Forecasting extremes of football players' performance in matches.

Michal Nowak

michal.nowak@rakow.com

Faculty of Physical Culture Sciences, Collegium Medicum. Dr. Wladyslaw Bieganski, Jan Dlugosz University in Czestochowa, 42-200 Czestochowa

Bartosz Bok

National Research Institute NASK, Warszawa

Artur Wilczek

National Research Institute NASK, Warszawa

Mariusz Kamola

National Research Institute NASK, Warszawa

Łukasz Oleksy

Faculty of Health Sciences, Department of Physiotherapy, Jagiellonian University Medical College, Kraków

Article

Keywords:

Posted Date: June 7th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4071433/v1

Additional Declarations: No competing interests reported.

Forecasting extremes of football players' performance in matches.

Michal Nowak[1,3,+], Bartosz Bok[2], Artur Wilczek[2], Łukasz Oleksy[4,5], and Mariusz Kamola[2,*]

[1]Jan Dlugosz University in Czestochowa, Collegium Medicum. Dr. Wladyslaw Bieganski; Faculty of Physical Culture Sciences, 42-200 Czestochowa, Poland
[2]National Research Institute NASK, 01-045 Warszawa, Poland
[3]Science Department of the RKS Rakow Czestochowa Academy, 42-200 Czestochowa, Poland
[4]Faculty of Health Sciences, Department of Physiotherapy, Jagiellonian University Medical College, Kraków, Poland
[5]Department of Orthopaedics, Traumatology and Hand Surgery, Faculty of Medicine, Wroclaw Medical University, 50-556 Wroclaw, Poland
[+]michal.nowak@rakow.com
[*]mariusz.kamola@nask.pl

ABSTRACT

Based on training history, accurate predictive modelling of an athlete's performance in competition can be a cornerstone for prior optimal planning of exercise mix and intensity. While universal in many sports, such a goal is challenging in football due to the complexity of factors leading to the final score and the complexity of the preceding training process. We developed and tested a range of models, the best of which were to forecast selected play performance indices with an accuracy of 10-20%. Such score applies to models run on raw player location data and aggregating performance indices developed with expert knowledge in the football training domain. Results show that individual player models perform better than collective ones and that more recent training data are better predictors. While we consider the accuracy of the models still of limited reliability, their transparency and present quality make them useful in the daily planning of training activities that impact player performance in the coming match. Additionally, observations of training parameters generated in short-term intervals are more effective and correlated with extreme match results than long-term dates. Specific training parameters may be key in predicting exceptional football player performance, but they may also vary from person to person.

Introduction

In recent years, the amount of data generated by the world of professional sports has grown significantly, which poses a challenge for coaches and analysts in interpreting this information, especially in the context of team games, where a large number of variables and interactions between players complicate the analysis[1]. These data include both quantitative and qualitative statistics, including, for example, distances run by competitors in different speed or intensity zones, number of accelerations, stops, and other parameters estimating external or internal loads[2]. Additionally, their position in the game is currently monitored using data from cameras, as well as more subjective assessments, such as the effectiveness of defensive play or creativity in attack, created by experts in given fields[3]. Adding to the challenges of interpreting this data, is the variety of data types, from basic measurements to advanced performance metrics. Focusing on football, one of the most popular sports in the world, it was considered how correlational models could identify potential parameters in sub-maximal zones, called extreme efforts for example Maximal Sprinting Speed[4]. Before the time that extreme results occur, e.g. during training or the competition itself, it is crucial to develop sports skills and prepare athletes to achieve excellent performance. An analysis of such an approach is presented in the work[5] concerning 100 m run competition. During this period, it focuses on intensive training, specific exercises aimed at improving physical condition, technique and tactics[6]. Proper preparation supported by analysis of training data and individualization of training programs enables athletes to achieve their maximum potential and manifest these extreme results at the time of competition[7]. In the world of professional sports, where a small advantage can determine the final result, understanding and predicting maximum efforts at the right moment becomes crucial for coaches, analysts and players themselves[8]. Applying statistical models and machine learning to match data allows for the identification of patterns and variables that contribute most to sports success[9,10]. Such approaches have been used in other disciplines, such as baseball, where advanced statistics and sabermetrics analyses are revolutionizing the approach to the game[11,12], or in National Hockey League (NHL), National Basketball Association (NBA), where spatial analysis and advanced metrics influence strategies and coaching decisions[13], they try to predict the results of matches also in live conditions[14–16]. New ways of interacting with fans are being explored to involve them directly in the game. A visual and spatial analysis system for the NBA that changes the

way fans watch basketball games. It offers viewers a unique experience thanks to the use of various AR (augmented reality) technologies and real-time data analysis. One study used these methods to analyze the shooting performance of each NBA player, as well as to determine which players exhibited the most powerful shooting behavior in space. In one of the modes, the algorithm indicated the probability of making an accurate shot.[17]. In our study, we aim to understand how these approaches can be adapted and used in the context of football to predict extreme events that can change the course of a match as well as to optimize the period of preparation for a match in the training process or to optimize match lineups or substitution management tactics for maintaining the most effective combination.

Scientists also deal on a daily basis with the challenge of limited data and high complexity of match result factors, which often constitute an obstacle to creating accurate predictive models[16]. By analyzing various modelling methods, from simple ones, linear models, to more complex numerics-based machine learning and ultimately large language models (LLMs), we look for optimal solutions for predicting match extremes[18]. It should be emphasized that the scientific literature so far lacks guidelines specifying precisely the parameter set or specially created index that should be observed in order to correctly predict the future in the short or long term, unlike the methods used to calculate the weather forecasts[19]. Our study highlights the importance of individualizing predictive models. Recognizing that each player and team is unique, tailoring models to individual characteristics can significantly increase the accuracy of predictions. Integrating models tailored to personal characteristics in sports data analysis significantly improves performance prediction accuracy. They highlight the effectiveness of algorithms in sports analysis, predicting various aspects such as player position, shooting performance and number of shots during matches, achieving high accuracy[20–22]. Data analytics in sports opens up new opportunities for sports research and practice, promoting a better understanding of how various factors influence performance. Integrating these methods into daily practice can lead to discoveries and strategies that, in turn, can transform approaches to training, team management and decision-making in sports at all levels. The idea of extreme value modelling has been present in environmental sciences for decades[23], where models of events of extraordinary magnitude were based on parametric tuning of one of three heavy-tailed distributions: Frechét, Weibull or Gumbel. In particular, Russell et al.[24] initially applied such an approach for extreme ozone concentration forecasting in cities by finding relevant factors like temperature, sunshine, and humidity — as well as their relevant thresholds, thus mapping a combination of extreme conditions onto a sort of extreme result. Russell extended the approach on the case of drafting prospective football players[8,25], coming up with the composition of battery tests that predict being drafted best. Therefore, applying extreme modelling for match performance forecasting seems promising, yet we are not in a position to decide which of the three classes of distributions is suitable. Moreover, we consider match performance subject to many more unobservable factors than in the above-mentioned cases — that is why we have chosen to start with a direct, linear relationship between extreme metrics observed in a match vs extreme metrics observed in preceding training history.

We start our study with discussion of available data, from simple global positioning system (GPS) players' locations, to specialized indices calculated automatically but based on discipline-specific expert knowledge. Next, we proceed in Models section to additional, sometimes extensive preprocessing of such data, before feeding them into relatively simple linear predictive models. Results section discusses quality of predictions from GPS data, against those derived from domain-specific preprocessing tool. Finally, we conclude reaults and outline prospective future work areas.

