©Journal of Sports Science and Medicine (2025) **24,** 764-778
http://www.jssm.org DOI: https://doi.org/10.52082/jssm.2025.764

Research article

Classifying Soccer Players Based on Physical Capacities and Match-Specific Running Performance Using Machine Learning

Michel de Haan [1], Stephan van der Zwaard [1,2], Jurrit Sanders [3], Peter J. Beek [1] and Richard T. Jaspers [1]✉

[1] Department of Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, Amsterdam, Netherlands; [2] Department of Cardiology, Amsterdam University Medical Center, location AMC, University of Amsterdam, Amsterdam, Netherlands; [3] PSV Eindhoven, Eindhoven, Netherlands

Abstract

Sprint and endurance capacities seem to be mutually exclusive or at least at odds with each other. However, this relationship has not been investigated in soccer, which appeals to both well-developed sprint and endurance capacities. This study explores the potential of machine learning to identify soccer players based on their unique combinations of sprint and endurance capacities and sprint and endurance match-specific running performance. In this context, the relationships between sprint and endurance capacities and between physical capacities and match-specific running performance are examined in detail. Match-specific running data were collected from 31 young elite male soccer players over two consecutive seasons. Additionally, these participants underwent exercise testing, consisting of a 20-meter sprint test and an incremental treadmill test to measure maximal oxygen uptake ($\dot{V}O_{2max}$). Subgroups were identified using *k*-means clustering and subgroup discovery, based on players' sprint and endurance capacities, sprint and endurance match-specific running performance, and playing position. Three distinct subgroups were identified using machine learning: players with high sprint capacity and sprinted meters ($n = 4$), players with high endurance capacity and meters ran at moderate and high intensities ($n = 6$), and players without high physical capacities or matching match-specific running performance ($n = 14$). Across all players, there was no significant relationship between 20-meter sprint speed and normalized $\dot{V}O_{2max}$ ($R^2 = 0.085$, $P = 0.17$), although 20-meter sprint speed was positively related to average match sprint distance ($R^2 = 0.168$, $P = 0.03$) and normalized $\dot{V}O_{2max}$ to average match distance at moderate and high intensities ($R^2 = 0.151$, $P = 0.04$). In young elite soccer players, sprint and endurance capacities show positive, moderate, relationships with corresponding match-specific running performance, but those capacities do not appear to be mutually exclusive or opposing. Clustering allows for identification of players who may benefit from alternative strategic roles during matches, are at risk of overuse, or could benefit from individualized training. This method can assist coaches in designing tailored training programs and optimizing overall match strategy.

Key words: Clustering, football, sprint speed, $\dot{V}O_{2max}$, MAS, MSS.

Introduction

Sprint and endurance capacity are key determinants of performance in many sports (van der Zwaard et al., 2018a). However, combining these two traits is complicated due to the interference effect that occurs with concurrent strength and endurance training. Incorporating endurance exercise can inhibit strength-related adaptations, such as muscle hypertrophy, strength, and power (Huiberts et al., 2024; Lundberg et al., 2022; Schumann et al., 2022). Conversely, integrating strength training can diminish endurance-related adaptations, like maximal oxygen uptake ($\dot{V}O_{2max}$), in certain conditions (Huiberts et al., 2024). These reciprocal effects make it difficult to optimize both traits simultaneously, indicating that sprint and endurance capacities might be mutually exclusive or at least at odds with each other. This is further underscored by the inverse relationship between sprint and endurance capacity that was found at both the muscle fiber and whole-body level in humans and other species (van der Zwaard et al., 2018a; van Wessel et al., 2010). This relationship has been well demonstrated in cyclic sports like rowing and cycling (van der Zwaard et al., 2018a; 2018b). The presence of such an inverse relationship suggests that optimizing these two physical traits simultaneously poses a challenge in general and particularly in elite athletes seeking to maximize both traits. Therefore, it is crucial to find the optimal balance between sprint and endurance capacities based on the specific demands of the sport or sport discipline of interest.

A team sport like soccer differs significantly from cyclic sports such as rowing and cycling, but also demands well-developed sprint and endurance capacities. Soccer is a dynamic sport, which requires a high endurance capacity, as elite (i.e. professional) soccer players typically cover 10 to 12 kilometers during a 90-minute game (Stølen et al., 2005). In the context of this endurance performance, the players perform a wide array of explosive activities and actions involving changes in pace and direction, such as sprinting, jumping, kicking, tackling, and turning (Stølen et al., 2005). Studies have estimated that the average work rate during a match is approximately 70% of $\dot{V}O_{2max}$, highlighting the importance of an efficient aerobic energy system to support sustained performance. This demands not only a high capacity for oxygen uptake and delivery to the muscles but also effective muscular oxygen utilization. However, match play also places significant demands on the anaerobic energy system, as evidenced by average lactate concentrations of 3 - 9 mM, with peaks often exceeding 10 mM (Bangsbo, 1994). Combining these two systems can be challenging as the muscle and cellular adaptations required for each can be antagonistic. Aerobic development typically favors increases in mitochondrial density, capillarization, and oxidative enzyme activity and typically decreases diffusion distances for substrates and gas

Received: 21 Jan 2025 / Accepted: 11 September 2025 / Published (online): 01 December 2025

exchange (Hawley, 2002); however, these adaptations might reduce maximal force generating capacity (van Wessel et al., 2010; Wilson et al., 2012). In contrast, anaerobic development relies more on increasing muscle cross-sectional area, and hypertrophy of fast-twitch fibers, which may reduce endurance efficiency (van der Zwaard et al., 2018a). While the reciprocal relationship between sprint and endurance is well established in cyclic sports, it has thus far not been investigated in team sports like soccer. Knowledge of the relationship between the sprint and endurance capacity is crucial for designing and optimizing training programs for individual athletes, also in team sports.

To understand this relationship, it is essential to accurately assess each capacity. Endurance capacity can be assessed using the $\dot{V}O_{2max}$, which is considered the gold standard for measuring endurance capacity and a critical determinant of endurance performance (van der Zwaard et al., 2021). Sprint capacity, on the other hand, can be measured using a sprint test over a fixed distance. A 20-meter sprint is representative for soccer sprints during competition and provides a relevant and practical indicator of the player's capacity to generate explosive power in match-specific contexts (Nikolaidis et al., 2016).

Additionally, in soccer, the individual match tasks can differ significantly per player based on their position in the field, playing style, physical capabilities, and coaches' tactical considerations and decisions (Buchheit et al., 2010; de Haan et al., 2025; Metaxas, 2021). This influences the physical load within the match for each individual player. To tailor training regimens to the individual player and optimize their physical performance, it is essential to understand the individual load of the match and how this relates to the physical capabilities of players. Match load can be quantified through various methods, broadly categorized into internal and external load metrics. Internal load represents the psychophysiological stress experienced by the player and is often measured using heartrate or rating of perceived exertion, while external load represents the dose performed (Campos-Vazquez et al., 2015; Jaspers et al., 2018). External match load can be quantified through match-specific running performance, defined as the combination of distance covered and speeds achieved.

While the individualized training is ideally tailored to each player's specific capacities and match-specific load, practical constraints often make this goal unattainable. A viable, but not yet explored, alternative is to categorize players into distinct subgroups based on their physical capacities and match-specific running performance. This grouping strategy allows for identification of clusters of players exhibiting similar combinations of sprint and endurance capacities, and accompanying match-specific running performance, which enables coaches to develop training strategies that specifically target players' individual strengths and needs. Such clustering analysis can be performed using unsupervised machine learning, for instance by applying the *k*-means algorithm. This technique has previously been successful in identifying subgroups with distinct physical characteristics in professional soccer players (Novack et al., 2013), as well as in categorizing young elite soccer players based on match-specific running performance (de Haan et al., 2025). It has also been successful in identifying distinct subgroups in other sports, such as elite cyclists based on their anthropometry (van der Zwaard et al., 2019) and NCAA Division 1 American football players based on their match demands (Shelly et al., 2020).

This study explored the potential of machine learning (*k*-means clustering and subgroup discovery) to identify distinct groups of players based on unique combinations of their sprint and endurance capacities, as well as their sprint and endurance match-specific running performance. In this context, we conducted an in-depth analysis of the relationship between sprint and endurance capacities across the entire cohort and examined how these physical traits correlated with their corresponding match-specific running performance.

We hypothesized that by using unsupervised machine learning (*k*-means clustering) we could identify clusters of players with unique combinations of sprint capacity, endurance capacity, and match-specific running performance. Moreover, we expected that the supervised machine learning method, subgroup discovery, would confirm the relevance of these clusters and provide additional insight into the role of playing position in match-specific running performance. Additionally, we expected sprint and endurance capacities to be inversely related in soccer, as in cycling and rowing, as well as a strong correlation between physical capacities and corresponding match-specific running performance.

Methods

Participants

This study included 31 young male elite soccer players at a professional football club of international caliber (U18 & U21, age = 18.0 ± 0.9 years, height = 1.79 ± 0.06 m, weight = 70.5 ± 6.9 kg; mean ± standard deviation). U18 played in the highest league for their age group in the Netherlands (Eredivisie) and U21 played in the second highest professional league in the Netherlands (the so-called Keuken Kampioen Divisie or KKD). The sample included 9 forwards, 8 attacking midfielders, 5 defending midfielders, 5 backs and 4 central defenders. Playing position was determined as assigned by the coach during matches (see Appendix A for a schematic overview of the playing positions). Goalkeepers were excluded, because they have a distinctly different match-specific running performance compared to outfield players. Match-specific running data were collected over two full consecutive seasons from all 31 players, with participants undergoing exercise testing at the start of the second season.

Ethical statement

The study was conducted in full compliance with the Declaration of Helsinki (2013) and approved by The Scientific and Ethical Review Board (VCWE-2023-054) of the Faculty of Behavioural and Movement Sciences of the Vrije Universiteit Amsterdam. Participants provided written informed consent. Participants were instructed to avoid strenuous exercise for 30 hours leading up to the exercise testing.

