Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.5855

A heuristic Data-Centric AI approach to predict non-contact injuries in elite football players

Matheus Santos Melo [**Federal Center of Technological Education Celso Suckow da Fonseca** | ***matheus.melo@aluno.cefet-rj.br***]
Juliano Spineti [**Fluminense Football Club** | ***juliano.spineti@fluminense.com.br***]
Diego Nunes Brandão [**Federal Center of Technological Education Celso Suckow da Fonseca** | ***diego.brandao@cefet-rj.br***]
Lucas Giusti Tavares [**Federal Center of Technological Education Celso Suckow da Fonseca** | ***lucas.giusti@aluno.cefet-rj.br***]
Jorge de Abreu Soares [**Federal Center of Technological Education Celso Suckow da Fonseca** | ***jorge.soares@cefet-rj.br***]

Computer Science Department, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (Cefet/RJ), Av. Maracanã, 229, Maracanã, Rio de Janeiro, RJ, 20271-110, Brazil

Physiology Department, Fluminense Football Club, Rua Álvaro Chaves, 41, Laranjeiras, Rio de Janeiro, RJ, 22231-220, Brazil

Received: 07 April 2025 • **Published:** 13 March 2026

Abstract

Preventing non-contact injuries in professional soccer is critical for safeguarding athlete health and minimizing disruptions to team performance and financial stability. This study investigates predictive modeling strategies for forecasting non-contact traumatic injuries during a microcycle of male professional players from Fluminense Football Club, integrating Data-Centric AI (DCAI) principles with machine learning algorithms. Building upon previous work, we extend the Regressive Multi-dimensional Model Selection (RMMS) methodology through new experiments that incorporate alternative class balancing strategies, hyperparameter tuning, and feature selection methods in place of Principal Component Analysis (PCA). Among the tested models, tree-based algorithms—particularly XGBoost—achieved the highest AUC-ROC (74.1%), though this result remained below the 79.8% baseline obtained with a Decision Tree in earlier research. Undersampling with a 70/30 ratio of non-injury to injury cases emerged as the most effective balancing approach, reinforcing prior findings. SHAP (SHapley Additive exPlanations) analysis identified Adaboost as the most positively impactful model, while feature selection and hyperparameter optimization yielded adverse effects on performance. These results suggest that PCA continues to be a more effective dimensionality reduction technique for this dataset. Future research should incorporate additional training seasons, match-related data, and broader athlete characteristics beyond GPS metrics—such as biochemical markers and perceived exertion—to improve model robustness and predictive accuracy.

Keywords: Professional soccer, Injury prediction, Machine learning, Sports injuries, Data Science

1 Introduction

According to the UEFA model, injuries—whether contact-related or not—are defined as any tissue damage sustained by a player that results in at least one day of absence from training or competition [Hägglund *et al.*, 2005; Rossi *et al.*, 2018]. The high prevalence of injuries in professional sports and their detrimental effects have drawn growing attention from researchers, coaching staff, and medical teams, prompting the development of technologies and methodologies to improve injury prevention strategies [Rossi *et al.*, 2018]. Beyond their physical toll on athletes, injuries can significantly impact team performance and compromise the financial stability of sports organizations [Rossi *et al.*, 2022].

The prevalence of injuries in sports and their associated negative effects have drawn increasing attention from researchers, team staff, and medical professionals, driving the development of studies and technologies aimed at effective injury prevention [Rossi *et al.*, 2018]. These injuries considerably impact the sports industry, affecting both team performance and the financial stability of sports organizations [Rossi *et al.*, 2022].

Injuries in professional sports lead to several consequences. Hägglund *et al.* [2013] conducted an 11-year study on UEFA Champions League teams and demonstrated that player absences due to injuries can negatively influence team performance. From a financial perspective, Fernández Cuevas *et al.* [2010] reported that, in Spain, injuries account for approximately 16% of player absences per season, resulting in an estimated financial loss of around 188 million euros annually. Additionally, the frequency and recovery duration of injuries are critical aspects. Specifically, Pfirrmann *et al.* [2016] found that professional soccer players sustain between 2.5 and 9.4 injuries per 1,000 hours of activity, while Fiscutean [2021] noted that most injuries require about a week of recovery, with the most frequent cases (accounting for 15% of the total) demanding a longer rehabilitation period.

In response to these challenges, injury prevention has become a key focus in sports medicine. Preventive strategies, supported by automated decision-making systems, are increasingly used to assist coaches and medical staff in monitoring athlete health and reducing injury risk [Kirkendall and Dvorak, 2010]. A longitudinal study by Ekstrand *et al.* [2021], spanning 18 years, highlighted the effectiveness of such strategies by showing reduced injury rates during matches and training, as well as lower recurrence rates—results attributed to the continuous refinement of preventive protocols.

