Data-driven approach to defining football styles in major leagues

Andres Chacoma*

*Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales,
Departamento de Física. Buenos Aires, Argentina and
CONICET - Universidad de Buenos Aires,
Instituto de Física Interdisciplinaria y Aplicada (INFINA). Buenos Aires, Argentina.*

Orlando V. Billoni

*Universidad Nacional de Córdoba. Facultad de Matemática, Astronomía,
Física y Computación. Grupo de Teoría de la Materia Condensada. Córdoba, Argentina and
Consejo Nacional de Investigaciones Científicas y Técnicas,
CONICET, IFEG. Córdoba, Argentina.*

Abstract

This study proposes a data-driven methodology to define and compare styles of play in football, with a focus on the top four teams from the English, French, German, Italian, and Spanish leagues during the 2017/2018 season. Using event-based metrics derived from possession intervals, we constructed a feature matrix representing tactical behaviors at the match level. A Principal Component Analysis, followed by Varimax rotation, revealed four interpretable and distinct emergent playing styles. By projecting matches onto this style-based representation, we evaluated stylistic differences across leagues. A one-way Anova test confirmed significant inter-league variation in style prevalence. Furthermore, a random forest classifier successfully identified leagues based on the style representation, and a game-theoretic feature importance analysis uncovered consistent associations between specific styles and leagues. These findings provide a robust, reproducible framework for empirically analyzing football playing styles across competitive contexts.

I. INTRODUCTION

In recent years, the ability to collect data from sports competitions has increased significantly. The integration of sensors on athletes, alongside the use of intelligent systems for the automated analysis of video recordings, combined with advanced statistical tools, has become a key strategy for many coaches seeking to optimize athletic performance. This data explosion has also opened new avenues for analysis from the perspective of complex systems physics, enabling the characterization of emergent phenomena of interest to both academic research [1–9] and the professional sports domain in its ongoing pursuit of performance enhancement [10–12].

In the context of the world's most popular sport, efforts to characterize team playing styles are ubiquitous. In what follows, we briefly describe the findings of some of the most relevant studies in the existing literature. From a qualitative perspective, Sarmento et al. [13] conducted semi-structured interviews with eight expert coaches from the Portuguese top-tier league, aiming to identify the factors that, in their view, differentiate the playing styles of the three major European football leagues (English Premier League, Spanish La Liga, and Italian Serie A). Through content analysis, the study explored cultural, strate-

[*] achacoma@df.uba.ar

gic, tactical, and technical aspects, as well as individual player characteristics and coaching philosophies, to understand how these elements shape the playing identities of each league and of iconic clubs such as Barcelona, Inter Milan, and Manchester United. In contrast, Fernández-Navarro et al. [14] examined the influence of contextual variables—match status, home advantage, and opponent quality—on different football playing styles. The analysis was based on data from 380 matches in the English Premier League during the 2015–2016 season. Linear mixed models were applied to assess how these variables affected style adherence scores across eight playing styles: Direct Play, Counterattack, Maintenance, Build-up, Sustained Threat, High Tempo, Crossing, and High Pressure. The results revealed significant effects of match status on all eight styles, of home advantage on all but Counterattack and Maintenance, and of opponent quality on all but Counterattack. Moreover, significant interaction effects among contextual variables were observed for several styles. These findings highlight the crucial role of context in shaping tactical choices and offer practical implications for coaches and performance analysts. Using a different approach, González – Rodenas et al. [15] analyzed differences in playing style and technical performance among professional teams in the Spanish league, based on their final ranking position. The study included 38 teams and a total of 4,940 matches played between the 2008/09 and 2020/21 seasons. Teams were grouped into five categories according to their final standing: champions, Champions League qualifiers, Europa League qualifiers, mid-table teams, and relegated teams. Linear mixed models were used to evaluate the effect of ranking on variables related to playing style and technical performance, with effect sizes estimated using the F^2 statistic. This approach allowed for the identification of distinct patterns in offensive organization and technical attributes associated with different levels of team competitiveness. In another study, Casal et al. [16] applied principal component analysis (PCA) to reduce the dimensionality of a large data matrix related to team performance in the Spanish league during the 2015/16 to 2017/18 seasons, thereby facilitating the interpretation of the involved variables. The resulting components were then used in multiple linear regression to conduct a comparative analysis between top-performing and lower-performing teams. This approach enabled the identification of key performance indicators (KPIs) that differentiate the two groups, highlighting relevant offensive and defensive features that contribute to success or failure in football performance.

As is well known, football is a cultural expression that can vary significantly from one country to another. Consequently, it is reasonable to expect that such differences manifest in distinct playing styles that can be observed and analyzed. Within this framework, the aim of our study is to investigate whether specific styles characterize the main European leagues and to quantify, using statistical methods, the extent to which these styles differ from one another. To this end, we propose a data-driven approach to characterize the playing style of the top-performing teams in the English, French, German, Italian, and Spanish leagues.

The article is organized as follows. In Section II, we describe the dataset used and define the tactical metrics on which our style analysis is based. Section III presents our main findings. In Section III.A, we provide a general statistical characterization of the selected metrics. In Section III.B, we explain how these metrics are used to identify emerging playing styles. Sections III.C and III.D focus on analyzing which styles best represent each league. Finally, in the last section, we summarize the key results of our work.

II. DATA

For this work, we used the event-based dataset provided by L. Pappalardo et al. in [17]. In that article, the authors visualize all matches from the 2017–2018 season of the top European football leagues: Premier League (England), Ligue 1 (France), Bundesliga (Germany), Serie A (Italy), and La Liga (Spain). For each match, they detect, classify, and locate in time and space all events: passes, shots, corners, etc. In the reference frame they employ, t denotes the time elapsed since the start of the match, the coordinate x indicates the distance from the goal defended by the team responsible for the event, and the coordinate y indicates the distance from the right sideline. The spatial coordinates are expressed as percentages of the field length and width, so that for example $x = 0$, $x = 50$, and $x = 100$ correspond respectively to the defending goal line, the midfield line, and the opponent's goal line.

Below we describe how we extracted from the database the metrics used in this study. First, we define a Ball Possession Interval (BPI) as the set of events forming a continuous sequence generated by a single team. It is important to note that each BPI contains information from only one team. Within this framework, we collected all BPIs from the top four teams of each league and extracted from them several metrics designed to identify the tactical resources deployed by the teams during each temporal window of the match. The metrics used in our analysis are based on those proposed by J. Fernández-Navarro in [14]. We describe each of them in detail below:

1. *Direct play.* For each pass or free kick in a BPI, we compute the average speed in the attacking direction, defined as the ratio between the distance traveled by the ball along the x-axis and the elapsed time. We retain the maximum value from each BPI. This allows us to assess how directly the team moves the ball toward the opponent's goal.

2. *Counterattack.* Given two consecutive events in a BPI, if the first occurs at $x_1 < 40$ and the second at $x_2 > 60$ within a time difference Δt, we report the speed as $v = \frac{x_2 - x_1}{\Delta t}$. Otherwise, we report 0. This metric quantifies how quickly a team transitions from a defensive to an offensive position.

3. *Maintenance.* For a given BPI, we compute the average x-position of all events. If $\bar{x} < 40$, i.e., if the events mostly take place in the team's defensive zone, we report the total possession time; otherwise, we report 0. This captures the extent to which a team maintains possession and builds play from its own half.

4. *Build up.* If $\bar{x} > 60$ within a BPI, indicating that the possession takes place primarily in the opponent's half, we report the total possession time. Otherwise, we report 0. This metric captures possession time in offensive contexts.

