from a unidimensional to a multidimensional approach.

Although the CART revealed acceptable discrimination and allowed to understand the complex interactions between features, the application of the model within a real context needs some considerations. First of all, the model was able to correctly predict 14 of the 20 non-contact injuries included in the test set, but at the same time it produced a high number of false positives (low precision), as shown in Table 2. Although from a clinical point of view the false negatives may produce a worse health impact (Petticrew, Sowden, Lister-Sharp, & Wright, 2000), at the same time a high number of false positives may lead a coach to 'stop' a young player several times, increasing the time-loss from sport participation.

Moreover, the CART was tested on a sample of only twenty-three U14 young soccer players and PHV, a specific biological indicator for adolescent athletes, was employed in the model.

Therefore, the results may not be generalized to older soccer players. Future studies should strengthen the model increasing the sample size, involving players of different ages and in multiple seasons, in order to improve the predictive ability of the model.

Predicting an injury with high accuracy continues to be a complex task due to its multifactorial nature. Despite the limitations previously reported, the following study allowed to investigate the complex interactions between workload, modifiable and non-modifiable risk factors, and to move from a unidimensional to a multidimensional approach.

Conclusions and Future Directions

The classification tree model developed in the present study was able to predict with acceptable discrimination non-contact injuries in young soccer players. Data mining allowed to investigate

how workload, modifiable and non-modifiable risk factors, interacting with each other, modify predisposition to injuries. The model could be used to identify players at risk of injury and consequently to promote prevention strategies. However, to further improve our prediction results, as future work we aim to collect more parameters (e.g., external load, heart rate and sleep quality data), and to increase sample size. Moreover, we aim to acquire more training data (during multiple seasons), for example by exploiting the various semantic technologies proposed in the computer science literature (Cima, 2017; Cima, Lenzerini, & Poggi, 2017), which allow obtaining high-quality data without (or, with little) manual intervention.

Acknowledgements

The authors would like to thank the club Pro Calcio Tor Sapienza (including contact persons, medical staff, coaching staff, and all players) for their participation in the study.

References

Andrade, R., Wik, E. H., Rebelo-Marques, A., Blanch, P., Whiteley, R., Espregueira-Mendes, J., & Gabbett, T. J. (2020). Is the acute: Chronic workload ratio (ACWR) associated with risk of time-loss injury in professional team sports? A systematic review of methodology, variables and injury risk in practical situations. *Sports medicine*, 1–23.

Ayala, F., López-Valenciano, A., Jose, A., De Ste Croix, M. B., Vera-García, F., García-Vaquero, M., … Myer, G. (2019). A preventive model for hamstring injuries in professional soccer: Learning algorithms. *International journal of sports medicine*, *40*(5), 344–353.

Bacon, C. S., & Mauger, A. R. (2017). Prediction of overuse injuries in professional u18-u21 footballers using metrics of training distance and intensity. *The Journal of Strength & Conditioning Research*, *31*(11), 3067–3076.

Bahr, R., & Holme, I. (2003). Risk factors for sports injuries—A methodological approach. *British journal of sports medicine*, *37*(5), 384–392.

Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

Bittencourt, N. F. N., Meeuwisse, W. H., Mendonça, L. D., Nettel-Aguirre, A., Ocarino, J. M., & Fonseca, S. T. (2016). Complex systems approach for sports injuries: Moving from risk factor identification to injury pattern recognition—Narrative review and new concept. *British journal of sports medicine*, *50*(21), 1309–1314.

Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., … Gregson, W. (2017). Monitoring athlete training loads: Consensus statement. *International journal of sports physiology and performance*, *12*(s2), S2-161-S2-170.

Bowen, L., Gross, A. S., Gimpel, M., & Li, F.-X. (2017). Accumulated workloads and the acute: Chronic workload ratio relate to injury risk in elite youth football players. *British journal of sports medicine*, *51*(5), 452–459.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brink, M. S., Visscher, C., Arends, S., Zwerver, J., Post, W. J., & Lemmink, K. A. (2010). Monitoring stress and recovery: New insights for the prevention of injuries and illnesses in elite youth soccer players. *British journal of sports medicine*, *44*(11), 809–815.