Sprint capacity

Sprint capacity was measured using an all-out linear sprint test over 20 meters on an artificial grass surface. Before the sprint test, participants underwent a standard warm-up routine for soccer practice designed by the physical training staff of the team. This routine consisted of dynamic stretching, running exercises and footwork drills. Participants were instructed to cover these 20 meters as fast as possible from a static start. They performed this test twice, with the fastest time being used for further analysis. Positional data were obtained using LPM (Inmotio, Zeist, the Netherlands; Inmotio GPS; Insiders, Lausanne, Switzerland) and integrated over time to determine the average sprint speed over 20 meters. In addition, maximal sprint speed (MSS) over 20 meters was assessed; see Appendix B for details. Players were familiar with the testing procedure as it is part of a regular testing battery performed by the professional sports team.

Endurance capacity

In this study, endurance capacity was quantified using the $\dot{V}O_{2max}$ obtained during a maximal incremental treadmill test (Kemi et al., 2003). Speed of the motorized treadmill (H/P/COSMOS - Pulsar 3P, Samcon bvba, Melle, Belgium) started at 8.5 km/h and was incrementally increased by 1.5 km/h every 2 minutes. The measurement concluded either when the player was unable to run at the treadmill speed and voluntarily stepped off or when they fell and were suspended by the safety harness. Breath-by-breath gas exchange analysis (Vyntus CPX, Jaeger-CareFusion, UK) was used to measure $\dot{V}O_{2max}$. Calibration was performed according to the manufacturer's instructions. The gas analyzer was calibrated using automatic reference gas calibration (15% O2, 5% CO2, 80% N2) and volume transducer was calibrated using the automatic integrated blower. Breath-by-breath data were smoothed, and $\dot{V}O_{2max}$ was calculated as the highest 30-s value. $\dot{V}O_{2max}$ was normalized to lean body mass$^{2/3}$ to eliminate the influence of body size in accordance with isometric scaling (McCann and Adams, 2002; van der Zwaard et al., 2018a). Note that no normalization for body size was performed for sprint capacity, as the physical dimension of speed is already size-independent, dividing distance by time (LT^{-1}). To determine running-specific endurance, we assessed maximal aerobic speed (MAS), a measure commonly used in soccer literature that is derived from $\dot{V}O_{2max}$ and also depends on running efficiency (see Appendix B).

Match-specific running performance

Over two consecutive seasons, match-specific running performance of the U18 and U21 teams was collected using multiple positional tracking systems (Inmotio Local Position Measurement (LPM); Inmotio, Zeist, the Netherlands; Inmotio GPS; Insiders, Lausanne, Switzerland, and SciSports Optical tracking; Panoris, Brno, Czech Republic). The LPM system measures with an overall sample frequency of 1,000 Hz, divided by the number of active transponders on the field. The average measurement frequency per active transponder varied from 40 to 80 Hz over the matches. The LPM system has been demonstrated to be an accurate and valid tool for tracking player movements in football, showing a mean difference from the actual distance of maximally -1.6% (Frencken et al., 2010; Ogris et al., 2012), and an accuracy of 10 cm according to the manufacturer. Inmotio GPS from Insiders measures with a frequency of 10 Hz and an accuracy of 30 cm, according to the manufacturer. SciSports Optical tracking with Panoris has a measurement rate of 25 Hz. After data collection, all data were processed using imoClient software (Inmotio, Zeist, the Netherlands). Data were available from all home and away games, totaling 619 matches (mean per player: 20.0 ± 11.3). Data obtained during friendly matches were not included in the dataset, as their physical performance in these games could differ from actual match conditions (Modric et al., 2019). Furthermore, a minimum of 80 minutes of playing time was required for a match to be included. Positional data were filtered using the same method for all tracking systems, using 100% weighted Gaussian average filter and a 500-ms speed frame interval. The position data were then integrated over time and categorized into multiple speed ranges. Running distances were classified as follows: Low Intensity Running (LIR) below 14 km/h, Moderate Intensity Running (MIR) between 14 and 19 km/h, High Intensity Running (HIR) between 19 and 24 km/h, and sprinting higher than 24 km/h. These zones are based on threshold values established by the professional soccer team and closely match those described in the literature (Gualtieri et al., 2023; Vieira et al., 2019).

Machine learning analysis

Unsupervised machine learning was employed using the *k*-means algorithm to cluster players into subgroups based on similar combinations of sprint and endurance capacities, and match-specific running performance. A detailed description of this method can be found in previous studies (Hartigan and Wong, 1979; van der Zwaard et al., 2019). During multiple iterations, data points were assigned to the most nearby centroid based on their Euclidean distance. Initial centroid positions were obtained at random. At each iteration, the centroid's location was recalculated as the average position of all assigned data points in that cluster. This process was repeated until the total within sum of square was minimized, and the location of the centroids stabilized. A maximum of 100 iterations was used and optimization was performed using 15 random starting partitions to enhance cluster stability.

As input variables, we included measures for sprint and endurance capacities (average 20-meter sprint speed and $\dot{V}O_{2max}$) as well as sprint and endurance match-specific running performance (sprint distance and combined distance traveled at moderate and high intensity (MIR + HIR)), resulting in a total of four input variables. All values were normalized to *Z*-scores before being entered into the algorithm. The optimal number of clusters for the present data was determined to be three, based on visual inspection of the elbow and silhouette plots. The stability of the cluster was evaluated by repeating the *k*-means algorithm 1,000 times and evaluating whether cluster assignment was consistent over these 1,000 runs. After this machine-learning analysis, differences in sprint, endurance and match-specific running performance were evaluated between the identified clusters. The clusters were visualized in two

dimensions by applying Principal Component Analysis (PCA) to reduce the high-dimensional dataset to its two most informative components.

Subsequently, we applied the supervised machine learning method of subgroup discovery (de Leeuw et al., 2022a; 2022b; Knobbe et al., 2017) to gain additional insight into how the match-specific running performance could be explained by a combination of 20-meter sprint speed and $\dot{V}O_{2max}$, player position and identified clusters obtained from *k*-means clustering. Subgroups were enforced to have corresponding sizes between 10% and 90% of the size of the entire data collection. We considered (inverted) *Z*-scores as quality measure to specify the difference between the subgroup and the entire data collection, with a positive sign reflecting better match-specific running performance and a negative sign reflecting worse performance. Finally, subgroups that are described by conditions on one or two predictors were investigated. The analysis was performed for both sprint and endurance match-specific running performance (i.e. sprint distance and combined distance traveled at moderate and high intensity [MIR + HIR]).

Statistical analysis

Data were presented as mean (*M*) ± standard deviation (*SD*) unless stated otherwise. First, the Shapiro-Wilk test and visual inspection of data distribution plots were used to verify normality of the data. After *k*-means clustering, one-way ANOVA tests were employed to compare the subgroups on 20-meter sprint speed, normalized $\dot{V}O_{2max}$, average match sprint distance, and average match distance ran at moderate and high intensity. When significant, Bonferroni post-hoc tests were performed to identify group differences between clusters. Additional cluster characteristics, including age, height, body mass, and BMI, were compared using a one-way ANOVA. BMI was calculated as weight divided by height squared (BMI = weight [kg] / height² [m]).

Linear regression analysis was used to identify the relationship between sprint capacity and endurance capacity. This analysis was performed for 20-meter sprint speed vs $\dot{V}O_{2max}$ normalized to $LBM^{2/3}$. The relationships between these physical traits were quantified in terms of explained variance (R^2). Subsequently, a Deming regression was employed after normalizing the physical traits to *Z*-scores, in accordance with previous literature (van der Zwaard et al., 2018a). This method models the relationship between these traits while accounting for errors in both variables, rather than only in the dependent variable, as is common in simple linear regression analysis.

To examine the relationship between sprint and endurance capacities with their respective match-specific running performance, the average match-specific sprint performance was plotted against the average sprint speed over 20 meters, while the distance covered at moderate and high intensity (MIR+HIR) during an average match was plotted against $\dot{V}O_{2max}$ normalized to $LBM^{2/3}$. Linear regression was used to examine these relationships and R^2 was used to quantify the explained variance of physical capacity on match-specific running performance.

The R^2 values were interpreted according to Cohen (1998) as follows: $R^2 < 0.02$ indicates a very weak relationship, $0.02 \leq R^2 < 0.13$ is weak, $0.13 \leq R^2 < 0.26$ is moderate, and $R^2 \geq 0.26$ is substantial. Results were deemed statistically significant if the *P*-value was ≤ 0.05 (α). Significance of the subgroup discovery results was assessed by performing 1,000-fold swap-randomization to determine the probability that observed differences in match-specific running performance between subgroups and the entire data collection was a true finding or a false discovery by testing many hypotheses (de Leeuw et al., 2022a; 2022b). Subgroups were considered significant if the probability that the observed difference was a false discovery was <5%.

Results

Sprint and endurance capacities

The 20-meter sprint test was completed by 27 players and the maximal incremental treadmill test by 28 players, with 24 players successfully completing both assessments. The average sprint speed of these 27 players on the 20-meter sprint test was 24.36 ± 0.66 km/h, with values ranging from 23.08 to 26.17 km/h. $\dot{V}O_{2max}$ relative to body weight of these 28 players was 57.93 ± 3.91 mL/kg/min. When normalized to $LBM^{2/3}$, $\dot{V}O_{2max}$ was 257.27 ± 14.51 $mL/kg_{LBM}^{2/3}/min$, with values ranging from 232.30 to 295.90 $mL/kg_{LBM}^{2/3}/min$. For average values of MAS and MSS, see Appendix B. For all 31 players match-specific running data were available, totaling 619 matches (mean per player: 20.0 ± 11.3); these data are shown in aggregated form together with the physical capacities in Table 1.

Clustering based on combinations of physical capacities and match-specific running performance

Clustering analysis was performed on the 24 players who completed both the sprint and endurance tests. They all had available match data, comprising a total of 458 individual match observations collected across two full competitive seasons, with an average of approximately 19 matches per player. Using *k*-means clustering, we identified three distinct non-overlapping clusters of players with specific combinations of sprint and endurance capacities, and sprint and endurance match-specific running performance (Figure 1). These three subgroups were compared on the four

Table 1. Physical capacity and match-specific running performance.

