Data-Centric AI for predicting non-contact injuries in professional soccer players

Matheus Melo[1], Matheus Maia[1], Gabriel Padrão[1], Diego Brandão[1], Eduardo Bezerra[1], Juliano Spineti[2], Lucas Giusti[1], Jorge Soares[1]

[1]Computer Science Department – Federal Center of Technological Education Celso Suckow da Fonseca (Cefet/RJ) – Rio de Janeiro, RJ – Brazil
[2]Physiology Department – Fluminense Football Club – Rio de Janeiro, RJ – Brazil

{matheus.melo,matheus.vieira.2,gabriel.padrao}@aluno.cefet-rj.br,

{diego.brandao,ebezerra}@cefet-rj.br,juliano.spineti@fluminense.com.br,

lucas.giusti@aluno.cefet-rj.br, jorge.soares@cefet-rj.br

Abstract. *One big concern in soccer professional teams is to search for preventive measures to reduce the frequency of harmful episodes in their athletes since these episodes greatly impact the sports industry and affect both the team's performance and the association's economic situation. Thus, the present work proposes a methodology to predict non-contact injury episodes that may affect them in a microcycle through Data-centric AI concepts. The prediction model is trained using a dataset related to professional soccer athletes. The most interesting result were with AUC-ROC of 79,8%. About the performance improvement strategies applied, the best undersampling ratio was 70/30, PCA with one or two principal components did best, and the Decision Tree algorithm excelled.*

1. Introduction

The existence of injuries in sports scenarios and their corresponding negative consequences have attracted the growing interest of researchers, managers, and coaches in studies and technologies aimed at appropriate actions to prevent them [Rossi et al. 2018]. These incidents have a significant impact on the sports industry, affecting both team performance and the association's economic situation [Rossi et al. 2022].

In general, the incidence of an injury in the sports environment results in multiple repercussions. Hägglund et al. [2013] monitored the impact of injuries on the performance of UEFA Champions League teams for 11 years and pointed out that an athlete, being out of the team, can have a significant negative influence on the team's performance [Hägglund et al. 2013]. In terms of financial impact, Cuevas et al. [2021] demonstrated that injuries in Spain, for example, causes about 16% of absences in the season of professional soccer players, corresponding to a cost of about 188 million euros per season [Fernández Cuevas et al. 2010]. Furthermore, the injury frequency and recovery time are also relevant. Specifically, Pfirrmann et al. [2016] showed that professional soccer players suffer between 2.5 to 9.4 injuries per 1000 hours of effort, while Fiscutean [2021] reported that most of them last around a week, with the most recurrent ones (corresponding to 15% of the total) requiring a longer rest period [Fiscutean 2021, Pfirrmann et al. 2016].

On the other hand, preventive measures have been increasingly adopted in sports medicine to provide automated support for coaches and medical teams in decision-making to prevent untimely injuries [Kirkendall and Dvorak 2010]. In this context, a study spanning 18 years demonstrated a decline in injury incidence during matches and training, as well as lower recurrence rates, which appear to be related to the gradual and effective enhancement of injury prevention activities [Ekstrand et al. 2021].

In the current sports context, there are several researches (already completed or ongoing) that use tools as a basis for data collection and analysis, such as wearable outfits with GPS technologies monitored by software [Rossi et al. 2018, Vallance et al. 2020, Pilka et al. 2023, Rossi et al. 2022]. This type of technology has shown beneficial results to sports teams, such as Toronto Raptors, which implemented wearable devices and soft tissue monitoring to improve team performance [Studnicka 2020]. With the highest number of injuries in the 2012 NBA, the Raptors achieved, in conjunction with this technology and better management of the collected data, a record of one of the lowest injury rates among 2014 NBA teams [Studnicka 2020].

In the context of Artificial Intelligence, Data-Centric AI is an emerging concept that emphasizes the importance of handling and adding value to the data considered in models. This approach introduces a potential alternative to improving the performance of predictive models, complementing the application of algorithms and their instantiations [Jarrahi et al. 2023]. Thus, the main objective of this work is to evaluate different modeling alternatives to predict non-contact traumatic injuries within a microcycle of male professional soccer athletes from Fluminense Football Club. Our approach uses an association of Data-Centric AI concepts [Jarrahi et al. 2023] and machine learning algorithms. To enhance predictive performance, data management methods were applied to the model from different perspectives, such as class balancing with subsampling, Principal Component Analysis (PCA), and concepts related to multicollinearity, as well as the use of machine learning algorithms based on the systematic search conducted. The main contributions of this work consists of developing a robust pre-processing stage (aligned with the context of Data-Centric AI) and validating the best classification model alternatives, based on a Regressive Multi-dimensional Model Selection (RMMS) approach.