Bult, H. J., Barendrecht, M., & Tak, I. J. R. (2018). Injury risk and injury burden are related to age group and peak height velocity among talented male youth soccer players. *Orthopaedic journal of sports medicine*, *6*(12), 2325967118811042.

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

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In O. Maimon & L. Rokach (A c. Di), *Data Mining and Knowledge Discovery Handbook* (pagg. 853–867). Boston, MA: Springer US. https://doi.org/10.1007/0-387-25465-X_40

Cima, G. (2017). Preliminary results on ontology-based open data publishing. In A. Artale, B. Glimm, & R. Kontchakov (A c. Di), *Proceedings of the 30th international workshop on description logics, montpellier, france, july 18-21, 2017*. CEUR-WS.org. Recuperato da http://ceur-ws.org/Vol-1879/paper24.pdf

Cima, G., Lenzerini, M., & Poggi, A. (2017). Semantic technology for open data publishing. *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, 1–1.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1–17.

De Ridder, R., Witvrouw, E., Dolphens, M., Roosen, P., & Van Ginckel, A. (2017). Hip strength as an intrinsic risk factor for lateral ankle sprains in youth soccer players: A 3-season prospective study. *The American journal of sports medicine*, *45*(2), 410–416.

Delecroix, B., Mccall, A., Dawson, B., Berthoin, S., & Dupont, G. (2019). Workload monotony, strain and non-contact injury incidence in professional football players. *Science and Medicine in Football*, *3*(2), 105–108.

Fanchini, M., Rampinini, E., Riggio, M., Coutts, A. J., Pecci, C., & McCall, A. (2018). Despite association, the acute: Chronic work load ratio does not predict non-contact injury in elite footballers. *Science and Medicine in Football*, *2*(2), 108–114.

Foster, C. (1998). Monitoring training in athletes with reference to overtraining syndrome. *Medicine and Science in Sports and Exercise*, *30*(7), 1164–1168. https://doi.org/10.1097/00005768-199807000-00023

Foster, C., Florhaug, J. A., Franklin, J., Gottschall, L., Hrovatin, L. A., Parker, S., … Dodge, C. (2001). A new approach to monitoring exercise training. *The Journal of Strength & Conditioning Research*, *15*(1), 109–115.

Fuller, C. W., Ekstrand, J., Junge, A., Andersen, T. E., Bahr, R., Dvorak, J., … Meeuwisse, W. H. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Scandinavian journal of medicine & science in sports*, *16*(2), 83–92.

Gabbett, T. J. (2016). The training—Injury prevention paradox: Should athletes be training smarter and harder? *British journal of sports medicine*, *50*(5), 273–280.

Gjaka, M., Tschan, H., Francioni, F. M., Tishkuaj, F., & Tessitore, A. (2016). MONITORING OF LOADS AND RECOVERY PERCEIVED DURING WEEKS WITH DIFFERENT SCHEDULE IN YOUNG SOCCER PLAYERS. *Kinesiologia Slovenica*, *22*(1).

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hulin, B. T., Gabbett, T. J., Blanch, P., Chapman, P., Bailey, D., & Orchard, J. W. (2014). Spikes in acute workload are associated with increased injury risk in elite cricket fast bowlers. *British journal of sports medicine*, *48*(8), 708–712.

Impellizzeri, F. M., Rampinini, E., Coutts, A. J., Sassi, A., & Marcora, S. M. (2004). Use of RPE-based training load in soccer. *Medicine & Science in sports & exercise*, *36*(6), 1042–1047.

Impellizzeri, F. M., Woodcock, S., Coutts, A. J., Fanchini, M., McCall, A., & Vigotsky, A. D. (2021). What Role Do Chronic Workloads Play in the Acute to Chronic Workload Ratio? Time to Dismiss ACWR and Its Underlying Theory. *Sports Medicine*, *51*(3), 581–592. https://doi.org/10.1007/s40279-020-01378-6

Jaspers, A., Kuyvenhoven, J. P., Staes, F., Frencken, W. G., Helsen, W. F., & Brink, M. S. (2018). Examination of the external and internal load indicators' association with overuse injuries in professional soccer players. *Journal of science and medicine in sport*, *21*(6), 579–585.