5. *Midfield play.* If $\bar{x}$ satisfies $40 \leq \bar{x} \leq 60$, indicating that possession occurs mainly in the middle third of the pitch, we report the total possession time; otherwise, we report 0. This metric quantifies how much time a team spends in the central area of the field.

6. *Flow rate.* In each BPI where $\bar{x} \geq 50$, we compute the time differences between all events and calculate their average, $\bar{dt}$. The metric is then defined as $1/\bar{dt}$. This provides a measure of how quickly the team moves the ball in the opponent's half.

7. *Crossing.* If a crossing event occurs within a BPI, we report 1; otherwise, 0. This metric counts the number of aerial attempts into the box.

8. *Pressure point.* From each BPI we take the first event and extract its x-position, representing where possession begins. This allows us to assess whether the team recovers the ball in its own half, the midfield, or the opponent's half.

9. *Pressure loss.* If a BPI begins with an event at $x > 40$, we report the total possession time of the opposing team from the immediately preceding BPI. This metric indicates whether the team is relaxing or increasing pressure in the middle and final thirds.

10. *Shots.* If a "Shot" event occurs within a BPI, we report 1; otherwise, 0. This metric counts the number of shots on goal.

It is important to note that for our analysis, we discarded all BPIs with fewer than 3 events and total duration less than 2 seconds. The goal was to retain only consolidated possessions and discard brief, transient recoveries. From this preprocessing, we obtained a total of 51833 BPIs.

After computing the metric values for each BPI, we analyzed their distributions. We observed that most metrics followed a log-normal-like distribution. Consequently, we applied a transformation $x \to \log(1 + x)$ to work with normal distributions. Once transformed, we grouped the data by match and by team, summing the values obtained for each metric. For instance, the *Maintenance* feature then represents the total maintenance time per match, and *Shots* counts the number of shots taken by the team during the match. This process yields a data matrix where each row corresponds to one team's performance in one match and contains all computed metrics. From now on, we refer to this matrix as DS. Note that in the 2017/2018 season, teams in the Spanish, English, French, and Italian leagues played 38 matches each. Therefore, using data from the top four teams in each of those leagues yields $38 \times 4 = 152$ samples per league. In the German league, where fewer teams play, there were 34 matches, resulting in 136 samples. Consequently, DS is a matrix with 744 rows and 10 columns. Finally, in a separate dataset, we collected metadata associated with each sample, including the team identity, the league, and the final standing in the league table, which will be used in subsequent analyses.

III. RESULTS

A. General characterization of the extracted features

In this section, we present an initial exploration of the data obtained during the collection process. Fig. 1 (a) shows the correlation between the extracted features, $Corr(i, j)$. The average correlation value is ~ 0.36 with a standard deviation of ~ 0.32. The maximum observed value is 0.8 and the minimum is 0.02. Next, and as an example, we discuss some of the cases with extreme correlation values. The highest observed correlation is found between the pair *Build up - Flow rate*, indicating that when the time of offensive build-up increases, the frequency between events also increases. Another strong correlation is seen in the pair *Direct play - Pressure point*, meaning that the average speed of the ball in the attacking direction increases when the ball is recovered closer to the opponent's goal. This is consistent with the previous observation: the game accelerates when the ball is near the goal line. For the same reason, we observe high correlations in the pairs *Flow rate - Pressure point* and *Build up - Pressure point*. We also find a strong correlation in the pair *Direct play - Midfield Play*, indicating that the speed in the attacking direction tends to increase when the build-up time in the middle third of the field increases. On the other hand, the feature *Maintenance* presents an interesting case, showing negative correlation in 6 out of its 9 possible pairs. This is because this feature describes a tactical development that is entirely different from the others. When a team increases its possession time in the first third of the field (an increase in *Maintenance*), it inevitably tends to decrease its possession times in the second and third thirds (a decrease in *Midfield play* and *Build up*), a reduction in the speed with which teams move the ball is observed (a decrease in *Flow rate*), and the opponent's possession time drops (a decrease in *Pressure loss*). We also observe strong correlations of the *Crossing* feature with *Direct play*, *Build up*, and *Pressure point*, indicating that the amount of aerial play increases with the ball circulation speed and with the proximity of ball recovery points to the opponent's goal.

We are now interested in analyzing whether the leagues exhibit significant differences in the feature values. To this end, we grouped the data by league and calculated the average value per feature, $\overline{F_{i,L}}$. In this notation, the index i corresponds to the feature and L to the league. In Fig. 1 (b) we show the results. The x-axis indicates the different features and the bars represent the average value observed for each league. The error bars indicate the standard error. Note that for each feature, the values are sorted in descending order; that is, the leagues with higher values appear first. Since for each feature the data are standardized using the global mean and standard deviation of the feature, grouping the data by different league subsets reveals deviations from the global mean. In general, we can say that the deviations are small; for instance, in the extreme case of the Italian league in the *Pressure* *loss* feature, the deviation does not exceed half the standard deviation of the whole set (which is equal to 1 due to standardization). Thus, the data are quite concentrated around the mean, though with some slight but significant deviations, as quantified by the error bars. It is worth highlighting the case of the Italian league, which shows above-average values in all features, or the case of the German league, which shows all values below the mean except for the *Maintenance* feature. We also observe that the Spanish league shows values very close to the global mean, although generally slightly below. Similarly, the English league exhibits marked variations both above and below the mean.

B. Detection of Playing Styles

The aim of this section is to use the extracted features to detect different playing styles. In this framework, a playing style associated with a team can be thought of as a subset of features. As shown in the previous section, the correlation between features is statistically significant. For this reason, we propose applying a dimensionality reduction method to the dataset in order to identify subsets of features that define the playing styles of teams. With this in mind, we performed a Principal Component Analysis (PCA) [18] on dataset DS.

We computed the covariance matrix, Cov, which is closely related to the correlation between each pair of features. Note that this matrix satisfies $Cov \in \mathbb{R}^{10\times 10}$. Next, we computed the eigenvalues and eigenvectors of Cov, $\lambda_i \in \mathbb{R}$ and $v_i = (w_1^i, ..., w_{10}^i) \in \mathbb{R}^{10}$, where $i = 1, ..., 10$ indexes both the features and the corresponding eigenvalues and eigenvectors. In this framework, each eigenvalue is proportional to the amount of variance that the corresponding eigenvector (principal component) represents in the data, which is referred to as explained variance. Note also that the values w_j^i, known as loadings, weigh the importance of each original feature in each principal component.

Fig. 2 (a) shows the explained variance per component obtained from the PCA. We can see that the first component accounts for over 50% of the variance. The second and third account for approximately $\sim 10\%$ each, and the remaining components are below the average. Fig. 2 (b) shows the cumulative explained variance as a function of the number of components. We observe that the first four principal components account for more than 80% of the total explained variance. In other words, more than 80% of the information is captured by these four components.