Modern sports environments have further benefited from technological advancements in data collection and analysis. Wearable devices with integrated GPS tracking, combined with analytical software platforms, enable real-time monitoring of training loads and player conditions [Rossi *et al.*, 2018; Vallance *et al.*, 2020; Rossi *et al.*, 2022; Pilka *et al.*, 2023]. These technologies have led to demonstrable improvements; for instance, after leading the NBA in injury incidence in 2012, the Toronto Raptors implemented wearable monitoring and soft tissue management systems. By 2014, they had reduced injury rates to one of the lowest in the league, largely due to these innovations and enhanced data management practices [Studnicka, 2020].

Within this context, the present study aims to evaluate different modeling approaches for predicting non-contact traumatic injuries in male professional soccer players from Fluminense Football Club, focusing on a single microcycle. We incorporate the principles of Data-Centric AI (DCAI), an emerging paradigm that emphasizes the quality and management of training data as a critical factor for improving model performance [Jarrahi *et al.*, 2023].

This study presents an extension of the methodology introduced in Melo *et al.* [2024], incorporating alternative modeling components to assess their impact on injury prediction performance. The previous approach employed undersampling and Principal Component Analysis (PCA) for handling class imbalance and dimensionality reduction, respectively. In contrast, the present work evaluates the effectiveness of ADASYN for oversampling, Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection, and GridSearchCV for hyperparameter optimization. The set of classifiers was also expanded to include Adaboost, XGBoost, and LinearSVC, complementing the previously utilized Decision Tree, Random Forest, and Logistic Regression models. Additionally, the Regressive Multi-dimensional Model Selection (RMMS) pipeline was refined through a full factorial experimental design, enabling a systematic evaluation of the interaction effects among six key factors: algorithm, feature subset (with multicollinearity control), class balancing method, injury case proportion, feature selection strategy, and hyperparameter tuning. A total of 4,320 models were trained and evaluated, providing a robust basis for empirical comparison.

However, the best predictive performance obtained in the current study (AUC-ROC = 74.1%) was inferior to the prior study (AUC-ROC = 79.8%). This result indicates that, under the specific conditions evaluated, conventional techniques such as PCA and undersampling may offer superior performance, particularly in highly imbalanced classification tasks. SHAP-based model interpretability further revealed that Adaboost contributed more significantly to predictive outcomes in the current setting, whereas Decision Trees had greater influence in the original study. Furthermore, the application of RFECV and hyperparameter optimization via GridSearchCV were found to negatively affect performance in this context. Also, 70/30 undersampling ratio consistently outperforms ADASYN, PCA remains a robust dimensionality reduction technique, and default hyperparameter configurations can in some cases yield better results than extensively tuned models. Thus, the results contribute to a more nuanced understanding of modeling choices in sports injury prediction and offer practical guidance for the development of interpretable and generalizable models within a data-centric framework.

This article is organized as follows: Section 2 presents related work derived from an updated systematic literature search, while Section 3 outlines the primary methodology developed, and Section 4 describes the new experiments conducted using the proposed methodology. The results are presented in Section 5, followed by a discussion of the findings in Section 6 and a comparative analysis with our previous work. Finally, Section 7 provides this work's final considerations and suggests future research directions.

2 Related Work

To explore existing studies on injury prediction in professional soccer, a systematic literature review was carried out following key principles of the PRISMA methodology (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [Page *et al.*, 2021]. The search was conducted using the PubMed[1] and Scopus Elsevier[2] databases. Initially, 150 articles were identified. After applying inclusion and exclusion criteria and removing duplicates from both databases, the selection was narrowed down to 33 articles. Among these, priority was given to studies focusing on the prediction of non-contact traumatic injuries using data from professional training sessions and official matches combined with machine learning techniques, resulting in a final selection of 12 articles. In comparison to our previous work, four new high-priority studies were added to provide insights aligned with the objectives of the present study.

Based on the studies listed in Table 1, the literature on injury prediction using machine learning algorithms appears to be expanding, with increasing evidence supporting the accuracy of these methods in predicting injury occurrences. However, it remains a challenging task due to the multifactorial nature of the problem, which includes the diversity of player characteristics, individual biological differences, physical predispositions, psychophysical conditions, metabolic metrics, and previous injuries affecting different body parts [Pilka *et al.*, 2023]. Another critical issue is the low frequency of injury events due to their rarity during training sessions and matches, leading to an imbalance between non-contact traumatic injury events and noninjury events. This imbalance causes the model to learn predominantly from the negative class, resulting in a high rate of false negatives and reduced sensitivity to rare injury occurrences. Oversampling strategies such as ADASYN and SMOTE were primarily employed in the studies to address this issue.