Johnson, D. M., Williams, S., Bradley, B., Sayer, S., Murray Fisher, J., & Cumming, S. (2020). Growing pains: Maturity associated variation in injury risk in academy football. *European journal of sport science*, *20*(4), 544–552.

Johnson, L. L., Borkowf, C., & Albert, P. (2007). *An Introduction to Biostatistics: Randomization, Hypothesis Testing, and Sample Size Estimation*.

Kenttä, G., & Hassmén, P. (1998). Overtraining and recovery. *Sports medicine*, *26*(1), 1–16.

Ko, J., Rosen, A. B., & Brown, C. N. (2018). Functional performance tests identify lateral ankle sprain risk: A prospective pilot study in adolescent soccer players. *Scandinavian Journal of Medicine & Science in Sports*, *28*(12), 2611–2616.

Kofotolis, N. (2014). Ankle sprain injuries in soccer players aged 7-15 years during a one-year season. *Biology of exercise*, *10*(2).

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Malina, R. M., Bouchard, C., & Bar-Or, O. (2004). *Growth, maturation, and physical activity*. Human kinetics.

Malone, S., Owen, A., Newton, M., Mendes, B., Collins, K. D., & Gabbett, T. J. (2017). The acute: Chonic workload ratio in relation to injury risk in professional soccer. *Journal of science and medicine in sport*, *20*(6), 561–565.

Marshall, D. A., Lopatina, E., Lacny, S., & Emery, C. A. (2016). Economic impact study: Neuromuscular training reduces the burden of injuries and costs compared to standard warm-up in youth soccer. *British journal of sports medicine*, *50*(22), 1388–1393.

McCall, A., Dupont, G., & Ekstrand, J. (2016). Injury prevention strategies, coach compliance and player adherence of 33 of the UEFA Elite Club Injury Study teams: A survey of teams' head medical officers. *British journal of sports medicine*, *50*(12), 725–730.

McCall, A., Dupont, G., & Ekstrand, J. (2018). Internal workload and non-contact injury: A one-season study of five teams from the UEFA Elite Club Injury Study. *British journal of sports medicine*, *52*(23), 1517–1522.

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

Mirwald, R. L., Baxter-Jones, A. D., Bailey, D. A., & BEUNEN, G. P. (2002). An assessment of maturity from anthropometric measurements. *Medicine & science in sports & exercise*, *34*(4), 689–694.

Montella, A., de Oña, R., Mauriello, F., Riccardi, M. R., & Silvestro, G. (2020). A data mining approach to investigate patterns of powered two-wheeler crashes in Spain. *Accident Analysis & Prevention*, *134*, 105251.

Oliver, J. L., Ayala, F., Croix, M. B. D. S., Lloyd, R. S., Myer, G. D., & Read, P. J. (2020). Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *Journal of science and medicine in sport*, *23*(11), 1044–1048.

Petticrew, M. P., Sowden, A. J., Lister-Sharp, D., & Wright, K. (2000). False-negative results in screening programmes: Systematic review of impact and implications. *Health technology assessment (Winchester, England)*, *4*(5), 1–120.

Philippaerts, R. M., Vaeyens, R., Janssens, M., Van Renterghem, B., Matthys, D., Craen, R., … Malina, R. M. (2006). The relationship between peak height velocity and physical performance in youth soccer players. *Journal of sports sciences*, *24*(3), 221–230.

Polinder, S., Haagsma, J., Panneman, M., Scholten, A., Brugmans, M., & Van Beeck, E. (2016). The economic burden of injury: Health care and productivity costs of injuries in the Netherlands. *Accident Analysis & Prevention*, *93*, 92–100.

