

Position-Aware Machine Learning Analysis of Football Shooting Performance

Marouane Baadi^{1,*}, Zakaria Khoudi¹, Lekbir Afraites¹, Soufiane Lyaqini²

¹LM2I, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni Mellal 23000, Morocco

²LAMDA Laboratory, National School of Applied Sciences, Sultan Moulay Slimane University, 25000, Khouribga, Morocco

Abstract Measuring individual shooting performance in professional football remains a challenging task because players differ considerably in tactical roles, playing time, and offensive responsibilities. Conventional indicators such as goals scored or total shots provide only partial assessments of shooting performance and often ignore positional expectations and playing-time normalization. Existing advanced metrics frequently rely on event-based or tracking data, limiting their accessibility for many practitioners.

In this study, we propose the Position-Aware Shooting Impact Score (PASIS), an interpretable composite metric that integrates shooting efficiency, shooting accuracy, shot volume, positional role, and playing time into a unified performance indicator using publicly available player statistics. Rather than replacing the analytical PASIS formulation, machine learning is employed to investigate the nonlinear relationships embedded within the metric, compare the predictive capabilities of several regression algorithms, and identify the variables that most strongly influence shooting performance.

A comparative evaluation was conducted using player-level shooting statistics from the English Premier League 2024–2025 season. Linear Regression, Ridge Regression, Lasso Regression, K-Nearest Neighbors, Gradient Boosting, and both baseline and hyperparameter-optimized Random Forest models were evaluated using coefficient of determination (R^2), mean absolute error (MAE), and mean squared error (MSE). Model interpretability was further investigated using SHAP analysis.

The experimental results indicate that ensemble learning methods outperform linear and distance-based models in approximating the nonlinear relationships underlying PASIS, with the tuned Random Forest achieving the highest predictive performance. The analysis further confirms that shooting efficiency, shooting accuracy, and shot volume are the dominant contributors to the metric. These findings demonstrate that PASIS provides an interpretable framework for analysing individual shooting performance while offering practical value for player evaluation, scouting, and football performance analysis.

Keywords Sports analytics; Machine learning; Football performance analysis; Shooting efficiency; Regression modeling

DOI: 10.19139/soic-2310-5070-4186

1. Introduction

Football analytics has experienced remarkable growth over the past decade owing to the increasing availability of detailed player- and match-level data, together with rapid advances in statistical modelling and machine learning. These developments have enabled researchers and practitioners to analyse player performance at an unprecedented level of detail, supporting tactical decision-making, player recruitment, and long-term performance evaluation [7, 9, 17, 13, 21]. Modern machine learning techniques are particularly well suited for modelling complex and nonlinear relationships commonly observed in football data [4].

Among the different aspects of football performance, shooting effectiveness remains one of the most important indicators of attacking success. Traditional statistics such as goals scored, total shots, or shots on target are intuitive and widely used by coaches and analysts. Nevertheless, these indicators evaluate only isolated aspects of

*Correspondence to: Marouane Baadi (Email: baadi.mar1@gmail.com)

offensive performance and often ignore contextual factors including positional role, playing time, and differences in tactical responsibilities. For example, defenders and midfielders are naturally expected to produce fewer shooting opportunities than forwards, whereas players with limited playing time may exhibit artificially high efficiency values that are not representative of their overall contribution throughout an entire season [15, 17].

To overcome these limitations, numerous advanced football analytics metrics have been proposed. Expected Goals (xG) and related metrics estimate the probability of scoring based on shot characteristics, whereas more recent frameworks such as VAEP and action-based models evaluate broader player contributions using event and tracking data [15, 19, 9]. Although these approaches provide valuable insights, they often require detailed event-level or spatial information that is not always publicly available, thereby limiting their applicability for many clubs, researchers, and practitioners.

Several studies have also proposed composite performance indicators that aggregate multiple performance dimensions into a single interpretable score [3, 2]. Such indicators facilitate player comparison and support decision-making by summarizing complex information into a single metric. However, many existing composite metrics either neglect positional expectations or rely on complex data sources that are unavailable in many practical settings. Consequently, there remains a need for a transparent, interpretable, and position-aware shooting metric that can be computed using publicly available statistics.

To address this research gap, we propose the *Position-Aware Shooting Impact Score (PASIS)*, a composite metric designed to evaluate individual shooting performance while accounting for shooting efficiency, shooting accuracy, shot volume, positional role, and playing time. Unlike advanced tracking-based metrics, PASIS is constructed exclusively from publicly available player statistics, making it easy to compute, reproduce, and interpret. The objective is not to replace established metrics such as Expected Goals, but rather to provide an accessible complementary indicator that facilitates fair comparisons among players occupying different tactical roles.

From a methodological perspective, PASIS is formulated as a continuous performance indicator suitable for regression-based analysis. In this work, machine learning is not intended to replace the analytical PASIS formulation. Instead, regression models are employed to investigate the nonlinear relationships underlying the proposed metric, compare the predictive behaviour of different learning algorithms, evaluate model robustness, and identify the variables that contribute most strongly to shooting performance through interpretable machine learning techniques [12]. This perspective provides additional insights into the structural characteristics of the proposed metric while supporting its practical application in football performance analysis.

The main contributions of this study are summarized as follows:

- We propose the Position-Aware Shooting Impact Score (PASIS), an interpretable composite metric for evaluating individual football shooting performance while accounting for positional role and playing time.
- We conduct a comprehensive comparative analysis of several regression algorithms, including Linear Regression, Ridge Regression, Lasso Regression, K-Nearest Neighbours, Gradient Boosting, and both baseline and hyperparameter-optimised Random Forest models.
- We investigate the nonlinear relationships between shooting efficiency, shooting accuracy, shooting volume, and positional role through interpretable machine learning techniques.
- We analyse model robustness using hyperparameter optimisation, cross-validation, and feature importance analysis based on SHAP values.
- We discuss the practical applicability of PASIS for player evaluation, scouting, and football performance analysis using publicly available data.