We then applied a Varimax rotation [19] to the set of principal components. This procedure redistributes the loadings within each principal component while preserving the explained variance, with the goal of obtaining rotated components that have high loadings in a few factors and low loadings in others. This facilitates the association between principal components and subsets of features, as well as the interpretation of the results. Fig. 3 presents a heatmap of the loadings for the first four rotated principal components S_1, S_2, S_3, and S_4. In S_1, we observe that the highest absolute loading is associated with the feature *Maintenance*, showing a value of -0.88. The next highest absolute loadings are for *Build up* and *Flow rate*, with values of 0.29 and 0.23, respectively. Below these, we find *Shots* and *Midfield play*, with values of -0.17 and 0.15. Note that in this component, the remaining features have loadings close to zero. Within this framework, we can define the subset of features with the highest loadings as a playing style. In the case of S_1, it represents a style focused on sustained ball possession across all three thirds of the pitch—that is, a team displaying this style aims for total control of the ball throughout the field. The style associated with S_2 is primarily related to the subset of features *Midfield play*, *Pressure point*, and *Direct play*, and to a lesser extent to *Counterattack*, *Flow rate*, and *Shots*. This defines a style characterized by extended possession in midfield, ball recovery near the opponent's goal, and rapid ball progression toward the rival goal. It also indicates a fast-paced game in the opponent's half and a strong generation of shots on goal. Note that this style differs significantly from that associated with component S_1. The style defined by S_2 prioritizes fast ball movement, high pressing, and shooting, whereas the style defined by S_1 prioritizes ball possession and control of the game. Regarding the style associated with S_3, we observe a strong relationship with the features *Pressure loss* and *Counterattack*, indicating that teams associated with this playing style tend to concede ball possession and rely on counterattacks, trying to exploit vulnerabilities that the opponent could expose while attacking. This style may reflect the approach a team would take when facing a clearly superior opponent. Finally, the style associated with S_4 shows a strong presence of the features *Shots*, *Crossing*, and *Build up*, suggesting that teams playing in this style prioritize ball possession near the opponent's goal, delivering crosses, and shooting.

As we can see, the dimensionality reduction technique allows us to adopt a data-driven approach to define playing styles. In the next section, we use this information as a basis to determine the playing styles associated with each of the five leagues analyzed.

C. Which styles best represent each league?

The aim of this section is to analyze whether it is possible to associate specific playing styles with each of the five European leagues. To this end, we first project each row in DS onto the basis defined by the principal components S_1, S_2, S_3, and S_4, which we will henceforth refer to as *styles*. This procedure allows us to rewrite DS into a new dataset consisting of 744 rows and 4 columns, which we denote as $\tilde{DS}$. In this new dataset, each row still represents a match played by a team in a given league, but each column now quantifies the extent to which that team exhibited each style in a particular match.

For each style in this new dataset, we first performed a one-way Anova analysis [20]. This allows us to determine whether there are statistically significant differences between the average values of the five leagues. Our goal is to understand whether the categories (leagues) are well separated in the space of the rotated principal components (styles). In Table I, we report the value of the test statistic and the p-value obtained from the analysis. We observe that in all cases, the p-value is less than 10^{-6}, which leads us to reject the null hypothesis that all classes are equal, confirming the presence of statistically significant differences in the subsets of data associated with each league.

This result serves as initial evidence that the different leagues may be classifiable based on the information provided by the quantification of styles. To test this hypothesis, we employed a Random Forest classifier. Our primary objective with this tool is to classify each row of $\tilde{DS}$ according to the league to which the corresponding match belongs. We aim to determine whether the styles contain enough information to distinguish between leagues. Since this classifier includes random components in its algorithm, to obtain statistically robust results, we ran the algorithm 100 times using different random seeds. In Fig. 4 (a), we display the confusion matrix computed from the classification results. This matrix shows the percentage of times the data were classified correctly and incorrectly. We observe that the classifier performs very well in all cases, exhibiting a low error rate. To quantify the classifier's performance, we used the confusion matrix to compute the *F1-Score* metric [21]. In this metric, values close to 1 indicate strong performance. The results are presented in Table II. We can see that the classifier performs very well in distinguishing each of the five leagues analyzed.

So far, we have established that styles, as represented in our framework, enable the classification of leagues. We now aim to extract information linking specific styles to specific leagues—that is, we seek to identify which styles best represent each league. To this end, we employed a method that allows us to weight the contribution of each style in the classification process. Given that we are working with a non-linear classifier, an appealing option in this context is the use of *Shapley Additive Explanations* (SHAP) [22], which assigns to each input style a value representing its contribution to a specific prediction, thereby ensuring interpretability. This approach, grounded in game theory, aims to fairly distribute the contribution of each player to a collective outcome by considering all possible coalitions that can be formed. In the context of predictive models, such as ours, each feature is treated as a player, and the Shapley value assigns an average importance to each feature by computing its marginal contribution across all possible permutations. This procedure guarantees desirable properties such as efficiency, symmetry, null player, and additivity, allowing for a consistent interpretation of the importance of each feature in the predictions generated by a complex model.

Formally, the Shapley value for a feature i in a predictive model is defined as,

$$\phi_i = \sum_{M \subseteq N \setminus \{i\}} \frac{|M|!(|N| - |M| - 1)!}{|N|!} \left[f(M \cup \{i\}) - f(M) \right], \tag{1}$$

where N denotes the set of all features, M is a subset of N not including i, the bars $|\cdot|$ indicate the number of elements, and $f(M)$ is the model prediction considering only the features in M. The term $\frac{|M|!(|N|-|M|-1)!}{|N|!}$ corresponds to the weight assigned to each possible coalition, ensuring a weighted average over all possible permutations of the features. This computation yields a fair measure of the marginal contribution of each feature to the model, in accordance with the fundamental properties of the Shapley value. Moreover, the predicted probability for class c, p_c, can be decomposed as the sum of a base value ϕ_0 and the Shapley values associated with each of the n_c features,

$$p_c = \phi_{c,0} + \sum_{i=1}^{n_c} \phi_{c,i}. \tag{2}$$

Note that $\phi_{c,0}$ is the expected value of the prediction when no features are present.

Thus, this method allows the classification probability to be interpreted as the sum of the individual contributions of each feature, enabling a clear interpretation. In our case, for each of the 100 runs of the Random Forest algorithm, we computed the Shapley values associated with the probability of predicting the league (class) to which each match (sample) belongs. It is important to note that each match is associated with four different Shapley values, each corresponding to the weight assigned to one of the styles.

In Fig. 4 (b), we present a histogram of the Shapley values obtained across the 100 classification instances. We observe a bell-shaped, slightly asymmetric distribution skewed towards positive values, with a mean of 0.083 and a standard deviation of 0.057. Regarding the base values, $\phi_{c,0}$, unlike the $\phi_{c,i}$, these are computed globally. From each classification run, we obtained 4 values of $\phi_{c,0}$, one for each class. Averaging over the instances, we observe that for all classes, $\phi_{c,0} \sim 0.2$, which corresponds to the probability of assigning each match randomly to one of the 5 leagues.

On the other hand, using the values $\phi_{c,0}$ and $\phi_{c,i}$, we calculated the probability p_c, as given by Eq. 2, for all matches in all classification instances. In Fig. 4 (c), we show a histogram of the obtained values. The distribution appears roughly symmetric and bell-shaped, with a mean of 0.53 and a standard deviation of 0.12. This indicates that, on average, the algorithm assigns the correct class a probability greater than 50% out of the 5 possible classes, which constitutes a strong result.

We are now interested in using the information provided by the Shapley values to analyze which styles best represent each league. To do so, we first separate the data by league, then for each sample we identify the style corresponding to the maximum Shapley value. Note that the maximum Shapley value for a sample indicates which style contributed the most to the classification of that sample's league. Therefore, we can assume that this value is characteristic of that league. Using this procedure, we generate data subsets $\{S_{i,L} \mid \phi = \phi_M\}$, where for example the subset $S_{1,EN}$ contains all values of style S_1 observed in the English league when the Shapley value is maximal for that sample. In Fig. 4 (d), we plot the mean value of these subsets for each league. The error bars indicate the standard error. Since the style values have been standardized with respect to the mean and standard deviation, the means reflect the deviations from the global average for each subset. At first glance, we can observe that most of these deviations are on the order of 1 standard deviation, which suggests statistical significance. For style S_1, we observe that the English and Spanish leagues are above the mean, while the other leagues are considerably below. This indicates that high values of this style are heavily weighted in the classification process for the English and Spanish leagues and thus can be considered characteristic of these. Conversely, it can be suggested that this style is not characteristic of the other leagues, at least in comparison with the first two. Regarding style S_2, the French league, and especially the Italian league, show values above the mean. This suggests a strong presence of style S_2 in the Italian league. The remaining leagues show values below the mean, with the German league being the least aligned with this style. For style S_3, values above the mean are observed only for the Italian league, indicating a strong presence of this style in that league. The other leagues show low values, particularly the French and Spanish leagues, which appear to be less representative of this style. Finally, for style S_4, we observe that the French and German leagues exhibit high values, making them the most representative of this style, whereas the Italian league appears to express it the least.