This study is organized into six more sections, beside this introduction. Section 2 presents related work obtained through a systematic literature search. Section 3 details the primary methodology, including the dataset used, feature engineering, feature selection, and the model selection approach employed. A description of the computational environment and how the experiments were conducted is presented in Section 4. Section 5 presents the results obtained by the applied techniques, followed by a discussion of the findings in Section 6. Finally, Section 7 points out the final considerations of this work and proposes directions for future research.

2. Related Work

To identify studies and research related to injury prediction in professional soccer, a systematic search was conducted following some guidelines from the PRISMA methodology (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)

[Page et al. 2021] using the PubMed[1] and Scopus Elsevier[2] databases. Initially, 121 articles were identified. After applying inclusion and exclusion criteria, as well as removing duplicates found in both databases, the number was reduced to 23 articles. These were then prioritized, with the highest priority given to studies focusing on the prediction of non-contact traumatic injuries using training and/or game data in conjunction with machine learning algorithms, resulting in a total of nine articles. A descriptive summary is provided in Table 1.

Table 1. Descriptive characteristics of the models in the base studies.

Studies	Attributes	Algorithms
[Rossi et al. 2018]	Body composition, GPS training load, playing time, injury history	DT, RF, LR
[Pilka et al. 2023]	Position, injury history, GPS training/game load	XGB
[Vallance et al. 2020]	Body composition, GPS training load, and intrinsic factors through questionnaires	KNN, LDA, LR, Ridge, GNB, DT, RF, SVM, MLP, XGB
[Eetvelde et al. 2021]	Injury history, training load, and body composition are the most repeated in the review articles	Trees (mainly DT), SVM, ANN
[Kolodziej et al. 2023]	Neuromuscular and biomechanical	LASSO
[Jauhiainen et al. 2022]	Demographic, neuromuscular, biomechanical, anatomical, and genetic	LR, RF, SVM
[Rossi et al. 2022]	GPS training/game load and blood samples	DT, XGB
[Martins et al. 2022]	Body composition and physical fitness tests	LASSO, SF, OLS, Ridge, ENET
[Dandrieux et al. 2023]	30-Meter Sprint and injury history	LR, RF, AdaBoost

Based on the search results and studies included in the systematic review by Eetvelde et al. [2021], the literature on injury prediction using machine learning algorithms appears to be expanding, with growing evidence supporting the accuracy of these methods in predicting injury episodes [Eetvelde et al. 2021]. Machine learning is particularly relevant in this context due to its ability to effectively and flexibly handle large datasets with numerous attributes [Majumdar et al. 2022]. However, predicting injuries remains challenging due to the diverse characteristics of players, including individual biological differences, physical predispositions, and psychophysical conditions [Pilka et al. 2023].

Among the studies listed in Table 1, Decision Tree machine learning algorithms were the most frequently used, appearing in the work of Rossi et al. [2018], Vallance et al. [2020], Eetvelde et al. [2021] and Rossi et al. [2022]. These studies primarily focused on common variables related to athlete training and/or game load, as well as some subjective anthropometric characteristics.

[1]`https://pubmed.ncbi.nlm.nih.gov`
[2]`https://www.scopus.com`

Rossi et al. [2018], Vallance et al. [2020], and Piłka et al. [2023] share some notable similarities that served as inspiration for the present research. All three studies combined GPS data with machine learning algorithms, primarily Decision Trees, to predict non-contact traumatic injuries among professional players. For model validation, they employed cross-validation techniques. They also addressed the issue of imbalanced target data using oversampling strategies such as SMOTE and ADASYN.

3. Methodology

The methodology of this work was described in four steps: 1) Data Collection and Cleaning, 2) Feature Engineering, 3) Potential Injury Risk Factors Selection and 4) Regressive Multi-dimensional Model Selection (RMMS). Step 1 (Section 3.1) comprises informations about the datasets provided and how it was cleaned. Step 2 (Section 3.2) explains the creation of the features to introduce within the prediction models. Step 3 (Section 3.3) consists of strategies adopted to look for features or combinations of them that can be potential injury risks. Finally, Step 4 (Section 3.4) shows how the models were developed, trained and evaluated.