Materials and Methods

Available data

For the purpose of models construction and experimentation, we used solely training and match data collected by Raków Football Academy. The data analysis was carried out based on training and match information of players of the professional football academy of the Extra League Club RKS Raków Częstochowa (first level of the competition) collected from January to June 2023. The study covered 29 professional players reserve teams, including 5 goalkeepers, 8 defenders, 14 midfielders and 2 forwards. The mean age was 18.3 +/- 3.3 years, the mean height was 185.4 +/- 5.4 cm, and the mean weight was 78.2 +/- 2.7 kg. The analysis included data from training and matches (control and championship, resulting in the average number of 9,000 rows of data per month. A standard working week at the Academy they consisted of 5 training units, 1 match and 1 day off.

Privacy details. Athlete's heart rate and GPS location were the raw data collected in real time; weight was measured periodically and with player's general consent stated in the contract with the Academy. All data were collected passively, which means that our experiments did not affect players' activity in any way. The authors declare that they will provide a representative and anonymized subset of the data upon the express request of interested parties. The person responsible for this matter is the corresponding author.

Ethics statements.

- The authors declare that all methods presented in this study are compliant with relevant guidelines and regulations of their affiliated institutions.

- The procedures used in this study adhere to the tenets of the Declaration of Helsinki. Approval was obtained from the bioethics committee at the District Medical Chamber in Krakow (approval number: No. 35/KBL/OIL/2024; approval date: 24 April 2024).

- A general informed consent has been obtained from all study participants and/or their legal guardians, at the time of their recruitment to Rakow Academy, to utilize personal data for analytics purposes.

GPS traces

Raw GPS data get collected for all considered players, and during all training and competition activities, with Apex Pro Series, STATSports, Premium System 2023, Sonra 4.0, Northern Ireland system. Position and instantaneous acceleration are sampled with 10 Hz and 100 Hz frequency, respectively, by the biometric vests. Let us define a sample taken at time t by a tuple

$$s = (t, t+\tau, A, \mathbf{q}) \, . \tag{1}$$

Each such measurement spans a period τ, i.e. contains all information for interval $[t, t+\tau)$. Each measurement record possesses a set of attributes A that carry any extra information about the measurement: player ID, player role, activity type etc. Set A together with $(t, t+\tau)$ identify each sample uniquely. Vector $\mathbf{q}$ contains the measurements proper. In our work we basically use instantaneous velocity at t, denoted q_v, and utilizing position and acceleration only for measurement validation purposes.

Apex Pro Series metrics

In parallel, GPS measurements undergo vast processing by Apex Pro Series software, resulting in a database of records of structure essentially identical as in (1) but with much richer measurement vector $\mathbf{q}$, containing various aggregations of player's activity over a period of τ, which can be configurable and span many time scales, according to user's wish.[26]. Typically, aggregated metrics which are of interest to coaches, cover activities performed with high metabolic load. Also, plain extreme values, e.g. of speed over a period τ, can be found useful by individual coaches, depending on their personal approach.

Considering the current practice at Raków Academy, we can point out measurements of the highest utility to our models as follows:

- Total Loading — Using accelerometer data alone gives a total of the forces on the player over the entire session without any weightings being applied.

- Total Acceleration/Deceleration — The ratio of the total number of accelerations to decelerations.

- Metabolic Time — Time parameters - estimating the instantaneous running time and its metabolic power requirement for any athlete.[27,28]

- HML Distance — the distance covered with high metabolic load;

- HML Efforts — count of HML episodes;

- HML Time — time spent in HML state (with possibility of breaking up HML state into zones, e.g. 4, 5 and 6, standing for increasing effort);

- Distance Zone — the distance covered in a specific zone (e.g. 4, 5 and 6, standing for increasing speed).

The contents of $\mathbf{q}$ is up to Apex Pro Series user and can be expandable and configurable. The number of all isolated measurements was 94, cf. **Appendix A** and the sample database analyzed is in **Appendix B**.

Additionally, the original measurements from Apex Pro are annotated with the type of Filtration used in the "drill title" column were taken in, e.g. Sprint Training, Small Games, Game Fragments, Supporting Game or Match Day. We refer to the specific performance type whenever necessary while presenting results. Moreover, selected measurements get divided by the distance covered in $[t, t+\tau)$, in order to address cases when the player was active only partly in the time covered by a sample. In such cases, we append the measurement time with "Per Distance" suffix.

Models used

Throughout this work we resort steadily to a linear regression model with positive coefficients. The reasons are a few, but first and foremost, it is due to the lack of data. Strangely as it may seem, the prediction model of player's performance in the match ultimately involves as many samples as that player's matches — regardless of how frequent his activity is performed. And, naturally, a linear model is the most indulgent to scarcity of samples. Apart from that, it is explainable, thus making it possible to map a success in a play onto particular samples (and, in fact, training exercises) which contributed most. Finally, we have not seen so far evident patterns in data in favor of modeling kernels other than the linear one. However, we bear in mind that non-linear and negative impact of the training effort on performance in competition is a established fact[29] and that it will have to be accounted for by the model as soon as enough data become available.

Our linear multivariate model can be written as

$$\hat{y} = a_0 + \mathbf{a}^{\mathrm{T}}\mathbf{x} \tag{2}$$

with $\hat{y}$ being the predicted performance index during a match, a_0 is the model intercept value, and $\mathbf{a} \geq \mathbf{0}, \mathbf{x} \geq \mathbf{0}$ are non-negative model weights and inputs, respectively.

Finding prospective models as presented in (2) in the collected data is one of our main contributions. It boils down to the problem of transforming the set of samples $\{s\}$ into a dataset of valid inputs $X = \{\mathbf{x}\}$ and outputs $Y = \{y\}$. It can be done in a number of ways; we describe the two general approaches in the following sections.

Pre-match metrics aggregation

Here, X and Y are constructed solely from Apex Pro Series metrics, with general aim to use training samples from H days preceding a match day d to form a single input vector $\mathbf{x}_d$ that predicts y_d — a performance index in a match on day d, made of samples from that match only. Both procedures can be written down formally

$$\mathbf{x}_d = \mathscr{T}(s : d - H \leq t < d)\,, \quad y_d = \mathscr{M}(s : t = d)\,. \tag{3}$$

Transformation $\mathscr{M}$ consists in selection of one match metrics of interest, q_i, whose values get aggregated, usually by summing them up or calculating the maximum for a given match day. As an option, such aggregates can undergo normalization to duration of player activity in the match or a distance covered. Occasionally, we also aggregate two or more metrics that are complementary, e.g. HML time which has been split by Apex Pro Series across predefined intensity levels.

Transformation $\mathscr{T}$ is done in the following stages:

1. Averaging metrics of repeated activity in a microcycle.

   In base regular training schedule, there is only one sample with unique attributes A in a microcycle (which is, typically, a week). However, due to frequent exceptions, players complete more than one training session or match of the same type in a week. Straight adding up the metrics in such case would be misleading because the higher overall value does not result from better results, but simply from the player's longer exposure. Therefore, the results, i.e. samples, for a specific training or match are averaged in a week-long moving window.