In terms of sports tactics, from the obtained data we can observe that the English league prioritizes a style based on long ball possession times across all areas of the field. Similarly, this league shows less affinity for a style that relinquishes ball possession and focuses on counterattacks. The Spanish league also tends to exhibit a style based on long ball possession times in the lower, middle, and upper areas of the field. We observe that this league shows less affinity for a style based on high pressing and fast, vertical movements. On the other hand, teams in the Italian league tend to play with a style strongly based on and performing fast and vertical movements. It also shows a tendency to cede possession to the opponent, trying to exploit vulnerabilities through counterattacks, being a major creator of shots on goal. It shows less affinity for long possession times.

In the case of the French league, it seems to express a style strongly based on prioritizing ball control near the opponent's goal to make crosses and generate shots on goal. It also seems to have a tendency for high pressing and fast, vertical movements. Like the Italian league, it appears less inclined towards a style based on long possessions.

Finally, the German league, like the French league, tends to prioritize a style based on building play near the opponent's area to send crosses and generate shots on goal. It also seems less inclined towards a game based on long possessions and total dominance. Unlike the French league, the German league does not appear to favor a style based on fast and vertical movements.

D. Differentiating Styles Between Leagues

Fig. 4 (d) shows us that a comparative analysis of the values $\overline{S_{i,L}}$ is useful to describe general differences in the playing styles exhibited by the leagues. At first glance, some interesting patterns can also be observed, such as the similarity between the English and Spanish leagues, which show values above the mean in style S_1 and values below the mean in the other styles. Note that in this case it is not possible to clearly determine if there is a significant difference beyond the small differences between these values. Thus, to better understand the differences between style values in each league, in Fig. 5 we show the distributions $P(S_{i,L}|\phi = \phi_M)$. For better visualization, the curves are plotted as a smoothed approximation using the Kernel Density Estimation (KDE) method. Note that in each panel of the figure we compare the distributions of all leagues with respect to a particular style.

In all styles, we can observe the presence of bimodal, sharp, flattened distributions, or with smaller secondary peaks. From the perspective offered by these distributions, differences and similarities between leagues can be better observed. For example, in style S_1, we had seen in Fig. 4 (d) that the German, French, and Italian leagues showed similar mean values below the mean, however, from the perspective of Fig. 5 (a) we can observe notable differences; for instance, we see the appearance of a small peak above the mean for the German and French leagues but not for the Italian. In the curves associated with style S_2 shown in Fig. 5 (b), we can see, for example, a bimodality in the French league, with a large peak centered slightly below the mean and another centered above the mean. In Fig. 5 (c) we show the curves associated with style S_3. For this style, from Fig. 4 (d) it follows that the English and French leagues behave similarly; here we can observe that although the mean values are similar, the curve associated with the English league is sharper, while the curve associated with the French league is more dispersed. Also, in Fig. 5 (d) we see significant differences in the curves of the French and German leagues, despite showing similar mean values (see Fig. 4 (d)).

Of all possible cases, the apparent similarity between the Spanish and English leagues shown in Fig. 4 (d) is what caught our attention the most. In the following lines, we use the information presented in Fig. 5 to try to elucidate differences. For the case of style S_1, in Fig. 5 (a) we see that the distributions associated with these leagues have a similar mean value, but they also exhibit a non-trivial bimodal behavior. Focusing on those curves, we can observe that the highest peak of the distribution associated with the Spanish league is at $S_{1,SP} \approx 1$, and around that region, the distribution associated with the English league shows a valley, which marks a clear difference between the leagues.

A naturally arising question is why a bimodal behavior appears in some distributions. In the case of the English league, in the curve associated with S_1, we see that the peaks are almost of the same height. The total curve seems composed of two bell shapes, one around $S_{1,EN} \approx 0$ and another around $S_{1,EN} \approx 2$. One hypothesis is that the style S_1 in the English league may appear in different contexts, that is, combined differently with the other styles. To investigate this idea further, we take the data from this league and divide it into two sets using a threshold value, $s_0 = 1$. Thus, one set contains all samples where $S_{1,EN} < s_0$ and the other set contains samples where $S_{1,EN} > s_0$. In this framework, in Fig. 6 (a) we show for the English league the distributions $P(S_{i,EN}|\phi = \phi_M)$; to visualize the distributions of all styles in the league in a single graph, in Fig. 6 (b) we show the distributions $P(S_{i,EN}|\phi = \phi_M, S_{i,EN} < s_0)$, i.e., the distributions that the rest of the styles acquire under the condition that $S_{i,EN}$ is below the threshold, and in Fig. 6 (c) we show the distributions $P(S_{i,EN}|\phi = \phi_M, S_{i,EN} > s_0)$, i.e., the distributions associated with the different styles conditioned on the values of $S_{i,EN}$ being greater than the threshold. Some interesting differences are observed in the distributions associated with one subset compared to the distributions associated with the other; we describe them in the following. We can observe that when $S_{1,EN} < s_0$, the distribution associated with $S_{2,EN}$ shifts slightly toward negative values, while the others remain around zero. On the other hand, when $S_{1,EN} > s_0$, we see that $S_{2,EN}$ shifts slightly toward positive values, $S_{3,EN}$ stays centered around the mean, and $S_{4,EN}$ moves toward the negatives. To perform a tactical analysis of these observations, recall that, broadly speaking, style S_1 was associated with a form of play that prioritizes controlling ball possession times across the entire field, style S_2 prioritizes fast ball movement, style S_3 defensive play centered on counterattacks, and style S_4 is associated with crossing and shooting. In this framework, we can say that when the team centers its style on total control around the global mean, the style associated with fast ball movements decreases slightly. On the other hand, when the team centers its style on total control above the mean, the style associated with fast movement increases and the style associated with crosses and shots decreases strongly. In Fig. 6 (d), (e), and (f) we perform the same analysis to study the bimodality observed in $S_{1,SP}$ in the Spanish league. In this case, we take $s_0 = 0.5$. For the subset associated with $S_{1,SP} < s_0$, we observe that the distributions do not change significantly and remain centered around the mean. For the case $S_{1,SP} > s_0$, we also observe distributions centered around the global mean but with a slight increase in the standard deviation. This marks a significant difference compared to what is observed in the English league. In the Spanish league, the bimodality does not seem related to how the other styles combine, as we see it does in the English league. This suggests that the main difference between the English and Spanish leagues seems to lie in how they combine the different playing styles. Spanish playing style seems to be intrinsically characterized by S_1 whereas in the English case, their style is a mix of two variants of play.

Interestingly, our data-driven analysis aligns, to some extent, with expert opinions [13]. Italian football is renowned for its tactical approach, where both defensive and offensive duties are executed with remarkable precision. This disciplined style, marked by rigorous defense and swift transitions to counter-attacks, is commonly known as "Catenaccio." This observation is reflected in the prominence of styles S_2 and S_3 identified in our analysis. In contrast, English football is characterized by its dynamic and direct play, often described as "Kick and Rush". The English game places less emphasis on tactical and strategic elements, instead prioritizing physicality and straightforward action. These attributes correspond to the components of S_1 in our findings. Spanish football, meanwhile, occupies a middle ground between the Italian and English styles. It is less open and direct than the English game, featuring more structure and tactical sophistication. Compared to the Italian approach, Spanish football is more aesthetically pleasing and emphasizes ball possession heavily. While it shares some tactical organization with Italy, Spanish football is closer to the English style in terms of fluidity and technical skill. The ball is constantly circulated in all areas of the pitch, and this style is well represented by the S_1 components in our analysis.