	Average 20-m speed (km/h)	Normalized $\dot{V}O_{2max}$ ($mL/kg_{lbm}^{2/3}/min$)	LIR (m)	MIR (m)	HIR (m)	Sprint (m)	TD (m)	MIR+HIR (m)
Mean± SD	24.36 ± 0.65	257.27 ± 14.25	8130 ± 380	1773 ± 312	678 ± 123	201 ± 78	10782 ± 658	2451 ± 406

Mean and standard deviations for the physical capacity measures (average 20-meter sprint speed and $\dot{V}O_{2max}$ normalized to $LBM^{2/3}$) and match-specific running performance variables (low-intensity running [LIR], moderate-intensity running [MIR], high-intensity running [HIR], sprint distance [Sprint], total distance covered [TD] and combined moderate- and high-intensity running [MIR + HIR]). 27 players completed the 20-meter sprint test, and 28 players completed the maximal incremental exercise test, providing $\dot{V}O_{2max}$ estimates. Match-specific running performance was available for all 31 players, totaling 619 matches (mean per player: 20.0 ± 11.3).

input variables: A) 20-meter sprint speed, B) average match sprint distance, C) normalized $\dot{V}O_{2max}$, and D) average match distance ran at moderate and high intensity (Figure 2). Clusters included one subgroup of players with high sprint capacity and a high number of meters sprinted during the matches (SPR, $n = 4$), another subgroup of players with high endurance capacity and a high number of meters ran at moderate and high intensities (END, $n = 6$), and a large group of players without high physical capacity or matching match-specific running performance (AVG, $n = 14$).

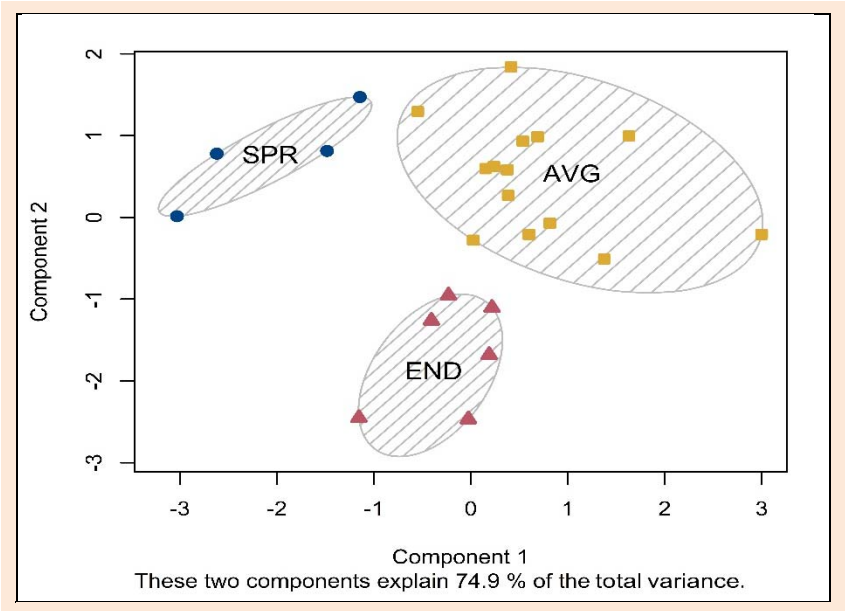


Figure 1. Cluster plot with a two-dimensional representation of the three identified clusters. Clusters are displayed in the two most important principal component dimensions (together explaining 74.9 % of the total variance). The dimensions in question are based on the combined distance traveled at moderate and high intensity (MIR + HIR), match-specific sprint distance, normalized $\dot{V}O_{2max}$, and average 20-meter sprint speed. Individual values and spanning ellipses of clusters are presented for SPR, which is a cluster with high sprint capacity and a large number of meters sprinted during the match (n = 4), for END, which is a group of players with high endurance capacity and a large number of meters ran at moderate and high intensity (n = 6), and for AVG, which is a group of players who either do not possess a high physical capacity or do not have a matching high match-specific running performance (n = 14). See text for further details.

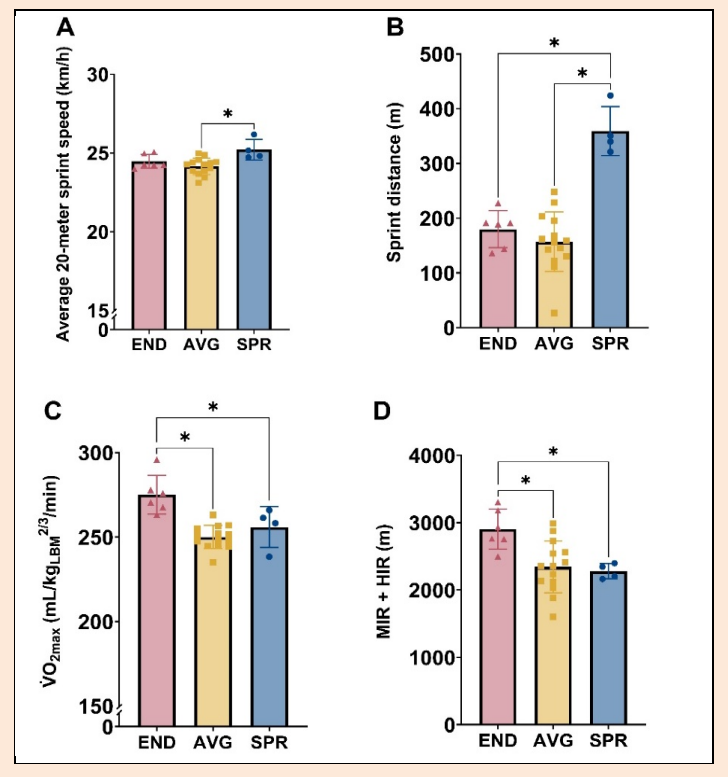


Figure 2. Group differences of the three clusters for their average match sprint distance, average 20-meter sprint speed, average match distance covered at moderate and high intensity, normalized $\dot{V}O_{2max}$. See the caption of Figure 1 and the text for the characteristics of the three clusters: SPR, END and AVG.

The SPR group had significantly higher sprint capacity than AVG (25.22 ± 0.66 vs 24.46 ± 0.44 km/h; mean difference = 1.07, 95% CI [0.29, 1.85], $P < 0.01$) and the SPR group covered a significantly greater average sprint distance during matches (359 ± 45 m) compared to both the AVG group (157 ± 55 m; mean difference = 202, 95% CI [130, 274], $P < 0.01$) and the END group (180 ± 34 m; mean difference = 179, 95% CI [97, 262], $P < 0.01$). Additionally, The END group had significantly higher normalized $\dot{V}O_{2max}$ (275.2 ± 11.5 mL/kg$_{LBM}^{2/3}$/min) compared to both the AVG group (250.0 ± 6.9 mL/kg$_{LBM}^{2/3}$/min; mean difference = 25.2, 95% CI [13.7, 36.7], $P < 0.01$) and the SPR group (256.0 ± 12.2 mL/kg$_{LBM}^{2/3}$/min; mean difference = 19.2, 95% CI [4.0, 34.4], $P = 0.01$). The END group also covered significantly more distance at moderate and high intensities (2905 ± 298 m) than both the AVG group (2344 ± 383 m; mean difference = 562 m, 95% CI [134, 989], $P < 0.01$) and the SPR group (2278 ± 112; mean difference = 628 m, 95% CI [62, 1193], $P = 0.03$). Moreover, the END group had a significantly higher MAS than the AVG group and there was a trend for increased MSS of the SPR group compared to END group (Appendix B. Figure 5).

Additional cluster characteristics, including playing position and competitive level, are displayed in Table 2. One-way ANOVA revealed no significant differences in age, height, body mass or BMI between the three clusters ($P \geq 0.19$). Considering these distinct sprint and endurance cluster groups, one might wonder if sprint and endurance are also mutually exclusive in these soccer players, as was expected from previous findings in cyclic sports.

Relationship between sprint and endurance capacities

To establish the relationship between sprint and endurance capacity, the $\dot{V}O_{2max}$ normalized to LBM$^{2/3}$ was plotted against the 20-meter sprint speed (Figure 3). Contrary to our expectations, no (negative) relationship between 20-meter sprint speed and $\dot{V}O_{2max}$ normalized to LBM$^{2/3}$ ($R^2 = 0.086$, $P = 0.16$) was observed. The Deming regressions revealed similar results ($P = 0.16$). Similarly, there was no (negative) relationship between MAS and MSS (Appendix B. Figure 6).

Relationship between physical capacity and match-specific running performance

Physical capacities were plotted against their corresponding match-specific running performance (Figure 4). A moderate positive relationship was identified between 20-meter sprint speed and average match sprint distance ($R^2 = 0.168$, $P = 0.03$). Similarly, there was also a moderate, significant positive relationship between average sprint distance during a match and the MSS (Appendix B. Figure 7a), indicating that faster players tend to cover more sprint distance during matches.

Table 2. Cluster characteristics.

Cluster	SPR	AVG	END
Playing Position	3 F | 1 B	3 F | 4 AM | 2 DM | 2 B | 3 CD	1 F | 2 AM | 2 DM | 1 B
Competitive level	2 U21 | 2 U18	14 U18	5 U21 | 1 U18
Age (years)	17.71 ± 1.16	17.72 ± 0.69	18.20 ± 1.00
Height (cm)	174.5 ± 4.7	178.9 ± 4.9	176.3 ± 6.3
Body mass (kg)	70.5 ± 2.3	68.9 ± 7.4	69.4 ± 5.8
BMI (kg/m^2)	23.2 ± 1.5	21.5 ± 1.8	22.3 ± 1.2

The playing position, competitive level, and the M ± SD for age, height, body mass, and BMI across the three identified clusters. Playing positions are abbreviated as F = Forward, AM = Attacking Midfielder, DM = Defensive Midfielder, B = Back, and CD = Central Defender.