3.1. Data Collection and Cleaning

For this research, data were collected from 182 professional players of Fluminense Football Club during the 2021 and 2022 seasons. The information was obtained from two sources: (1) Workload data, automatically collected through GPS-integrated wearable vests, and (2) Injury history provided by the club's medical staff. By combining these two sources, an initial dataset, *ATHLETES_DATA* was created, where each entry corresponds to a period, defined as a subdivision of the data collected on a training or match activity day for each athlete. For example, a match activity could be divided into three periods: (i) Warm-up, (ii) First half, and (iii) Second half. Consequently, *ATHLETES_DATA* consisted of 44,354 rows and a set of 1,715 features, along with a binary injury label indicating whether an injury had occurred. In total, 39 non-contact traumatic injuries were recorded among 22 players, with five players sustaining three injuries, seven sustaining two injuries, and ten sustaining one injury each.

To ensure the quality and relevance of the data, three cleaning steps were performed: (i) removal of goalkeeper data as they had specific training metrics and patterns, (ii) removal of data from athletes with less than 12 activities (two weeks and two rest days) as they had few information, (iii) completely null or zero columns were removed, while the columns with some nulls present were imputed by the average. With these steps, *ATHLETES_DATA* reduced to 41,109 rows and 801 features among 79 athletes.

3.2. Feature Engineering

The *MC_ATHLETES_DATA* dataset was derived from *ATHLETES_DATA*, focusing on specific features for model input. *MC_ATHLETES_DATA* was created by aggregating certain *ATHLETES_DATA* variables into microcycles, defined as all training activities and the subsequent match, resetting with the next training activity for each athlete. This approach was intended to reduce class imbalance and concentrate on state-of-the-art features, rather than utilizing all 1,715 columns from *ATHLETES_DATA*. The work of Vallance et al. [2020], Rossi et al. [2018] and Piłka et al. [2023], along with insights from

expert club physiologists, guided this process. In addition to the aggregated features derived from *ATHLETES_DATA* GPS variables, three new key features were created for the model: two to account for injury recurrences derived from the target label and one to track the duration of each microcycle in days, as suggested by the references mentioned. Ultimately, *MC_ATHLETES_DATA* comprised 147 microcycles, 4,326 rows, 26 independent variables, and one injury label with 39 injury cases, as shown in Table 2.

Table 2. *MC_ATHLETES_DATA* dataset variables to input the models.

Variables	Description
mc_field_time	Sum of field time
mc_tot_dist	Sum of distance in meters covered
mc_tot_dist_min	Sum of distance in meters covered divided by sum of field time
mc_vel1	Sum of distances covered between 0 and 1 km/h
mc_vel2	Sum of distances covered between 1.1 and 7 km/h
mc_vel3	Sum of distances covered between 7.2 and 14.4 km/h
mc_vel4	Sum of distances covered between 14.4 and 19.8 km/h
mc_vel5	Sum of distances covered between 19.8 and 25 km/h
mc_vel6	Sum of distances covered more than 19.8 km/h
mc_vel6_min	Sum of distances covered more than 19.8 km/h divided by sum of field time
mc_vel7	Sum of distances covered more than 25.2 km/h
mc_vel7_min	Sum of distances covered more than 25.2 km/h divided by sum of field time
mc_acel+desacel_high	Sum of high intensity inertial motion analysis for acceleration and deceleration
mc_acel+desacel_high_min	Sum of high intensity inertial motion analysis for acceleration and deceleration divided by sum of field time
mc_acel+desacel_>2ms	Sum of accelerations and decelerations above 2m/s²
mc_acel+desacel_>3ms	Sum of accelerations and decelerations above 3m/s²
mc_tot_load	Sum of total player load
mc_rhies	Sum of repeated high-intensity efforts
mc_rhies_min	Sum of repeated high-intensity efforts divided by sum of field time
mc_dir_changes	Sum of total changes of direction
mc_jumps	Sum of total number of jumps
mc_max_vel	Maximum speed reached
mc_max_acel	Maximum acceleration achieved
mc_duration	Count of number of days present in microcycle
injury_target	1— Yes, 0— No, if the injury occurred within a microcycle
binary_reincidence	1— Yes, 0— No, for injury reincidences
accumulated_reincidence	Cumulative sum of injury reincidences

3.3. Potential Injury Risk Factors Selection

Before creating multi-dimensional models, the Mann-Whitney U test was performed to select potential injury risk factors through a bivariate analytical comparison. With the obtained statistical calculations, the relevance of each of the 26 variables was seen by the measured p-value according to the designated significance level (alpha).

Afterward, with the 26 features created, different combinations of strategies were considered for removing multicollinear variables (Table 3) in an attempt to improve the performance of the created models from different perspectives. Thus, 30 different feature combinations were filled, divided by three different strategies. All combinations consisted of using only one of the two reincidence variables (*accumulated_reincidence* or *binary_reincidence*) at a time because using them together is redundant due to strong correlation and using them together is also ambiguous.