The remainder of this paper is organised as follows. Section 2 presents the problem formulation and modelling framework. Section 3 describes the dataset and preprocessing procedures. Section 4 reports the experimental results and discussion. Finally, the paper concludes with the main findings, limitations, and future research directions.

2. Setting of the Problem

The evaluation of individual shooting performance in professional football remains a challenging problem because offensive output depends on multiple interacting factors rather than a single statistic. Goals scored, shots attempted, shooting accuracy, playing time, and positional role jointly influence a player's shooting effectiveness, while tactical responsibilities vary considerably across forwards, midfielders, and defenders. Consequently, shooting performance cannot be adequately represented using isolated statistics, motivating the development of composite performance indicators capable of integrating these complementary dimensions.

The Position-Aware Shooting Impact Score (PASIS) was introduced in this study to provide an interpretable and position-aware representation of shooting performance based exclusively on publicly available player statistics. PASIS combines shooting efficiency, shooting accuracy, shot volume, positional expectations, and playing time into a single continuous indicator that enables fairer comparisons among players with different tactical responsibilities.

Because PASIS is formulated as a nonlinear mathematical function of several interacting variables, understanding how these variables collectively influence the metric is not straightforward. Rather than replacing the analytical PASIS equation, the objective of the present work is to investigate whether different machine learning algorithms can successfully approximate these nonlinear relationships, evaluate their robustness, and identify the variables that contribute most strongly to the proposed metric. This perspective allows machine learning to serve as an analytical and interpretability tool instead of merely reproducing the analytical formula.

Among the available regression techniques, Random Forest regression was selected as one of the principal benchmark models because of its ability to capture nonlinear interactions, accommodate correlated predictors, and provide reliable estimates of feature importance [6]. In addition to Random Forest, several linear, regularized, distance-based, and boosting algorithms are evaluated to provide a comprehensive comparison of regression behaviour across different modelling paradigms.

Since the predictive performance of ensemble learning algorithms strongly depends on appropriate hyperparameter selection, Random Forest models were evaluated under both default and optimized configurations. Hyperparameter optimization was performed using randomized search with CV, which has been shown to provide efficient exploration of large hyperparameter spaces while requiring substantially lower computational cost than exhaustive grid search [5, 1].

Accordingly, the objectives of this study are:

- To develop an interpretable Position-Aware Shooting Impact Score (PASIS) for evaluating individual football shooting performance.
- To investigate the nonlinear relationships underlying PASIS using several regression-based machine learning algorithms.
- To compare the predictive behaviour and robustness of linear, regularized, distance-based, boosting, and ensemble regression models.
- To analyse the effect of hyperparameter optimization on Random Forest regression performance.
- To interpret the contribution of shooting-related variables through feature importance analysis and discuss the practical implications of PASIS for football performance analysis and scouting.

2.1. Random forest regression mathematical formulation

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the training dataset, where $\mathbf{x}_i$ represents the predictor variables (shooting volume, shooting efficiency, shooting accuracy, and playing time), and y_i denotes the corresponding PASIS value. Random Forest regression estimates the nonlinear mapping

$$f : \mathbf{x} \rightarrow y$$

where $\mathbf{x}$ denotes the predictor variables and y represents the corresponding PASIS value. The final prediction is obtained by averaging the outputs of an ensemble of regression trees.

$$\hat{y}_i = \frac{1}{T} \sum_{t=1}^T h_t(\mathbf{x}_i) \quad (1)$$

where each $h_t(\cdot)$ is a regression tree—a type of decision algorithm that divides the data into branches—trained on a bootstrap sample (randomly collected data with replacement) of the original dataset.

At each split of a tree, only a random subset of features are used, which decorrelates individual trees and increases generalisation. By combining predictions from these diverse models, Random Forests can capture complex, non-linear relationships while remaining robust to noise and overfitting.

2.2. Hyperparameter Space

The predictive capability and generalization performance of Random Forest regression depend on several hyperparameters that control the complexity and diversity of the ensemble. To ensure a fair comparison between the baseline and optimized models, the following hyperparameters were considered during the optimization procedure:

- `n_estimators`: number of trees in the ensemble
- `max_depth`: maximum depth of each tree
- `min_samples_split`: minimum number of samples required to split an internal node
- `min_samples_leaf`: minimum number of samples required at a leaf node
- `max_features`: number of features considered at each split
- `bootstrap`: whether bootstrap sampling is applied

In this study, `RandomizedSearchCV` was employed to optimize the Random Forest regression model by efficiently exploring predefined hyperparameter distributions. Compared with exhaustive grid search, randomized search provides a computationally efficient alternative while maintaining competitive optimization performance for high-dimensional hyperparameter spaces [5, 1]. The optimization procedure evaluated 40 randomly sampled hyperparameter combinations using five-fold CV and the Negative Mean Squared Error (`neg_mean_squared_error`) as the optimization criterion. Table 1 summarizes the search space adopted in this study.

Table 1 summarises the search space adopted for Random Forest hyperparameter optimisation. The selected ranges were chosen based on commonly recommended values in the machine learning literature and provide an appropriate balance between model complexity, predictive performance, and computational efficiency.

2.3. Randomized Hyperparameter Search with Cross-Validation

Table 1. Hyperparameter search space used for Random Forest optimization.

Hyperparameter	Search Values
<code>n_estimators</code>	{200, 300, 400, 600, 800}
<code>max_depth</code>	{None, 5, 10, 15, 20}
<code>min_samples_split</code>	{2, 5, 10}
<code>min_samples_leaf</code>	{1, 2, 4}
<code>max_features</code>	{sqrt, log2, None}
<code>bootstrap</code>	{True, False}
<code>n_iter</code>	40
CV	5-fold
Scoring metric	Negative Mean Squared Error (<code>neg_mean_squared_error</code>)
Random state	42

Algorithm 1 Randomised hyperparameter search for Random Forest regression

```

1: Initialize  $S^* \leftarrow +\infty, \theta^* \leftarrow \emptyset$ 
2: for  $t = 1$  to  $N_{\text{iter}}$  do
3:   Sample hyperparameters  $\theta_t \sim \mathcal{P}$ 
4:   Perform  $k$ -fold CV on training data
5:   Compute mean validation MSE
6:   if  $\text{MSE} < S^*$  then
7:      $S^* \leftarrow \text{MSE}$ 
8:      $\theta^* \leftarrow \theta_t$ 
9:   end if
10: end for
11: return  $\theta^*, S^*$ 

```

Let θ denote a vector of Random Forest hyperparameters, and let $\mathcal{P}$ represent the joint distribution over the hyperparameter space. The optimal configuration is defined as:

$$\theta^* = \arg \min_{\theta} \text{CV-MSE}(\theta), \quad (2)$$

where $\text{CV-MSE}(\theta)$ denotes the mean squared error estimated using k -fold cross-validation (CV).

