



## OPEN Forecasting extremes of football players' performance in matches

Michał Nowak<sup>1,3</sup>, Bartosz Bok<sup>2</sup>, Artur Wilczek<sup>2</sup>, Łukasz Oleksy<sup>4,5</sup> & Mariusz Kamola<sup>2</sup>

This study evaluates the use of simple linear or piecewise linear predictive models to predict extreme performance metrics in soccer matches, based on historical training and to match data of soccer players from RKS Raków Częstochowa football club. The data were collected from January to June 2023. The collected training and matched data average is 9000 records per month. A standard workweek at the RKS Academy consisted of 5 training units and at least 1 match. The best individual models found predict selected game performance metrics with a relative error of 2.3%, suggesting an excellent model fit between prediction and the actual value. This is illustrated by input data metric called “Metabolic Time Zone 5 and 6 Per Distance”, and output data by “Decelerations Total Distance in Zone 5 and 6 Per Distance”—calculated for in 3 min sliding window and characterized by the highest value of the generated parameter based on High Metabolic Load Distance (HMLD). The result concerns models run on aggregated performance metrics developed in APEX-PRO system using expert knowledge in soccer training, while raw GPS location-based models performing worse but still acceptably. Although we believe that the accuracy of the models still has limited reliability, their clarity and up-to-date quality make them useful in the daily planning of training activities and the management of workloads that affect player performance in the upcoming match, as well as the tactical decisions of the coach. More accurate predictions are given by individual models compared to aggregated models (player position), but there are exceptions where group models also perform very well. Adding a second metric to the input did not show a significant difference in the analyzed examples (the results are very similar). Our findings indicate that the model based on metrics from the last match also effectively predict extreme motor performances occurring in the game. In the case of the analyzed player, it was at the input “Accelerations Total Time Per Distance in Zone 6” at the output “Distance in Zone 6”. Specific training or match parameters can be key in predicting exceptional soccer performance, but they can also vary depending on the analyzed player. This confirms the need for further analysis of this issue.

**Keywords** Extreme, Game, Maximal, Model, Monitoring, Peak, Performance, Prediction, Soccer, Sub-maximal, Training

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<sup>1</sup>. 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<sup>2,3</sup>. 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<sup>4</sup>. Adding to the challenges of interpreting this data, is the variety of data types, from basic measurements to advanced performance metrics. One of the main challenges of contemporary sports analysis is the lack of standards in the approach to predicting extreme match events. Most existing studies focus on predicting classic indicators, such as the number of goals, ball possession, or the distance covered by players, omitting more dynamic extreme variables that are crucial for changing the performance of players and, consequently, the final result of the match. Currently, there is a lack of studies that would connect these extreme parameters with tactical–training decisions in an easy and practical way, indicating directions and suggestions

<sup>1</sup>Faculty of Physical Culture Sciences, Jan Długosz University in Częstochowa, Collegium Medicum. Dr. Władysław Bieganski, 42-200 Częstochowa, Poland. <sup>2</sup>National Research Institute NASK, 01-045 Warszawa, Poland. <sup>3</sup>Science Department of the RKS Raków Częstochowa Academy, 42-200 Częstochowa, Poland. <sup>4</sup>Department of Physiotherapy, Faculty of Health Sciences, Jagiellonian University Medical College, Kraków, Poland. <sup>5</sup>Department of Orthopaedics, Traumatology and Hand Surgery, Faculty of Medicine, Wrocław Medical University, 50-556 Wrocław, Poland. ✉email: [michal.nowak@rakow.com](mailto:michal.nowak@rakow.com); [mariusz.kamola@nask.pl](mailto:mariusz.kamola@nask.pl)

for implementation both in planning and during live observation, which is a clear gap in the scientific literature. Focusing on football, one of the most popular sports in the world, the studies considered so far how correlational models could identify potential parameters in sub-maximal zones, called extreme efforts for example Maximal Sprinting Speed<sup>5,6</sup>. The use of Extreme Value Theory (EVT) to analyze these extreme moments describing a player's game in football matches is a pioneering approach that can provide new tools for predicting key periods of the game in which a player, after making extreme efforts, potentially exposes himself to a temporary decrease in the quality of his game (the effect of temporary fatigue) or finding opportunities against the opponent, taking into account the above-mentioned intensive sprints in critical moments of the match, e.g. end of the first or second half. Using this method results from the need to understand better and plan the extreme efforts of players—under the tactical assumption that the implementation of the planned scenario can only be successful if each player acts in optimal conditions exertion for themselves, thus enabling correct decision-making in the ideal time of the game—which can not only improve the team's results but also prevent injuries and optimize fatigue management. In a sport where small differences determine success, such analysis has a huge potential for use by coaches and players themselves. This is confirmed by the work<sup>7</sup> where EVT has proven to be an effective tool for predicting extreme values in sports physiology based on extreme VO<sub>2</sub>max values. EVT was chosen over conventional prediction methods because it allows for the analysis of extreme values, which are crucial in the context of sports performance. Extreme, maximum values, such as speed, strength, or endurance, often determine the outcome of sports competition. Only performances close to the maximum can give an advantage over the opponent, bringing real benefits during competition and deciding about victory. Conventional methods focus on average trends, not considering these key extreme moments determining success. 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<sup>8</sup> concerning 100 m run competition. During this period, it focuses on intensive training, specific exercises aimed at improving physical condition, technique and tactics<sup>9</sup>. 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<sup>10</sup>. 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<sup>11</sup>. Applying statistical models and machine learning to match data allows for the identification of patterns and variables that contribute most to sports success<sup>12,13</sup>. Such approaches have been used in other disciplines, such as baseball, where advanced statistics and sabermetrics analyses are revolutionizing the approach to the game<sup>14,15</sup>, or in National Hockey League (NHL), National Basketball Association (NBA), where spatial analysis and advanced metrics influence strategies and coaching decisions<sup>16</sup> they try to predict the results of matches also in live conditions<sup>17–19</sup>. 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<sup>20,21</sup>. 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<sup>19</sup>. By analyzing various modeling 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<sup>22</sup>. 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<sup>23</sup>. 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<sup>24–26</sup>. Many studies have used player position as a determinant of tactical context and have analyzed various physical metrics by position, including e.g. the occurrence of explosive sprints (with a fast acceleration phase)<sup>27,28</sup>, as well as the overall team performance by game formation<sup>29</sup>. In other studies<sup>30</sup>, they identified significant effects of high-intensity running and sprinting results on fatigue and the periods following them, but according to the authors, they adopted too long time intervals for analysis (5 min), and the zone defining sprint speed was assumed as > 25.2 km/h. This value is the scientifically accepted gold standard, unfortunately general for all. Consequently, it may differ significantly for the extreme values of individual players. In our case, the model itself defines the extreme parameters at exit and entry, and the coach himself can choose which explaining variables from the available set is most reasonable for him in terms of assessing the training and match work of a given player. 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. By using simple linear models, we assume that they have a more significant application dimension, showing that understanding such models is easier and more instructive for a much larger population of coaches or players versus using large and complex predictive models such as those described in the study by Rahimian et al.<sup>31</sup> Graph Neural Networks in Soccer (GNN) used there to model the interactions

between players and the ball to predict events on the pitch (passes or dribbling) as well as Transformer-Based Action Prediction Models based on transformers, show excellent abilities to predict long-term relationships and interactions between players, analyze both spatial and temporal data from soccer matches, used to predict future actions, (passes, shots or defense interventions). Reinforcement Learning (RL) finds many applications in soccer analysis, including valuation (ranking) of actions and optimization of decisions. RL models, such as Conservative Q-Learning, optimize decisions made by players to maximize the chances of scoring a goal. However, the models presented above often call for human competencies that trainers do not possess.