[1] https://pubmed.ncbi.nlm.nih.gov

[2] https://www.scopus.com

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

Studies	Attributes	Machine Learning Algorithms
Rossi *et al*. [2018]	Body composition, GPS training load, playing time, injury history	DT, RF, LR
Vallance *et al*. [2020]	Body composition, GPS training load, and intrinsic factors through questionnaires	KNN, LDA, LR, Ridge, GNB, DT, RF, SVM, MLP, XGBoost
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
Jauhiainen *et al*. [2022]	Demographic, neuromuscular, biomechanical, anatomical, and genetic	LR, RF, SVM
Pilka *et al*. [2023]	Position, injury history, GPS training/game load	XGBoost
Haller *et al*. [2023]	Physical fitness tests, intrinsic factors through questionnaires, blood samples, neuromuscular	LinearSVM
Kolodziej *et al*. [2023]	Neuromuscular and biomechanical	LASSO
Dandrieux *et al*. [2023]	30-Meter Sprint and injury history	LR, RF, AdaBoost
Fang and Xiang [2024]	GPS, metabolic and physical characteristics, injury history	DT, RF
Freitas *et al*. [2025]	GPS training/match	SVN, FNN, AdaBoost
Saberisani *et al*. [2025]	GPS training/match	DT

Machine learning is particularly relevant because it can efficiently and flexibly process large datasets with numerous attributes [Majumdar *et al*., 2022]. For the prediction task, both linear and non-linear algorithms were employed. Among the nonlinear approaches, tree-based models were the most commonly used across studies, including Decision Tree (DT), Random Forest (RF), XGBoost, and AdaBoost. Stratified cross-validation was the most frequently applied technique for model validation due to the low number of injury events in the validation sets.

Regarding data collection, wearable sensors, particularly GPS-equipped devices, have become the primary method for gathering data because they provide reliable and credible measurements during both open-air training sessions and matches [Pilka *et al*., 2023]. This is crucial since it is important in managerial and coaching decisions. Consequently, the success of soccer clubs depends on what is measured and the accuracy of those measurements in generating high-quality predictions [Rossi *et al*., 2018]. Rossi *et al*. [2018] pioneered using GPS data for injury prediction, proposing a multidimensional approach based on external load data collected through GPS tracking. He trained decision trees to predict whether a player would likely sustain an injury in the next match or training session to achieve this. These non-linear models applied solely to external training load, demonstrated superior performance compared to traditional statistical methods in assessing injury risk. However, their predictive performance remains suboptimal, achieving 50% precision and 80% recall. This limitation has inspired subsequent research efforts to explore new approaches that integrate additional data sources and techniques alongside GPS data to enhance predictive accuracy [Vallance *et al*., 2020; Rossi *et al*., 2022; Pilka *et al*., 2023; Fang and Xiang, 2024; Freitas *et al*., 2025; Saberisani *et al*., 2025].

Based on these studies, Pilka *et al*. [2023] achieved the best results with a multidimensional model using XGBoost and training and match GPS workload, along with additional variables such as player position and injury history within a microcycle to predict injuries in the upcoming microcycle, reaching a performance of 92.4% precision, 96.5% recall and an f1-score of 94.4%.

3 Methodology

The methodology of this study was described in four steps: 1) Data Collection and Cleaning, 2) Feature Engineering, 3) Multicollinearity Removal Strategy, and 4) Regressive Multi-dimensional Model Selection (RMMS). Step 1 (Section 3.1) details the datasets used and the cleaning procedures applied. Step 2 (Section 3.2) outlines the development of features incorporated into the predictive models. Step 3 (Section 3.3) discusses the approaches taken to create groups of features based on a multicollinearity removal strategy. Finally, Step 4 (Section 3.4) describes how the models were constructed, trained, and evaluated.

3.1 Data Collection and Cleaning

For this study, data were collected from 182 professional players of Fluminense Football Club during the 2021 and 2022 seasons. The dataset was built from two primary sources: (1) workload metrics automatically captured by GPS-integrated wearable vests, and (2) injury records provided by the club's medical staff. By integrating these two sources, the initial dataset, *ATHLETES_DATA*, was constructed. Each entry in this dataset corresponds to a specific period with a defined segment of a training or match day for each athlete. For example, a match might be divided into three distinct periods: (i) warm-up, (ii) first half, and (iii) second half. The resulting *ATHLETES_DATA* dataset consisted of 44,354 rows and 1,715 features, including a binary label indicating whether a non-contact injury occurred within that period.

In total, 39 acute non-contact injuries were recorded among 22 players. Of these, five players sustained three injuries, seven sustained two, and ten sustained one injury each. The injuries were classified into 34 muscular injuries, 4 joint-related injuries, and 1 ligament injury, with the majority involving muscle strains and sprains affecting various segments of the lower body.

To enhance data quality and relevance, a three-step cleaning process was applied:

(i) Goalkeepers' data were removed due to their distinct training metrics and movement patterns.

(ii) Data from athletes with fewer than 12 recorded activities (equivalent to two weeks of training, including two rest days) were excluded due to insufficient information.

(iii) Columns containing only null or zero values were eliminated, while columns with partial missing values were imputed using the mean. The final dataset, *ATHLETES_DATA*, was reduced to 41,109 rows and 801 features, representing data from 79 athletes.