Randomized hyperparameter optimization enables efficient exploration of the search space while reducing computational cost compared with exhaustive grid search. The optimized Random Forest model was subsequently compared with the baseline configuration to evaluate the effect of hyperparameter optimization on predictive performance and robustness.

3. Material and Methods

3.1. Dataset Description

Player-level shooting statistics were collected from FBref, which provides standardized football statistics derived from official match reports. The dataset contains observations from the English Premier League 2024–2025 season, ensuring temporal consistency across all analysed players.

The dataset includes shooting related performance metrics for every outfield player who appeared in at least one Premier League game during the season. Goalkeepers were not included, because shooting statistics do not fully reflect their contributions on the field. The rows are dedicated to individual players and their consolidated seasonal stats.

We have the following collection of shooting metrics available for each shooter:

- Total shots attempted (Sh)
- Shots on target (SoT)
- Goals scored (Gls)
- Per-90-minute shooting rates (Sh/90, SoT/90)
- Finishing effectiveness ratios (G/Sh & G/SoT)
- Playing time expressed as the number of complete 90-minute equivalents (90s)
- Position played (Pos): Forward (FW), Midfielder (MF), or Defender (DF)

Table 2. Descriptive statistics of shooting variables

Variable	Mean	Std	Min	Max
Sh	21.32	23.60	1.00	130.00
SoT	7.47	9.54	0.00	61.00
Gls	2.42	3.88	0.00	29.00
Sh/90	1.74	1.78	0.05	18.00
SoT/90	0.66	1.24	0.00	18.00
G/Sh	0.093	0.136	0.00	1.00
G/SoT	0.268	0.257	0.00	1.00
90s	15.67	12.27	0.10	45.90

This player-centred visualisation allows detailed comparison of individual shooting performance while controlling for variation in both playing time and role specification. After the data had been consolidated and cleaned, all team-level datasets were joined in an overall master table containing all eligible 2024-2025 Premier League players.

3.2. Preprocessing

Several preprocessing steps were performed to validate the quality of data and ensure methodological rigour prior to analysis. We filtered out the goalkeepers by their positional tags. Furthermore, players who had not taken any shots were discarded to prevent undefined efficiency ratios and keep the analysis focused on active shooters.

Because playing time has been shown to affect unadjusted prognostic metrics, all rate-based predictors were scaled to a per-90-minute basis. Moreover, player positions were explicitly included in the modelling to address differences between forwards, midfielders and defenders with regards to their tactical roles.

PASIS scoring was not normalised because the final score posed scale-invariance in terms of playing time. However, for machine learning methods sensitive to feature scaling, input variables were standardised with z-score normalisation.

A visual inspection alongside summary statistics during exploratory data analysis confirmed no anomalous values were present. Outliers weren't rejected, as extreme shooting performances can give valuable information about analyzing football performance.

The resulting dataset provides a robust basis for regression-based analysis of shooting performance and for characterizing different player shooting profiles.

The analyses were implemented in Python using the Scikit-learn machine learning library together with NumPy, Pandas, Matplotlib, and SHAP. A fixed random state (42) was used throughout the experiments to ensure reproducibility.

3.3. Descriptive Statistical Analysis

Then, to better understand the distribution of the shooting-related variables we calculated standard descriptive statistics (mean; standard deviation; minimum; maximum).

There is a large degree of variation between players for many of the shot volume and efficiency metrics, as demonstrated by the descriptive stats table. This further confirms the heterogeneous nature of shooting performance among tactical roles.

3.4. Correlation Analysis

A Pearson correlation matrix was computed to examine the relationships among the shooting-related variables.

The strongest positive correlation was observed between total shots (Sh) and shots on target (SoT) ($r = 0.95$), followed by goals scored (Gls) and shots on target (SoT) ($r = 0.94$). Goals also exhibited a strong positive

correlation with total shots ($r = 0.86$), indicating that players generating more shooting opportunities generally scored more goals.

The finishing efficiency indicators (G/Sh and G/SoT) were strongly correlated ($r = 0.81$), suggesting that both variables capture complementary aspects of shooting quality.

Playing time (90s) was moderately correlated with shot volume ($r = 0.63$), indicating that players with greater playing time generally attempted more shots throughout the season.

Overall, these relationships demonstrate that shooting performance depends jointly on shooting volume and finishing efficiency, supporting the rationale for integrating both components into the proposed PASIS metric.

3.5. Feature of the Target

The dependent variable in this study is the Player Adjusted Shooting Impact Score (PASIS) which is a continuous performance variable of a player's shooting impact normalized across different positions and minutes played.

PASIS is derived by a single number that combines shooting efficiency (value of shots), shooting accuracy (hit percentage), shot volume, and minutes played into one number. The PASIS, however, is a normalised (in that it accounts for the varying amounts of time different players have to play in games) and role-specific estimate of shooting impact which allows you to compare shooting on a fair basis beyond what should be realistic or unrealistic for an individual player based on his tactical engagement.

Formally, PASIS is defined as:

$$\text{PASIS}_i = \left(\alpha_{p(i)} \cdot \frac{\text{Gls}_i}{\text{Sh}_i} + \beta_{p(i)} \cdot \frac{\text{SoT}_i}{\text{Sh}_i} \right) \cdot \frac{\log(1 + \text{Sh}_i)}{\sqrt{90\text{s}_i}}$$

where Gls_i = goals, Sh_i = shots, SoT_i = shots on target, and 90s_i = full matches played by player i . $\alpha_{p(i)}$ and $\beta_{p(i)}$ are position-dependent weights, indicating various tactical quality expectations of forwards, midfielders and defenders.