IV. CONCLUSION

The main goal of this study was to propose a data-driven approach to define styles of play in football and to compare the playing styles of the most prominent European football leagues. To this end, we used a publicly available dataset containing event data for every match played during the 2017/2018 season of the English, French, German, Italian, and Spanish football leagues. In order to obtain a representative sample of European football, we selected from each league the teams that finished in the top four positions of the final standings—those that qualified for the UEFA Champions League. Therefore, these teams can be considered among the best in Europe for that season. For each of the selected teams and for each match, we extracted various standard metrics associated with possession intervals. This allowed us to build a data matrix in which each row corresponds to a match played by a team of a given league, and each column represents the relevance of a specific metric during that match. The columns of this matrix, or features, quantify certain relevant tactical aspects displayed by the teams throughout the games. By conducting a correlation analysis between the different features, we observed that they do not behave independently, but rather form highly correlated groups. This is expected, as in any sport, the increase or decrease in certain tactical aspects often requires the simultaneous adjustment of others. Based on these groups of correlated features, we proposed performing a dimensionality reduction of the system. For this purpose, we applied a Principal Component Analysis (PCA), through which we found that four components capture more than 80% of the explained variance. Additionally, a Varimax rotation was applied to the principal components in order to redistribute the loadings, encouraging high loadings on only a few factors and low loadings on the others. This facilitates the interpretation of the results. In doing so, we observed that the rotated components are combinations of features that represent distinct and complementary tactical aspects. This allowed us to define four different styles of play, each based on subsets of the original metrics, which emerge naturally from the data analysis.

Having observed the emergence of the styles, we asked whether the different European leagues are associated with distinct styles of play. To address this, we first transformed the original data matrix using the style basis, resulting in a new matrix in which each row represents a match played by a team from a specific league, and each column reflects the expression of a particular style in that match. Following this transformation, we performed a one-way Anova test to determine whether each league exhibited style values that could be statistically distinguished from those of the other leagues. The test yielded significant results, leading us to conclude that the leagues could indeed be classified based on the quantification of styles observed in each match. Having established this, we proceeded to implement a non-linear classifier of the random forest type to classify the leagues using the style values. This approach produced satisfactory results. We were also interested in analyzing stylistic differences in the five leagues studied. To this end, during the classification process, we employed a game-theoretic algorithm that quantifies the relative importance of each style in determining the classification of each sample. This analysis enabled us to establish a clear association between leagues and styles, thereby identifying statistically significant differences in the emergent styles across the various leagues.

Compared with previous studies that analyze football playing styles, our work proposes a quantitative, automated, and data-driven methodology to empirically identify emerging styles based on objective metrics derived from possession events. This approach enables a non-arbitrary assignment of styles, advancing from a theoretical perspective toward a concrete and quantifiable operationalization. Moreover, the developed methodology is reproducible and allows for statistical testing of the prevalence of these styles in different competitive contexts. Additionally, our analysis encompasses the five major European leagues, providing a broader and more robust comparative perspective relative to other studies in the literature. As a future line of research, it would be valuable to enrich the present study by incorporating additional contextual information. In particular, the proposal by González et al. [15] to include variables such as home advantage and the competitive level of the opponent would allow for an assessment of whether teams adjust their playing styles when playing at home versus away, as well as detecting tactical variations according to the opponent's strength. Furthermore, it is relevant to explore the relationship between the prevalence of specific emergent styles and team performance. For instance, in [12], we documented links between performance and certain structures in the passing network; in this regard, a natural extension of our analysis would be to determine whether such structures correspond to specific playing styles. A study in this direction could shed light on how emergent styles influence sporting outcomes and support the development of more precise tactical recommendations.

ACKNOWLEDGEMENT

This work was partially supported by CONICET under Grant No. PIP 112 20200 101100 and SeCyT-UNC (Argentina).

[1] Johann H Martínez, David Garrido, José L Herrera-Diestra, Javier Busquets, Ricardo Sevilla-Escoboza, and Javier M Buldú. Spatial and temporal entropies in the Spanish football league: A network science perspective. Entropy, 22(2):172, 2020.

[2] Haroldo V Ribeiro, Satyam Mukherjee, and Xiao Han T Zeng. Anomalous diffusion and long-range correlations in the score evolution of the game of cricket. Physical Review E, 86(2):022102, 2012.

[3] A Clauset, M Kogan, and S Redner. Safe leads and lead changes in competitive team sports. Physical Review E, 91(6):062815, 2015.

[4] Ken Yamamoto, Seiya Uezu, Keiichiro Kagawa, Yoshihiro Yamazaki, and Takuma Narizuka. Theory and data analysis of player and team ball possession time in football. Physical Review E, 109(1):014305, 2024.

[5] A Chacoma, Nahuel Almeira, Juan Ignacio Perotti, and Orlando Vito Billoni. Modeling ball possession dynamics in the game of football. Physical Review E, 102(4):042120, 2020.

[6] A Chacoma, OV Billoni, and MN Kuperman. Complexity emerges in measures of the marking dynamics in football games. Physical Review E, 106(4):044308, 2022.

[7] Andrés Chacoma and Orlando V Billoni. Simple mechanism rules the dynamics of volleyball. Journal of Physics: Complexity, 3(3):035006, 2022.

[8] Andrés Chacoma and Orlando V Billoni. Probabilistic model for padel games dynamics. Chaos, Solitons & Fractals, 174:113784, 2023.

[9] Ming-Xia Li, Li-Gong Xu, and Wei-Xing Zhou. Motif analysis and passing behavior in football passing networks. Chaos, Solitons & Fractals, 190:115750, 2025.

[10] A Chacoma, N Almeira, JI Perotti, and OV Billoni. Stochastic model for football's collective dynamics. Physical Review E, 104(2):024110, 2021.

[11] Andrés Chacoma and Orlando V Billoni. Emergent complexity in the decision-making process of chess players. Scientific Reports, 15(1):23234, 2025.

[12] Andrés Chacoma. Identification and optimization of high-performance passing networks in football. Physical Review E, 111(4):044313, 2025.

[13] Hugo Sarmento, Antonino Pereira, Nuno Matos, Jorge Campaniço, T Maria Anguera, and José Leitão. English premier league, spaińs la liga and italỳs seriés a – what's different? International Journal of Performance Analysis in Sport, 13(3):773–789, 2013.

[14] Javier Fernandez-Navarro, Luis Fradua, Asier Zubillaga, and Allistair P McRobert. Influence of contextual variables on styles of play in soccer. International Journal of Performance Analysis in Sport, 18(3):423–436, 2018.

[15] Joaquín González-Rodenas, Jordi Ferrandis, Víctor Moreno-Pérez, Roberto López-Del Campo, Ricardo Resta, and Juan Del Coso. Differences in playing style and technical performance according to the team ranking in the spanish football laliga. a thirteen seasons study. Plos one, 18(10):e0293095, 2023.

[16] Claudio A Casal, José L Losada, Daniel Barreira, and Rubén Maneiro. Multivariate exploratory comparative analysis of laliga teams: Principal component analysis. International Journal of Environmental Research and Public Health, 18(6):3176, 2021.

[17] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. Scientific data, 6(1):236, 2019.