The idea of extreme value modelling has been present in environmental sciences for decades<sup>32</sup>, where models of events of extraordinary magnitude were based on parametric tuning of one of three heavy-tailed distributions: Fréchet, Weibull or Gumbel. In particular, Russell et al.<sup>33</sup> 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<sup>11,34</sup>, 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. Finally, the number of individual training samples is equal to the number of matches played, that is, very small w.r.t. contemporary models' requirements. That is why we have chosen to examine simple, linear or piecewise 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 results and outline prospective future work areas.

## Materials and methods

### The dataset used

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  $18.3 \pm 3.3$  years, the mean height was  $185 \pm 5.4$  cm, and the mean weight was  $78.2 \pm 2.7$  kg. The analysis included data from training and matches (control and championship, resulting in the average number of 9000 rows of data per month. A standard working week at the Academy consisted of 5 training units, 1 match (abbrev. MD) and 1 day off. The strength and conditioning coach (level ASCA2) working with the team was responsible for pre-processing of the data. This involved identifying and removing all incorrect data points, such as an incomplete path or outliers resulting from a temporary loss of GPS signal. The decision was made based on knowledge of the number of available satellites and the signal quality (such information was available in the system). In inconclusive situations, a consultation with the science department head trainer took place. At the same time, the total training time was determined that was subject to further analysis, as well as additional types of exercises (drills) in order to stimulate the model adequately. The collected raw data consisted finally of the following information: Latitude, Longitude, Height above sea level, Timestamp, Speed, Course/Heading, Accuracy: Horizontal Dilution of Precision—HDOP, Vertical Dilution of Precision—VDOP, Number of visible satellites, Satellite ID. The files were exported to the database in CSV format and were subject to further analysis by the scientific team. Samples of the collected datasets have been made available online (part B); see “Code and data availability” statement below.

### 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 Raków 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. The data, before being used in our experiments, get either handled completely by STATSports software, or taken fed raw into the models. In the latter case, they must be handled with care, because GPS data are commonly known to be noisy and usually need some sort of pre-processing, e.g. Kalman filtration.<sup>35</sup> We have examined our data that have been collected exclusively in one training field and with good horizon visibility. Basic verification was performed of how the reported acceleration relates to acceleration estimated by velocity difference quotient, cf. Figure 1. The graph shows almost perfect linear correlation, with typical cutout value of  $6 \text{ m/s}^2$ , and without any apparent impact of instantaneous number of visible satellites. Therefore we decided to take GPS data as is—especially that precise player location in the pitch is not needed by the considered models. Let us therefore define a sample taken at time  $t$  by a tuple

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

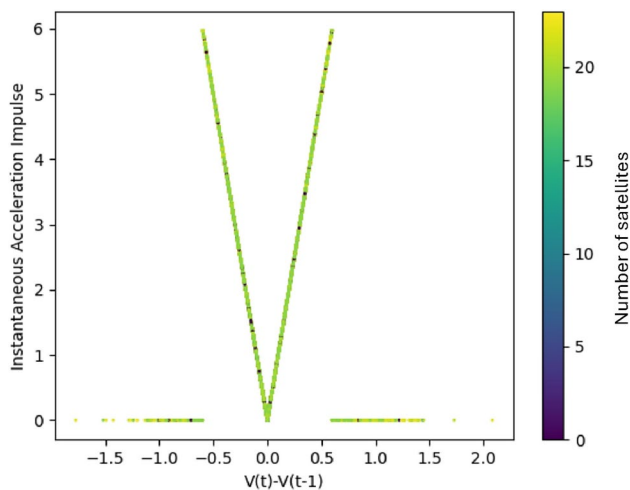
Each such measurement spans a period  $\tau$ , i.e. contains all information for interval  $[t, t + \tau)$ . Each measurement record possesses a set of attributes  $A$  that carry any extra information about the measurement: player ID, player role, activity type etc. Set  $A$  together with  $(t, t + \tau)$  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

The main task of GPS monitoring systems is to create measures that define the size of the athlete's load. Analyses based on such data allow for optimizing player and team management. 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  $\tau$ , which can be configurable and span many time scales, according to user's wishes.<sup>36</sup> 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  $\tau$ , 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—Estimating the instantaneous running time and its metabolic power requirement for an athlete.<sup>37,38</sup>
- HML Distance—the distance covered with high metabolic load;
- 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);



**Fig. 1.** Estimated versus directly reported acceleration, for measurements gathered from a biometric vest. Points represent data samples collected for one athlete during a single training session. Point color denote the actual number of visible GPS satellites. Horizontal axis: acceleration estimated from speed by difference quotient ( $0.1 \text{ m/s}^2$ ). Vertical axis: acceleration as read directly from the vest ( $\text{m/s}^2$ ).

- Distance Zone — the distance covered in a specific zone (e.g. 4, 5 and 6, standing for increasing speed). The contents of  $q$  is up to Apex Pro Series user and can be expandable and configurable. The number of all isolated measurements was 94, see the online material part A, also see the sample databases there in part B. The ranges defined as (zones 4–6) refer to higher intensities during the game. As a result, observing them better reflects the dynamic nature of various tactical and technical exercises and allows you to understand the game’s requirements. From the point of view of physiology, the development of the athlete’s body’s efficiency is aimed at leveling it to a higher level. Maximal and sub-maximal efforts allow for appropriate stimulation and force adaptations. Additionally, only values close to the maximum have a decisive influence on the final result because they give a real advantage over the opponent, especially in key moments of the game.

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 normalized w.r.t. covered distance, with “Per Distance” suffix.

**Models used**

Throughout this work we base our analyses on plain or moderately evolved linear regression models with positive slope 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, ultimately, the training exercises) which contributed most. Whenever the data volume permits, we also verify effectiveness of partially linear models which can handle e.g. saturation phenomenon. We bear in mind that non-linear and negative impact of the training effort on performance in competition is an established fact<sup>39</sup> and that it will have to be accounted for by the model as soon as enough data become available.

Our basic linear multivariate model can be written as

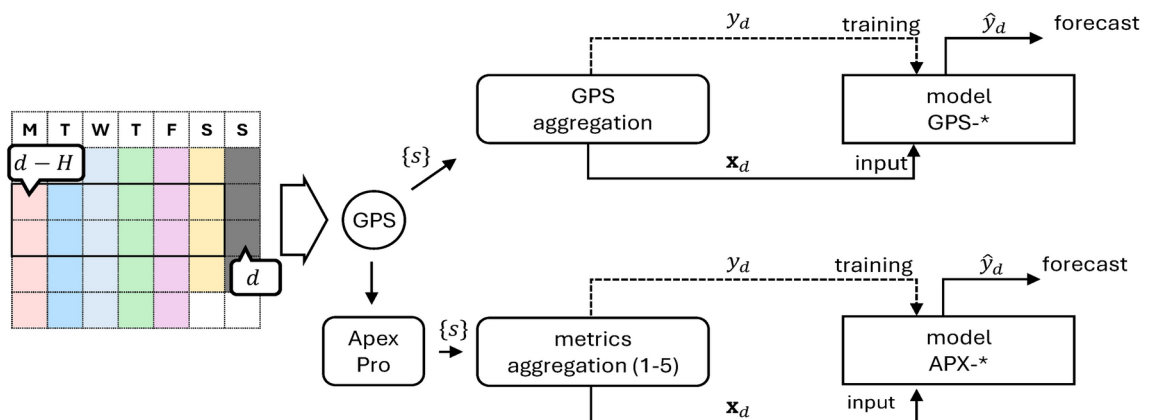
$$\hat{y} = a_0 + \mathbf{a}^T \mathbf{x} \tag{2}$$

with  $\hat{y}$  being the predicted performance index during a match,  $a_0$  is the model intercept value, and  $\mathbf{a} \geq 0, \mathbf{x} \geq 0$  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 and show the corresponding data flows comprehensively on a diagram in Fig. 2.