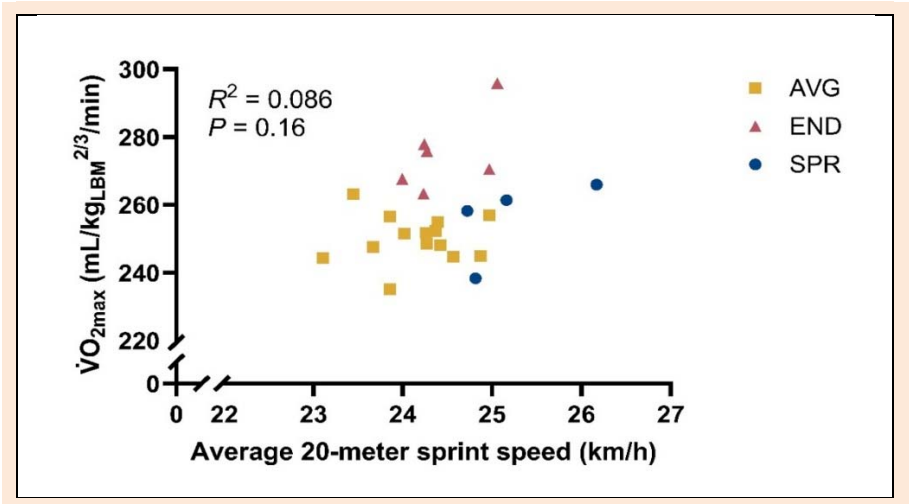


Figure 3. Normalized $\dot{V}O_{2max}$ plotted against the average 20-meter sprint speed. Each point reflects the data of a single young elite soccer player, for a total of 24 players. A linear regression revealed no significant relationship between the depicted variables. Clusters are indicated as follows: yellow square (AVG, average), red triangle (END, endurance-oriented), and blue circle (SPR, sprint-oriented).

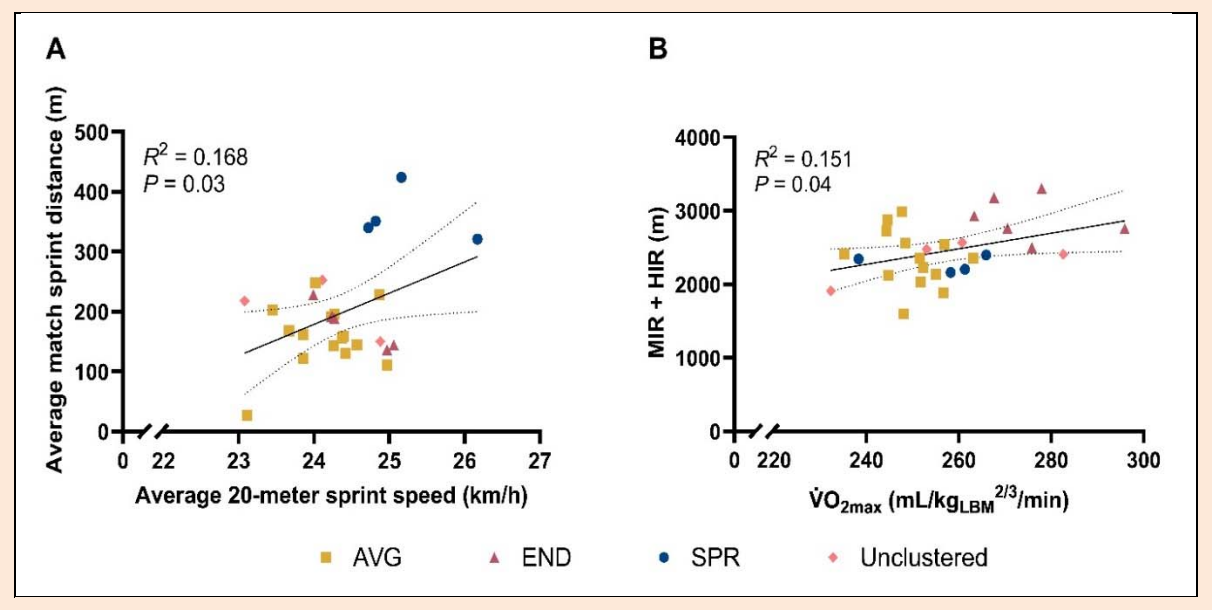


Figure 4. (A) Average sprint distance during a match plotted against the average sprint speed over a 20-meter sprint. Each point represents the data of a single young elite soccer player, for a total of 27 players. Linear regression (black line) shows a moderate, significant positive relationship between these variables. Clusters are indicated as follows: yellow square (AVG, average), red triangle (END, endurance-oriented), blue circle (SPR, sprint-oriented), and pink rhombus (insufficient data for clustering). (B) Average match distance at moderate and high intensity (14-24 km/h) plotted against normalized $\dot{V}O_{2max}$ for 28 young elite soccer players. As in (A), a moderate, significant positive relationship was found.

A moderate but significant positive relationship was also observed between normalized $\dot{V}O_{2max}$ and average match distance at moderate and high intensity ($R^2 = 0.151$, $P = 0.04$), Moreover, a substantial, significant positive relationship between average match distance at moderate and high intensity and MAS was found (Appendix B. Figure 7b), suggesting that higher endurance capacity translates to greater distances covered between 14 and 24 km/h. So, sprint and endurance capacities show positive relationships with corresponding match-specific running performance, but do not seem to be mutually exclusive.

The supervised subgroup discovery analysis showed that belonging to the SPR cluster was the most important predictor for increased sprint distances during matches ($P < 0.01$). Similarly, belonging to the END cluster, along with a lower average 20-meter sprint speed of ≤ 24.24 km/h, showed a trend towards increased moderate- and high-intensity running distance during matches ($P < 0.10$). Furthermore, belonging to the AVG cluster, with players who played as central defenders in particular showed significantly reduced moderate- and high-intensity running distances during matches ($P < 0.05$).

Discussion

The present study set out to evaluate the potential of machine learning to identify players with unique combinations of sprint and endurance capacities, and their match-specific running performance. *k*-means clustering identified three distinct subgroups of soccer players, including a subgroup of players with high sprint capacity and a high number of meters sprinted during the matches (SPR), another subgroup of players with high endurance capacity and a high number of meters ran at moderate and high intensities (END), and a large group of players without high physical capacities or matching match-specific running performance (AVG). This suggests that while many players do not optimize for either sprint or endurance, certain individuals develop in one area and leverage these abilities during matches. This was further supported by the subgroup discovery analysis, which identified classification in the SPR cluster as a significant predictor of greater sprint distances during matches ($P < 0.01$). Similarly, classification in the END cluster, combined with an average 20-meter sprint speed of ≤ 24.24 km/h, showed a trend toward greater moderate- and high-intensity running distances during match play ($P < 0.10$). These sprint and endurance matched groups, as identified by *k*-means clustering, could be an indication that, as we hypothesized, sprint and endurance are mutually exclusive in soccer players, akin to what has been observed in cyclic sports. However, even though physical capacities were positively related to the corresponding match-specific running performance ($R^2 = 0.168$, $P = 0.03$ and $R^2 = 0.151$, $P = 0.04$ for sprint and endurance respectively), we did not observe a significant (negative) relationship between sprint and endurance capacities, i.e. between 20-meter sprint speed and $\dot{V}O_{2max}$ normalized to $LBM^{2/3}$ ($R^2 = 0.086$, $P = 0.16$). Thus, sprint and endurance capacity do not appear to be mutually exclusive or opposing in young elite soccer players.

The potential of unsupervised machine learning in identifying player clusters

In this study, unsupervised machine learning, in particular *k*-means clustering, was introduced as an innovative method for grouping soccer players based on their combinations of sprint and endurance capacity, and their sprint and endurance match-specific running performance. By applying this clustering method, we identified three subgroups: SPR (sprint-oriented), END (endurance-oriented),

and AVG (average). The SPR and END groups were relatively small, comprising 4 and 6 players respectively, and contained players whose high physical capacities aligned well with their corresponding high match-specific running performance. The AVG group consisted of a comparatively large group of players ($n = 14$) without high physical capacity or matching match-specific running performance. This suggests that while many players do not optimize for either sprint or endurance, certain individuals develop in one area and leverage these abilities during matches. This was further supported by the subgroup discovery analysis, which identified classification in the SPR cluster as a significant predictor of greater sprint distances during matches ($P < 0.01$). Similarly, classification in the END cluster, combined with an average 20-meter sprint speed of ≤ 24.24 km/h, showed a trend toward greater moderate- and high-intensity running distances during match play ($P < 0.10$).

One possible explanation for these results is that players with high sprint or endurance capacities are recognized by coaches and strategically utilized in matches in accordance with those strategies. For these players, focusing on their specific playing style and incorporating tailored sprint or endurance training might be of great importance. Conversely, as shown in Figure 4, some players in the AVG cluster possess similar sprint capacity to those in the SPR group and endurance capacity compared to those in the END groups but do not use these capacities to full extent during match play, which raises the question whether these players could not be utilized more strategically based on their physical capabilities. Conversely, players in the AVG cluster with relatively higher match-specific running performance but lower physical capacities may be at risk of overuse or could benefit from targeted sprint or endurance training to enhance their performance and durability.

Cluster analysis using unsupervised machine learning allows for identification of players who may benefit from alternative strategic roles during matches, are at risk of overuse, or could benefit from individualized training. This information can assist coaches in designing tailored training programs for individual athletes and optimizing overall match strategy.