3.2 Feature Engineering

The *MC_ATHLETES_DATA* dataset was derived from *ATHLETES_DATA*, focusing on a selected set of features for model input. *MC_ATHLETES_DATA* was constructed by aggregating specific variables from *ATHLETES_DATA* into microcycles, defined as all training activities leading up to and including a match, with a reset occurring at the start of the next training session for each athlete. This approach aimed to mitigate class imbalance and prioritize state-of-the-art features rather than incorporating all 1,715 columns from *ATHLETES_DATA*.

The selection of common features was informed by GPS-based studies presented in related work (especially, Rossi *et al.* [2018]; Vallance *et al.* [2020]; Pilka *et al.* [2023]) alongside insights from expert club physiologists, who guided the feature engineering process. In addition to the aggregated features derived from GPS variables in *ATHLETES_DATA*, three new key features were introduced: two capturing injury recurrence patterns based on the target label and one measuring the duration of each microcycle in days, as suggested by the referenced studies. Consequently, *MC_ATHLETES_DATA* consisted of 147 microcycles, 4,326 rows, 26 independent variables, and one injury label, with 39 recorded injury cases, as detailed in Table 2.

3.3 Multicollinearity Removal Strategy

Before creating multi-dimensional models, with the 26 features created, different combinations of strategies were considered for removing multicollinear variables (Table 3) 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.

- **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).

3.4 Regressive Multi-dimensional Model Selection (RMMS)

We implemented a methodology called RMMS [Giusti *et al.*, 2022; Melo *et al.*, 2024] to assess the impact of different modeling alternatives on performance. Its first step involves defining a main function that generates and validates various multidimensional injury prediction models based on a group of specified parameters, producing a range of results by parameter combinations. The main function is presented in Algorithm 1, with its key parameters detailed in Table 5.

Algorithm 1: Methodology to Create and Validate the Predictive Model

```
function classification_predictions($df, features, target, ml, strategy, n, test\_size, grid\_search, fs$):
    $df\_processed \leftarrow copy(df, features, target)$
    $train\_X, test\_X, train\_y, test\_y \leftarrow Hold\text{-}out(df\_processed, test\_size)$
    if $strategy \leftarrow Undersampling$ then
        $sampler \leftarrow RandomUnderSampler(n)$

    if $strategy \leftarrow Oversampling$ then
        $sampler \leftarrow ADASYN(n)$
    if $fs$ then
        $pipeline\_rfecv \leftarrow Pipeline(sampler, StandardScaler, RFECV(ml))$
        $pipeline\_rfecv.fit(train\_X, train\_y)$
        $selected\_features \leftarrow get\_best\_features(pipeline\_rfecv)$
    else
        $selected\_features \leftarrow features$
    $train\_X \leftarrow train\_X(selected\_features)$
    $test\_X \leftarrow test\_X(selected\_features)$
    $pipeline\_final \leftarrow Pipeline(sampler, StandardScaler)$
    if $grid\_search$ then
        $pipeline\_final.steps.append(GridsearchCV(ml))$
    else
        $pipeline\_final.steps.append(ml)$
    $model \leftarrow pipeline\_final.fit(train\_X, train\_y)$
    $pred\_y \leftarrow model.predict(test\_X)$
    $methods \leftarrow [accuracy, precision, f1, recall, AUC\text{-}ROC]$
    foreach $method \in methods$ do
        $results \leftarrow results \cup method(test\_y, pred\_y)$
return $results$
```

The function processes the data frame using a specific set of features and the target variable, defined by the parameters $features$ and $target$, respectively. Validation is then performed using the stratified hold-out technique,

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 the 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_class	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

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%

where the $test\_size$ parameter determines the portion allocated for validation, while the remaining data is used for model training. Given the low number of injury occurrences, stratification was applied to preserve the proportional distribution of injury cases in both sets. Next, a conditional procedure determines the appropriate class balancing strategy, $grid_search$selecting between random undersampling or ADASYN (an oversampling technique used in literature).

To assess the potential benefits of hyperparameter tuning and feature selection in injury prediction models, we included both RFECV[3] and GridSearchCV[4] as optional components within the RMMS methodology. RFECV (Recursive Feature Elimination with Cross-Validation) is designed to iteratively remove less informative features while validating model performance at each step, thereby aiming to reduce dimensionality and improve generalization. GridSearchCV, on the other hand, is a widely adopted method for systematically exploring the hyperparameter space of machine learning algorithms, aiming to enhance model performance by identifying optimal configurations.

With this in mind, the next step of the classification function considers the use of RFECV if the parameter fs was set to True, integrating this method into a Scikit-Learn pipeline[5] to select the most important features according to the machine learning algorithm (ml) used, through stratified cross-validation. The pipeline ensures that class balancing, following the chosen strategy, and data normalization using StandardScaler are applied only to the training set within the K-folds of stratified cross-validation. Once completed, the designated machine learning model is trained using the selected features, which are then assigned to $train\_X$ and $test\_X$. If the parameter fs was set to False, the feature selection process does not occur, and the selected features remain as originally provided as parameter $features$.