Read, P. J., Oliver, J. L., De Ste Croix, M. B. A., Myer, G. D., & 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.

Richardson, A., Clarsen, B., Verhagen, E., & Stubbe, J. H. (2017). High prevalence of self-reported injuries and illnesses in talented female athletes. *BMJ open sport & exercise medicine*, *3*(1), e000199.

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., … Witvrouw, E. (2020). A machine learning approach to assess injury risk in elite youth football players. *Medicine and science in sports and exercise*, *52*(8), 1745–1751.

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

Ruddy, J., Shield, A., Maniar, N., Williams, M., Duhig, S., Timmins, R., … Opar, D. (2018). Predictive modeling of hamstring strain injuries in elite Australian footballers. *Medicine and science in sports and exercise*, *50*(5), 906–914.

Sansone, P., Tschan, H., Foster, C., & Tessitore, A. (2020). Monitoring training load and perceived recovery in female basketball: Implications for training design. *The Journal of Strength & Conditioning Research*.

Seshadri, D. R., Thom, M. L., Harlow, E. R., Gabbett, T. J., Geletka, B. J., Hsu, J. J., … Voos, J. E. (2021). Wearable technology and analytics as a complementary toolkit to optimize workload and to reduce injury burden. *Frontiers in sports and active living*, *2*, 228.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

Singh, S., & Gupta, P. (2014). Comparative study ID3, cart and C4. 5 decision tree algorithm: A survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, *27*(27), 97–103.

Timpka, T., Risto, O., & Björmsjö, M. (2008). Boys soccer league injuries: A community-based study of time-loss from sports participation and long-term sequelae. *European journal of public health*, *18*(1), 19–24.

Towlson, C., Salter, J., Ade, J. D., Enright, K., Harper, L. D., Page, R. M., & Malone, J. J. (2020). Maturity-associated considerations for training load, injury risk, and physical performance within youth soccer: One size does not fit all. *Journal of Sport and Health Science*.

Vallance, E., Sutton-Charani, N., Imoussaten, A., Montmain, J., & Perrey, S. (2020). Combining Internal-and External-Training-Loads to Predict Non-Contact Injuries in Soccer. *Applied Sciences*, *10*(15), 5261.

van der Sluis, A., Elferink-Gemser, M. T., Coelho-e-Silva, M. J., Nijboer, J. A., Brink, M. S., & Visscher, C. (2014). Sport injuries aligned to peak height velocity in talented pubertal soccer players. *International journal of sports medicine*, *35*(04), 351–355.

Vänttinen, T., Blomqvist, M., Nyman, K., & Häkkinen, K. (2011). Changes in body composition, hormonal status, and physical fitness in 11-, 13-, and 15-year-old Finnish regional youth soccer players during a two-year follow-up. *The Journal of Strength & Conditioning Research*, *25*(12), 3342–3351.

Venturelli, M., Schena, F., Zanolla, L., & Bishop, D. (2011). Injury risk factors in young soccer players detected by a multivariate survival model. *Journal of science and medicine in sport*, *14*(4), 293–298.

Wang, C., Stokes, T., Steele, R., Wedderkopp, N., & Shrier, I. (2020). Injury risk increases minimally over a large range of the acute: Chronic workload ratio in children. *arXiv preprint arXiv:2010.02952*.

Watson, A., Brickson, S., Brooks, A., & Dunn, W. (2017). Subjective well-being and training load predict in-season injury and illness risk in female youth soccer players. *British journal of sports medicine*, *51*(3), 194–199.

Windt, J., & Gabbett, T. J. (2017). How do training and competition workloads relate to injury? The workload—Injury aetiology model. *British Journal of Sports Medicine*, *51*(5), 428–435.

Winter, E. M., & Maughan, R. J. (2009). Requirements for ethics approvals. *Journal of sports sciences*, *27*(10), 985.

Zouhal, H., Boullosa, D., Ramirez-Campillo, R., Ali, A., & Granacher, U. (2021). Acute: Chronic Workload Ratio: Is There Scientific Evidence? *Frontiers in Physiology*, *12*.