Table 3. Multicollinearity between features with correlation above 95%.

Feature 1	Feature 2	Correlation
mc_vel6	mc_vel5	99.1%
mc_tot_load	mc_tot_dist	98.9%
mc_tot_dist	mc_vel2	96.8%
mc_vel2	mc_field_time	96.6%
mc_tot_load	mc_vel2	95.7%
mc_tot_dist	mc_field_time	95.7%
mc_tot_load	mc_field_time	95.7%

- **Strategy 1 (2 combinations)**: Keep all features from *MC_ATHLETES_DATA*, without removal of multicollinear variables;
- **Strategy 2 (12 combinations)**: Keep all features from *MC_ATHLETES_DATA* except one of the six multicollinear variables;
- **Strategy 3 (16 combinations)**: Keep all features from *MC_ATHLETES_DATA* with only one of the four repeating multicollinear variables (mc_carga_tot, mc_tot_dist, mc_field_time, and mc_vel2) along with one of the two non-repeating ones (mc_vel6 or mc_vel5).

Table 4. Feature combinations divided for *MC_ATHLETES_DATA* by three different multicollinearity removal strategy. AR = With accumulated_reincidence feature; BR = With binary_reincidence feature.

Combination Names	Multicollinearity removal strategy
All_BR and All_AR	Keep all features
mcr1_BR and mcr1_AR	Keep all except mc_vel6
mcr2_BR and mcr2_AR	Keep all except mc_carga_tot
mcr3_BR and mcr3_AR	Keep all except mc_tot_dist
mcr4_BR and mcr4_AR	Keep all except mc_vel2
mcr5_BR and mcr5_AR	Keep all except mc_vel5
mcr6_BR and mcr6_AR	Keep all except mc_field_time
mcr7_BR and mcr7_AR	Keep all except mc_vel6, mc_carga_tot, mc_tot_dist, mc_vel2
mcr8_BR and mcr8_AR	Keep all except mc_vel5, mc_carga_tot, mc_tot_dist, mc_vel2
mcr9_BR and mcr9_AR	Keep all except mc_vel6, mc_field_time, mc_tot_dist, mc_vel2
mcr10_BR and mcr10_AR	Keep all except mc_vel5, mc_field_time, mc_tot_dist, mc_vel2
mcr11_BR and mcr11_AR	Keep all except mc_vel6, mc_field_time, mc_carga_tot, mc_vel2
mcr12_BR and mcr12_AR	Keep all except mc_vel5, mc_field_time, mc_carga_tot, mc_vel2
mcr13_BR and mcr13_AR	Keep all except mc_vel6, mc_field_time, mc_carga_tot, mc_tot_dist
mcr14_BR and mcr14_AR	Keep all except mc_vel5, mc_field_time, mc_carga_tot, mc_tot_dist

3.4. Regressive Multi-dimensional Model Selection (RMMS)

To objectively understand the impact of different modeling alternatives on predictive performance, we implemented the RMMS technique [Giusti et al. 2022]. Accordingly, the first step was to develop a function that creates and validates multi-dimensional predictive models based on the values inserted as parameters (seven in total, as seen in Table 5). Algorithm 1 shows its methodology.

Table 5. Description of the parameters used in Algorithm 1.

df	Dataset to be used in the models
$features$	df specific combinations of features selected
$target$	df target label
ml	Machine Learning algorithms chosen
n	Proportion of negative and positive case samples for class balancing with undersampling
$test\_size$	Proportion of test data for stratified hold-out validation
pca	Number of principal components with PCA

Algorithm 1: Methodology to Create and Validate the Predictive Model

```
1  function classification_predictions(df, features, target, ml, n, test_size, pca):
2      df_processed ← copy(df, features, target)
3      train_X, test_X, train_y, test_y ← Hold-out(df_processed, test_size)
4      if pca > 0 then
5          train_X ← StandardScaler.fit_transform(train_X)
6          test_X ← StandardScaler.transform(test_X)
7          train_X ← PCA.fit_transform(train_X, pca)
8          test_X ← PCA.transform(test_X, pca)
9      end
10
11     train_X, train_y ← Undersampling(train_X, train_y, n)
12     ml.fit(train_X, train_y)
13     pred_y ← ml.predict(test_X)
14     methods ← [accuracy, precision, f1, recall, AUC-ROC]
15     foreach method ∈ methods do
16         results ← results ∪ method(test_y, pred_y)
17     end
18 return results
```

The function begins with the processing of the df using specific features combinations and the target variable, defined respectively by $features$ and $target$ parameters. Next, validation was performed using the stratified hold-out method with the division of test proportion setted by the parameter $test\_size$. After the hold-out division, if there is an intention to apply PCA to the $features$ used ($features$ embedding dimensionality is considered if pca equals to zero), data normalization is performed for both training and test data using $StandardScaler$, followed by the application of PCA into principal components. Then, undersampling is applied only to the training data, reducing the amount of data to the proportion defined in n. Finally, the model training and prediction stages occur, which are done with the machine learning algorithms specified on ml. To evaluate

the classifications, an iteration is performed through the $methods$ list to evaluate the result with each performance metric and return them to the $results$ variable. In the study, five metrics were considered [Majumdar et al. 2022]. The formula for each metric was showed on Table 6.