The position-dependent coefficients $\alpha_{p(i)}$ and $\beta_{p(i)}$ were selected to reflect the different offensive responsibilities associated with forwards, midfielders, and defenders. These coefficients ensure that the contribution of shooting efficiency and shooting accuracy is evaluated relative to the tactical role of each player, thereby enabling fairer comparisons across playing positions. The coefficient values were selected based on football domain knowledge regarding the offensive responsibilities associated with each playing position. Future studies may estimate these coefficients using data-driven optimization techniques.

The logarithmic transformation of shot volume provides diminishing returns for players with exceptionally high shooting frequency, preventing the score from being dominated solely by shot quantity. Likewise, dividing by the square root of playing time normalizes the score across different participation levels while avoiding excessive penalization of players with fewer minutes played.

PASIS is formulated as a continuous target variable, making it naturally suitable for regression analysis. Rather than replacing the analytical PASIS equation, the regression models are employed to investigate the nonlinear relationships underlying the proposed metric, compare different learning algorithms, and identify the contribution of the constituent variables through interpretable machine learning.

The tactics inference framework is also successfully employed to infer other tactical aspects of football, including predicting match results and supporting in-game decisions [8, 3].

Ridge Regression

Ridge regression adds a L_2 -norm regularisation term to make coefficient estimates more stable and lower model variance, especially when predictors are correlated. The optimisation problem is defined as follows: let $X \in \mathbb{R}^{n \times p}$ be a feature matrix and $y \in \mathbb{R}^n$ be a target vector (PASIS).

$$\hat{\beta}_{\text{ridge}} = \arg \min_{\beta} \{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \}, \quad \lambda \geq 0$$

Algorithm 2 Ridge Regression via Closed-Form Solution**Require:** Feature matrix $X \in \mathbb{R}^{n \times p}$, target vector $y \in \mathbb{R}^n$, regularization parameter λ **Ensure:** Ridge estimator $\hat{\beta}$

- 1: Compute $A \leftarrow X^\top X + \lambda I$
- 2: Compute $b \leftarrow X^\top y$
- 3: Solve the linear system $A\beta = b$
- 4: **return** $\hat{\beta} = \beta$

Algorithm 3 Ridge Regression via Gradient Descent**Require:** X, y, λ , learning rate η , iterations T **Ensure:** $\hat{\beta}$

- 1: Initialize $\beta^{(0)} = 0$
- 2: **for** $t = 0$ to $T - 1$ **do**
- 3: Compute gradient g
- 4: Update parameters $\beta^{(t+1)} = \beta^{(t)} - \eta g$
- 5: **end for**
- 6: **return** $\beta^{(T)}$

Closed-Form Solution

$$\hat{\beta}_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y$$

*Algorithm: Ridge Regression (Closed-Form)**Ridge Regression via Gradient Descent* The regularised loss function is defined as:

$$L(\beta) = \frac{1}{2n} \|y - X\beta\|_2^2 + \frac{\lambda}{2n} \|\beta\|_2^2$$

Its gradient with respect to β is

$$\nabla_{\beta} L = -\frac{1}{n} X^\top (y - X\beta) + \frac{\lambda}{n} \beta$$

The iterative update rule is given by:

$$\beta^{(t+1)} = \beta^{(t)} - \eta \nabla_{\beta} L$$

Lasso Regression Lasso regression employs an L_1 -norm penalty, promoting sparsity in the estimated coefficients and enabling implicit feature selection:

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \}, \quad \lambda \geq 0$$

The L_1 penalty is defined as:

$$\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

Due to the non-differentiability at zero, Lasso does not admit a closed-form solution and is typically optimized using coordinate descent.

Algorithm 4 Lasso Regression via Coordinate Descent**Require:** X, y, λ , iterations T **Ensure:** Lasso estimator $\hat{\beta}$

```

1: Initialize  $\beta = 0$ 
2: for  $t = 1$  to  $T$  do
3:   for  $j = 1$  to  $p$  do
4:     Compute partial residual.  $r_j$ 
5:     Compute  $z_j = \frac{1}{n} X_j^\top r_j$ 
6:     Update  $\beta_j$  via soft-thresholding
7:   end for
8: end for
9: return  $\beta$ 

```

Algorithm: Lasso Regression via Coordinate Descent For standardized features, the update rule for the coefficient j is:

$$\beta_j \leftarrow S \left(\frac{1}{n} \sum_{i=1}^n x_{ij} \left(y_i - \sum_{k \neq j} x_{ik} \beta_k \right), \frac{\lambda}{2n} \right)$$

where $S(\cdot)$ denotes the soft-thresholding operator:

$$S(z, \gamma) = \begin{cases} z - \gamma, & z > \gamma \\ 0, & |z| \leq \gamma \\ z + \gamma, & z < -\gamma \end{cases}$$

Random Forest Regression Random Forest regression is one of the most widely used ensemble learning methods because of its robustness, predictive accuracy and resistance to overfitting [6, 20].

Formally, for an input vector $\mathbf{x}$, the Random Forest prediction is given by:

$$\hat{y}_{\text{RF}}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T h_t(\mathbf{x})$$

where $h_{t=1}^T$ is the ensemble of T regression trees.

Two Random Forest configurations were evaluated: a baseline model using the default hyperparameters and an optimized model obtained through randomized hyperparameter search with cross-validation.

Gradient Boosting Regression Gradient Boosting Regression is an additive ensemble learning method originally introduced by Friedman [11]. Unlike Random Forests, which construct trees independently, Gradient Boosting sequentially adds weak learners to minimize a predefined loss function through gradient-based optimization.

The model is trained with the loss function $L(y, \hat{y})$ as:

$$\hat{y}(x) = \sum_{m=1}^M \gamma_m h_m(x)$$

where $h_m(x)$ is a shallow decision tree learned by fitting to the negative gradient of the loss function at iteration m , and γ_m is a learning rate which determines how much each tree contributes to the model.