2. Introduction of weekly HMLD totals.

   The modeling should also take into account the players' overall exposure to exercise over a microcycle. Therefore, it was decided to create secondary samples indicating the athlete's total effort. Based on expert knowledge, HMLD parameter was the metrics of the choice. Its value is summed up for the entire week, including every training and match activity.

3. Data aggregation with forgetting factor α.

   To predict match performance, data that precedes the match by four weeks is used. However, it is assumed that the older the data, the less impact it should have on the match values, which leads to the following weighted aggregation procedure $\mathscr{T}$ for each element i of input $\mathbf{x}$ on match day d

   $$x_{id} = \frac{\sum_{k=1\ldots4} \alpha^k q_n}{\sum_{k=1\ldots4} \alpha^k}\,, \tag{4}$$

   where q_n is a selected metrics of a sample with fixed attributes A, as in (1), and k iterates weeks preceding match day d, selecting samples only in current week. This allows to normalize the data and omit weeks in which the player did not complete a given training session or data are missing. We assumed the period to be 4 weeks, the usual macrocycle duration.

   If a given player has not completed any activity selected by (A, q_n) in the four weeks preceding the match, the most recent historical value is taken as substitute.

4. Selection of non-redundant model inputs.

   Data preparation described so far results in vector $\mathbf{q}$ in each sample containing 536 elements which can potentially be used as model inputs. This number of features is definitely too big w.r.t. the number of samples that can be used for model creation. So it was decided to find pairs of features that correlate with each other at a level above the threshold. In our case it was 0.99. The Mutual Information method was used for this purpose. For each such pair appearing in the dataset, one of the features was randomly removed.

5. Separate data scaling for each player.

   During the research, two modeling approaches were examined. One, called APX-Ind, is to make individual models for each player; the other, APX-Grp, is to devise a common model for a group of players (e.g. in a given position in the pitch). In the case of the common model, it was decided to standardize samples of each player separately, i.e. get model inputs referred as $\mathbf{x}$ in Eq. (2) as well as output y replaced by their transformed values, by applying so-called standard scaling procedure:

$$\mathbf{x}'_{ij} = \frac{\mathbf{x}_{ij} - \overline{\mathbf{x}}_j}{\sigma_j} , \quad y'_{ij} = \frac{y_{ij} - \overline{y}_j}{\varsigma_j} , \tag{5}$$

   with i indexing a sample for a player j. Symbols $\overline{\mathbf{x}}$, $\overline{y}$ with bars denote mean values for a player, σ and ς denote standard deviations. Such scaling can be performed even when only few samples are available per player, but collectively their number is sufficient to make a model for a group of players.

GPS traces aggregation

In case of GPS traces being taken to form model input, the overall processing scheme follows the one formulated in (3), but with a single sample consisting merely of velocity v measured at time t

$$s = (t, v) \tag{6}$$

because τ is constant and equal 0.1 s, and attributes A are absent except for player ID, which can be omitted since we make players' individual models. The general idea of aggregation $\mathscr{T}$ is to account for extreme effort events found in history window of H days prior to a match. Therefore, we search for intervals in history, i.e. a set of consecutive in time sub-sequences of $\{s\}$

$$S(t_{min}, v_{min}, d, H) = \{S_i\} \text{ where } S_i = \{(t,v) : d - H \le t < d, v \ge v_{min}\} \text{ and } \max_{S_i} t - \min_{S_i} t \ge t_{min} . \tag{7}$$

Such approach is a clear analogue to HML-related Apex Pro Series metrics, with the difference that we do not assume any specific threshold parameter values but leave them for exploration as parameters v_{min} of minimum continuous velocity maintained for interval of at least time t_{min}. Next, we calculate scalar model input x_d, consistently with (3) as the count of found training intervals

$$x_d = ||S(t_{min}, v_{min}, d, H)|| . \tag{8}$$

Transformation (8) can be applied over a grid of parameter values of interest, $\mathbf{t}_{min} \times \mathbf{v}_{min}$, in search for pairs that make good prediction of match metrics, much as it was described in the preceding section. Let us call the model approach with $\mathscr{T}$ defined as in (8) the base GPS model, GPS-Cnt, where 'Cnt' stands for 'count'.

GPS-Cnt applied over $\mathbf{t}_{min} \times \mathbf{v}_{min}$ provides, in fact, bivariate complementary cumulative distribution function (ccdf) for intervals over period $[d-H, d)$ for a given player. An example of such ccdf is drawn in Fig. 1a with contour plot, and as such carries much synthetic information valuable to a coach about player's spontaneous or forced ability to perform fast and long sprints defined by (t_{min}, v_{min}).

Probabilistic interpretation of training intervals makes it possible to build two extra $\mathscr{T}$ routines: instead of returning *count* of intervals for (t_{min}, v_{min}), we calculate

- GPS-Time: value of t_{min} corresponding to p-th percentile of samples for a given v_{min};
- GPS-Vel: value of v_{min} corresponding to p-th percentile of samples for a given t_{min}.

Exemplary contour plots of t_{min} and v_{min}, for $p \in \{.5, .75, .9, .95, .98, .99, .999\}$, are shown in Fig 1b and c, respectively. One can easily notice that far parts of the distribution tails contain only a couple of intervals, and hence are risky to base any reliable model upon.

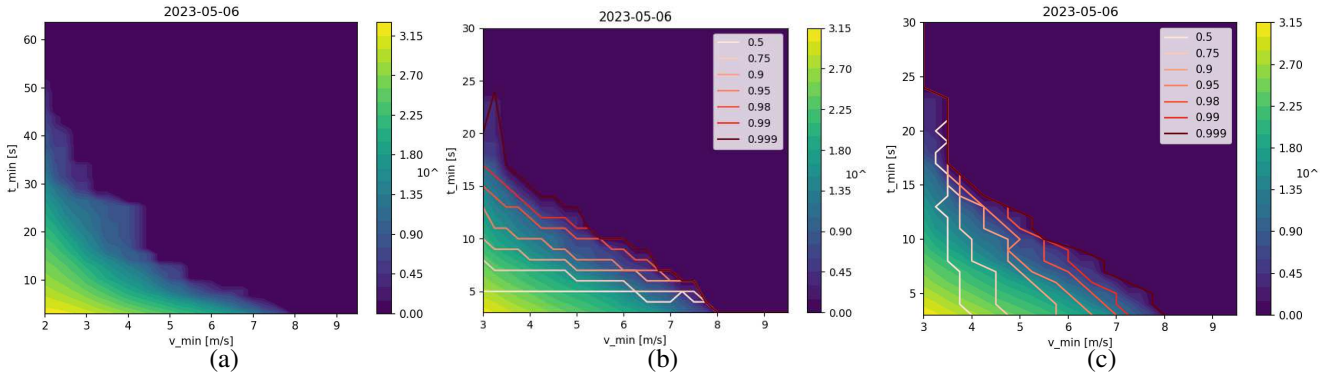


Figure 1. Sample ccdf for 3-week training period before match day on 2023-05-06 — (a) pure count of intervals, log scale, (b) with percentile contours of t_{min}, (c) with percentile contours of v_{min}.

Ultimately, we can apply the same approach as in (6)–(8) to match samples, which yields

$$y_d = ||S(t'_{min}, v'_{min}, d+1, 1)|| \ , \tag{9}$$

i.e. the count of intervals on match day w.r.t. an arbitrary (t'_{min}, v'_{min}) which in general can differ from (t_{min}, v_{min}), hence the prime notation. With y_d defined so, we check performance of models for input and output defined by a tuple $(t_{min}, v_{min}, t'_{min}, v'_{min})$. Let us call this approach GPS-X, for cross-checking (t_{min}, v_{min}) against (t'_{min}, v'_{min}).