Absence of an inverse relationship between sprint and endurance capacities in young elite soccer players

We hypothesized an inverse relationship between sprint and endurance capacity in soccer, corresponding to the distinct SPR and END groups revealed by the clustering and previous observations of such a relationship in cyclic sports (van der Zwaard et al., 2018a; 2018b). However, no correlation was found between 20-meter sprint speed and normalized $\dot{V}O_{2max}$. One plausible explanation is that in soccer every outfield player needs to combine both sprint and endurance capacity due to the repeated sprints nature of this team sport, whereas cycling and rowing include more distinct sprint and endurance specialists (van der Zwaard et al., 2018a; 2018b). When comparing the $\dot{V}O_{2max}$ of soccer players with that of rowers and cyclists, it is evident that our young male soccer players (257 ± 15 mL/kg$_{LBM}^{2/3}$/min) exhibit lower values compared to a group of 18 male elite rowers (320 ± 15 mL/kg$_{LBM}^{2/3}$/min) (van der Zwaard et al., 2018b) and a group of 28 male elite cyclists (287 ± 25 mL/kg$_{LBM}^{2/3}$/min) (van der Zwaard et al., 2018a). Using an independent samples *t*-test on the reported means, standard deviations, and sample sizes revealed that the difference in $\dot{V}O_{2max}$ between groups was statistically significant ($P < 0.0001$). This confirms that our young soccer players had a significantly lower endurance capacity compared to these groups of rowers and cyclists. In terms of sprint capacity, our soccer players revealed similar sprint speed over 20-meter (± 24 km/h) compared to normative values in soccer players of the same age (± 23 km/h, (Nikolaidis et al., 2018)). Interestingly, this previous study conducted both Wingate tests and 20-meter sprints, reporting normative Wingate peak power values of 10-13 W/kg in soccer players of similar age. This sprint capacity measured during a Wingate test, however, was much higher in elite rowers (16.5 ± 1.6 W/kg, range: 13.3-18.7) (van der Zwaard et al., 2018b) and elite cyclists (17.2 ± 1.2 W/kg, range: 14.4-19.6) (van der Zwaard et al., 2018a). The absence of an inverse relationship could be due to the lower sprint and endurance capacities in soccer players, as the incompatibility between sprint and endurance capacities are expected to be most prominent at the physiological extremes. It remains to be investigated whether an inverse relationship between sprint and endurance is also absent in elite soccer players competing at the highest level, demonstrating stronger physical sprint and endurance capacities.

Age and training status are likely important factors influencing the differences in the relationship between sprint and endurance capacities observed in different sports. The previously studied cyclists (25 ± 7 years) (van der Zwaard et al., 2018a) and rowers (27 ± 3 years) (van der Zwaard et al., 2018b) were significantly older than the young soccer players in the present sample of soccer players (18.0 ± 0.9 years). Additionally, the amount of time these athletes spend focusing on training their specific physical capacities could further impact the relationship between sprint and endurance capacities in these groups.

Lastly, we measured $\dot{V}O_{2max}$ and sprint speed at the whole-body level. In this context, it should be noted that both physical traits depend on multiple factors, and these whole-body measurements are not direct reflections of muscle fiber characteristics. For example, the ability to transport oxygen to the muscles has a large influence on whole-body $\dot{V}O_{2max}$ (Bassett and Howley, 2000) and factors like lower limb coordination and leg length could potentially influence sprint speed (Wang et al., 2023). The inverse relationship between sprint and endurance might exist at the muscle fiber level in our soccer players, but other critical factors for whole-body sprint and endurance capacity could obscure this relationship in soccer. In cycling and rowing, where an inverse relationship has been observed at the whole-body level, physical capacity is more directly linked to performance outcomes, which may explain why this relationship is more evident in those sports compared to soccer.

The relationship between physical capacities and corresponding match-specific running performance

Only a moderate correlation between measured sprint and

endurance capacities and match-specific running performance was found, with an explained variance of 17 and 15% for sprint and endurance, respectively. This suggests that factors other than physical capacities might be more crucial for determining match-specific running performance. The subgroup discovery analysis confirmed importance of the clusters for interpreting differences in match-specific running performance and showed that central defenders (who were all part of the AVG group) showed significantly reduced moderate- and high-intensity running distances during matches, demonstrating that playing position is a factor that can influence match-specific running performance. Central defenders have less opportunities for longer sprints because they operate in the pitch region which is typically densely packed with players. This spatial constraint and their primarily defense tactical application are likely reasons for their decrease in moderate- and high-intensity running (Bradley, 2023; Sarmento et al., 2024). Other potential factors that could influence match-specific running performance are technical skills with the ball, the ability to read the game and identify opportunities, and tactical considerations and corresponding instructions for the soccer player by the coach.

Another factor influencing this relationship is running economy, which refers to the efficiency with which players use oxygen during submaximal running. A better running economy may enable some players to sustain higher match running outputs without necessarily having a higher $\dot{V}O_{2max}$. A measure that does incorporate this running economy is the maximal aerobic speed (MAS). In this study, we primarily focused on $\dot{V}O_{2max}$ because it is more closely related to actual muscle-level aerobic capacity. However, MAS may be a practically relevant measure, as it reflects the endurance demands in a match context. The relationship between MAS and endurance match-specific running performance (Appendix B) showed that MAS indeed demonstrated a substantial relationship with moderate- and high-intensity running during matches ($R^2 = 0.409$, $P < 0.01$).

The moderate to substantial relationships between physical capacity measures and corresponding match-specific running performance imply that players with lower aerobic or sprint capacities can still exhibit greater running output than peers with higher physical capacities. However, increasing match-specific running performance might not always translate to increased match performance. Literature indicates that in won matches, wide midfielders and forwards significantly increase their overall distance covered, particularly at speeds above 21 km/h, while full-backs, central defenders, and central midfielders tend to cover significantly less distance, likely within the 17-24 km/h range (Chmura et al., 2018). Good positioning on the field or fewer passing errors, for example, may reduce the need for sprinting, thereby decreasing the demand for sprint capacity during the match.

Overall, our results indicate that many soccer players do not specifically optimize towards either sprint or endurance performance. Instead, there may be minimal threshold levels for these capacities that players must meet to perform at an elite level within the U18 and U21 age groups. For sprint capacity, this threshold appears to be an average speed above 23 km/h over a 20-meter sprint, while for endurance capacity, it is more than 230 ml/$kg_{lbm}^{2/3}$/min or 50 ml/kg/min. These values are lower than those reported for cycling and rowing, raising the question of whether these players could be further developed physiologically and how such development might influence their match performance.

Limitations and perspectives

A sample size of 24 young elite soccer players could be viewed as relatively small for machine learning applications, but unlike traditional statistical analyses, the statistical power in cluster analysis is largely driven by the degree of cluster separation rather than sample size (Dalmaijer et al., 2022). In this study, we were able to identify distinct clusters with clear separation, confirming the method's robustness. Furthermore, while the participant pool included 24 players, the dataset used for clustering was based on 458 individual match observations collected over two full competitive seasons, averaging approximately 19 matches per player. For each player, match-specific running performance was averaged across all available matches and combined with the outcomes of the physical capacity tests to generate a single representative data point. This averaging process reduces the influence of match-to-match variability and enhances the reliability of the clustering outcomes.

The results from the *k*-means clustering are inherently dependent on the input data, meaning the identified clusters are specific to this group of young soccer players. The match-specific running performance and physical capacity measures among the participants were rather homogenous, especially compared to previous observations in cycling (van der Zwaard et al., 2019). This distribution makes it more difficult to designate all players to a distinct subgroup, as reflected by the average silhouette score (0.31). Nonetheless, Figure 1 displays distinct and non-overlapping subgroups in the cluster plot using the two primary principal components. When applying this methodology to another team or club, a new cluster analysis must be conducted, which may, and is indeed likely to, result in different clusters depending on the team's or club's formation. Likewise, using a different clustering method could, and is likely to, result in different clusters. To assess the robustness of our findings, we compared the results obtained using *k*-means clustering to those derived from alternative methods, including *k*-means++, *k*-medoids, and hierarchical clustering, and observed very similar clusters (see Appendix C). While *k*-means clustering is widely used due to its interpretability, it does come with limitations. These include the assumption of equally sized, spherical clusters and sensitivity to initial centroid placement. To minimize this last limitation, we implemented 15 different starting positions and repeated the clustering analysis 1,000 times to establish the robustness of the clustering procedure. This yielded very similar clusters compared to *k*-means++, that selects cluster starting positions using sampling based on the points' contribution to overall inertia. Future research could also explore the use of alternative clustering methods, such as DBSCAN, mean shift or spectral clustering.

The inverse relationship between sprint and endurance capacity observed in cyclic sports was based on

Wingate power output as a sprint measure. In this study, we did not conduct Wingate tests but opted for a 20-meter sprint test to measure sprint capacity. This test is more soccer-specific and representative of in-game sprint performance. It is however a less direct measure of leg power compared to a Wingate test. Instead, it reflects a combination of leg power and other factors, such as running technique. This difference in methodology makes it difficult to draw direct comparisons to the inverse relationship found in cyclists and rowers, especially since previous literature has shown that the explained variance between 20-meter sprint speed and Wingate power was only 19% (Nikolaidis et al., 2018). This difference in method for measuring sprint capacity could have contributed to why the inverse relationship that was previously observed in cyclic sports was not found in our study of young soccer players.

The exercise tests were performed in pre-season. This timing could present limitations, as players may not have fully reached their peak physical condition. However, we selected this point in time because it falls between the two seasons from which we collected match-specific running performance data. Additionally, conducting the tests outside of the competitive season helps to minimize variation caused by differences in in-season training loads and match-related fatigue.

The question that remains is if coaches should encourage their players more to play according to their physical abilities. There are players that have certain well-developed physical capacities but are not identified within the SPR or END groups. This is because they are not utilizing these abilities in the match to the same extent as the players identified in these groups. Similarly, there are players who display high match-specific running performance despite having lower physical capacities. Future research should aim to determine if optimizing physical capacity to match running performance or employing players based on their physical capacity can lead to improved overall performance.

Cluster analysis based on physical capacity and match-specific running performance could be further enriched by incorporating player's morphological characteristics to better understand the physiological basis behind the identified clusters. Muscle properties, such as quadriceps muscle volume, fiber length, pennation angle, and physiological cross-sectional area, could shed light on why certain players excel in either sprint or endurance performance (Weide et al., 2017). For instance, larger muscle volume is associated with greater power production (O'Brien et al., 2009), which may explain enhanced sprinting capabilities. Additionally, longer muscle fibers might circumvent the negative effect of muscle fiber hypertrophy on oxygen diffusion and maximal oxygen consumption (van der Zwaard et al., 2018b) and thus explain enhanced endurance capacity. Investigating the muscle morphology of soccer players in future studies could provide valuable insights and may even explain why both capacities were not inversely related.

Moreover, mixed sprint and endurance exercises, like repeated sprints or shuttle run tests, could be collected, and might provide extra information about the nature of the players in each cluster group. Additionally, incorporating longer sprints could give a clearer view of the maximal sprint speed of the players. Furthermore, these clusters could be used to evaluate training effects, to possibly identify groups of players who respond well to specific training impulses.