Gradient Boosting excels in capturing subtle, non-linear relationships and interaction effects, making it well suited for modelling composite performance indicators like PASIS. It is a very flexible one, providing better predictive capability but in return making it highly sensitive to hyperparameter tuning.

K-Nearest Neighbors (KNN) Regression K-Nearest Neighbours Regression (k-NN) is a non-parametric, instance-based learning algorithm that predicts the target values based on their resemblance with observation. Instead of explicitly learning a model like the previous algorithms, KNN memorizes the training data and, for prediction, averages target values from k closest neighbours in feature space.

For a query point x , the KNN prediction is defined as:

$$\hat{y}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i$$

where $\mathcal{N}_k(x)$ means the set of k nearest neighbors of x computed using a distance metric of choice (most often euclidean).

KNN works well when the target variable is locally smooth in the feature space. This framework, when embedded in PASIS modelling, can learn locally around shooting performance between players with similar measures of efficiency, accuracy and time played. Feature scaling is very important for KNN, we had standardised all variables which were input before training.

4. Experimental Results and Discussion

The proposed Position-Aware Shooting Impact Score (PASIS) was evaluated using several regression-based machine learning models to investigate the nonlinear relationships underlying the proposed metric and compare their predictive performance. The primary objective was to evaluate the ability of different regression algorithms to approximate PASIS as a continuous target variable and to compare their predictive performance.

The analysis is per player-season and utilizes only shooting-related variables, which allows for consideration in the context of individual offensive contribution without ignoring how playing time/role can vary along position.

The need for interpretability of applied machine learning models has been well established in general as well as in sports analytics [18].

4.1. Experimental Setup

The dataset was randomly partitioned into 80% training data and 20% testing data using a fixed random state (42) to ensure reproducibility. Model training, hyperparameter optimization, and validation were performed exclusively on the training set, while the independent test set was reserved for final performance evaluation. All experiments were implemented in Python using the Scikit-learn library.

The predictor variables consisted of shooting volume, shooting accuracy, shooting efficiency, playing time, and player position. Continuous variables were standardized using z-score normalization for regression models sensitive to feature scaling, whereas tree-based ensemble models were trained using the original feature values.

PASIS was treated as a continuous target variable throughout the regression analysis. We performed regression analyses including Linear Regression, Ridge regression, Lasso Regression, K-Nearest Neighbours, gradient Boosting Regression and both baseline and tuned/random Forest models. All Random Forest hyperparameter optimisations were performed using randomised search with five-fold CV.

4.2. Evaluation Metrics

Three standard metrics were used to evaluate the regression performance:

- **Coefficient of Determination (R^2):**

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

which gauges the percentage of variance in PASIS that is accounted for by the model.

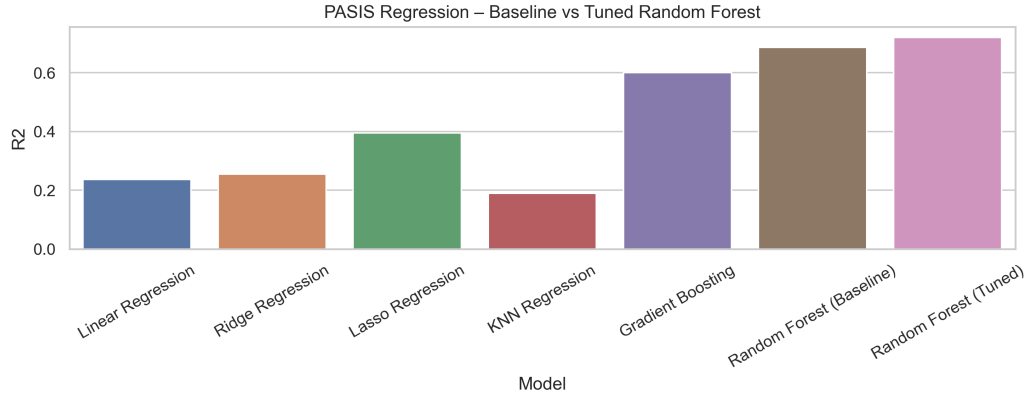


Figure 1. Coefficient of determination (R^2) of the regression models for PASIS prediction.



Figure 2. MAE comparison between regression models for PASIS prediction.

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

which represents the average scale of prediction errors.

- **Mean Squared Error (MSE):**

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

which punishes big mistakes and measures robustness of the model.

4.3. Regression Results

The regression results demonstrate clear differences in predictive performance across the evaluated algorithms. Linear Regression and Ridge Regression achieved only moderate predictive accuracy, indicating that the

Table 3. Performance comparison of regression models for PASIC prediction.

Model	MAE	R^2
Linear Regression	0.077	0.42
Ridge Regression	0.078	0.41
Lasso Regression	0.087	0.27
KNN Regression	0.083	0.26
Gradient Boosting	0.056	0.48
Random Forest (Baseline)	0.055	0.57
Random Forest (Tuned)	0.057	0.58