Next, another pipeline was implemented to perform preprocessing, including class balancing and scaling. However, hyperparameter optimization for the machine learning algorithm ml occurs only if $grid\_search$ is set to True, adding this step to the pipeline using GridSearchCV, based on the features defined in the previous step. This process determines the optimal hyperparameters, which are then assigned to the main model through the $model$ variable. Otherwise, the default hyperparameters train the main model via the

[3] `https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html`

[4] `https://scikit-learn.org/0.15/modules/grid_search.html`

[5] `https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html`

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

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
$strategy$	Class balancing strategy to be applied
n	Proportion of case samples for class balancing
$test\_size$	Proportion of test data for validation
$grid\_search$	Boolean to set the use or not of hyperparameters optimization
fs	Boolean to set the use or not of a feature selector

$model$ variable.

Finally, predictions are generated, followed by an iteration over the $methods$ list to evaluate the results using multiple performance metrics stored in the $results$ variable. In this study, five evaluation metrics were considered [Majumdar *et al.*, 2022]:

1. **Accuracy**: Proportion of correctly classified injuries and non-injuries to the total number of observed injuries and non-injuries.

$$\frac{TP+TN}{TP+TN+FP+FN}$$

2. **Precision**: Proportion of correctly classified injuries to the total number of injuries classified.

$$\frac{TP}{TP+FP}$$

3. **Recall**: Proportion of correctly classified injuries to the total number of injuries.

$$\frac{TP}{TP+FN}$$

4. **F1-score**: Harmonic mean between precision and recall.

$$2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$$

5. **AUC-ROC**: Area under the ROC curve that evaluates the relationship between true positive rates and false positive rates.

$$ROC = rate(\frac{TP}{FP})$$

Due to the significant class imbalance in the test set, AUC-ROC was chosen as the primary performance metric, ensuring a more reliable assessment of the positive classification results. The classification models were designed to enhance performance by iterating over the cross-product of the classification function parameters. This cross-product was performed for each combination of $features$, ml, $strategy$, n, $grid\_search$ and fs corresponding to F, M, S, N, G, FS, respectively.

Once the classification models were generated, a simultaneous validation of all results was carried out. This enabled the identification of parameters with the most significant positive or negative impact on performance, which is the core objective of the RMMS approach. The strategy employs a Random Forest regression model, where the features include all possible iterations of classification parameters (F, M, S, N, G, FS), and the target variable is the obtained AUC-ROC score. To facilitate regression, one-hot encoding was applied to accommodate categorical values, and evaluation was conducted using the Mean Absolute Percentage Error metric. With the regression model established, the classification parameter analysis was visualized through SHAP (SHapley Additive exPlanations) values[6], providing a clear interpretation of the positive or negative influence of these parameters on the predictive performance.

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 $3.60G$Hz and 12GB of installed RAM. The project was entirely implemented in Python version 3.12.4.

Considering the classification modeling part of the RMMS methodology, 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 the $features$ parameter and set as the *MC_ATHLETES_DATA* target vari-

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

able, respectively. Along with this, $test\_size$ was also fixed at 20% for validation.

The parameter ml consisted of non-linear algorithms such as Decision Tree (DT), Random Forest (RF), Adaboost (ADA), and XGBoost (XGB), as well as linear ones like Logistic Regression (LR) and LinearSVC (LSVC). The chosen machine learning algorithms were used in our previous work and also in related studies for comparison.

To mitigate class imbalance, we chose two strategies: one for undersampling and another for oversampling, both assigned to $strategy$. With this, n consisted of three possible proportions between negative and positive case samples: 70%/30%, 60%/40%, and 50%/50%. The parameter $features$ was filled with 30 different feature combinations, divided by three multicollinearity removal strategies (explained in Section 3.3).

Finally, $grid\_search$ and fs consisted of boolean values indicating whether or not to apply machine learning hyperparameter optimization and feature selection methods.

Algorithm 2: Experimental Setup and Parameters Cross Product

```
$df \leftarrow MC\_ATHLETES\_DATA$
$target \leftarrow MC\_ATHLETES\_DATA[injury\_class]$
$test\_size \leftarrow 0.2$
$F \leftarrow \{All\_BR, All\_AR, mcr1\_BR, ..., mcr14\_BR, mcr14\_AR\}$
$M \leftarrow \{DT, RF, LR, ADA, XGB, LSVC\}$
$S \leftarrow \{Undersampling, Oversampling\}$
$N \leftarrow \{70/30, 60/40, 50/50\}$
$G, FS \leftarrow \{True, False\}$
$results \leftarrow \emptyset$
foreach $features \in F, ml \in M, strategy \in S, n \in N, grid\_search \in G, fs \in FS$ do
    $results \leftarrow results \cup classification\_predictions(df, features, target, ml, strategy, n, test\_size, grid\_search, fs)$
```

5 Results

Applying the methodology from Algorithms 1 and 2 generated 4320 different models by combining all parameters. Table 6 compares the performance of the best model from our previous work [Melo *et al.*, 2024], used as a baseline B, with the best results obtained for each of the six machine learning algorithms applied and their respective parameter combinations.