Results

All the presented approaches have been tested in the same setting, in a group of players of age 17 to 21, in order to predict match performance in spring 2023. The number of matches per player with acceptable performance data was 10 to 16; out of which 75% was used for model training, and the rest of data for verification. Training data correlation was measured with R^2 score. Predictive quality of models was scored on the test data with mean absolute prediction error (MAPE).

Models based on Apex Pro Series metrics

APX-Ind

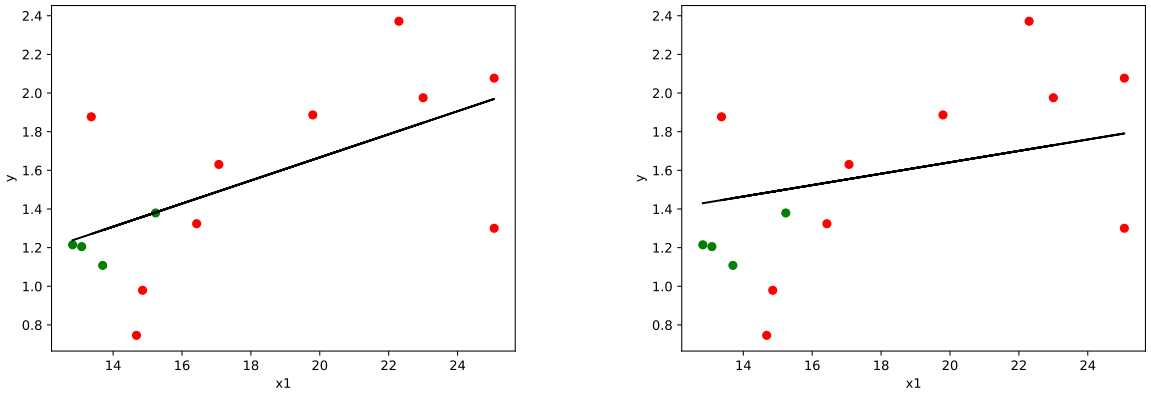
The best individual model that predicts one match performance index based on one training index is the model forecasting Total Deceleration Loading Per Distance with input as Total Loading Per Distance in Sprint Training, for player referenced here as A1. This model was obtained for α parameter of 0.5, which resulted in MAPE score of 0.137 and R^2 score of 0.374, see Fig. 2a. The individual player appears to be very stable — checking it against other values of α parameter resulted in MAPE values in the range of 0.209–0.286, cf. data in row 1 of Tab. 1.

The best model that predicts one match performance indicator based on two training indices is the above model but with extra input of Metabolic Time Per Distance of Small Games Training in Zone 6, cf. Fig. 3a. Such best model was also obtained for α of 0.5, which led to MAPE of 0.097 and R^2 of 0.445. This two-input model turns out to be less stable — checking it against other values of α resulted in MAPE in the range of 0.357–0.749, see Tab. 1, row 3.

		$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
1 input	APX-Ind	0.264	0.209	0.137	0.219	0.286
	APX-Grp	0.502	0.532	0.574	0.615	0.647
2 inputs	APX-Ind	0.749	0.357	0.097	0.445	0.720
	APX-Grp	0.502	0.532	0.574	0.615	0.647

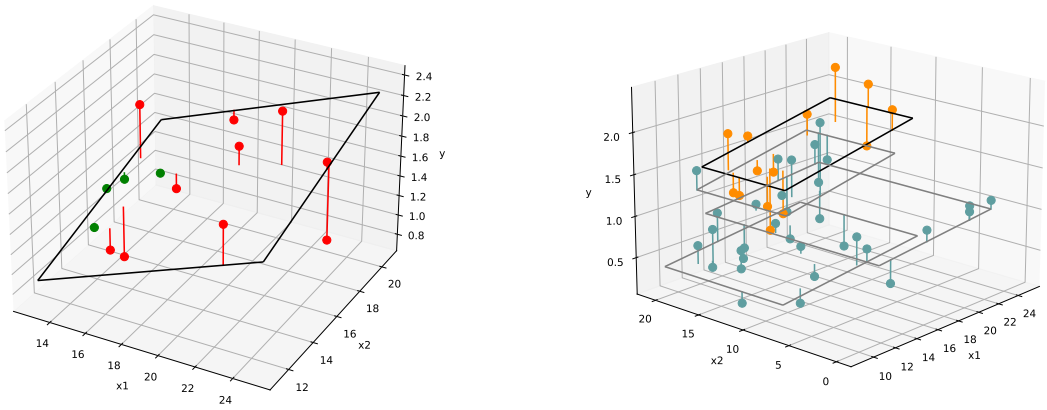
Table 1. MAPE values from the best model with one and two inputs for adjacent alpha values

The examination of complete space of possible modeled indices vs. possible model inputs reveals also good predictive models based on previous match performance rather than training history. Fig. 4 provides such one for player named A2. The model was obtained with α equal 0.1 and resulted in R^2 score of 0.84. This result is very good, moreover, it proves that big variability of match indices provided in Apex Pro Series is not due to noise but is caused by some deterministic, unobserved yet persistent in time window, factors. However comforting such claim may seem, it is of little practical utility for coaches because match physical performance, unlike the training, is not subject to systemic control and direct planning.



(a) (b)

Figure 2. Best models predicting one match performance index — y as a Total Deceleration Loading Per Distance, based on one training index — x_1 which is Total Loading Per Distance for Sprint Training, for a) individual modeling (APX-Ind) and b) group modeling (APX-Grp) for player A1. Training data are marked red, test data are marked green, with models represented by black lines.



(a) (b)

Figure 3. Best models predicting match performance index — y as Total Deceleration Loading Per Distance, based on two training indexes — x_1 as Total Loading Per Distance for Sprint Training, and x_2 as Metabolic Time Per Distance in Zone 6 for Small Games Training. Figure a) visualizes model in 2D in case of individual modeling (APX-Ind) of a player. The training data are marked in red, test data are marked in green and the model is represented by a sloped plane. Figure b) shows group modeling (APX-Grp), where A1 player's data are marked in orange with the model represented by a plane with black rim, and other players' data are marked in blue with models represented by the remaining planes. Data points are pinned to planes representing respective individual models.

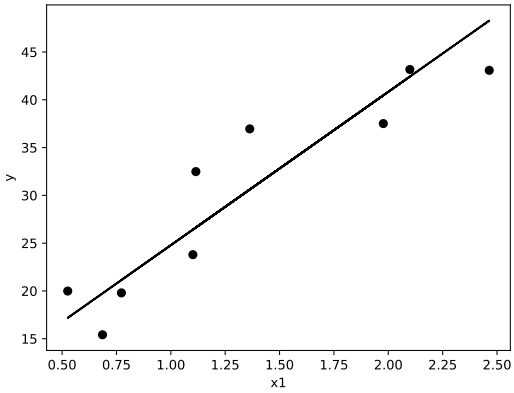


Figure 4. The best model predicting selected match performance index — y as a Distance in Zone 6, based on another match index for previous matches — x_1 as Accelerations Total Time Per Distance in Zone 6 . Data are marked as black dots with the model represented by a black line.