In the future, cluster analysis could serve as a foundational tool for predictive modeling, with potential applications in talent identification and injury prevention. Clustering by physical capacity and match-specific running performance could help predict and better discriminate which young athletes might excel in certain roles or which players are at higher risk of injury, allowing for more tailored training and development strategies.

Practical implications

The training staff identified two players who, based on their playing style, would be expected to align with those in the endurance cluster. It was surprising to them, however, to discover that these players fell significantly short in terms of endurance capacity compared to the cluster. This discrepancy could be an indication that these specific players would benefit from individualized endurance training to enhance their physical capabilities and optimize their performance in line with other endurance-oriented players. Additionally, the training staff identified players in the sprint cluster as those who leveraged their speed as a primary strength during matches. In contrast, players with similar sprint speeds but lower sprint distances tended to rely more on technical skill or tactical awareness during matches. These players were often positioned more centrally, which resulted in covering less sprint distance. Conversely, players in the sprint cluster, typically playing as wingers or backs, were able to maximize their sprint distance due to the nature of their roles, which demanded more movement and subsequently higher sprint workloads. This suggests that, in some cases, players with higher technical ability can perform effectively without relying on a large volume of sprint meters, while other players with superior physical capacities may potentially compensate for lower technical skills by being positioned in roles that allow them to fully exploit their speed. This distinction highlights that physical capacity is just one element of overall performance, and that a holistic assessment that includes technical and tactical abilities is crucial. By combining these factors, coaches could establish different benchmarks for players, such as those with high technical skills or sprint-oriented players, as identified by clustering. These insights could guide training adjustments, positioning strategies, or match strategy to maximize player performance.

Conclusion

Clustering revealed distinct subgroups of soccer players with unique combinations of sprint and endurance capacities, and sprint and endurance match-specific running performance. In young elite soccer players, sprint and endurance capacities showed positive, moderate, relationships with corresponding match-specific running performance but did not seem to be mutually exclusive or opposing. Derived clusters allow for identification of players who may benefit from alternative strategic roles during matches, are

at risk of overuse, or could gain from individualized training. This information can assist coaches in designing tailored training programs for individual athletes and optimizing overall match strategy.

Acknowledgements

We would like to thank all the players who participated in this study for their efforts and dedication. Moreover, we express our gratitude to the soccer club where the present study was conducted for allowing our testing procedure to be part of the yearly medical examination and their willingness to share the match data of their players. The authors thank Bart Vromans and Nicole Koopman-Verbaarschot for their great assistance with the data collection. We are thankful to Anita Loomans and Anna TopSupport for allowing us to utilize their exercise testing room, treadmill ergometer, and heart rate monitor during the testing week. Finally, we thank all physicians that supervised these measurements.

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy rules that are in place at the professional soccer club in question. However, anonymous data are available upon request from the corresponding author: Richard Jaspers, r.t.jaspers@vu.nl.

This work was supported by an NWO-ZonMW National Sportinnovator Grant (grant number: 538001045). Authors report no conflict of interest. The results of the present study do not constitute endorsement by ACSM. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

References

Bangsbo, J. (1994) Energy demands in competitive soccer. *Journal of Sports Sciences*, **12**(sup1), 5-12. https://doi.org/10.1080/02640414.1994.12059272

Bassett, D. R. and Howley, E. T. (2000) Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Medicine & Science in Sports & Exercise*, **32**(1), 70-84. https://doi.org/10.1097/00005768-200001000-00012

Bradley, P. S. (2023) 'Setting the benchmark' Part 1: The contextualised physical demands of positional roles in the FIFA World Cup Qatar 2022. *Biology of Sport*, **41**(1), 261-270. https://doi.org/10.5114/biolsport.2024.131090

Buchheit, M., Mendez-Villanueva, A., Simpson, B. M. and Bourdon, P. C. (2010) Match running performance and fitness in youth soccer. *International Journal of Sports Medicine*, **31**(11), 818-825. https://doi.org/10.1055/s-0030-1262838

Campos-Vazquez, M. A., Mendez-Villanueva, A., Gonzalez-Jurado, J. A., León-Prados, J. A., Santalla, A. and Suarez-Arrones, L. (2015) Relationships between rating-of-perceived-exertion and heart-rate-derived internal training load in professional soccer players: A comparison of on-field integrated training sessions. *International Journal of Sports Physiology and Performance*, **10**(5), 587-592. https://doi.org/10.1123/ijspp.2014-0294

Chmura, P., Konefał, M., Chmura, J., Kowalczuk, E., Zając, T., Rokita, A. and Andrzejewski, M. (2018) Match outcome and running performance in different intensity ranges among elite soccer players. *Biology of Sport*, **35**(2), 197-203. https://doi.org/10.5114/biolsport.2018.74196

Dalmaijer, E. S., Nord, C. L. and Astle, D. E. (2022) Statistical power for cluster analysis. *BMC Bioinformatics*, **23**(1), 205. https://doi.org/10.1186/s12859-022-04675-1

de Haan, M., van der Zwaard, S., Sanders, J., Beek, P. J. and Jaspers, R. T. (2025) Beyond playing positions: Categorizing soccer players based on match-specific running performance using machine learning. *Journal of Sports Science and Medicine*, **24**(3), 565-577. https://doi.org/10.52082/jssm.2025.565

de Leeuw, A.-W., van Baar, R., Knobbe, A. and van der Zwaard, S. (2022a) Modeling match performance in elite volleyball players: Importance of jump load and strength training characteristics. *Sensors*, **22**(20), 7996. https://doi.org/10.3390/s22207996

de Leeuw, A.-W., van der Zwaard, S., van Baar, R. and Knobbe, A. (2022b) Personalized machine learning approach to injury monitoring in elite volleyball players. *European Journal of Sport Science*, **22**(4), 511-520. https://doi.org/10.1080/17461391.2021.1887369

Frencken, W. G., Lemmink, K. A. and Delleman, N. J. (2010) Soccer-specific accuracy and validity of the local position measurement (LPM) system. *Journal of Science and Medicine in Sport*, **13**(6), 641-645. https://doi.org/10.1016/j.jsams.2010.04.003

Gualtieri, A., Rampinini, E., Dello Iacono, A. and Beato, M. (2023) High-speed running and sprinting in professional adult soccer: Current thresholds definition, match demands and training strategies: A systematic review. *Frontiers in Sports and Active Living*, **5**, 1116293. https://doi.org/10.3389/fspor.2023.1323440

Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **28**(1), 100-108. https://doi.org/10.2307/2346830

Hawley, J. A. (2002) Adaptations of skeletal muscle to prolonged, intense endurance training. *Clinical and Experimental Pharmacology and Physiology*, **29**(3), 218-222. https://doi.org/10.1046/j.1440-1681.2002.03623.x

Huiberts, R. O., Wüst, R. C. and van der Zwaard, S. (2024) Concurrent strength and endurance training: A systematic review and meta-analysis on the impact of sex and training status. *Sports Medicine*, **54**(2), 485-503. https://doi.org/10.1007/s40279-023-01943-9

Jaspers, A., De Beéck, T. O., Brink, M. S., Frencken, W. G., Staes, F., Davis, J. J. and Helsen, W. F. (2018) Relationships between the external and internal training load in professional soccer: What can we learn from machine learning? *International Journal of Sports Physiology and Performance*, **13**(5), 625-630. https://doi.org/10.1123/ijspp.2017-0299

Kemi, O., Hoff, J., Engen, L., Helgerud, J. and Wisløff, U. (2003) Soccer specific testing of maximal oxygen uptake. *Journal of Sports Medicine and Physical Fitness*, **43**(2), 139.

Knobbe, A., Orie, J., Hofman, N., Van der Burgh, B. and Cachucho, R. (2017) Sports analytics for professional speed skating. *Data Mining and Knowledge Discovery*, **31**(6), 1872-1902. https://doi.org/10.1007/s10618-017-0512-3

Lundberg, T. R., Feuerbacher, J. F., Sünkeler, M. and Schumann, M. (2022) The effects of concurrent aerobic and strength training on muscle fiber hypertrophy: A systematic review and meta-analysis. *Sports Medicine*, **52**(10), 2391-2403. https://doi.org/10.1007/s40279-022-01688-x

McCann, D. J. and Adams, W. C. (2002) A theory for normalizing resting VO_2 for differences in body size. *Medicine & Science in Sports & Exercise*, **34**(8), 1382-1390. https://doi.org/10.1097/00005768-200208000-00022

Metaxas, T. I. (2021) Match running performance of elite soccer players: $\dot{V}O_2$max and players position influences. *Journal of Strength and Conditioning Research*, **35**(1), 162-168. https://doi.org/10.1519/JSC.0000000000002646

Modric, T., Versic, S., Sekulic, D. and Liposek, S. (2019) Analysis of the association between running performance and game performance indicators in professional soccer players. *International Journal of Environmental Research and Public Health*, **16**(20), 4032. https://doi.org/10.3390/ijerph16204032

Nikolaidis, P. T., Matos, B., Clemente, F. M., Bezerra, P., Camões, M., Rosemann, T. and Knechtle, B. (2018) Normative data of the Wingate anaerobic test in 1-year age groups of male soccer players. *Frontiers in Physiology*, **9**, 1619. https://doi.org/10.3389/fphys.2018.01619

Nikolaidis, P. T., Ruano, M. A. G., de Oliveira, N. C., Portes, L. A., Freiwald, J., Leprêtre, P. M. and Knechtle, B. (2016) Who runs the fastest? Anthropometric and physiological correlates of 20 m sprint performance in male soccer players. *Research in Sports Medicine*, **24**(4), 341-351. https://doi.org/10.1080/15438627.2016.1222281

Novack, L. F., Nascimento, V. B., Salgueirosa, F. d. M., Carignano, L. F., Fornaziero, A., Gomes, E. B. and Osiecki, R. (2013) Subgroup distribution based on physiological responses in professional soccer players by k-means cluster technique. *Revista Brasileira de Medicina do Esporte*, **19**, 130-133. https://doi.org/10.1590/S1517-86922013000200012

O'Brien, T. D., Reeves, N. D., Baltzopoulos, V., Jones, D. A. and Maganaris, C. N. (2009) Strong relationships exist between muscle volume, joint power and whole-body external mechanical power in adults and children. *Experimental Physiology*, **94**(6), 731-738. https://doi.org/10.1113/expphysiol.2008.045062