*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



**Fig. 2.** Data processing diagram for Apex Pro and raw GPS modeling approaches. Various types of training are shown in the calendar with different colors, with Sundays (typically) considered as match days. Metrics and GPS aggregation boxes refer to their corresponding descriptions in “Models used” section. Model boxes refer to models developed and described in the Results section; GPS-\*stands for any model starting with “GPS”, and APX-\*—likewise.

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

Transformation  $\mathcal{M}$  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 4, 5 and 6.

Transformation  $\mathcal{T}$  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 persistence factor  $\alpha$ . 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 can have on the match values, which leads to the following weighted aggregation procedure  $\mathcal{T}$  with value decay  $\alpha^k$  applied for each element  $i$  of input  $\mathbf{x}$  on match day  $d$ :

$$x_{id} = \frac{\sum_{k=1..4} \alpha^k q_n}{\sum_{k=1..4} \alpha^k}, \quad (4)$$

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 derived from the 94 general metrics mentioned above 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:

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

with  $i$  indexing a sample for a player  $j$ . Symbols  $\bar{\mathbf{x}}, \bar{y}$  with bars denote mean values for a player,  $\sigma$  and  $\varsigma$  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. In case of the common model, linear regression fit takes place in the common and normalized space; however the final model evaluation is composed of errors evaluated in the original spaces of individual 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 a single sample consists merely of instantaneous velocity  $v$  measured at time  $t$

$$s = (t, v) \quad (6)$$

i.e., with  $\tau$  and  $A$  omitted w.r.t. (1) because  $\tau$  is constant and equal to 0.1 s and  $A$  is constant in data used to model an individual player. The general idea of aggregation  $\mathcal{T}$  is to account for extreme effort events found in

history window of  $H$  days prior to a match. Therefore, we search for high-effort intervals in history, i.e. a set of sub-sequences of  $\{s\}$ , consecutive in time, such that

$$S(t_{min}, v_{min}, d, H) = \{S_i\} \text{ where } S_i = \{(t, v) : d - H \leq t < d, v \geq v_{min}\} \text{ and } \max_{S_i} t - \min_{S_i} t \geq t_{min} \quad (7)$$

Such definition is a clear analogue to HML-related Apex Pro Series metrics, in terms of counting periods of player's performance happening only above a given speed threshold  $v_{min}$ . The technical difference is that in (7) no upper speed limit is assumed; however, we impose a lower limit on minimal duration  $t_{min}$  of every high-effort episode. Also, the tactical difference is that, unlike in APEX system, we do not assume any specific values for  $v_{min}$  or  $t_{min}$  but leave them to be explored systematically. With the given parameters, we can extract a scalar model input  $x_d$ , consistently with (3) as the count of found training intervals

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

Such parametrized data filtering procedure as in (8) can be applied over a grid of parameter values of interest,  $t_{min} \times v_{min}$ , in search for training statistics that make good prediction of match metrics, much as it was described in the preceding section. The choice of  $v_{min}$  and  $t_{min}$  range of interest should cover jointly the high-effort area, i.e. both very short but quick runs, as well as long periods of moderate but stable running. The approach proposed here is of exploratory nature and does not favor any combination of  $v_{min}$  or  $t_{min}$ . Let us name such model approach (with  $\mathcal{T}$  defined as in (8)) as base GPS model, GPS-Cnt, where 'Cnt' stands for 'count' (of high-effort episodes).

GPS-Cnt applied over  $t_{min} \times v_{min}$  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. 3a 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  $\mathcal{T}$  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 3b,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.

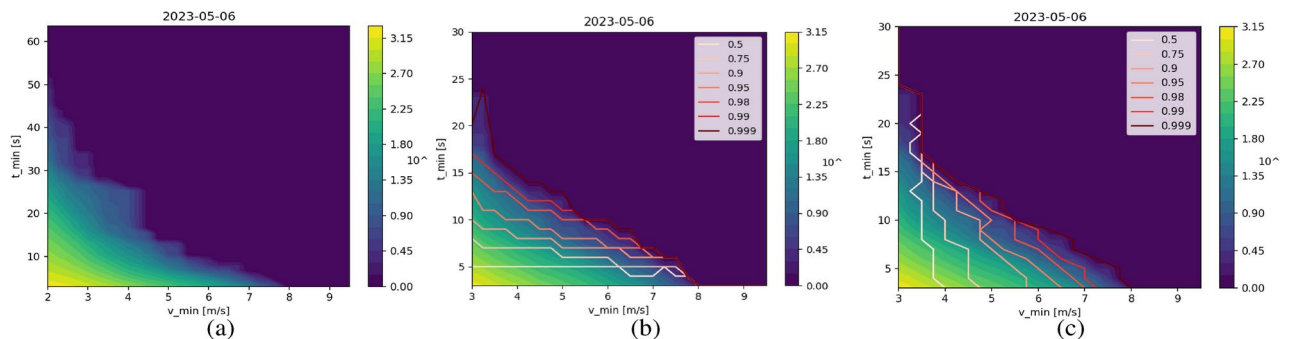
Ultimately, we can apply the same approach as in (6)–(8) to match samples, which yields

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

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

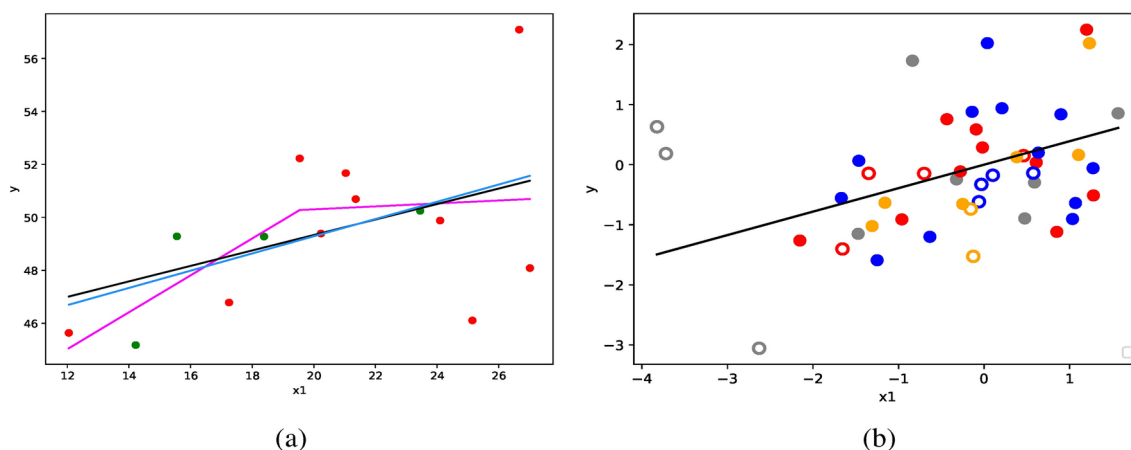
All the presented approaches have been tested in the same setting, in a group of players of age 17 to 21, in order to predict match performance in spring 2023. The number of matches per player with acceptable performance data was 10 to 16; out of which 75% was used for model training, and the rest of data for verification. Training



**Fig. 3.** Sample complementary cumulative distribution function (ccdf) The intervals are collected over a 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  $v_{min}$ .