	R^2			MAPE		
	$H = 14$	$H = 21$	$H = 28$	$H = 14$	$H = 21$	$H = 28$
GPS-Cnt	0.77	0.36	0.30	0.10	0.07	0.14
GPS-Time	0.56	0.32	0.22	0.16	0.11	0.11
GPS-Vel	0.53	0.45	0.37	0.14	0.09	0.12
GPS-X	0.45	0.35	0.31	0.15	0.15	0.15

Table 2. Comparison of performance for best GPS-based models, for different approaches and training history taken into account.

APX-Grp

With collective modeling approach, the common model eventually gets adjusted to individual range of data by means of standard scaling (5). This has been visualized in Figs. 2b and 3b by lines and planes sloped differently than corresponding individual models (Figs. 2a and 3a). Inevitably, such models performance is also inferior in terms of MAPE, cf. adjacent 'individual' and 'collective' rows in Tab. 1. However, such model can be used as last resort in cases when there are no match performance data for a player (e.g. a newly recruited one, or returning from recovery), yet prediction of any quality would be better than none.

Models based on GPS

Due to high variability of match metrics caused by many controlled factors (opponent's strength, own strategy etc.) that have not been captured by Apex Pro Series, and few match data it is impossible in the current stage of research to generalize on a predictive model performing well for all considered players. This is why we consistently resort to individual models and here present best models obtained for player named A3 who, not by chance, played in most matches and attended most training sessions of all players considered. Performance of the models have been compiled in Tab. 2. The overall results are acceptable in quantitative terms, showing that it is possible to point out decently correlated training and match data, as it is possible to find fine quality predictions. Observed values of R^2 score show that the shorter training history H is considered, the better alignment of GPS-based intervals with match metrics is obtained. However, this does not translate into prediction quality as indicated by MAPE — therefore, the both model assessment metrics are incoherent. The results are shown in detail in the sections that follow.

GPS-Cnt

Models fed with intervals created in range of $\mathbf{t}_{min} \times \mathbf{v}_{min}$ have been checked against selected Apex Pro Series match metrics representing well extremes of player performance, i.e.:

- HML Distance,
- HML Efforts Max Speed,

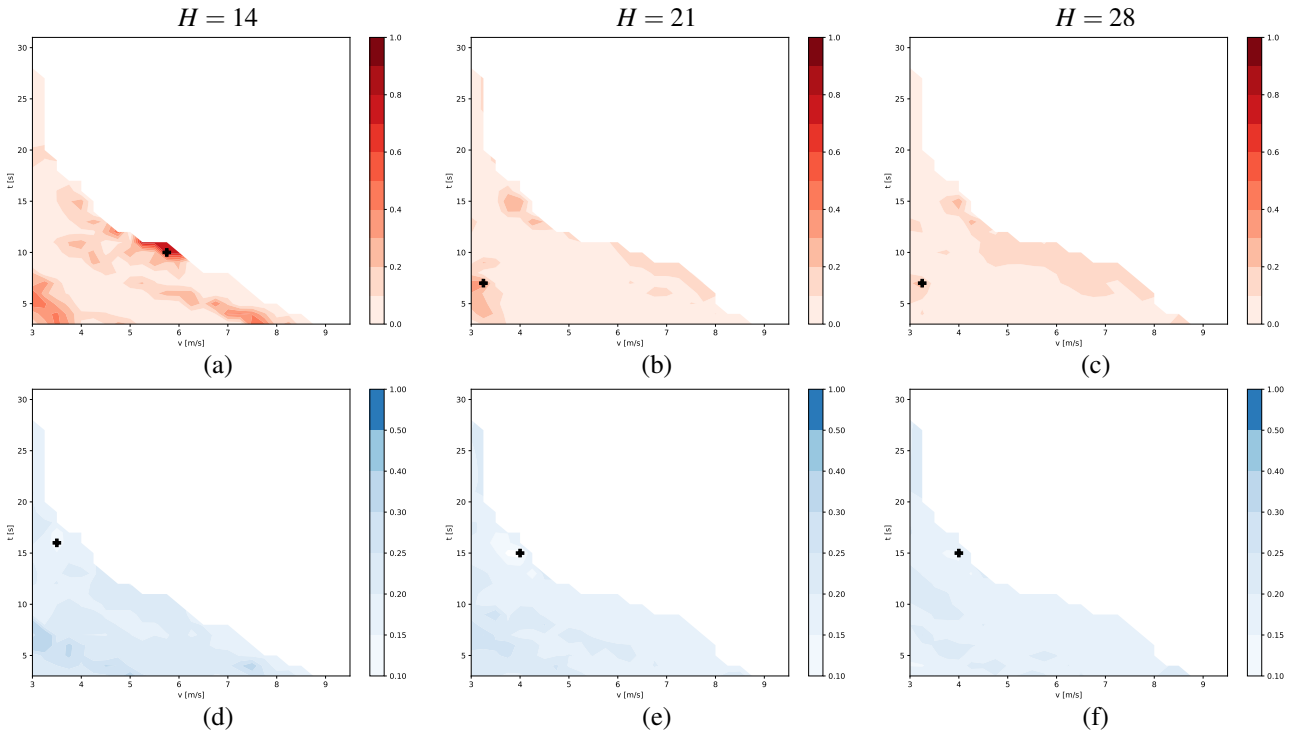


Figure 5. Best models found by approach GPS-Cnt: R^2 score in (a)–(c), MAPE score in (d)–(f). Interval parameter values yielding best models are indicated with small crosses.

- HML Time.

Match metrics HML Distance have turned out to yield the most promising model for $H = 14$, as shown in Fig. 5a, with $R^2 = 0.77$ for intervals defined by $v_{min} = 5.75\ m/s$ and $t_{min} = 10\ s$. Models for data from longer history, $H = 21$ and $H = 28$, are much less correlated with any of match metrics, yet such correlation is still considerably better for extreme velocities and interval durations (save for artifacts seen at plane origin). MAPE errors have been calculated for the last 20% of available samples, with the best prediction quality for intervals of much smaller, $v_{max} \approx 4m/s$, than for the best R^2 model. However, t_{max} for all MAPE models still remains extreme.

Out of the other two considered match metrics, HML Time results proved to be very similar to those presented in Fig. 5, yet HML Efforts Max Speed tends to be much less correlated with training history.

GPS-Time

The concept behind GPS-Time approach is a minor generalization of GPS-Cnt because the interval duration is not an absolute value anymore but gets expressed by percentile of t_{max} distribution for a given v_{max}. Assessment of models models found this way is provided in Fig. 6, resulting again in best results for modeling HML Distance with $H = 14$, the other models being substantially inferior. Interestingly, the highest R^2 is found for really maximal interval time at considerably high speed of 6 m/s. And the minimal prediction error, MAPE, is located not far from the above location — cf. Fig. 6a and d.

GPS-Vel

The exploration of time vs. speed percentile space gave most consistent findings, with all best R^2 and MAPE-related parameters located withing a small area of t_{max} from 12 to 17 seconds and percentile of v_{max} from 0.75 to 0.95 — cf. Fig. 7. The modeled match metrics is still HML Distance. Moreover, in the graphs there appears to be no alternative area of high correlations or good predictions, which leads to the conclusion that the phenomenon capture by the model is by no means an accidental one.