1. **Accuracy**: Proportion of correctly classified injuries and non-injuries to the total number of observed injuries and non-injuries.
2. **Precision**: Proportion of correctly classified injuries to the total number of injuries classified.
3. **Recall**: Proportion of correctly classified injuries to the total number of injuries.
4. **F1-score**: Harmonic mean between precision and recall.
5. **AUC-ROC**: Area under the ROC curve that evaluates the relationship between true positive rates and false positive rates.

Table 6. Formula for each evaluation metric used on the classification models.

(1)	(2)	(3)	(4)	(5)
$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$	$ROC = rate(\frac{TP}{FP})$

Due to the prevailing class imbalance in the test set, the main performance measure for the work was AUC-ROC, generating a more consistent analysis of the positive results obtained in the classification. Initially, the baseline results of the project consisted of two raw models to be taken as a comparison after the application of multicollinearity features removal strategies, PCA, undersampling and different machine learning algorithms. Thus, the classification models were created with the intention of improving performance through an iteration of the cross-product of the parameters of the classification function. The cross-product occurred for each possibility of $features$, ml, n, and pca, referring to F, M, N, P respectively.

After creating the classification models, a validation of all obtained results was conducted simultaneously. With this, it is possible to observe the parameters that had the greatest positive or negative impact on performance, which is the main goal of RMMS approach. The strategy consists of a Random Forest regression model, where the features are all possibilities of the iterated parameters in classification (F, M, N, P), and the target variable is the obtained AUC-ROC. To enable regression, one-hot encoding was applied due to the possibility of categorical values. The evaluation was done with the Mean Absolute Percentage Error metric. With the regression model created, the analysis of the classification parameters was interpreted in a graphical illustration with SHAP (SHapley Additive exPlanations[3]) values. This allowed for a clear explanation of the positive or negative impact of these parameters on the predictive outcome.

4. Experimental Setup

The computational environment used for the experiments consisted of a computer running Windows 10 Pro 64-bit version 22H2 with an Intel(R) Core(TM) i3-10100 processor clocked at 3.60GHz and 12GB of installed RAM. The project was entirely implemented in Python version 3.9.12.

[3]`https://shap.readthedocs.io/en/latest/`

Firstly, for the configuration of baseline models *B1* and *B2*, pca and n parameters were not considered, and $features$ consisted of all features from *MC_ATHLETES_DATA* with only one reincidence variable for each (first two combinations from Table 4). Additionally, the parameter ml was filled only by the Decision Tree, and $test\_size$ was fixed at 20%. For the main classification modeling, Algorithm 2 shows the instantiation of the parameters. Initially, among the seven parameters present, df and $target$ were fixed values, defined as *MC_ATHLETES_DATA* to be filtered by $features$ parameter and *MC_ATHLETES_DATA* target variable, respectively. Along with this, $test\_size$ also had a fixed value of 20%, as in the baseline. The parameter ml consisted of Decision Tree *DT*, Random Forest *RF*, and Logistic Regression *LR*, inspired from Rossi et al. [2018] algorithms used for comparison. The pca parameter had 21 different values, varying the possibility of the model being created with $features$ embedding dimensionality or dimensionality reduction with 1 to 20 principal components. To mitigate class imbalance, n consisted of three possibilities of proportions between data negative and positive case samples: 70%/30%, 60%/40%, and 50%/50%. Finally, $features$ was filled with 30 different combinations of features divided by three strategies of multicollinearity removal (explained in Section 3.3).