		Player	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
1 input	APX-Ind	A1	2.7	2.3	2.3	2.3	2.1
	APX-Grp	A1	2.6	2.3	2.3	2.2	2.1
	APX-Grp	A2	17.5	17.5	17.5	17.4	17.3
	APX-Grp	A3	2.1	2.3	2.6	2.8	3.1
	APX-Grp	A4	12.2	12.3	12.1	11.8	11.5
2 inputs	APX-Ind2	A1	2.2	2.2	2.2	2.2	2.2
	APX-Grp2	A1	2.3	2.1	2.2	2.4	2.7

**Table 1.** MAPE values (in percents) for the base model with one and two inputs, for increasing  $\alpha$  values and with comparison between players.



**Fig. 4.** Best single-input models predicting a match performance index STATsport metric Metabolic Time Zone 5 and 6 Per Distance is the modeled match value  $y$ . Decelerations Total Distance Zone 5 and 6 Per Distance—directly or scaled—is the model input  $x_1$ . **(a)** Player A1 samples (red dot—training, green dot—test) vs. lines representing models: blue line—linear model trained on individual data, purple line—individual piecewise linear model trained on individual data, black line—linear model trained on group data (cf. fig. b) but applied after scaling to A1’s data. **(b)** Scaled samples of players A1 to A4 (red, blue, orange and gray dots—respectively) playing in the same position in the pitch. Training data marked with filled circles, test data with empty circles. The black line represents collective model fitted to individually scaled input and output data, cf.  $x_1$  and  $y$  scale w.r.t. **(a)**.

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

Our systematic search for prospective model structures revealed many input/output pairs that yielded low error models. While the final selection of the ones for daily use is up to the coach, we start results discussion with one of them as the base and exemplary one. In subsequent paragraphs, APX-Ind to APX-Grp2 we start from the base model and test various model modifications. Numeric results are provided in Table 1 at the end of the section.

#### APX-Ind

The best individual and acceptable purely linear model that predicts one match performance index based on one training index is the model forecasting Metabolic Time Zone 5 and 6 Per Distance with Decelerations Total Distance Zone 5 and 6 Per Distance as input for player referenced here under identity A1. This model was obtained for  $\alpha$  parameter of 0.5, which resulted in MAPE score of 2.3% and  $R^2$  score of 0.52, see Fig. 4, the blue line. Further on, we consider inputs of model as illustrative and instructive enough for experimentation with the model structure itself. However, it is only one of many feasible and acceptable models found by sifting through pairs of match performance and possible explaining variable pairs.

There exist other linear models that surpass the presented one w.r.t. MAPE, reaching values as small as 0.8%—however, their inputs either are not desirable by experts (see the list in Apex Pro Series metrics’ section above) or have other deficiencies. For instance, the model with Total Acceleration Loading Per Distance and the model with Total Loading Per Distance as the output variable formally have an error of 2.3% but this is due to almost flat regression line, rendering such models useless in practice.



### APX-Ind piecewise linear

The main reason for application of pure linear model is the extremely low number of individual matches played, which maps directly to the number of available training samples. Such situation is intentional: by limiting data history to one season only, we prepare a kind of a snapshot of individual capacity, which we further treat as time-invariant. Should we expand the dataset time range, we would also have to account in the model for time-dependent factors such as the trend, which we otherwise have learned to be extremely difficult to capture in the field data.

The immediate effect is that application of currently widespread models e.g. neural networks to the task, is out of question. However, we have performed a test on a piecewise linear model where two linear regression tasks are solved simultaneously in order to fit two line segments so that mean square error of the predicted value is minimized. The model can be seen as an extension of a two-branch decision tree model, with the difference that the model output is not a constant but a linear function. Considering the nature of modeled phenomena, we claim such an extension feasible and reasonable.

For exemplary model structure of such model, cf. Figure 4a, the purple line. The first line segment is much steeper than the base model (blue line), but the second segment is almost flat, which indicates that intense training ( $x_1 \gtrsim 20$ ) does not result in consistently improved match performance. MAPE for the piecewise linear model is 1.9%, notably better than for the base model.

### APX-Ind2

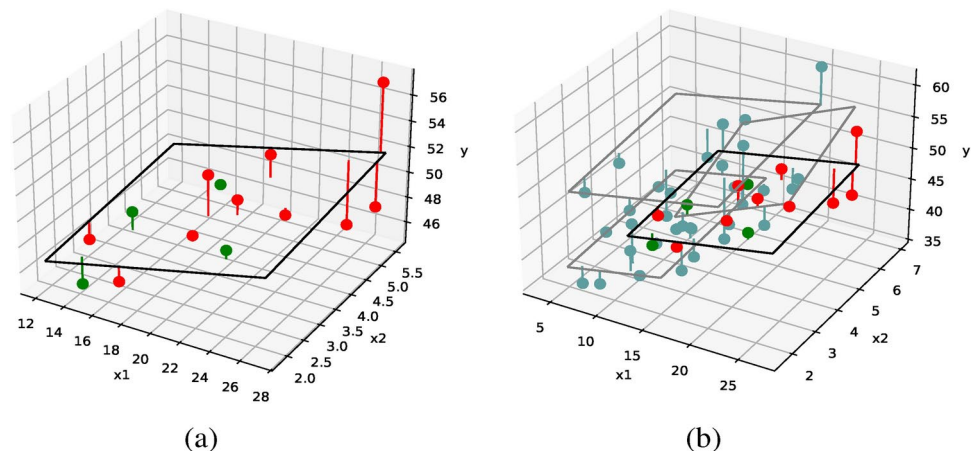
The best individual model being an extension of the base model to two input variables takes Speed Intensity Zone 6 (Relative) Per Distance as the extra input  $x_2$ , cf. Fig. 5a. Introduction of the additional input reduced MAPE by 0.1 pp., to the value of 2.2% and improvement of  $R^2$  by 0.051 to the value of 0.571 w.r.t. base model scores.

### APX-Grp

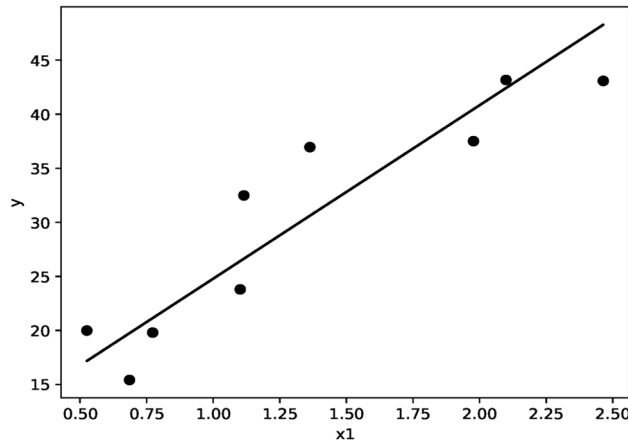
With the collective modeling approach, the common linear model eventually gets adjusted to the dataset of normalized individual samples by means of standard scaling (5). Such dataset and the group model is shown in Fig. 4b. Note the abstract axis values as compared with Fig. 4a; samples of different players get mixed in such rescaled space, enabling to find patterns of training/match performance that are universal. The line of linear regression goes clearly up; the line, after reverse scaling for player A1 is also shown in black in Fig. 4a, and in case of A1 performs as well as individual model in terms of MAPE. For players A2–A4, the error varies but it stays in acceptable range, cf. Table 1, the column for  $\alpha = 0.5$  versus the rows marked APX-Grp. The collective model can also be used as rough initial estimate 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.