GPS-X

Assessment of GPS-X, which is a freestyle approach with no expert knowledge provided via StatSport involved, should be done with caution. Fig. 8 presents how training and match intervals chosen w.r.t. (v_{min}, t_{min}) correlate. The arrows map particular (v_{min}, t_{min}) to (v'_{min}, t'_{min}) that result in highest R^2. Here one can see a clear shift between training samples used for model input and match samples used for model output, suggesting that extreme training intervals correlate well with match

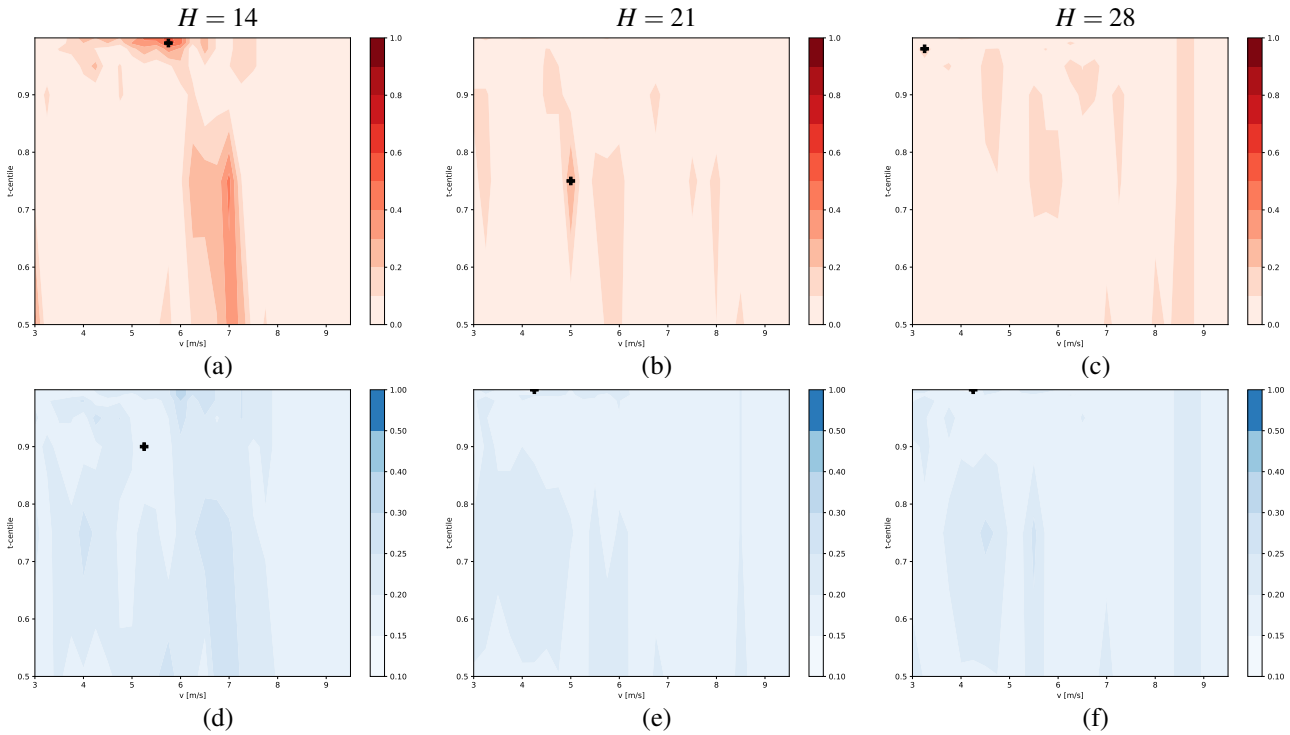


Figure 6. Best models found in approach GPS-Time.

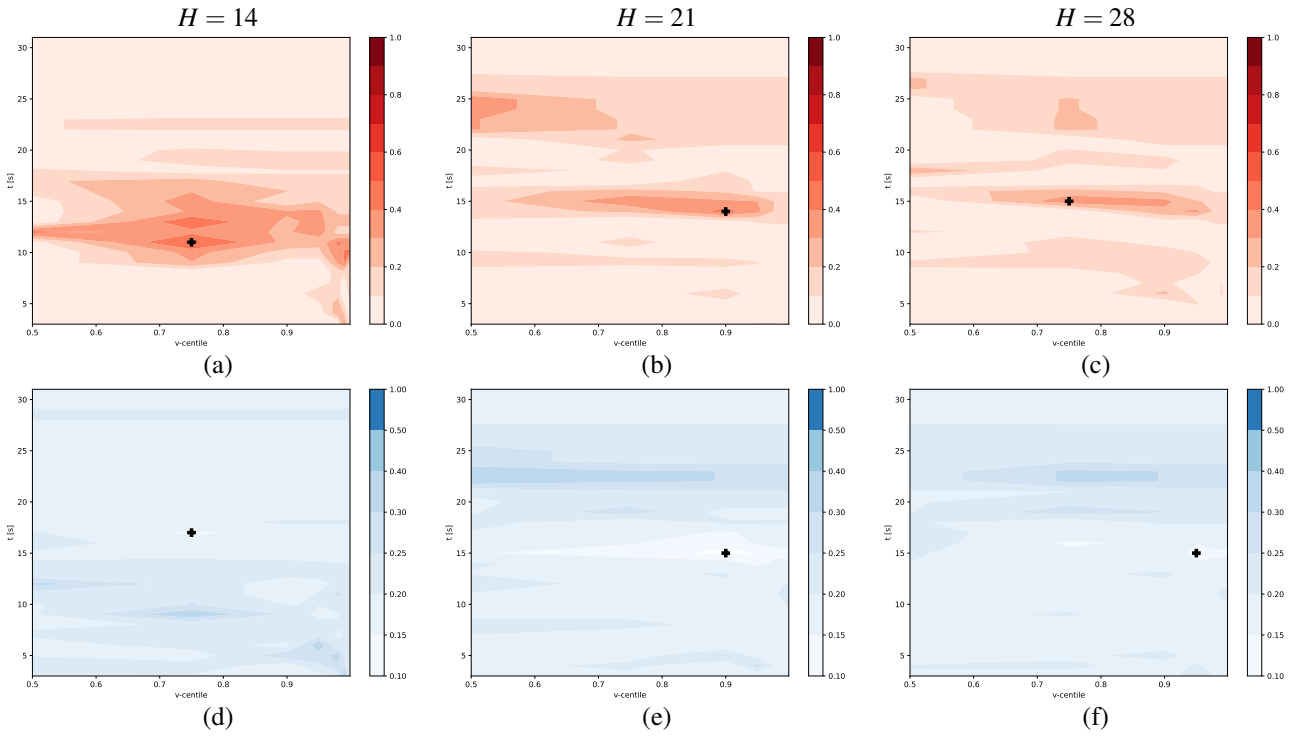


Figure 7. Best models found in approach GPS-Vel.

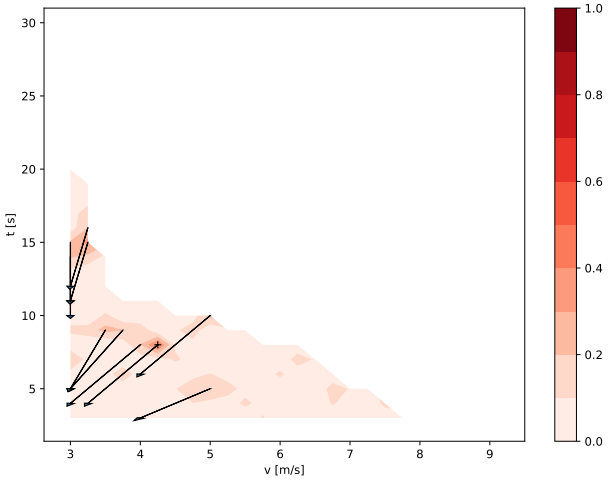


Figure 8. R^2 for linear models found by approach GPS-X.

intervals less intense by 0.7 m/s and by 3 s. Undoubtedly, the phenomenon should be verified both by domain experts, especially by mapping those intervals on particular match fragments, undergoing further scrutiny.