Algorithm 2: Experimental Setup and Parâmeters Cross Product

$df \leftarrow$ *MC_ATHLETES_DATA*
$target \leftarrow$ *MC_ATHLETES_DATA*$[injury\_target]$
$test\_size \leftarrow 0.2$
$F \leftarrow \{All\_BR, All\_AR, mcr1\_BR, mcr1\_AR, ..., mcr14\_BR, mcr14\_AR\}$
$M \leftarrow \{DT, RF, LR\}$
$N \leftarrow \{70/30, 60/40, 50/50\}$
$P \leftarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$
$results \leftarrow \emptyset$
foreach $features \in F, ml \in M, n \in N, pca \in P$ **do**
    $results \leftarrow results \cup$
    $classification\_predictions(df, features, target, ml, n, test\_size, pca)$
end

5. Results

The Mann-Whitney U test was calculated for each relationship between feature and target variables. The value defined for α was 0.05. Thus, among the 26 variables, only mc_rhies_min proved to be relevant with the application of the test ($p\text{-}value < \alpha$). This suggests that because most of the variables do not have a statistically significant association with the target label, the predictions of the models developed showed more difficulties to perform effectively.

With the application of the methodology of Algorithms 1 and 2, 5670 different models were obtained by the parameters combinations. Table 7 compares the test performance of the baseline models and the best results for each of the three algorithms used, based on the literature. Prioritizing the AUC-ROC metric, the best result was achieved with the *DT*, showing a decent recall and AUC-ROC. This is a significant impact with respect to the baselines *B1* and *B2*, demonstrating the application of the techniques dis-

Table 7. Classification results on test set for the baselines *B1* and *B2* and best *DT*, *RF* and *LR*, sorted by AUC-ROC.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
DT	72,2%	2,8%	87,5%	5,5%	79,8%
RF	52,3%	1.9%	100%	3.7%	75,9%
LR	84,3%	2,2%	37.5%	4,2%	61,1%
B1	98,3%	0%	0%	0%	49,6%
B2	98,3%	0%	0%	0%	49,6%

cussed in the methodology through the parameters. *RF* showed a recall of 100% compared to the DT and a similar AUC-ROC. However, both models exhibited much lower precision compared to recall. This explains the trade-off reflected in the F1-score, which resulted in low values. Finally, *LR* had a much inferior performance in terms of precision, recall, and AUC-ROC, compared to *DT* and *RF*.

For the simultaneous analysis of the parameters used to develop the 5670 models, SHAP values approach provided a clear explanation of the positive or negative impact on the output, according to the indicated magnitude. The predictive regression modeling with Random Forest Regressor had a 2.8% Mean Absolute Percentage Error, indicating a low percentage of error in the predicted value. Due to the fact that the attribute values are binary, the graph in Figure 1 shows only two colors, red indicating the use of the variable in the model and blue indicating non-usage. Thus, red to the right indicates a positive impact of its use, and to the left, a negative impact. Consequently, the blue color to the right signifies a positive impact of not using the variable, while to the left, it is the opposite.

6. Discussion

Firstly, considering the results in Table 7, accuracy is not a priority in this study due to its deceptively high scores in cases of class imbalance. This is because accuracy accounts for both correct classifications of the positive and negative classes among all predictions. Since injury cases are extremely rare in the test set after stratified hold-out (858/8), the accuracy remains high by correctly classifying most non-injury cases, while the primary interest lies in the correct classification of the positive class.

In a real-world scenario of predicting non-contact traumatic injuries among professional soccer players, it is crucial to avoid false alarms in both the positive and negative classes, which is reflected in high performance in both precision and recall. High recall is essential to prevent players at true risk of injury from entering the field due to a false-negative alarm, while high precision is crucial to avoid situations where athletes are unnecessarily spared due to a false-positive alarm. Both scenarios are detrimental to the team and/or the athletes involved. Despite this, the primary objective of the current study is to establish a methodology that objectively assesses the impact of different modeling parameters on predictive performance. Therefore, the AUC-ROC metric highlights the model's ability to robustly differentiate between positive and negative classes, even in the presence of imbalance, and primarily serves the study's purpose within the RMMS methodology.

Results in Table 7 mainly serve to indicate the best performances achieved through