### APX-Grp2

Adding the same extra input to the group model gave results similar to those described for individual case, APX-Ind2. MAPE was reduced marginally to 2.2% for player A1. Impact of model extension on players A2–A4 was not checked, yet visual inspection of model parameters for each of the players as provided in Fig. 5b shows



**Fig. 5.** Two-input models developed from the base model. STATSport metric Metabolic Time Zone 5 and 6 Per Distance is the modeled match value  $y$ . Decelerations Total Distance Zone 5 and 6 Per Distance is the model input  $x_1$ . Speed Intensity Zone 6 (Relative) Per Distance is the model input  $x_2$ . **(a)** Player A1 samples (red dots—training, green dots—test) vs. sloped plane representing the individual linear model APX-Ind2 parameters **(b)** Player A1 samples (red dots—training, green dots—test) vs player A2–A4’s samples (blue) used to construct the group model APX-Grp2. The model is not shown; instead, individual models derived from it by reverse scaling are displayed in the original input space, with A1 model marked with black rim. All graphs were generated with the Python Matplotlib package.



**Fig. 6.** The best model that predicts selected match performance index from an earlier match index. Accelerations Total Time Per Distance in Zone 6 (s/m) is the model input  $x_1$ . Distance in Zone 6 (m) is the model output  $y$ .

	$R^2$			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.

that most of the planes representing models are rather flat yet increasing w.r.t.  $x_2$ , thus validating the positive yet marginal effect it has.

Performance of models discussed so far is shown concisely in Table 1, column  $\alpha = 0.5$ , i.e. the value of historical data persistence factor for which the base factor was founded. In order to investigate impact of forgetting older training values, we provide model errors for other  $\alpha$  values in adjacent columns. Data show that, first, all models are stable w.r.t.  $\alpha$ , i.e. their errors do not fluctuate much. On the contrary, MAPE tends to depend little on the rate older data gets forgotten, with the biggest variation of one percent point in case of player A3. This case is an interesting one because it is the only one that clearly indicates that discounting importance of older data ( $\alpha = 0.1$ ) results in model that is better (2.1%) than in case when old data are as important as recent ones ( $\alpha = 0.9$  and MAPE = 3.1%). For other players, the rule is reversed and less pronounced.

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. Figure 6 provides such one for player A2. This linear model was obtained for  $\alpha$  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 factors, which are unobservable yet persistent. 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.

### 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 has been compiled in Table 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  $t_{min} \times v_{min}$  have been checked against selected Apex Pro Series match metrics representing well extremes of player performance, i.e.:

- HML Distance,
  - HML Efforts Max Speed,
  - HML Time.
- Match metrics HML Distance have turned out to yield the most promising model for  $H = 14$ , as shown in Fig. 7a, 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,  $v_{max} \approx 4$  m/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. 7, yet HML Efforts Max Speed tends to be much less correlated with training history.

### GPS-Time

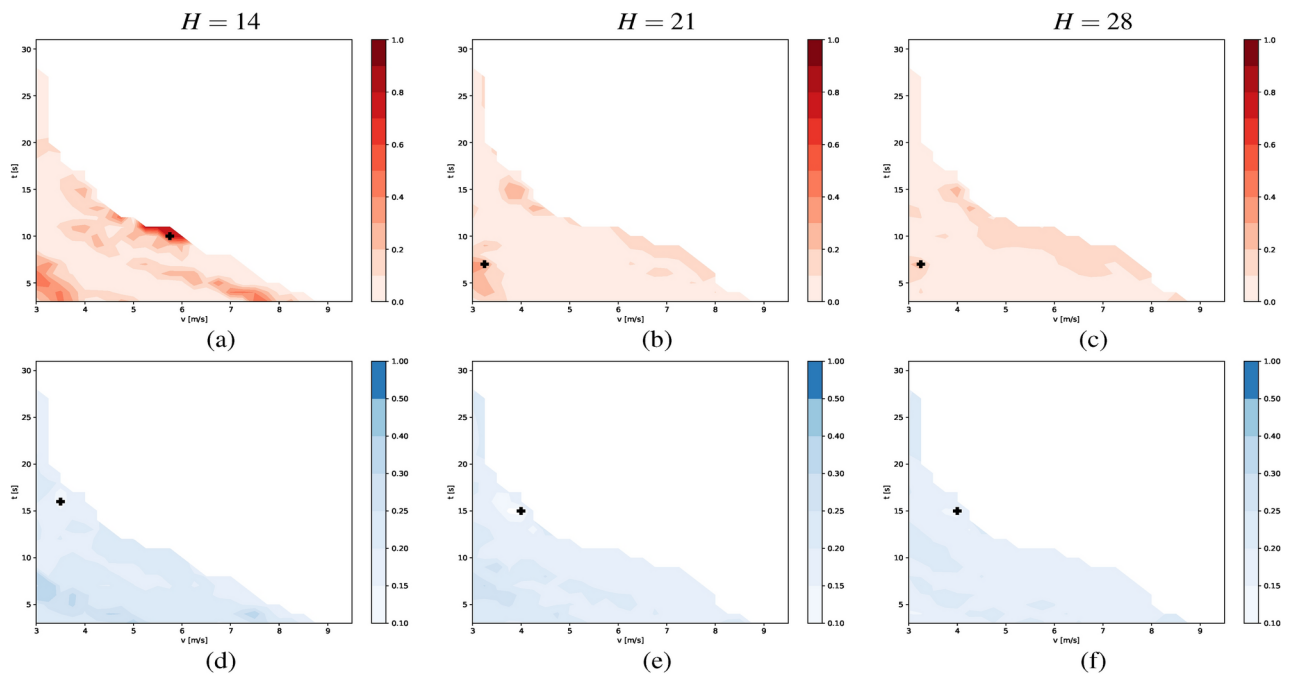
The concept behind GPS-Time approach is a minor generalization of GPS-Cnt because the interval duration is not an absolute value anymore but gets expressed by percentile of  $t_{max}$  distribution for a given  $v_{max}$ . Assessment of models found this way is provided in Fig. 8, resulting again in 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. Figure 8a,d.