Discussion

The presented study's main contribution is comparing match outcome prediction models for training data with different structures. It is divided into various and innovative methods of data preparation, also taking into account the specificity of training cycles particular to a given study group[30,31]. Also new is the automatic search for the best input type for GPS and pre-aggregated data. Our work led to the selection of models based on predefined measures developed in the APEX system and raw GPS data, which are stable in their parameter space, although their development required significant restrictions.

Choosing a linear model over more complex, nonlinear structures such as neural networks comes from several key factors. The first is the limited amount of data available. Despite collecting data for months, the dataset size is not large enough. Introducing more complex models for small amounts of data may lead to overfitting, which will negatively impact the model's ability to generalize[32]. Another problem is the high degree of match variability. Matches depend on many variables, such as the form of the team, game tactics, and above all, randomness, widely described in the literature[33], i.e. non-deterministic. Despite its simplicity, the linear model may prove to be more stable by minimizing the response to short-term fluctuations. A critical factor in predicting soccer match outcomes is successfully incorporating domain knowledge into the machine-learning modelling process[34]

Building a standard linear model for an entire soccer team or subgroup, as in the APX-Grp approach, can be misleading and should be used cautiously and only when not enough individual data is available. We have encountered situations where the relationships in unique models contradict those that reveal joint modelling for the group, perfectly illustrating Simpson's paradox[35,36]. The experience reminds us to be careful when interpreting soccer data because relationships observed for a group of players usually refer to momentary relationships between players, not temporal relationships in player performances. Therefore, the article focuses on individual models.

The quality of the obtained models strongly indicates a relationship between pre-match activity and in-match performance. Still, this relationship is not accurate enough to predict match performance rates accurately. The resulting models are, therefore, descriptive rather than predictive. However, they have considerable practical utility because they give the coach qualitative indications of which training periods (and thus which drills) contribute most to extreme match performance. Such knowledge enables conscious and aggressive experimentation with the training puzzle mix, ultimately leading to better fitting results and, last but not least, more variable and valuable training data, thus improving the model over the cycle. We believe that our study can also be applied to other sports. There are scientific reports on similar work in the field of predicting the final result in swimming competitions using known statistical methods[37], as well as reports related to direct correlation with the planned training work[38]. In the latter case, the lack of a description of the scheme or the algorithm's structure does not allow for a precise comparison[39]. The authors are aware that commercialising such a fully automatic solution with a self-learning option is desirable in fields such as the military or medicine. This may directly impact such restrictions related to access to technical details and prevent the copying of elitist solutions.

Limitations in research

The limitations identified in the analyzed work mainly concern the use of linear models, which may not be able to capture the entire complexity of match data. This situation results from the limited availability of comprehensive data sets, which results in the selection of more straightforward analytical methods. Another significant limitation is the high degree of variability in sports data, which makes it challenging to create generally applicable models. Additionally, there is a risk of overfitting when applying complex models to small data sets, leading to incorrect conclusions and predictions. Considering these limitations in further research is crucial to developing more effective analytical methods in sports. It is also not confirmed that collecting data about position (averaging) is the correct direction. Depending on the sorting level and individual character, they may differ.

Practical application

Practical applications of work results can significantly improve sports results thanks to the individualization of training requirements by finding individual weights of key parameters, giving mathematical chances for predicting key and easily measurable parameters (measures) related to load or intensity as well as the quality of work. In our research, these were some extreme values. These models can also help optimize match lineups and manage game strategies in real-time. Thanks to the ability to adapt models to various sports disciplines, their application may extend beyond the original research areas. Using advanced data analysis techniques in sports opens up new opportunities for coaches and analysts to explore previously unused strategies. In the long term, the practical application of these models may contribute to a revolution in preparing and conducting sports competitions. It also has inevitable consequences in terms of other requirements of people involved in sports analysis and a change in thinking from explaining based on historical data to assessing the probability of selected training methods in relation to their impact on the final effect and, in our case, the result of the sports competition.

Further research directions

Future research on the prediction of match extremes may focus on exploring more complex nonlinear models and neural networks that better reflect the dynamics of sports data. Extending the database with additional variables, such as weather conditions, players' physical condition or detailed match statistics, can significantly increase the precision and reliability of forecasts. Combining dynamically changing patterns of pairs, threes or fours into patterns corresponding to a specific positioning of the players in relation to the ball or the opponent will allow for additional analysis of the team's work efficiency in terms of the number of tactical errors (deviating from the definition, template, average value, etc.). Interdisciplinary approaches combining knowledge from sports psychology, biomechanics and computer science can open new perspectives for research. Working to improve data analysis methods can also benefit from a better understanding of how various factors influence athletic performance. Finally, future research directions should also focus on developing analytical tools that are easier to use for sports practitioners, enabling them to benefit directly from advanced data analysis methods. A critical element is the popularization of such tools so that an increasing number of coaches and players themselves can influence the evaluation of this type of approach that's easy to implement.

Conclusions

1. Observations of training parameters generated in short-term intervals are more effective and correlated with extreme match results than long-term dates.

2. The correct direction of analysis is to individualize prognostic models to more accurately predict sports results.

3. Specific training parameters may be key in predicting exceptional football player performance, but they may also vary from person to person.

4. The results suggest optimizing training strategies can significantly improve extreme match performance.

References

1. Wunderlich, F. & Memmert, D. A big data analysis of Twitter data during premier league matches: do tweets contain information valuable for in-play forecasting of goals in football? *Soc. network analysis mining* **12**, 1–15 (2022).

2. Ortiz, J. G., De Lucas, R. D., Teixeira, A. S., Mohr, P. A. & Guglielmo, L. G. A. Match-play running performance in professional male soccer players: The role of anaerobic speed reserve. *Res. Q. for Exerc. Sport* 1–8 (2024).

3. Dick, U. & Brefeld, U. Learning to rate player positioning in soccer. *Big data* **7**, 71–82 (2019).

4. Djaoui, L., Chamari, K., Owen, A. L. & Dellal, A. Maximal sprinting speed of elite soccer players during training and matches. *The J. Strength & Cond. Res.* **31**, 1509–1517 (2017).

5. Tam, C.-K. & Yao, Z.-F. Advancing 100m sprint performance prediction: A machine learning approach to velocity curve modeling and performance correlation, DOI: 10.31219/osf.io/rx5fz (2024).

6. Rein, R. & Memmert, D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus* **5**, 1–13 (2016).

7. Gentilin, A. The informative power of heart rate along with machine learning regression models to predict maximal oxygen consumption and maximal workload capacity. *Proc. Inst. Mech. Eng. Part P: J. Sports Eng. Technol.* DOI: 10.1177/17543371231213904 (2023).

8. Russell, B. T. & Hogan, P. Analyzing dependence matrices to investigate relationships between national football league combine event performances. *J. Quant. Analysis Sports* **14**, 201–212 (2018).