For the simultaneous analysis of the parameters used to develop the 4320 models, the SHAP values approach clearly explained the positive or negative impact on the output, according to the indicated magnitude. The predictive regression modeling with Random Forest Regressor had a 1,7% Mean Absolute Percentage Error, indicating a low error percentage in the predicted value. Since 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 to the right signifies a positive impact of not using the variable, while the opposite is on the left.

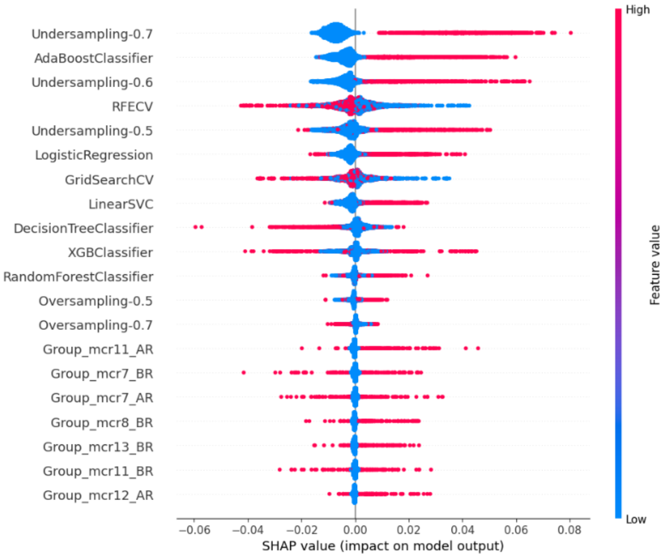

Figure 1. Graph of the positive or negative impact of the 20 most relevant SHAP values in relation to the parameters present in the regression model.

6 Discussions

As mentioned, this study aims to evaluate different modeling approaches for predicting non-contact traumatic injuries within a microcycle of male professional soccer players from Fluminense Football Club, combining Data-Centric AI (DCAI) concepts with machine learning algorithms while extending our previous work with new experiments using the RMMS methodology. As primary results, Table 6 compares the test performance of the baseline model from our previous work with the best results for each of the six algorithms explored. Prioritizing the AUC-ROC metric, the best-performing models were tree-based non-linear algorithms (XGB, RF, ADA, DT), which outperformed linear models ($LR, LSVC$). Other studies that utilized GPS data and machine learning, such as Vallance *et al.* [2020], Pilka *et al.* [2023], Dandrieux *et al.* [2023], Fang and Xiang [2024], Freitas *et al.* [2025], and Saberisani *et al.* [2025], also demonstrated strong predictive performance with tree-based algorithms like Decision Tree, Random Forest, Adaboost, and especially XGBoost. XGBoost is particularly relevant in football injury prediction, as it builds an ensemble of small classification trees (weak learners), improving them iteratively by correcting their errors. The highest reported performance in the literature was from Pilka *et al.* [2023], achieving 92.4% precision, 96.5% recall, and an F1-score of 94.4%. Similarly, in our study, XGBoost achieved the best results with an AUC-ROC of 74.1%. An important aspect is that it did not rely on GridSearchCV or RFECV, using its default parameters without an external feature selection process. Although these results were promising, they remained below the baseline obtained in our previous study, which achieved an AUC-ROC of 79.8% using the Decision Tree algorithm, 5.7% higher than the best model from the new experiments.

Table 6. Classification results on the test set for the baseline *B* (our previous work [Melo *et al.*, 2024]) and best performance for each machine learning algorithm and their respective parameter combinations, sorted by AUC-ROC. GSCV = GridSearchCV.

Model	Features	Sampling	GSCV	RFECV	Accuracy	Precision	Recall	F1-score	AUC-ROC
B	mcr14_AR	*Undersampling 70/30*	-	-	72,2%	2,8%	87,5%	5,5%	79,8%
XGB	mcr8_BR	*Undersampling 50/50*	False	False	73,2%	2,5%	75%	4,9%	74,1%
RF	mcr8_AR	*Undersampling 60/40*	True	True	71,2%	2,4%	75%	4,6%	73,1%
ADA	All_BR	*Undersampling 60/40*	True	False	66,2%	2%	75%	3,9%	70,5%
DT	mcr12_BR	*Undersampling 60/40*	True	True	77%	2,5%	62,5%	4,8%	69,8%
LR	mcr8_BR	*Undersampling 70/30*	False	True	71,1%	2%	62,5%	3,8%	66,9%
LSVC	mcr8_AR	*Undersampling 70/30*	False	True	71%	2%	62,5%	3,8%	66,8%