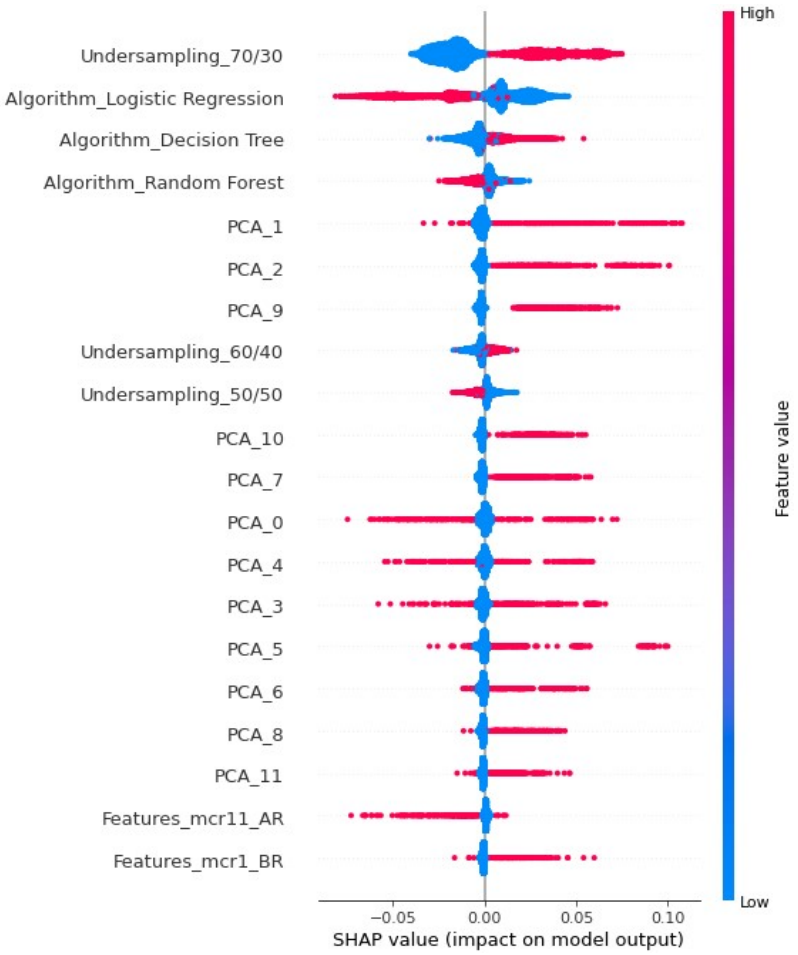

Figure 1. Graph of the positive or negative impact of the 20 most relevant SHAP values in relation to the parameters present in the regression model.

strategies that allowed potential performance improvements, highlighted for each algorithm and baseline. However, since the models were developed based on a combination of parameters, the SHAP values provided a comprehensive view of the parameters that had the greatest impact on model performance. Figure 1 illustrates the 20 most relevant parameters for the results, both positively and negatively, allowing for some observations. Among the techniques applied, the most prominent factors influencing the models were the different machine learning algorithms, undersampling proportions, and varying numbers of principal components with PCA. In contrast, strategies related to feature selection had the least impact on the models.

According to the SHAP values, *DT* demonstrated a positive impact when applied. In Rossi et al. [2018], it was possible to detect 80% of non-contact injuries with 50% precision using this algorithm, significantly reducing false alarms. The study demonstrated the use of both nonlinear and linear classifiers in a multidimensional context, which also inspired the present work to attempt the same. Despite the prominence of *DT*, using *LR* showed an opposite effect on performance, suggesting a nonlinear relationship between features and the target. Other studies that used GPS and machine learning, such as Vallance et al. [2020], also demonstrated the predictive power of tree-based algorithms,

achieving good results with tree models, random forests, and especially XGBoost. XGBoost is particularly interesting in the context of football injuries, as it aims to improve several small classification trees (also known as "weak learners") based on their errors. This technique has also been used with good results in Piłka et al. [2023]. XGBoost effectiveness on datasets with high class imbalance is noteworthy, as this is a common issue in all these studies due to the rarity of injuries in professional soccer. Considering class imbalance, this study employed undersampling to balance the classes in the training set, unlike Rossi et al. [2018] and Piłka et al. [2023], who used oversampling to add positive cases to the training set instead of removing negative cases. The effectiveness of undersampling was most evident with a sample proportion of 70% non-injury cases and 30% injury cases.

One important application was PCA, which had a positive impact by reducing the features into different principal components (mainly one or two), compared to using the features in their embedded dimensionality. The challenge with using PCA in this scenario lies in the inconclusiveness regarding the relevance of the features transformed and reduced to principal components. However, focusing on predictive performance rather than feature interpretability, PCA remains an interesting tool. Despite this, the advantage of using RMMS in conjunction with SHAP is the ability to simultaneously observe the set of features selected within the applied PCA. Thus, some combinations of features for multicollinearity removal, such as mcr11_AR and mcr1_BR, proved to be relevant.

7. Final Thoughts

The purpose of the current study was to demonstrate the functionality of the Regressive Multi-dimensional Model Selection (RMMS) methodology for predicting non-contact injuries in professional soccer players, by evaluating different modeling alternatives on predictive performance. In this context, considering Data-Centric AI principles to modify the data through various strategies, 30 different feature combinations were explored, accounting for multicollinearity removal, three different undersampling proportions, and 1 to 20 principal components, in addition to embedding dimensionality. Beyond prediction, an important aspect of the methodology is the understanding and interpretation of the parameters used for model construction, facilitated by SHAP values, which revealed the positive or negative relevance of the top 20 parameters. Among all the models developed, the best AUC-ROC metric achieved was 79.8% using *DT*, which also showed a major positive impact in the SHAP analysis, as well as in other studies.