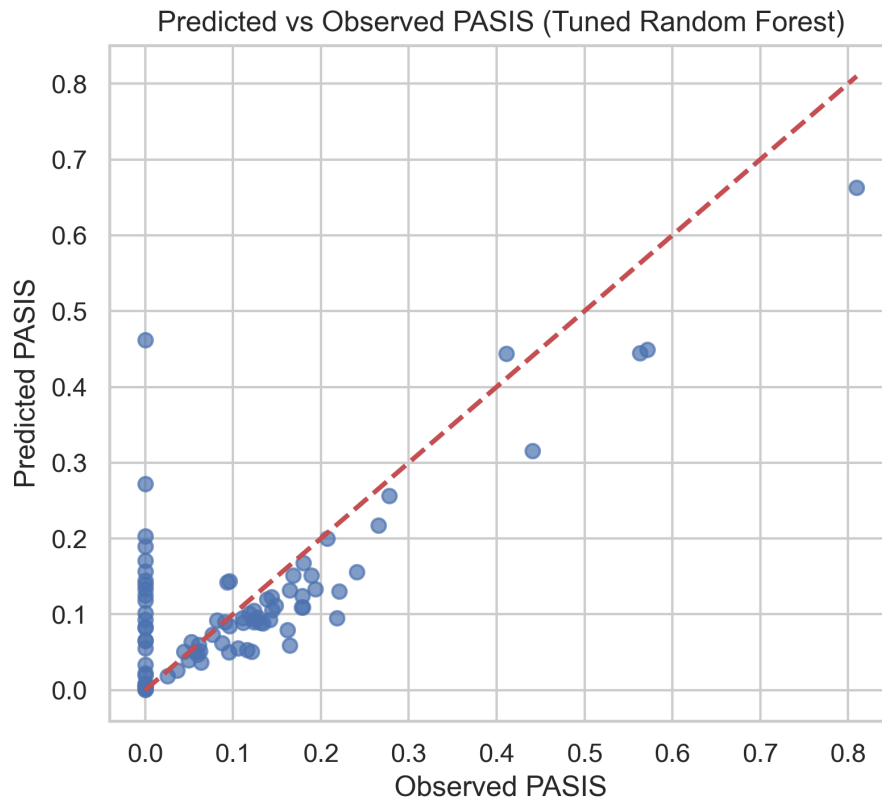


Figure 3. Predicted versus observed PASIC values for the tuned Random Forest regression model. The dashed line indicates perfect prediction.

relationship between the predictor variables and PASIC cannot be adequately represented using purely linear models. Lasso Regression and K-Nearest Neighbours produced lower predictive performance, whereas the ensemble-based methods consistently achieved superior results. These findings indicate that nonlinear learning algorithms are more capable of approximating the complex relationships underlying PASIC than linear or distance-based regression models.

The performance of all regression models measured by explained variance (R^2) and mean absolute error (MAE) are visualized in Figure 1 (the left side group of figures) and Figure 2, respectively. Table 3 shows the

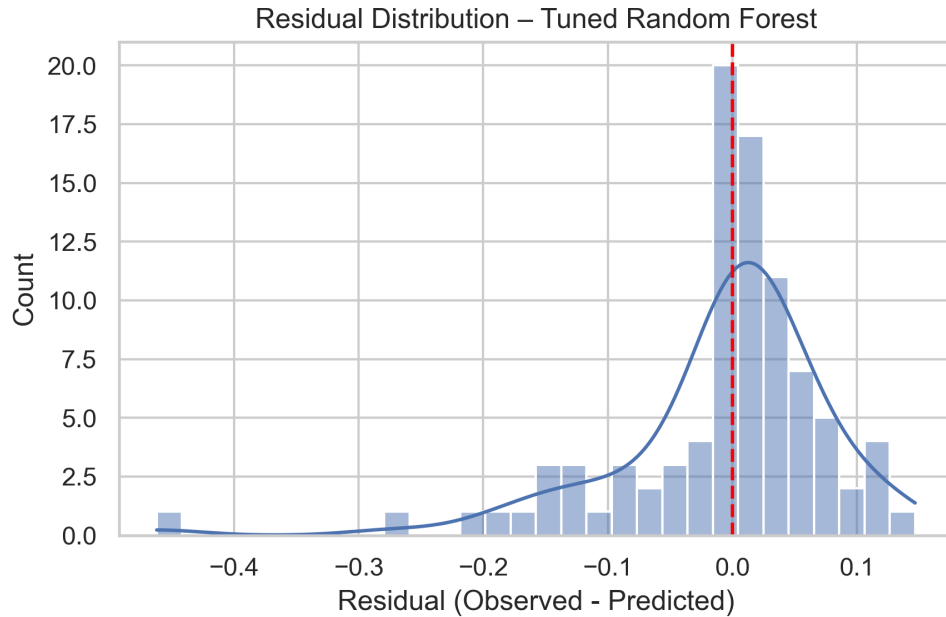


Figure 4. Distribution of prediction residuals for the tuned Random Forest regression model.

associated numerical results, enabling model comparisons directly. Among all evaluated models, the ensemble-based approaches, particularly the Random Forest regressors, achieved the highest predictive accuracy while maintaining the lowest prediction errors.

Among all evaluated algorithms, the Random Forest models achieved the highest predictive performance. The tuned Random Forest obtained the highest coefficient of determination (R^2), whereas the baseline model produced a slightly lower mean absolute error (MAE). Overall, hyperparameter optimization resulted in a modest improvement in predictive performance while maintaining comparable prediction errors, demonstrating the benefit of systematic parameter tuning.

Observed vs predicted PASIS values for the tuned Random Forest model are shown in Figure 3 with evident alignment along the identity line and very little systematic bias. The residuals distribution as presented in Figure 4 is mostly centered around 0 with no apparent skewness. Residuals as a function of predicted PASIS are shown in Figure 5 and reveal no discernable patterns indicative of heteroscedasticity or structural misspecification.

Overall, these findings demonstrate that nonlinear ensemble learning models are better suited for approximating the nonlinear structure of PASIS than conventional linear regression approaches. The improved predictive performance obtained with Random Forest confirms its suitability for modelling complex interactions among shooting efficiency, shooting accuracy, shot volume, and playing time.

4.3.1. Statistical Significance of Features To further investigate the contribution of individual predictors, feature significance was examined using linear regression coefficients together with ensemble-based feature importance measures. The results consistently indicate that shooting efficiency (G/Sh) and shooting accuracy (SoT/90) contribute most strongly to PASIS, whereas shot volume (Sh/90) provides additional explanatory information. Playing time (90s) contributes mainly through the normalization component of PASIS rather than acting as a dominant predictor. These findings agree with the mathematical formulation of PASIS and confirm that shooting quality is more influential than shooting quantity alone.