9. Grunz, A., Memmert, D. & Perl, J. Tactical pattern recognition in soccer games by means of special self-organizing maps. *Hum. movement science* **31**, 334–343 (2012).

10. Garganta, J. Trends of tactical performance analysis in team sports: bridging the gap between research, training and competition. *Revista Portuguesa de Ciencias do desporto* **9** (2009).

11. Sun, H.-C., Lin, T.-Y. & Tsai, Y.-L. Performance prediction in major league baseball by long short-term memory networks. *Int. J. Data Sci. Anal.* **15**, 93–104 (2023).

12. Albert, J. Sabermetrics: The past, the present, and the future. *Math. sports* **43**, 15 (2010).

13. Noel, J. T. P., Prado da Fonseca, V. & Soares, A. A comprehensive data pipeline for comparing the effects of momentum on sports leagues. *Data* **9**, 29 (2024).

14. Thabtah, F., Zhang, L. & Abdelhamid, N. NBA game result prediction using feature analysis and machine learning. *Annals Data Sci.* **6**, 103–116 (2019).

15. Pischedda, G. Predicting NHL match outcomes with ML models. *Int. J. Comput. Appl.* **101** (2014).

16. Horvat, T., Job, J., Logozar, R. & Livada, C. A data-driven machine learning algorithm for predicting the outcomes of NBA games. *Symmetry* **15**, 798 (2023).

17. Goldsberry, K. Courtvision: New visual and spatial analytics for the NBA. In *2012 MIT Sloan sports analytics conference*, vol. 9, 12–15 (2012).

18. Washif, J., Pagaduan, J., James, C., Dergaa, I. & Beaven, C. Artificial intelligence in sport: Exploring the potential of using ChatGPT in resistance training prescription. *Biol. Sport* **41**, 209–220 (2023).

19. Coscia, M. Which sport is becoming more predictable? A cross-discipline analysis of predictability in team sports. *EPJ Data Sci.* **13**, 8 (2024).

20. Apostolou, K. & Tjortjis, C. Sports analytics algorithms for performance prediction. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–4 (IEEE, 2019).

21. Bunker, R. & Thabtah, F. A machine learning framework for sport result prediction. *Appl. Comput. Informatics* DOI: 10.1016/J.ACI.2017.09.005 (2019).

22. Jain, P., Quamer, W. & Pamula, R. Sports result prediction using data mining techniques in comparison with base line model. *OPSEARCH* **58**, 54–70, DOI: 10.1007/s12597-020-00470-9 (2020).

23. Cooley, D., Hunter, B. D. & Smith, R. L. Univariate and multivariate extremes for the environmental sciences. *Handb. environmental ecological statistics* 153–180 (2019).

24. Russell, B. T., Cooley, D. S., Porter, W. C., Reich, B. J. & Heald, C. L. Data mining to investigate the meteorological drivers for extreme ground level ozone events. *The Annals Appl. Stat.* **10**, 1673 – 1698, DOI: 10.1214/16-AOAS954 (2016).

25. Wunderlich, F. & Memmert, D. Forecasting the outcomes of sports events: A review. *Eur. J. Sport Sci.* **21**, 944–957 (2021).

26. Beato, M., Wren, C. & de Keijzer, K. L. The interunit reliability of global navigation satellite systems Apex (STATSports) metrics during a standardized intermittent running activity. *The J. Strength & Cond. Res.* 10–1519 (2022).

27. di Prampero, P. E. & Osgnach, C. Metabolic power in team sports-part 1: an update. *Int. journal sports medicine* **39**, 581–587 (2018).

28. Osgnach, C. & di Prampero, P. E. Metabolic power in team sports-part 2: aerobic and anaerobic energy yields. *Int. journal sports medicine* **39**, 588–595 (2018).

29. Simão, R. *et al.* Comparison between nonlinear and linear periodized resistance training: hypertrophic and strength effects. *The J. strength & conditioning research* **26**, 1389–1395 (2012).

30. Aquino, R. L. *et al.* Periodization training focused on technical-tactical ability in young soccer players positively affects biochemical markers and game performance. *The J. Strength & Cond. Res.* **30**, 2723–2732 (2016).

31. Szymanek-Pilarczyk, M., Nowak, M., Podstawski, R. & Wasik, J. Development of muscle power of the lower limbs as a result of training according to the model of modified tactical periodization in young soccer players. *Phys. Activity Rev.* **11** (2023).

32. Montesinos Lopez, O. A., Montesinos Lopez, A. & Crossa, J. Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction*, 109–139 (Springer, 2022).

33. Saribekyan, G. & Yarovoy, N. Football prediction model based on the teams' Elo ratings and scoring indicators. *Res. Sq.* DOI: https://doi.org/10.21203/rs.3.rs-3861295/v1 (2024).

34. Berrar, D., Lopes, P. & Dubitzky, W. Incorporating domain knowledge in machine learning for soccer outcome prediction. *Mach. learning* **108**, 97–126 (2019).

35. Antequera, D. R. *et al.* Asymmetries in football: The Pass-Goal paradox. *Symmetry* **12**, 1052 (2020).

36. Sarkar, S. Paradox of crosses in association football (soccer) – a game-theoretic explanation. *J. Quant. Analysis Sports* **14**, 25–36 (2018).

37. Mujika, I. *et al.* Next-generation models for predicting winning times in elite swimming events: Updated predictions for the Paris 2024 Olympic Games. *Int. J. Sports Physiol. Perform.* **1**, 1–6 (2023).

38. Eriksson, R., Nicander, J., Johansson, M. & Mattsson, C. M. Generating weekly training plans in the style of a professional swimming coach using genetic algorithms and random trees. In *International Conference on Security, Privacy, and Anonymity in Computation, Communication, and Storage*, 61–68 (Springer, 2021).

39. Mattsson, C. M. Silicon valley exercise analytics case study - Swedish swimming. https://svexa.com/case-studies/swedish-swimming/ (2020). Last access: March 9, 2024.

Acknowledgements

The authors would like to thank all the coaches working every day at the RKS Raków Częstochowa Academy and involved in implementing the project. Special thanks should be given to the people managing the Academy, i.e. Marek Śledz and Dariusz Grzegrzółka, for constantly searching for a sports advantage on the pitch using the potential of science.

Funding

The above research did not receive any funding. The authors performed the tasks in accordance with the tasks and time specified in normal working hours in accordance with their affiliations with universities or clubs. All authors can confirm no conflict of interest in the manuscript.

Author contributions statement

Conceptualization: M.N., M.K., B.B. Methodology: M.N., M.K., B.B. Validation and statistical analysis: B.B, A.W. Investigation: M.K., M.N. Resources: M.N., M.K, B.B. Data curation B.B., A.W., M.K. Ethics: Ł.O. Writing - original draft preparation: M.N., M.K., B.B., A.W. Writing - review and editing: M.N, M.K., Visualization: M.K, B.B. Supervision: M.N., Ł.O., M.K. Project administration: M.N., M.K., All authors have read and agreed to the published version of the manuscript.

Additional information

The authors declare that they will provide a representative and anonymized subset of the data upon the express request of interested parties. The person responsible for this matter is the corresponding author.

Attachments:

A List of parameters generated from the Apex Pro Series, STATSports, Premium System 2023, Sonra 4.0,

B Example table with data after anonymization.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixA.pdf
- AppendixB.xls