### GPS-Vel

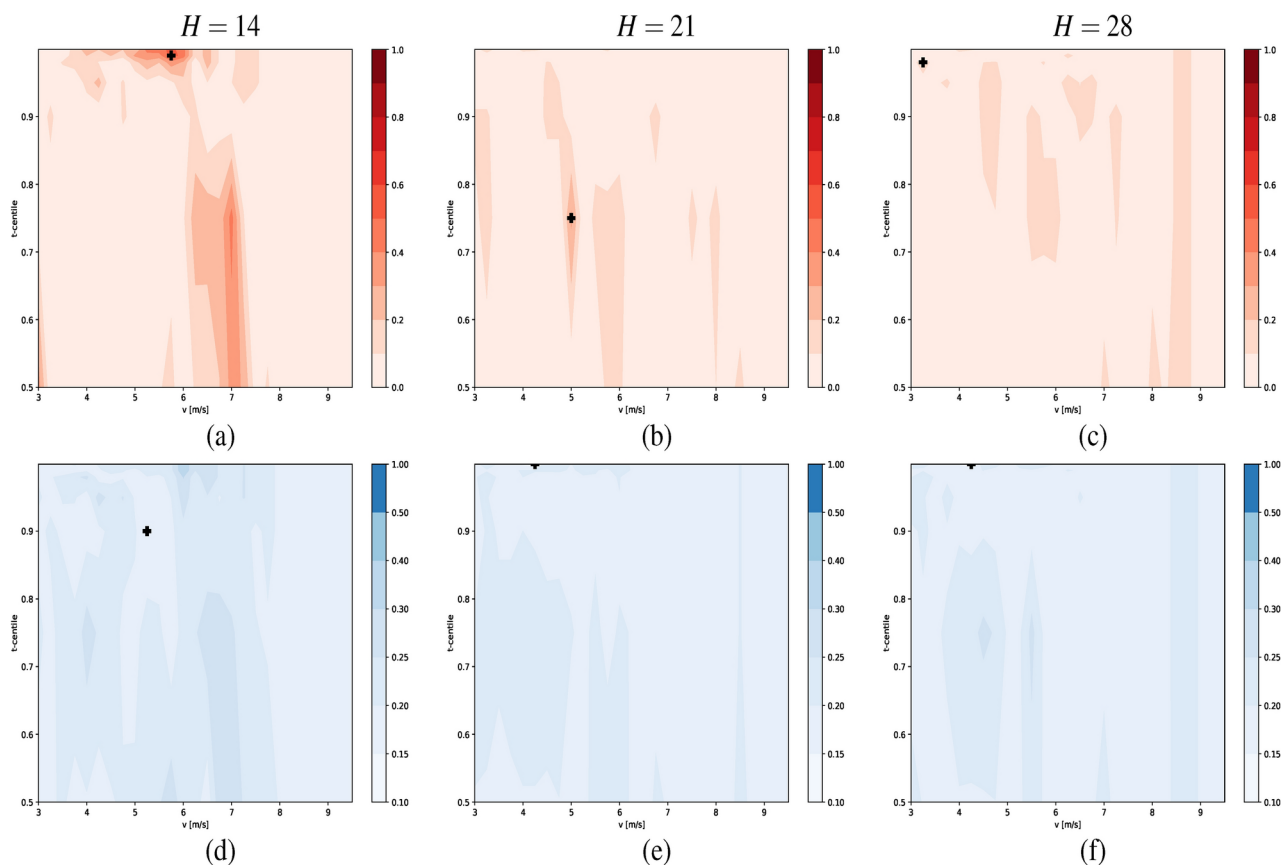
The exploration of time vs. speed percentile space gave the most consistent findings, with all best  $R^2$  and MAPE-related parameters located within a small area of  $t_{max}$  from 12 to 17 s and percentile of  $v_{max}$  from 0.75 to 0.95—cf. Figure 9. The modeled match metrics is still HML Distance. 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. Figure 10 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



**Fig. 7.** Best models found by approach GPS-Cnt:  $R^2$  score in (a–c), MAPE score in (d–f). The training periods before match day are 14, 21 and 28 days long (columnwise). Axes  $v$  and  $t$  denote interval parameter values  $v_{min}$  and  $t_{min}$ , respectively—as provided in Fig. 3 and formula (7). Interval parameter values yielding best models are indicated with small crosses.



**Fig. 8.** Best models found in approach GPS-Time:  $R^2$  score in (a–c), MAPE score in (d–f). The training periods before match day are 14, 21 and 28 days long (columnwise). Axis  $v$  denotes interval parameter values  $v_{min}$ . Axis  $t$ -centile denotes  $p$ -th percentile (w.r.t.  $t_{min}$ , cf. Figure 3b) of number of samples for a given  $v_{min}$ . Parameter values yielding best models are indicated with small crosses.

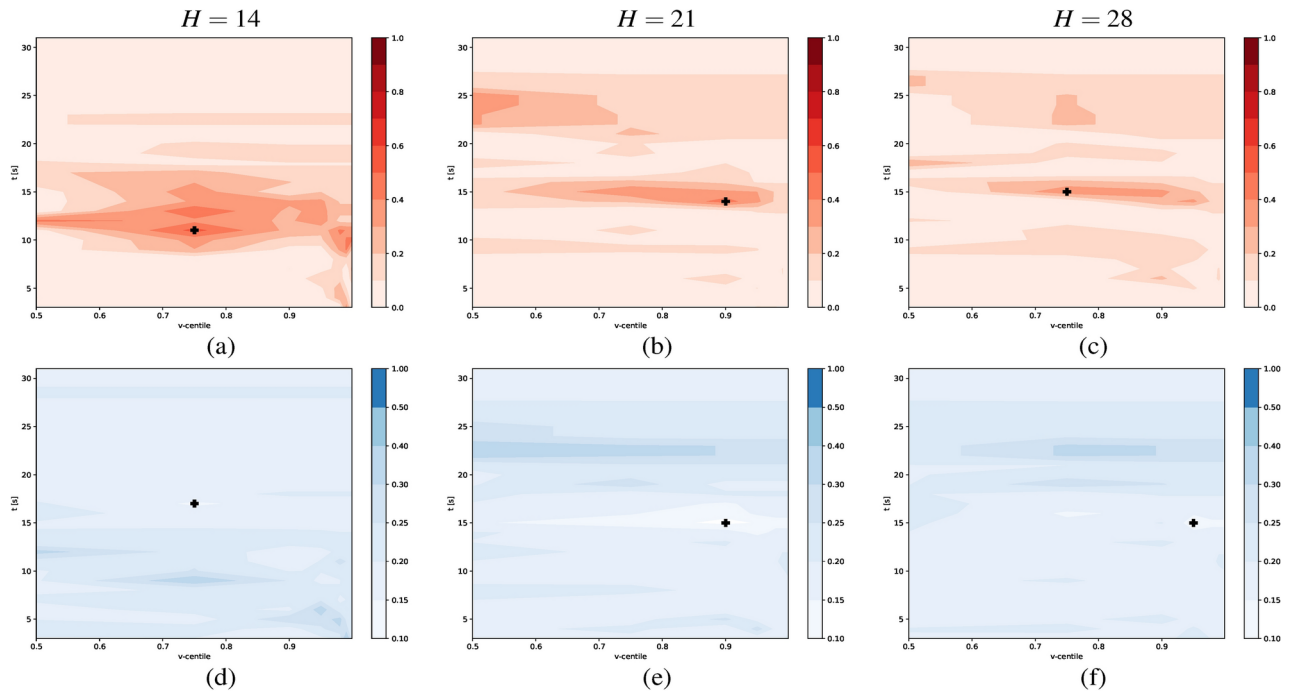
shift between training samples used for model input and match samples used for model output, suggesting that extreme training intervals correlate well with match intervals less intense by 0.7 m/s and by 3 s. Undoubtedly, the phenomenon should be verified both by domain experts, especially by mapping those intervals on particular match fragments, undergoing further scrutiny.

## Discussion

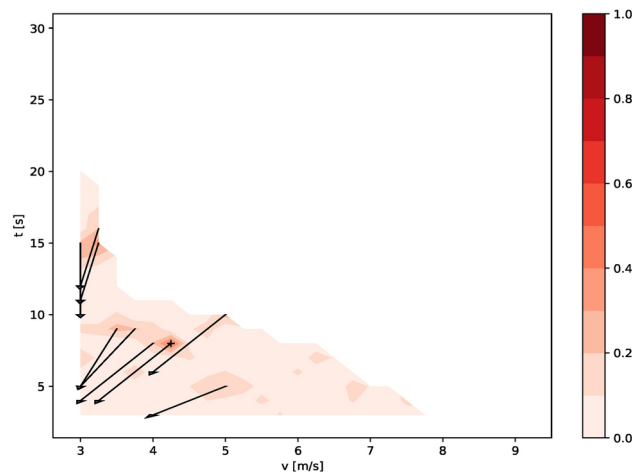
The presented study's main contribution is comparing match outcome prediction models for training data with different structures. It is divided into various and innovative methods of data preparation, also taking into account the specificity of training cycles particular to a given study group.<sup>40,41</sup> 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.