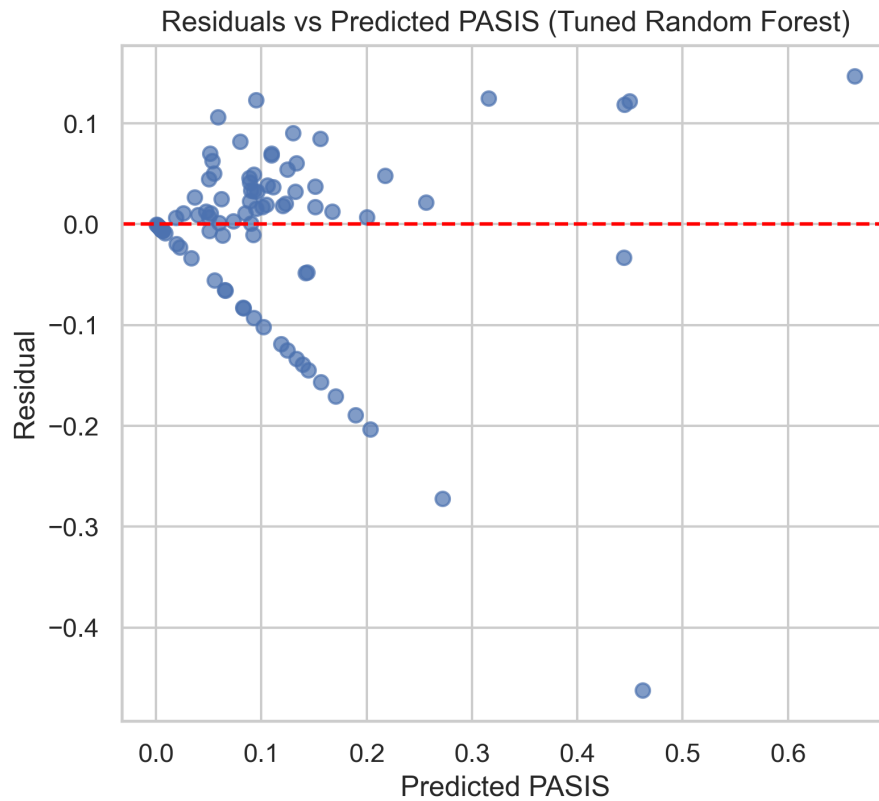


Figure 5. Residuals versus predicted PASIS values for the tuned Random Forest model.

4.4. Feature Importance

Feature importance analysis was performed using the Random Forest model to quantify the contribution of each predictor variable to PASIS estimation. The analysis consistently identified shooting efficiency (G/Sh) as the most influential predictor, followed by shooting accuracy (SoT/90). Shot volume (Sh/90) contributed moderately, whereas playing time (90s) mainly acted as a normalization factor that stabilized the final PASIS estimates. These results confirm that PASIS is primarily driven by shooting quality while still accounting for offensive involvement.

4.4.1. Residual Diagnostics Visual inspection of the residual distribution and the residual-versus-predicted plot revealed no evident systematic bias or heteroscedasticity. The residuals remained approximately centred around zero, indicating that the fitted model provides stable predictions across the range of PASIS values.

4.4.2. SHAP-Based Feature Importance For the Random Forest model, SHAP (SHapley Additive Explanations) values were computed to improve interpretability and quantify feature contributions [16, 12].

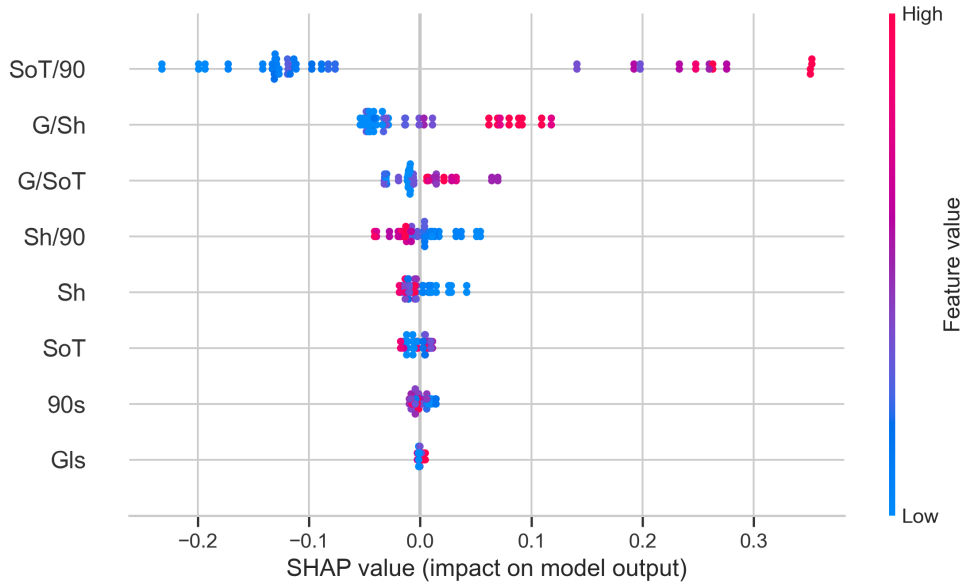


Figure 6. Summary plot of SHAP values for PASIS prediction

The SHAP analysis indicates that shooting efficiency (G/Sh) and shooting accuracy (SoT/90) contributed most strongly to PASIS prediction. Shot volume (Sh/90) exhibited a moderate influence, whereas playing time (90s) primarily stabilized the predicted values rather than acting as a dominant predictor. These findings are consistent with the feature importance analysis and further improve the interpretability of the proposed PASIS framework.

4.5. Discussion and Implications

The experimental evidence suggests that individual shooting performance in professional football is driven by complex nonlinear relationships among efficiency, accuracy, opportunity volume, playing time, and positional role. These findings are consistent with recent studies in football analytics that combine machine learning techniques with domain-specific performance metrics for player evaluation and tactical analysis [10, 3, 12]. The results indicate that ensemble learning models are more capable of capturing these complex interactions than linear or distance-based approaches [14].

It is important to emphasize that PASIS is a deterministic performance metric derived analytically from observable shooting statistics. Consequently, the objective of the regression models is not to replace the analytical PASIS equation but rather to investigate how different machine learning algorithms approximate its nonlinear structure and to identify the variables that contribute most strongly to the metric. Therefore, the reported predictive performance should be interpreted as evidence of each model's ability to learn the nonlinear relationships underlying PASIS rather than to estimate an unknown target variable.