[18] Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. Nature Reviews Methods Primers, 2(1):100, 2022.

[19] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23(3):187–200, 1958.

[20] Markus Janczyk and Roland Pfister. One-way analysis of variance (anova). In Understanding inferential statistics: From A for significance test to Z for confidence interval, pages 97–125. Springer, 2023.

[21] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. A review of evaluation metrics in machine learning algorithms. In Computer science on-line conference, pages 15–25. Springer, 2023.

[22] Meng Li, Hengyang Sun, Yanjun Huang, and Hong Chen. Shapley value: from cooperative game to explainable artificial intelligence. Autonomous Intelligent Systems, 4(1):2, 2024.

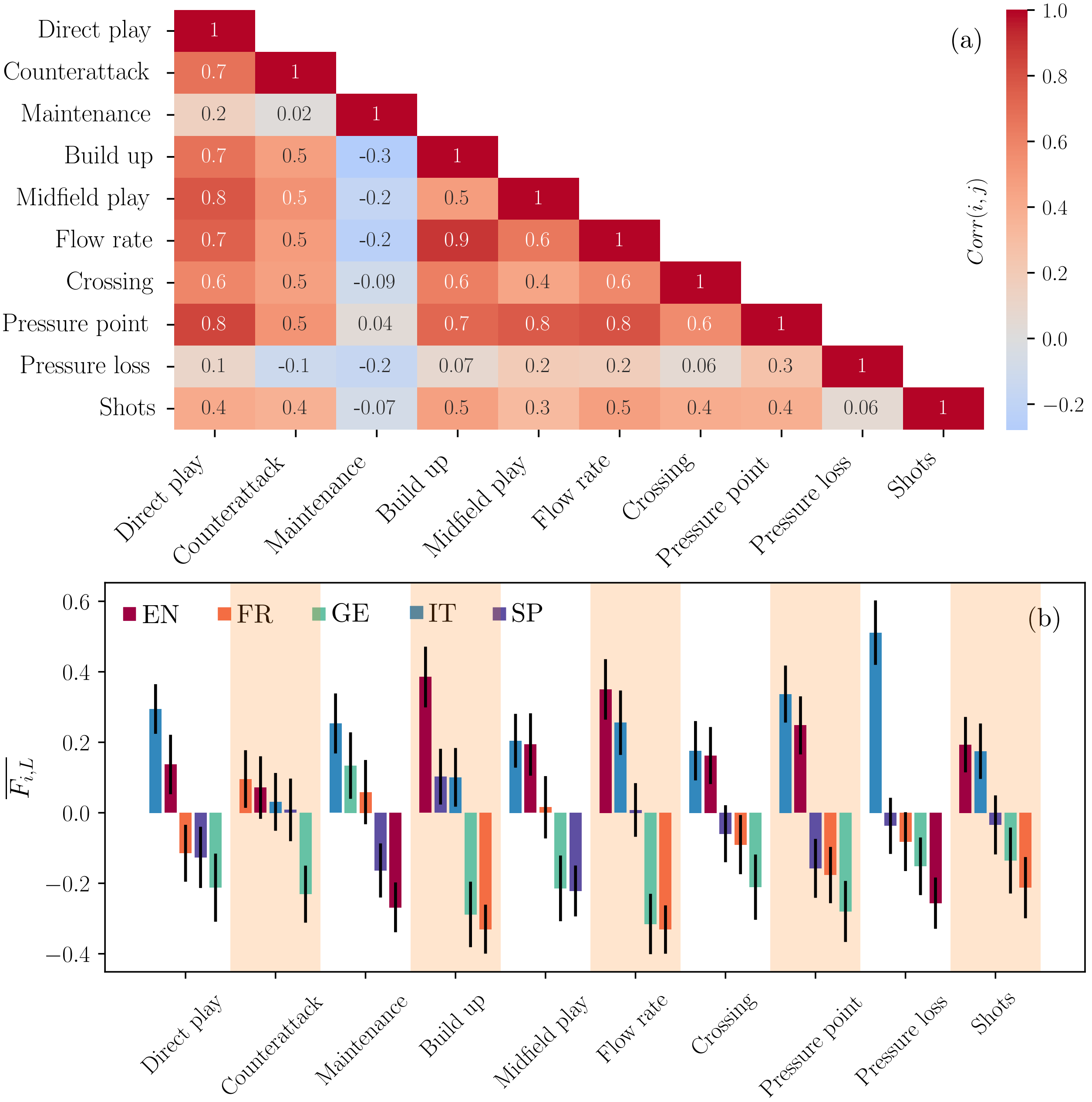

FIG. 1. Initial characterization of dataset DS. (a) Correlation among the different features in the dataset. (b) Mean value of each feature, separated by league and ordered from highest to lowest.

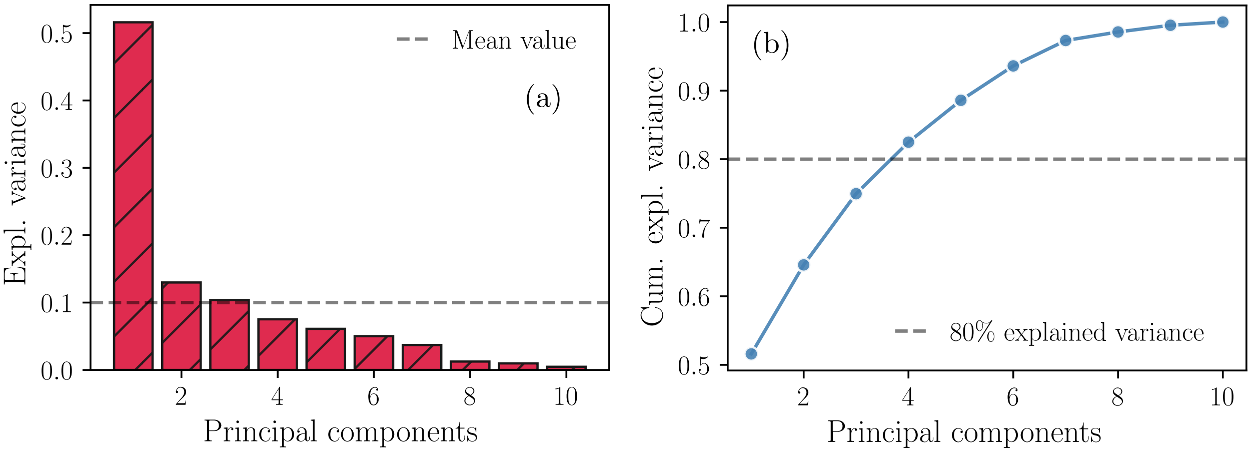

FIG. 2. PCA analysis. (a) Variance explained by each principal component. (b) Cumulative explained variance as the number of principal components increases.

Style	F-Statistic	p-value
S_1	9.63	1.3×10^{-7}
S_2	9.67	1.2×10^{-7}
S_3	8.66	7.8×10^{-7}
S_4	10.7	0.2×10^{-7}

TABLE I. F-statistic and p-value associated with the ANOVA analysis for the four styles.

League	F1-Score
EN	0.96
FR	0.99
GE	0.98
IT	0.96
SP	0.98

TABLE II. F1-score values for each league obtained from the Random Forest classification process.