Except for accuracy, recall, and AUC-ROC, the other metrics remained lower for all models due to class imbalance in the test set, which could not be transformed with techniques such as undersampling or oversampling to prevent data leakage. In cases of class imbalance, accuracy is not a reliable metric as it tends to have high scores by correctly classifying most non-injury cases, which are predominant in the test set (858 non-injury vs. 8 injury cases after stratified holdout). Since injury cases are rare, accuracy remains high, but the main focus should be on the correct classification of the positive class, that is injury cases. 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. High recall helps to prevent false-negative alarms, ensuring players at true risk of injury are not wrongly allowed to play. High precision avoids unnecessary restrictions on players by reducing false-positive alarms. Both situations are detrimental to the team and athletes involved. Despite this, the main goal of this study is to establish a methodology that objectively evaluates the impact of different modeling parameters on predictive performance. Therefore, the AUC-ROC metric highlights the model's capacity to differentiate between positive and negative classes, even when there is a class imbalance, aligning with the objectives of this study within the RMMS methodology. A perfect AUC value of 1 signifies optimal performance in class distinction. In contrast, a value of 0.5 reflects random predictions, and a value lower than 0.5 indicates performance worse than random, often resulting from errors in distinguishing between injury and non-injury cases [Kuhn and Johnson, 2013].

The results in Table 6 highlight the best performances achieved through strategies that enhanced model effectiveness, emphasizing each algorithm. However, since multiple parameters influenced performance, neither the machine learning algorithms nor individual parameters alone can fully explain the outcomes. For that matter, Figure 1 displays the 20 most relevant parameters, where those at the top had the greatest impact on AUC-ROC result variations. The distribution of SHAP values offers a deeper understanding of which factors contributed positively and negatively to model performance, suggesting alternative modeling approaches for better future classification models.

According to the SHAP values, Adaboost demonstrated the most significant positive modeling impact among the machine learning algorithms used. This is particularly interesting because it suggests that is a relatively new algorithm introduced in the literature for predicting injuries in professional football with GPS, as seen in the studies by Dandrieux *et al.* [2023] and Freitas *et al.* [2025], which achieved strong predictive results using this approach. Although XGBoost yielded the best results in our research and previous studies from the state of the art, it exhibited a balanced behavior between positive and negative impacts, suggesting that it did not perform well with certain modeling parameters. For instance, our best result was with this method without hyperparameter optimization (GridSearchCV) and feature selection (RFECV). Random Forest also had a positive impact. However, unlike in our previous work, the Decision Tree did not rank among the 20 most relevant features in the SHAP analysis, where it provided the best results and a positive impact on SHAP. This suggests that ensemble-based tree models perform better in this context than a single Decision Tree model. Finally, linear models such as Logistic Regression and LinearSVC, despite achieving lower overall results, had a positive impact according to the SHAP values. This contradicts our previous findings, where Logistic Regression significantly degraded performance.

Considering class imbalance, this study explored undersampling and oversampling techniques to balance the classes in the training set. In the literature, oversampling is commonly used to add positive cases to the training set rather than removing negative cases. This approach was adopted in studies such as Rossi *et al.* [2018, 2022] and Pilka *et al.* [2023], which utilized ADASYN and SMOTE for this purpose. However, oversampling in the current research did not significantly improve predictive performance, as can be seen both in the best results from Table 6 and in the overall analysis using SHAP Values. On the other hand, although undersampling has not been widely applied in related studies, it proved to be the most effective strategy, particularly when using a sample proportion of 70% non-injury cases and 30% injury cases. This aligns with our previous work, where the same undersampling proportion strongly impacted the results.

Hyperparameter optimization using GridSearchCV and feature selection with RFECV were new approaches introduced in this study for experimentation. However, the SHAP results indicate that their use had a strong negative impact, suggesting that avoiding them leads to better performance. In our previous study, we applied PCA as a dimensionality reduction technique, significantly affecting the results. Given that the datasets used in both studies are the same, with identical strategies for removing multicollinearity to form feature groups, when hyperparameter optimization and feature selection were not applied, the model closely resembled the one from the previous study, except for the absence of PCA and the inclusion of new machine learning algorithms. This suggests that PCA played a crucial role in improving model performance compared to other feature selection strategies and appears to be a good choice for further analysis. Specifically, RFECV assumes a sufficiently large sample size to iteratively validate feature utility, an assumption that may be limited in this context due to data sparsity, with only 39 injury cases, which could have contributed to the destabilizing effect observed when applying RFECV and GridSearchCV in this study.

Finally, some preselected feature combinations for multicollinearity removal proved relevant on SHAP Values analysis, particularly the feature mcr11_AR, which was also the most impactful in our previous study.

7 Conclusion

This study aimed to validate and extend the functionality of the Regressive Multi-dimensional Model Selection (RMMS) methodology for predicting non-contact injuries in professional soccer players while extending previous work Melo *et al.* [2024] with new experiments to evaluate different modeling alternatives.

By systematically exploring a range of machine learning models within a Data-Centric AI (DCAI) framework, the study reinforced the strong performance of tree-based nonlinear algorithms—particularly XGBoost, which achieved the highest AUC-ROC (74.1%). Despite this, the model still underperformed compared to the 79.8% baseline previously attained using a simpler Decision Tree, highlighting the continued relevance of more interpretable models in certain contexts.