These findings are in line with recent work in football analytics that combines interpretable machine learning models with domain-specific performance indicators [?, ?, ?, 12].

While the PASIS framework aims for accurate predictions under the assumption of model correctness and data integrity, the primary contribution of this work lies in identifying structural relationships among fundamental shot-related concepts. Future extensions could incorporate additional shot data with spatial components (such as shot locations and expected goals), contextual match variables, or longitudinal analyses to further enhance individual-level shooting assessment.

4.6. Limitations

This study has several limitations. First, the experiments were conducted using data from a single Premier League season (2024–2025), which may limit the generalizability of the findings across leagues and seasons. Second, only publicly available shooting statistics were considered, whereas contextual variables such as expected goals (xG), shot location, defensive pressure, and match state were not included. Finally, although the regression models successfully approximated the nonlinear structure of PASIS, external validation on independent datasets would further strengthen the robustness of the proposed framework.

5. Conclusion

This study introduced the Position-Aware Shooting Impact Score (PASIS), an interpretable composite performance metric that integrates shooting efficiency, shooting accuracy, shot volume, positional role, and playing time into a single continuous indicator for evaluating individual football shooting performance. Using player-level data from the English Premier League 2024–2025 season, PASIS was analysed through several regression-based machine learning models to investigate the nonlinear relationships underlying the proposed metric.

The experimental results demonstrated that ensemble learning methods consistently outperformed conventional linear and distance-based regression models. Among the evaluated approaches, the optimized Random Forest achieved the strongest overall predictive performance, confirming its ability to capture the complex interactions among the shooting-related variables composing PASIS. Furthermore, feature importance analysis showed that shooting efficiency (G/Sh) and shooting accuracy (SoT/90) were the most influential predictors, whereas shot volume and playing time provided complementary information for robust performance evaluation.

Rather than replacing the analytical PASIS formulation, the proposed machine learning framework provides an interpretable approach for analysing its nonlinear structure and identifying the relative contribution of the underlying performance variables. Consequently, the proposed methodology offers a useful decision-support framework for football performance analysis, player evaluation, and scouting applications.

Future research will focus on extending PASIS by incorporating richer contextual information, including expected goals (xG), shot location, defensive pressure, and match context. Additional validation across multiple leagues, seasons, and independent datasets will further assess the generalizability and robustness of the proposed framework.

Author Contributions

Marouane Baadi conceived the study, designed the methodology, implemented the machine learning models, analysed the data, and drafted the manuscript. Zakaria Khoudi contributed to the study design, interpretation of the results, and manuscript revision. Lekbir Afraites participated in methodological validation and critical revision of the manuscript. Soufiane Lyaqini supervised the research and contributed to the interpretation of the findings. All authors read and approved the final manuscript.

6. Statements and Declarations

- **Conflict of interest:** The authors declare no conflict of interest.
- **Availability of data and materials:** The data used in this study were obtained from open-access sources. Player-level shooting statistics for the English Premier League 2024–2025 season were retrieved from <https://fbref.com/en/>.
- **Code availability:** The Python source code used to reproduce the experiments is available from the corresponding author upon reasonable request.

REFERENCES

1. Sudirman Rizki Ariyanto, Bambang Suprianto, Warju, and Ata Syifa' Nugraha. Hybrid rsm-ga hyperparameter tuning of artificial neural networks for academic performance prediction. *Statistics, Optimization and Information Computing*, 16(2):1182–1201, 2026.
2. Marouane Baadi, Lekbir Afraites, and Soufiane Lyaqini. Machine learning prediction of long jump performance based on biomechanical factors. In Amine Laghrib and Abdelghani Ghazdali, editors, *Proceedings of the Third ICMDs'24: Machine Learning, Inverse Problems and Related Fields*, volume 1466 of *Lecture Notes in Networks and Systems*, pages 23–35. Springer, Cham, 2025.
3. Marouane Baadi, Zakaria Khoudi, Lekbir Afraites, and Soufiane Lyaqini. Machine learning for optimal player substitutions in soccer. *SN Computer Science*, 6:240, 2025.
4. Imane Bahammou, Soumia Ziti, and Karim El Bouchti. Machine learning for financial market prediction: A systematic literature review. *Statistics, Optimization and Information Computing*, 16(2):1697–1715, 2026.
5. James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.
6. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
7. Rory P. Bunker and Fadi Thabtah. A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1):27–33, 2019.
8. Anthony Constantinou. A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322–339, 2012.
9. Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1851–1861, 2019.
10. Javier Fernández and Luke Bornn. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *Proceedings of the 12th MIT Sloan Sports Analytics Conference*, Boston, MA, USA, 2018.
11. Jerome Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.
12. T. M. Mehrab Hasan, Rakib Hossen, Anichur Rahman, Mst Deoara Khushi, Md Masud Rana, and Jerin Akter. Intelligent investment predictor: An explainable machine learning framework for investment sector preference prediction. *Statistics, Optimization and Information Computing*, 16(2):1417–1446, 2026.
13. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009.
14. Teguh Herlambang, Mohamad Yusak Anshori, Bambang Suharto, Zuraini Othman, Mohd Sanusi Azmi, and Mochammad Romli Arief. A comparative statistical analysis of decision tree and adaboost ensemble for employee performance classification in the hospitality sector. *Statistics, Optimization and Information Computing*, 16(2):1813–1824, 2026.
15. Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, Stuart Morgan, and Iain Matthews. Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *Proceedings of the 8th MIT Sloan Sports Analytics Conference*, Boston, MA, USA, 2014.
16. Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
17. Ian G. McHale, Philip A. Scarf, and David E. Folker. On the development of a soccer player performance rating system for the english premier league. *Interfaces*, 42(4):339–351, 2012.
18. Christoph Molnar. *Interpretable Machine Learning*. 2020. Accessed July 2026.
19. Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6:236, 2019.
20. Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.
21. Robert Rein and Daniel Memmert. Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus*, 5(1):1410, 2016.