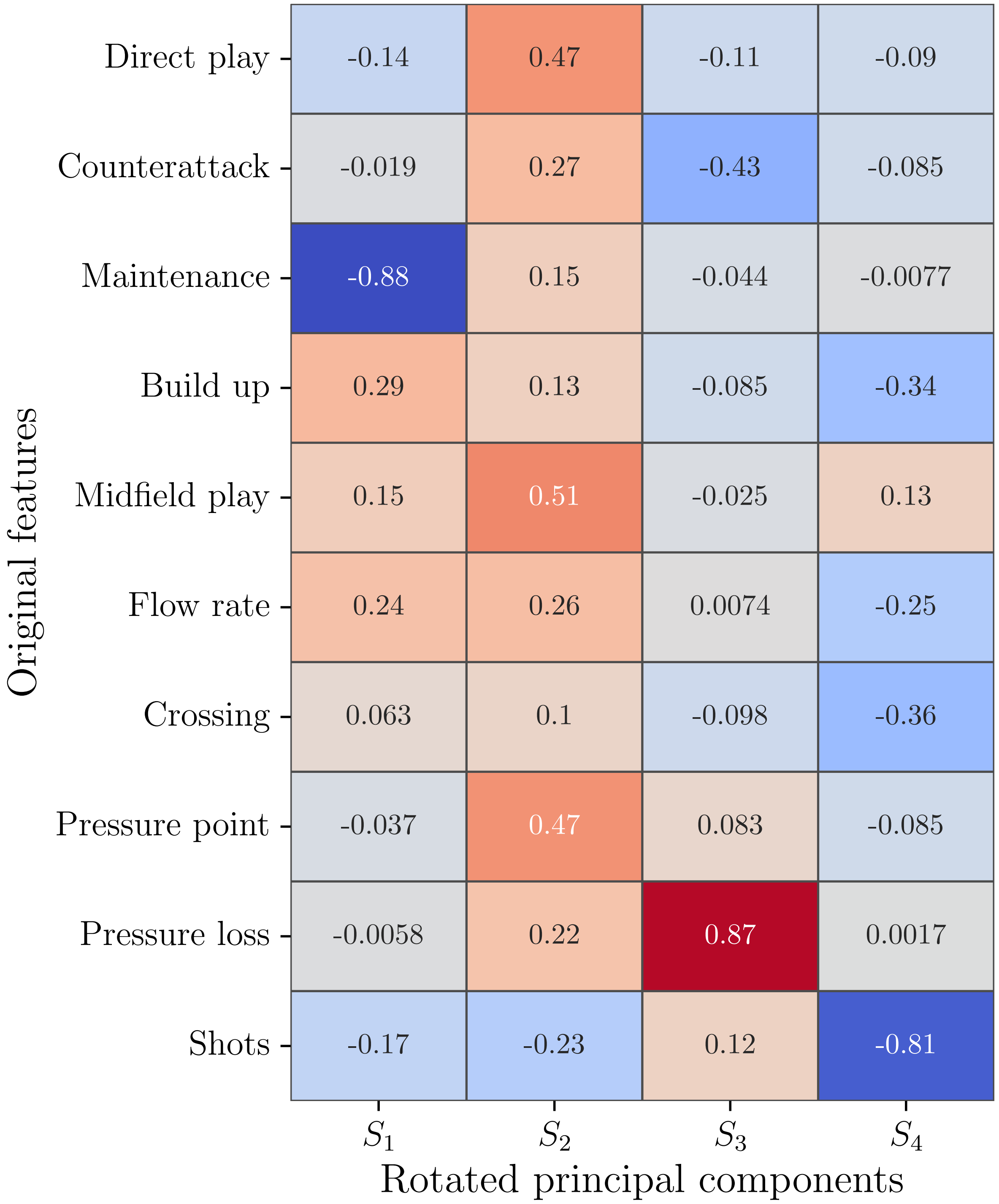

FIG. 3. Varimax rotation. Loading values of each feature associated with each of the rotated components (styles).

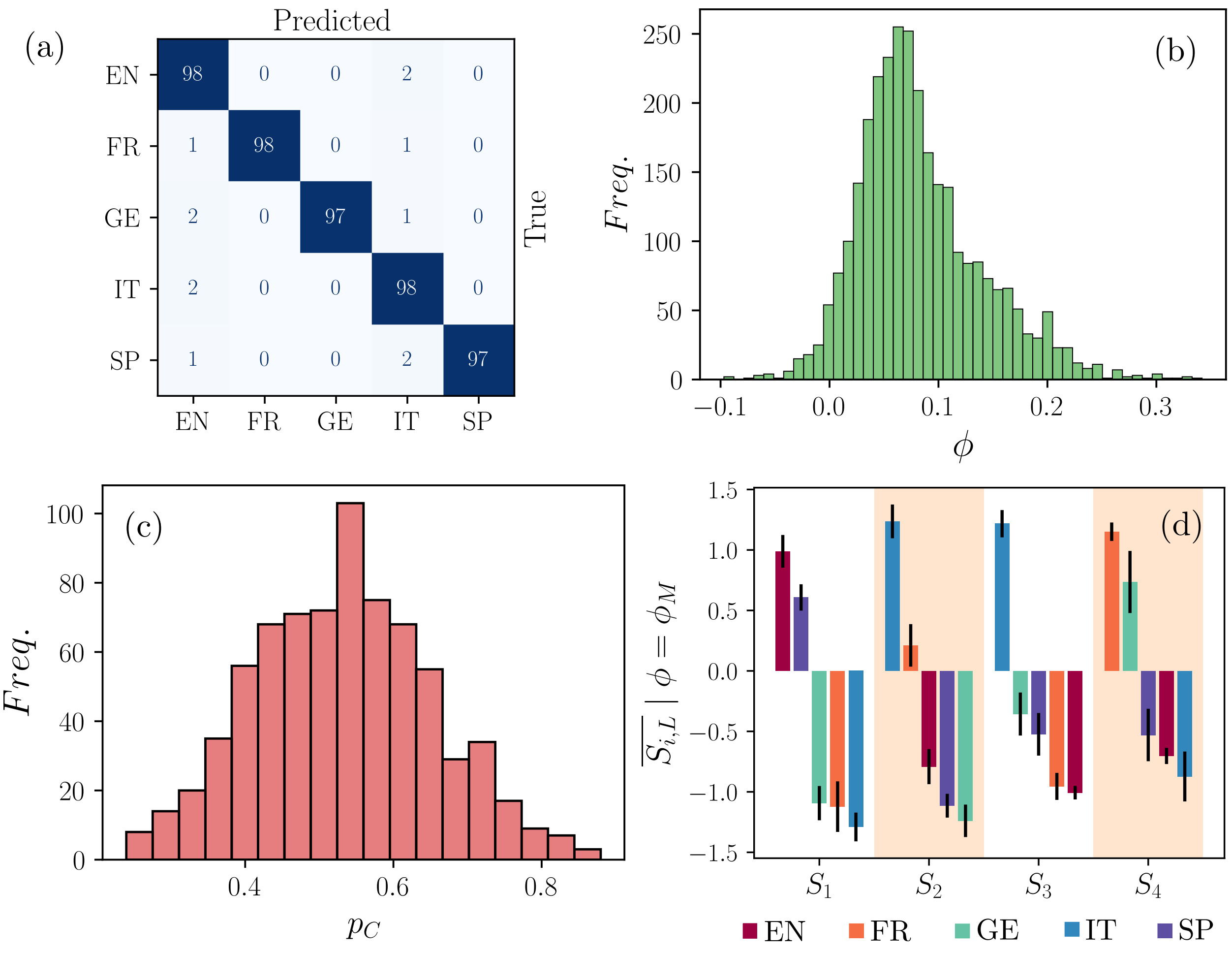

FIG. 4. Results of the classification using Random Forest and associated Shapley values. (a) Confusion matrix for the classification instances. Values are shown as percentages. (b) Histogram of the Shapley values obtained across the classification instances. (c) Histogram of the probability values p_c obtained in each classification instance. (d) Mean value of the style quantifiers associated with each league when the Shapley value is maximal in the sample.