Regarding future work, several ideas could enhance the project. The RMMS approach employs various parameters to demonstrate the potential for building models from different perspectives simultaneously. However, it would be beneficial to carefully select these parameters to avoid complicating processing time for larger datasets. For instance, the number of principal components in PCA could be estimated using strategies like Scree-Plot or Broken-Stick. Additionally, implementing a more robust grid search strategy could help identify the best features for each model. To address class imbalance, other techniques, such as oversampling or adjusting class weights, could be explored. Finally, incorporating new athlete features beyond GPS data, such as biochemical tests and perceived exertion, would further enrich the models.

References

Dandrieux, P.-E., Tondut, J., Nagahara, R., Mendiguchia, J., Morin, J.-B., Lahti, J., Ley, C., Edouard, P., and Navarro, L. (2023). Prédiction des blessures des ischiojambiers en football à l'aide d'apprentissage automatique: étude préliminaire sur 284footballeurs. *Journal de Traumatologie du Sport*, 40(2):69–73.

Eetvelde, H., De Michelis Mendonça, L., Ley, C., Seil, R., and Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics*, 8(1).

Ekstrand, J., Spreco, A., Bengtsson, H., and Bahr, R. (2021). Injury rates decreased in men's professional football: An 18-year prospective cohort study of almost 12 000 injuries sustained during 1.8 million hours of play. *British Journal of Sports Medicine*, 55(19):1084–1091.

Fernández Cuevas, I., Carmona, P., Quintana, M., Salces, J., Arnaiz-Lastras, J., and Barrón, A. (2010). Economic costs estimation of soccer injuries in first and second spanish division professional teams. In *Proceedings of the 15th Annual Congress of the European College of Sport Sciences (ECSS)*.

Fiscutean, A. (2021). Data scientists are predicting sports injuries with an algorithm. *Nature*, 592(7852):S10–S11.

Giusti, L., Carvalho, L., Gomes, A. T. A., Coutinho, R., de Abreu Soares, J., and Ogasawara, E. S. (2022). Analyzing flight delay prediction under concept drift. *Evolving Systems*, (0123456789).

Hägglund, M., Waldén, M., Hedevik, 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.

Jarrahi, M. H., Memariani, A., and Guha, S. (2023). The Principles of Data-Centric AI. *Communications of the ACM*, 66(8):84–92.

Jauhiainen, S., Kauppi, J.-P., Krosshaug, T., Bahr, R., Bartsch, J., and Äyrämö, S. (2022). Predicting ACL Injury Using Machine Learning on Data From an Extensive Screening Test Battery of 880 Female Elite Athletes. *American Journal of Sports Medicine*, 50(11):2917–2924.

Kirkendall, D. T. and Dvorak, J. (2010). Effective injury prevention in soccer. *Physician and Sportsmedicine*, 38(1):147–157.

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 and Science in Sports*, (February 2022):1–13.

Majumdar, A., Bakirov, R., Hodges, D., Scott, S., and Rees, T. (2022). Machine Learning for Understanding and Predicting Injuries in Football. *Sports Medicine - Open*, 8(1).

Martins, F., Przednowek, K., França, C., Lopes, H., Nascimento, M., Sarmento, H., Marques, A., Ihle, A., Henriques, J., and Gouveia, E. (2022). Predictive Modeling of

Injury Risk Based on Body Composition and Selected Physical Fitness Tests for Elite Football Players. *Journal of Clinical Medicine*, 11(16).

Page, M., Mckenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., Shamseer, L., Tetzlaff, J., Akl, E., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M., Li, T., Loder, E., Mayo-Wilson, E., Mcdonald, S., and Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, 372.

Pfirrmann, D., Herbst, M., Ingelfinger, P., Simon, P., and Botzenhardt, 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.

Pilka, T., Grzelak, B., Aleksandra, S., Górecki, T., and Dyczkowski, K. (2023). Predicting injuries in football based on data collected from gps-based wearable sensors. *Sensors*, 23(3).

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

Rossi, A., Pappalardo, L., Filetti, C., and Cintia, P. (2022). Blood sample profile helps to injury forecasting in elite soccer players. *Sport Sciences for Health*, 19(1):285–296.

Studnicka, A. (2020). The emergence of wearable technology and the legal implications for athletes, teams, leagues and other sports organizations across amateur and professional athletics. *DePaul J. Sports L.*, 16:i.

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