Ogris, G., Leser, R., Horsak, B., Kornfeind, P., Heller, M. and Baca, A. (2012) Accuracy of the LPM tracking system considering

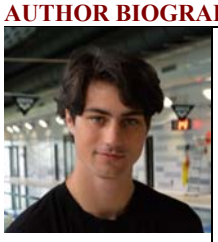
dynamic position changes. *Journal of Sports Sciences*, **30**(14), 1503-1511. https://doi.org/10.1080/02640414.2012.712712
Sarmento, H., Martinho, D. V., Gouveia, É. R., Afonso, J., Chmura, P., Field, A., Savedra, N. O., Oliveira, R., Praça, G. and Silva, R. (2024) The influence of playing position on physical, physiological, and technical demands in adult male soccer matches: A systematic scoping review with evidence gap map. *Sports Medicine*, **54**(11), 2841-2864. https://doi.org/10.1007/s40279-024-02088-z
Schumann, M., Feuerbacher, J. F., Sünkeler, M., Freitag, N., Rønnestad, B. R., Doma, K. and Lundberg, T. R. (2022) Compatibility of concurrent aerobic and strength training for skeletal muscle size and function: An updated systematic review and meta-analysis. *Sports Medicine*, **52**(3), 601-612. https://doi.org/10.1007/s40279-021-01587-7
Shelly, Z., Burch, R. F., Tian, W., Strawderman, L., Piroli, A. and Bichey, C. (2020) Using k-means clustering to create training groups for elite American football student-athletes based on game demands. *International Journal of Kinesiology and Sports Science*, **8**(2), 47-63. https://doi.org/10.7575//aiac.ijkss.v.8n.2p.47
Stølen, T., Chamari, K., Castagna, C. and Wisløff, U. (2005) Physiology of soccer: An update. *Sports Medicine*, **35**, 501-536. https://doi.org/10.2165/00007256-200535060-00004
van der Zwaard, S., Brocherie, F. and Jaspers, R. T. (2021) Under the hood: skeletal muscle determinants of endurance performance. *Fspor*, **3**, 719434. https://doi.org/10.3389/fspor.2021.719434
van der Zwaard, S., de Ruiter, C. J., Jaspers, R. T. and de Koning, J. J. (2019) Anthropometric clusters of competitive cyclists and their sprint and endurance performance. *Frontiers in Physiology*, **10**, 1276. https://doi.org/10.3389/fphys.2019.01276
van der Zwaard, S., van der Laarse, W. J., Weide, G., Bloemers, F. W., Hofmijster, M. J., Levels, K., Noordhof, D. A., de Koning, J. J., de Ruiter, C. J. and Jaspers, R. T. (2018a) Critical determinants of combined sprint and endurance performance: An integrative analysis from muscle fiber to the human body. *FASEB Journal*, **32**(4), 2110-2123. https://doi.org/10.1096/fj.201700827R
van der Zwaard, S., Weide, G., Levels, K., Eikelboom, M. R. I., Noordhof, D. A., Hofmijster, M. J., van der Laarse, W. J., de Koning, J. J., de Ruiter, C. J. and Jaspers, R. T. (2018b) Muscle morphology of the vastus lateralis is strongly related to ergometer performance, sprint capacity and endurance capacity in Olympic rowers. *Journal of Sports Sciences*, **36**(18), 2111-2120. https://doi.org/10.1080/02640414.2018.1439434
van Wessel, T., De Haan, A., Van Der Laarse, W. and Jaspers, R. (2010) The muscle fiber type-fiber size paradox: Hypertrophy or oxidative metabolism? *European Journal of Applied Physiology*, **110**, 665-694. https://doi.org/10.1007/s00421-010-1545-0
Vieira, L. H. P., Christopher, C., Barbieri, F. A., Aquino, R. and Santiago, P. R. P. (2019) Match running performance in young soccer players: A systematic review. *Journal of Sports Medicine*, **49**, 289-318. https://doi.org/10.1007/s40279-018-01048-8
Wang, L., Wang, W., Li, S. and Zhang, H. (2023) Stride length mediates the correlation between movement coordination and sprint velocity. *Journal of Sports Sciences*, **41**(1), 72-79. https://doi.org/10.1080/02640414.2023.2197523
Weide, G., Van Der Zwaard, S., Huijing, P. A., Jaspers, R. T. and Harlaar, J. (2017) 3D ultrasound imaging: Fast and cost-effective morphometry of musculoskeletal tissue. *Journal of Visualized Experiments*, **2017**(129), 1-10. https://doi.org/10.3791/55943
Wilson, J. M., Marin, P. J., Rhea, M. R., Wilson, S. M., Loenneke, J. P. and Anderson, J. C. (2012) Concurrent training: A meta-analysis examining interference of aerobic and resistance exercises. *Journal of Strength and Conditioning Research*, **26**(8), 2293-2307. https://doi.org/10.1519/JSC.0b013e31823a3e2d

✉ **Richard Jaspers**
Vrije Universiteit Amsterdam, De Boelelaan 1108, 1081 HZ Amsterdam, Netherlands

Key points

- Sprint and endurance capacities showed positive, moderate, relationships with corresponding match-specific running performance but did not seem to be mutually exclusive or opposing.
- Clustering revealed distinct subgroups of soccer players with unique combinations of sprint and endurance capacities, and sprint and endurance match-specific running performance.
- Clustering allows for identification of players who may benefit from alternative strategic roles during matches. are at risk of overuse, or could benefit from individualized training.

AUTHOR BIOGRAPHY



Michel de HAAN
Employment
Department of Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, Amsterdam, Netherlands
Degree
MSc, PhD student
Research interests
Exercise physiology, muscle physiology, sports science, data science
E-mail: m.h.de.haan@vu.nl

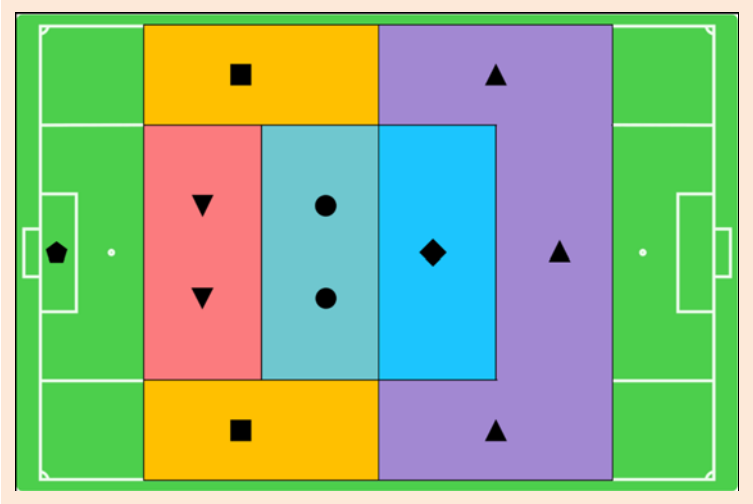
Stephan van der ZWAARD
Employment
Department of Cardiology, Amsterdam University Medical Center, location AMC, University of Amsterdam, Amsterdam, Netherlands
Degree
Assistant professor
Research interests
Sports science, exercise physiology, data science
E-mail: s.vanderzwaard@amsterdamumc.nl

Jurrit SANDERS
Employment
PSV Eindhoven, Eindhoven, Netherlands
Degree
MSc
Research interests
Soccer, applied physiology, data science, talent development
E-mail: j.sanders@psv.nl

Peter BEEK
Employment
Department of Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, Amsterdam, Netherlands
Degree
Full professor
Research interests
Coordination dynamics, perceptual-motor skills, sports science, rehabilitation
E-mail: p.j.beek@vu.nl

Richard JASPERS
Employment
Department of Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, Amsterdam, Netherlands
Degree
Full professor
Research interests
Exercise physiology, muscle physiology, molecular physiology, sports science
E-mail: r.t.jaspers@vu.nl

Appendix A: Schematic representation of the playing positions.



Schematic representation of the playing positions. Forwards (upright triangles in purple area), attacking midfielders (diamond in blue area), defending midfielders (circles in green area), backs (squares in yellow area) and central defenders (inverted triangles in red area). The goalkeeper is depicted as a pentagon.

Appendix B: MAS and MSS

Introduction

Maximal aerobic speed (MAS) and maximal sprinting speed (MSS) are two alternative measures of endurance and sprint capacity commonly used in soccer literature (Mendez-Villanueva et al., 2010; Sandford et al., 2021). MAS reflects running-specific endurance and is derived from $\dot{V}O_{2max}$ but is also influenced by running efficiency, which depends on anthropometric and biomechanical factors. In this study, we primarily focused on $\dot{V}O_{2max}$ because it is more closely related to actual muscle-level aerobic capacity. However, MAS may be a practically relevant measure, as it reflects the endurance demands in a match context. To address this, we compared MAS and MSS over the clusters identified by unsupervised machine learning (SPR, END, AVG), and examined how MAS and MSS relate to each other as well as to corresponding match-specific running performance.

Methods

MAS was defined as the running speed at which $\dot{V}O_{2max}$ was attained. Since $\dot{V}O_{2max}$ was calculated using a 30-s rolling average, the midpoint of this window was used to determine the time at which $\dot{V}O_{2max}$ occurred. To derive MAS as a continuous variable rather than a discrete one, the running speed was interpolated based on the elapsed time within the final stage, proportionally assigning a speed between the current and subsequent stage of the test. The MSS was defined as the highest speed recorded during the 20-meter sprint. This likely underestimates the players' true maximum speed because a longer distance, closer to 40 meters, is typically required to reach top speed. However, it still provides a useful indication of maximal sprinting ability within a soccer-specific context.

MAS and MSS were compared over the clusters identified by unsupervised machine learning (SPR, END, AVG). Additionally, linear regression analysis was used to identify the relationship between MAS and MSS. The relationships between these physical traits were quantified in terms of explained variance (R^2). Subsequently, a Deming regression was employed after normalizing the physical traits to *Z*-scores in accordance with previous literature (van der Zwaard et al., 2018) to fit the relationship between these traits accounting for errors in both traits rather than that of the dependent variable only (as is common in simple linear regression analysis). Moreover, the average match-specific sprint performance was plotted against the MSS, while the distance covered at moderate and high intensity (MIR+HIR) during an average match was plotted against the MAS. Linear regression was used to examine these relationships and R^2 was used to quantify the explained variance of physical capacity on match-specific running performance.