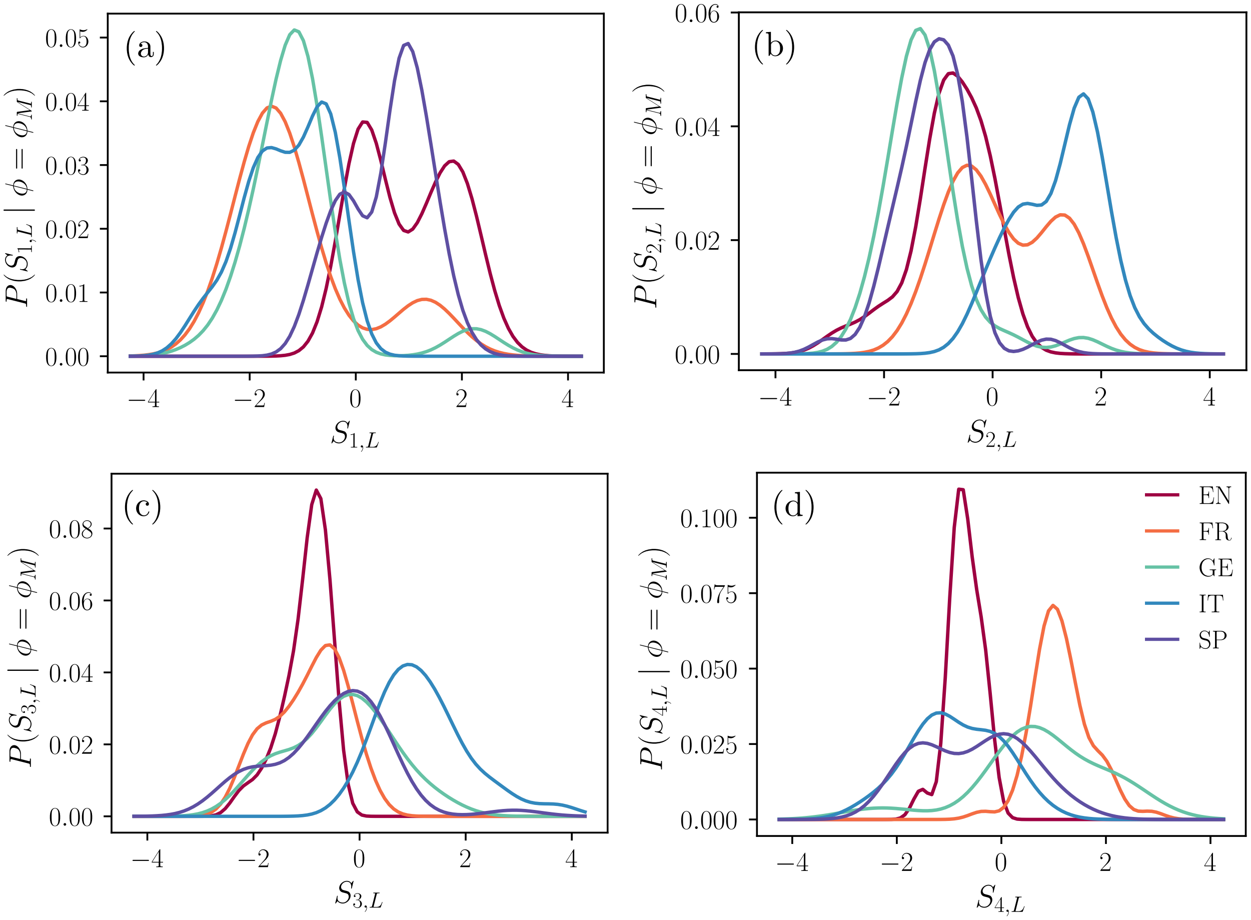

FIG. 5. Distribution of the values of the different style quantifiers, separated by league. (a) Style S_1. (b) Style S_2. (c) Style S_3. (d) Style S_4.

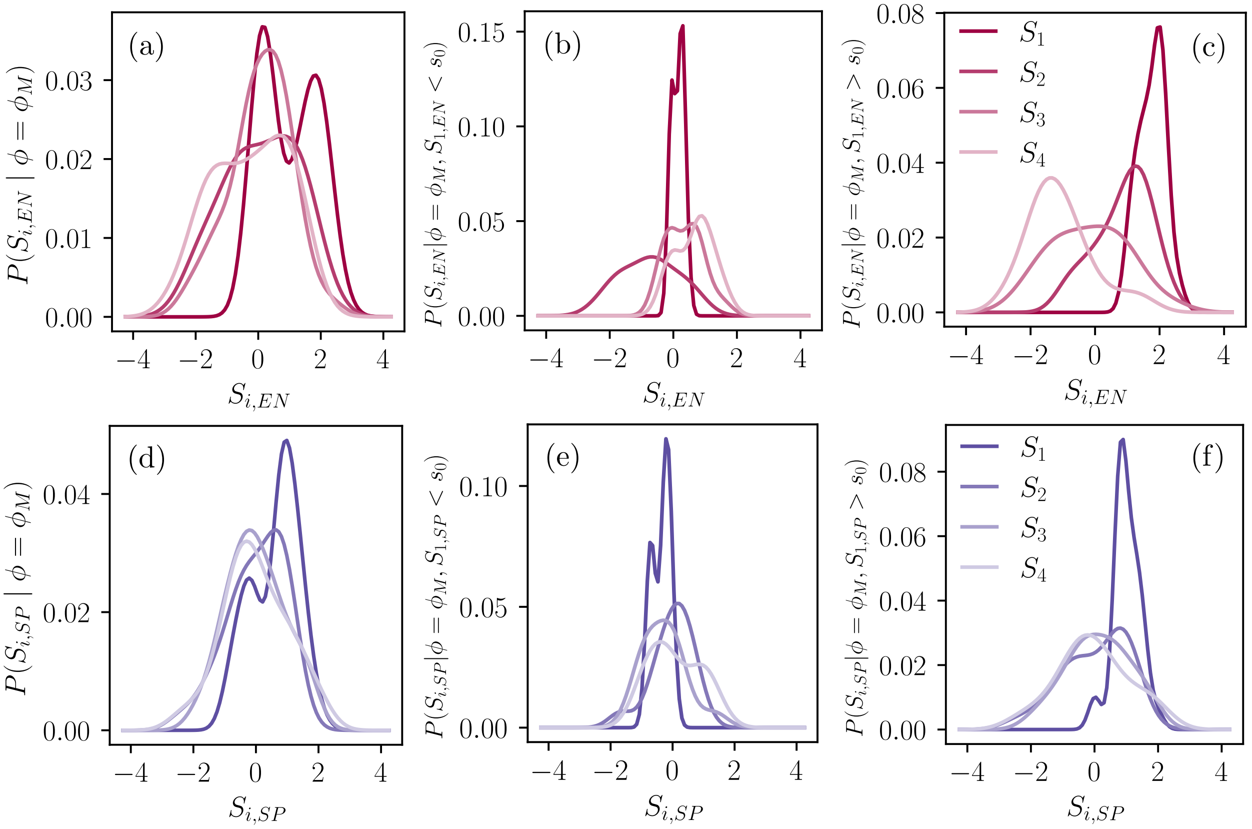

FIG. 6. Study of the bimodality in the style distributions associated with the English and Spanish leagues. (a) Distribution of the quantifier values for the four styles in the English league. (b) Distribution of the quantifier values for the four styles when the values of S_1 are below a threshold s_0. (c) Distribution of the quantifier values for the four styles when the values of S_1 are above the threshold s_0. Panels (d), (e), and (f) are analogous to the first three panels but correspond to the Spanish league.