The experiments also provided important insights into class balancing strategies. Undersampling, especially with a 70%-30% distribution, consistently outperformed ADASYN, a oversampling method. This finding aligns with our previous work results and offers practical guidance for similar predictive tasks involving rare events.

SHAP value analysis revealed Adaboost as the most positively impactful algorithm in the overall modeling process. In contrast, methods typically associated with improved performance — such as feature selection via RFECV and hyperparameter tuning with GridSearchCV — were found to degrade results in this setting. These outcomes suggest that Principal Component Analysis (PCA), used in previous work, remains a more effective dimensionality reduction technique for this dataset.

Overall, the study contributes both methodological and empirical advances by offering a robust experimental framework for evaluating predictive modeling alternatives in injury forecasting. For future research, expanding the dataset to include additional seasons, match data, and broader athlete metrics—such as biochemical markers and subjective exertion measures—may further improve model generalization and practical utility.

Acknowledgements

We thank Fluminense Football Club for providing the necessary data to conduct the experiments. We also thank Cefet/RJ and Capes for partially funding this research.

Authors' Contributions

Matheus Santos Melo contributed to the introduction, systematic literature search, conceptualization of the methodology, experimental procedures, results analysis, and manuscript writing. **Juliano Spineti** provided expertise as a physiologist, assisting in the selection of GPS features based on the literature, and supplied the club's data. **Diego Nunes Brandão** contributed to enhancing the methodological proposal and manuscript writing (review and editing). **Lucas Giusti Tavares** contributed to the conceptualization of the methodology, supervision, and manuscript writing (review and editing). **Jorge de Abreu Soares** contributed to the conceptualization of the methodology, supervised the research, and participated in manuscript writing (review and editing).

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. DOI: https://doi.org/10.1016/j.jts.2023.04.003.

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. DOI: 10.1136/bjsports-2020-103159.

Fang, J. and Xiang, T. (2024). Medical Decision Support for Football Players Based on Machine Learning Historical Injury Data. *Revista Internacional de Medicina y Ciencias de la Actividad Fisica y del Deporte*, 24(96):479–489. DOI: 10.15366/rimcafd2024.96.029.

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. DOI: 10.1038/d41586-021-00818-1.

Freitas, D. N., Mostafa, S. S., Caldeira, R., Santos, F., Fermé, E., Gouveia, É. R., and Morgado-Dias, F. (2025). Predicting noncontact injuries of professional football players using machine learning. *PLoS ONE*, 20(1):1–21. DOI: 10.1371/journal.pone.0315481.

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. *Evolv-*

ing Systems, (0123456789). DOI: 10.1007/s12530-021-09415-z.

Hägglund, M., Waldén, M., Bahr, R., and Ekstrand, J. (2005). Methods for epidemiological study of injuries to professional football players: Developing the UEFA model. *British Journal of Sports Medicine*, 39(6):340–346. DOI: 10.1136/bjsm.2005.018267.

Haller, N., Kranzinger, S., Kranzinger, C., Blumkaitis, J. C., Strepp, T., Simon, P., Tomaskovic, A., O'brien, J., Düring, M., and Stöggl, T. (2023). Predicting Injury and Illness with Machine Learning in Elite Youth Soccer: A Comprehensive Monitoring Approach over 3 Months. *Journal of Sports Science and Medicine*, 22(3):475–486. DOI: 10.52082/jssm.2023.475.

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. DOI: 10.1136/bjsports-2013-092215.

Jarrahi, M. H., Memariani, A., and Guha, S. (2023). The Principles of Data-Centric AI. *Communications of the ACM*, 66(8):84–92. DOI: 10.1145/3571724.

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. DOI: 10.1177/03635465221112095.

Kirkendall, D. T. and Dvorak, J. (2010). Effective injury prevention in soccer. *Physician and Sportsmedicine*, 38(1):147–157. DOI: 10.3810/psm.2010.04.1772.

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. DOI: 10.1111/sms.14322.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. DOI: 10.1007/978-1-4614-6849-3.

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). DOI: 10.1186/s40798-022-00465-4.

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). DOI: 10.3390/jcm11164923.

Melo, M., Maia, M., Padrão, G., Brandão, D., Bezerra, E., Spineti, J., Giusti, L., and Soares, J. (2024). Data-centric ai for predicting non-contact injuries in professional soccer players. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 167–180, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2024.240518.

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. DOI: 10.1136/bmj.n71.

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. DOI: 10.4085/1062-6050-51.6.03.

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). DOI: 10.3390/s23031227.

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. DOI: 10.1007/s11332-022-00932-1.

Saberisani, R., Barati, A. H., Zarei, M., Santos, P., Gorouhi, A., Ardigò, L. P., and Nobari, H. (2025). Prediction of football injuries using GPS-based data in Iranian professional football players: a machine learning approach. *Frontiers in Sports and Active Living*, 7(January):1–9. DOI: 10.3389/fspor.2025.1425180.

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). DOI: 10.3390/APP10155261.