Comparison of match performance prediction models for variously structured training data is the main contribution of the presented study. It breaks down into different and novel methods of data preparation, also accounting for specifics of training cycles. Automated lookup for best input type, both for GPS and pre-aggregated data, can also be considered a novelty. Our work led to selection of Apex Pro Series and GPS-based models that are stable in their parameter space, although the models had to be developed under considerable constraints.

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<sup>42</sup>. 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<sup>43</sup>, 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 modeling process.<sup>44</sup> It should be emphasized that the limited number of samples per an athlete, defined as train-then-play performance pairs, is inherently limited. A player performs at a dozen or so matches in a season,



**Fig. 9.** Best models found in approach GPS-Vel:  $R^2$  score in (a–c), MAPE score in (d–f). The training periods before match day are 14, 21 and 28 days long (columnwise). Axis  $t$  denotes interval parameter values  $t_{min}$ . Axis  $v$ -centile denotes  $p$ -th percentile (w.r.t.  $v_{min}$ , cf. Fig. 3c) of number of samples for a given  $t_{min}$ . Parameter values yielding best models are indicated with small crosses.



**Fig. 10.**  $R^2$  for linear models found by approach GPS-X.

which determines the number of samples. Collecting samples for longer period would introduce non-stationary phenomena into the data that the model is incapable of handling.

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 (not shown in the paper) where the relationships in unique models contradict those that reveal joint modeling for the group, perfectly illustrating Simpson's paradox.<sup>45,46</sup> 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.

On the other hand, more data samples are available for group models, with two consequences. First, the models themselves can be more complex, capturing common phenomena (e.g. saturation) as well as the domain-specific ones. Second, multiple groups can be arranged simultaneously by various criteria (e.g. by playing position, career time, body features) with aim to use the best model found or to be fused into a joint model.

The quality of the obtained models strongly indicates a relationship between pre-match activity and in-match performance. Still, considering the small amount of test samples, we consider this relationship 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,<sup>47</sup> as well as reports related to direct correlation with the planned training work.<sup>48</sup> In the latter case, the lack of a description of the scheme or the algorithm's structure does not allow for a precise comparison.<sup>49</sup> 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.

The ability of predictive models based on GPS data is a novel approach. Therefore, it faces some problems, such as large variability in data between individual athletes (individual and group models). The work of McMillan et al.<sup>50</sup> showed promising correlations between training activities and match results, which confirms the potential applications of these methods on a large scale. Individual differences between players indicate the need for further refinement and adaptation of predictive models to individual athletes' specific needs and characteristics. Therefore, additional research and development of more advanced algorithms are crucial to increase the effectiveness of prediction and ensure the reliability of analytical results obtained from GPS or other data.

### Limitations in research

Pilot studies are associated with certain limitations that may affect the interpretation of results. The small sample size may limit the generalizability of results to a wider population. The specifics of the research environment, including controlled training conditions and limitations related to access to more diverse data, may affect the representativeness of the results. The limitations identified in the reviewed work mainly concern the use of linear models, which may not be able to capture the full complexity of match data. This situation results from the limited availability of comprehensive data sets, which results in the selection of more direct analytical methods. Another significant limitation is the high degree of variability of sports data, which makes it challenging to create models with general applicability. In addition, there is a risk of overfitting when applying slightly complex models to small data sets, which leads to erroneous conclusions and predictions. In addition, the limited number of samples in terms of training data and match results increases the risk of random error and may reduce the accuracy of predictive models. The standard error of measurements also poses a challenge, especially in the context of using advanced technologies, such as GPS systems, which may be susceptible to interference. Considering these limitations in further research is crucial to developing more effective analytical methods in sports. It has also not been confirmed that collecting position data (averaging) is the right direction. Depending on the level of sorting and the individual nature, they may differ.

### Further research directions

Future research on predicting match extremes may focus on exploring more complex nonlinear models and neural networks that better reflect the dynamics of sports data. They must take into account time, place, and game conditions to predict scenarios of subsequent actions, e.g., percentage of success or take into account the error. Extending the database with additional variables, such as weather conditions, physical condition of players, or detailed match statistics (defined club DNA), can significantly increase the precision and reliability of predictions.

As the already used inputs, plus the extra considered ones, offer varying data granularity, the information in the modeling system will have to be fused into a single match prediction or training decision output. Classically, such fusion can be done either at input level (by data preprocessing—such as subsampling—so that all model inputs have equal data samples), or at output level (by weighted averaging of scores of separate sub-models serving separate inputs). While both approaches prove to be successful, modern neural models offer extended possibilities of merging information somewhere in the middle of modelling workflow—cf. eg. model distillation approach. Adoption of such concepts presents as one of interesting future research directions.

Combining dynamically changing patterns of players' behavior in pairs, threes, or fours into patterns corresponding to a specific positioning 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 (deviations from the definition, template, average value, etc.). Interdisciplinary approaches combining knowledge from the field of sports psychology, biomechanics and computer science can open new perspectives for this type of research. Work on improving data analysis methods should also benefit from a better understanding of how different factors affect sports results (evaluation of efficiency and the impact of threshold values). Finally, future research directions should also focus on developing tools that are easier for athletes and coaches to use, enabling them to directly use advanced data analysis methods equipped with preliminary interpretations and future suggestions. The key element is to popularize such tools so that more and more coaches and players themselves can influence the evaluation of this type of product. The use of the method based on extremes will directly affect the creation of training conditions—individual exercises—so that they meet the assumed goals in the context of suggestions for the athlete to develop the minimum threshold values suggested by the algorithm. The parameter should be variable, and the values of these parameters should be individualized on a scale of, for example, a microcycle. It should be emphasized that at this stage of work, expert knowledge in this aspect cannot be excluded. The

work should go from the general (global) to the specific (local). 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. In addition to its typical applications, GPS data modeling can be used to monitor and adjust rehabilitation plans in real time. This allows for precise tracking of patient progress and optimization of recovery processes and motor skills. In uniformed services such as the police, fire department, or military, predictive models can significantly improve operational efficiency, which is crucial for ensuring public safety and increasing the effectiveness of rescue operations.