Results

For our sample of young elite soccer players, the MSS over 20 meters was 29.58 ± 0.71 km/h with values ranging from 27.73 to 31.20 km/h. MAS was 19.50 ± 1.05 km/h with values ranging from 16.49 to 21.30 km/h. The three clusters identified by *k*-means clustering were compared for differences in MAS and MSS (Figure 5). END had a significantly higher MAS than AVG (20.35 vs 18.86 km/h; mean difference = 1.49, 95%CI [0.35, 2.63], $P < 0.01$) and there was a trend for increased MSS of SPR compared to END (30.30 vs 29.18 km/h; mean difference = 1.12; 95%CI [0.07, 2.32], $P = 0.07$).

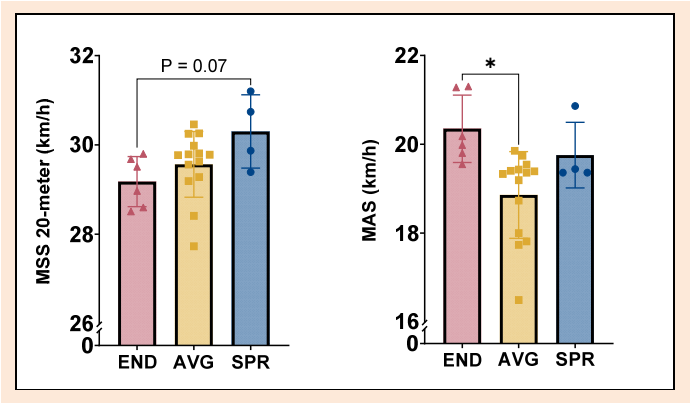


Figure 5. Group differences of the three clusters for their maximal 20-meter sprint speed (MSS 20-meter) and maximal aerobic speed (MAS). See the caption of Figure 1 and the text for the characteristics of the three clusters: SPR, END and AVG.

Similar to the comparison between 20-meter sprint speed and $\dot{V}O_{2max}$ normalized to $LBM^{2/3}$ ($R^2 = 0.086$, $P = 0.16$), no (negative) relationship between MAS and MSS ($R^2 = 0.009$, $P = 0.66$) was observed (Figure 6). The Deming regressions revealed similar results in both cases ($P = 0.16$ and 0.66, respectively).

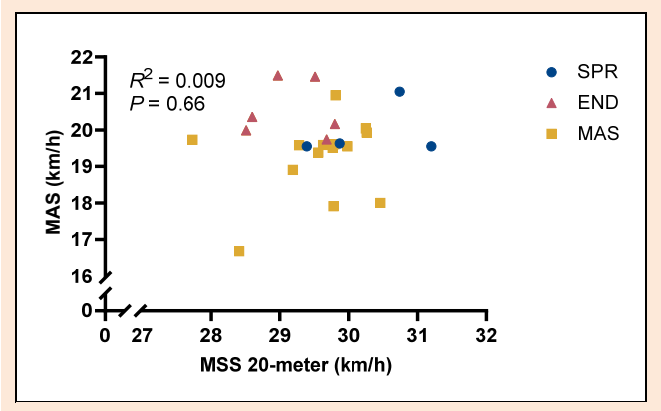


Figure 6. Maximal aerobic speed (MAS) plotted against the maximal 20-meter sprint speed (MSS 20-meter). Each point reflects the data of a single young elite soccer player, for a total of 24 players. A linear regression revealed no significant relationship between the depicted variables. Clusters are indicated as follows: yellow square (AVG, average), red triangle (END, endurance-oriented), and blue circle (SPR, sprint-oriented).

Similar to the comparison between average sprint speed and average sprint distance during a match, the linear regression showed a moderate, significant positive relationship between average sprint distance during a match and the MSS ($R^2 = 0.230$, $P = 0.01$). Moreover, linear regression revealed a substantial, significant positive relationship between average match distance at moderate and high intensity and MAS ($R^2 = 0.409$, $P < 0.01$) (Figure 7).

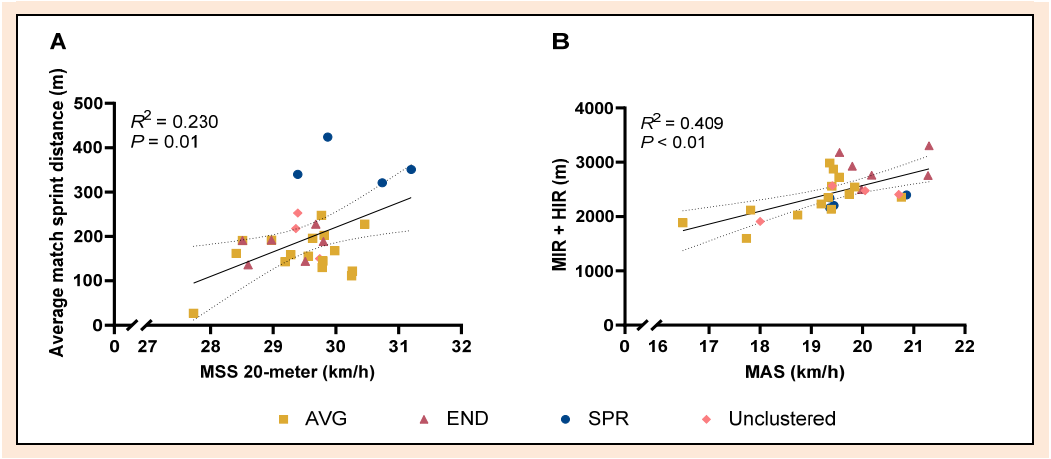


Figure 7. (A) Average sprint distance during a match plotted against the maximal sprint speed over a 20-meter sprint (MSS 20-meter). Each point represents the data of a single young elite soccer player, for a total of 27 players. Linear regression (black line) shows a moderate, significant positive relationship between these variables. (B) Average match distance at moderate and high intensity (14-24 km/h) plotted against maximal aerobic speed (MAS) for 28 young elite soccer players. A substantial, significant positive relationship was found. Clusters are indicated as follows: yellow square (AVG, average), red triangle (END, endurance-oriented), blue circle (SPR, sprint-oriented), and pink rhombus (insufficient data for clustering).

Appendix C: Alternative clustering methods

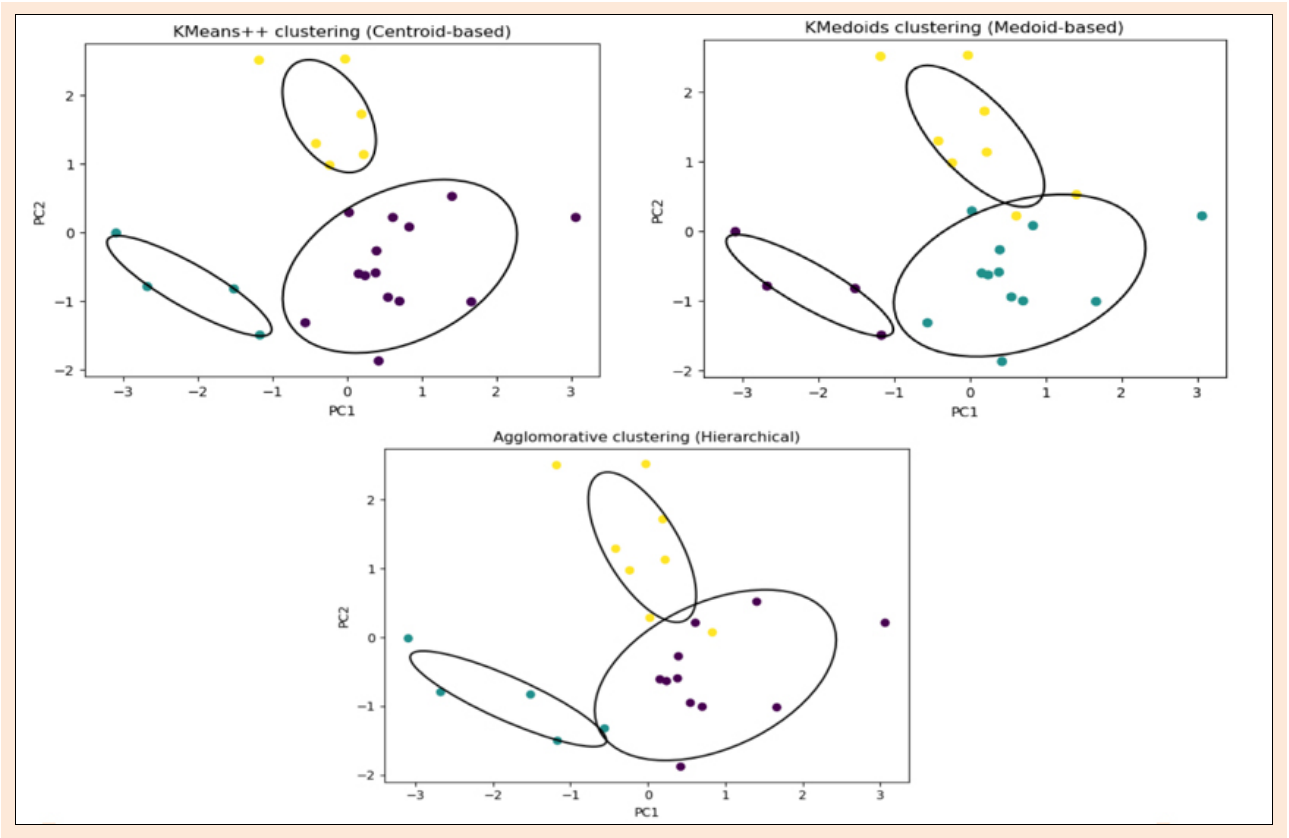


Figure 8. *k*-means++, *k*-medoids, and hierarchical clustering as alternative clustering methods to *k*-means. All three alternative methods show very similar clusters compared to *k*-means.