### Practical application

Using our results in work with athletes will help build progress or maintain stability in specific parameters that most burden the athlete (maximally) during competitions and training. Filtering and removing noise in the data allows us to find weighting values for these parameters, giving mathematical chances to predict their occurrence in match conditions. This concerns extreme values generated in the week preceding the competition. They determine the starting disposition of footballers. These models can also help to optimize match lineups and manage game strategies in real-time. Thanks to the possibility of adapting models to different sports disciplines, their application can go beyond the original research areas. Using advanced data analysis techniques in sports opens up new possibilities for coaches and analysts to explore previously unused strategies. In the long term, the practical application of these models can contribute to a revolution in preparing and managing sports competitions. This has inevitable consequences in the context of other requirements placed on people involved in sports analysis and a change in thinking from an explanation based on historical data to an assessment of the probability of selected training methods about their impact on the final effect and in our case on the result of sports competitions through the results of individual athletes. Looking at the daily work of a coach, implementing a model with such features will help to focus only on the most important tips. A specific filtration process will allow us to determine the most critical parameter for a given player. This will transfer into easy planning and ongoing monitoring of the implementation of training tasks. If not performing a specified number of exercises, the coach will be able to initiate an additional exercise or, otherwise, early cessation of activities of high intensity. A good example is the management of a parameter rarely described in the literature, such as time spent in the 6th metabolic power zone. The model built with that kind of parameter could support the coach during the game in making tactical decisions by rotating the lineup based on fatigue forecasts (achieving the planned target). Better changes of players in this aspect can have a positive impact on the team's match results. The possible reduced risk of injury related to overload is also worth mentioning. Individual models, such as APX-Ind, show high effectiveness in working with individual players; however, it is necessary to consider a combined approach in the future. This can be done at the level of several similar players playing in the same position or entire teams in different age groups working with the same methodology. Modeling such as APX-Grp for the whole team is particularly useful when access to individual data is limited, such as in the case of new players in the team (with no history in the data). Calibrating models to account for a position on the pitch, team style of play, or specific tactical factors enables the broad application of modeling while retaining key, individualized information needed for accurate predictive analysis. This can benefit individual teams and entire leagues, enabling more comprehensive and consistent use of advanced sports analytics in professional football or other sports.

### Conclusions

Observations of extreme training and match parameters are good tools for modeling and prediction. It has not been clearly established what effect the preceding training periods of the studied training have on the prediction model, taking into account 14, 21, or 28 days back. However, logic suggests that the last seven days are good predictors (in this aspect, further research is needed). The right direction of analysis is to individualize predictive models to more accurately predict individual maximum and submaximal athletes' results (according to the principle from general to specific). In this case, specific training parameters can be crucial in predicting the maximum results for football players, but they can also vary depending on the person being analyzed. Optimization of training strategies based on an adapted model can significantly improve the performance of the player, giving an advantage in maintaining high and consequently affecting the final results of the match. A very well-adapted predictive model is based on the maximum values generated in the previous match of a given player.

### Data availability

Sample datasets and executable code have been published on Zenodo service under DOI: <https://doi.org/10.5281/zenodo.13825386>. The online auxiliary material has been divided into three parts: A) a comparison of models

found, B) sampleraw datasets, C) source code and extra data for a selection of models presented here. See data description and license description therein for the details. More information can be obtained from the first author Phd Michał Nowak upon personal request at [michal.nowak@rakow.com](mailto:michal.nowak@rakow.com).

Received: 11 March 2024; Accepted: 4 November 2024

Published online: 09 November 2024

## 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. Netw. Anal. Min.* **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. Quart. Exerc. Sport* 1–8 (2024).
3. Castellano, J., Lázpez-Del Campo, R. & Hílano, R. Tell me how much your opponent team runs and iA will tell you how much you should run. *Biol. Sport* **41**, 275–283. <https://doi.org/10.5114/biolsport.2024.132984> (2024).
4. Dick, U. & Brefeld, U. Learning to rate player positioning in soccer. *Big Data* **7**, 71–82 (2019).
5. Djaoui, L., Chamari, K., Owen, A. L. & Dellal, A. Maximal sprinting speed of elite soccer players during training and matches. *J. Strength Cond. Res.* **31**, 1509–1517 (2017).
6. Gregory, S., Robertson, S., Aughey, R., Spencer, B. & Alexander, J. Assigning goal-probability value to high intensity runs in football. *PLoS ONE* **19**, 1–27 (2024).
7. Vicente, S., Alves, M. F. & Gomes, M. Extreme value theory and sports: The maximal oxygen uptake. In *Symposium on Recent Advances in Extreme Value Theory: Book of Abstracts, CEAUL Editions*, 111–114 (2013).
8. Tam, C.-K. & Yao, Z.-F. Advancing 100m sprint performance prediction: A machine learning approach to velocity curve modeling and performance correlation. *Plus One*. <https://doi.org/10.31219/osf.io/rx5fz> (2024).
9. Rein, R. & Memmert, D. Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus* **5**, 1–13 (2016).
10. 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.* [SPACE] <https://doi.org/10.1177/17543371231213904> (2023).
11. Russell, B. T. & Hogan, P. Analyzing dependence matrices to investigate relationships between national football league combine event performances. *J. Quant. Anal. Sports* **14**, 201–212 (2018).
12. Grunz, A., Memmert, D. & Perl, J. Tactical pattern recognition in soccer games by means of special self-organizing maps. *Hum. Mov. Sci.* **31**, 334–343 (2012).
13. Garganta, J. Trends of tactical performance analysis in team sports: Bridging the gap between research, training and competition. *Revista Portuguesa de Ciências do desporto* **9** (2009).
14. 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).
15. Albert, J. Sabermetrics: The past, the present, and the future. *Math. Sports* **43**, 15 (2010).
16. 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).
17. Thabtah, F., Zhang, L. & Abdelhamid, N. NBA game result prediction using feature analysis and machine learning. *Ann. Data Sci.* **6**, 103–116 (2019).
18. Pischedda, G. Predicting NHL match outcomes with ML models. *Int. J. Comput. Appl.* **101** (2014).
19. 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).
20. Goldsberry, K. Courtvision: New visual and spatial analytics for the NBA. in *2012 MIT Sloan Sports Analytics Conference*, vol. 9, 12–15 (2012).
21. Cervone, D., D'Amour, A., Bornn, L. & Goldsberry, K. Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. In *Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA*, vol. 28 (2014).
22. 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).
23. Coscia, M. Which sport is becoming more predictable? A cross-discipline analysis of predictability in team sports. *EPJ Data Sci.* **13**, 8 (2024).
24. 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).
25. Bunker, R. & Thabtah, F. A machine learning framework for sport result prediction. *Appl. Comput. Inf.* [SPACE] <https://doi.org/10.1016/J.ACI.2017.09.005> (2019).
26. Jain, P., Quamer, W. & Pamula, R. Sports result prediction using data mining techniques in comparison with base line model. *Opsearch* **58**, 54–70. <https://doi.org/10.1007/s12597-020-00470-9> (2020).
27. Bloomfield, J., Polman, R. & O'Donoghue, P. Physical demands of different positions in fa premier league soccer. *J. Sports Sci. Med.* **6**, 63 (2007).
28. Di Salvo, V. et al. Sprinting analysis of elite soccer players during European champions league and UEFA cup matches. *J. Sports Sci.* **28**, 1489–1494 (2010).
29. Bradley, P. S. et al. The effect of playing formation on high-intensity running and technical profiles in English fa premier league soccer matches. *J. Sports Sci.* **29**, 821–830 (2011).
30. Bradley, P. S. & Noakes, T. D. Match running performance fluctuations in elite soccer: Indicative of fatigue, pacing or situational influences? *J. Sports Sci.* **31**, 1627–1638 (2013).
31. Rahimian, P., Mihalyi, B. M. & Toka, L. In-game soccer outcome prediction with offline reinforcement learning. *Mach. Learn.* 1–27 (2024).
32. Cooley, D., Hunter, B. D. & Smith, R. L. Univariate and multivariate extremes for the environmental sciences. In *Handbook of Environmental and Ecological Statistics* 153–180 (2019).
33. 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. *Ann. Appl. Stat.* **10**, 1673–1698. <https://doi.org/10.1214/16-AOAS954> (2016).
34. Wunderlich, F. & Memmert, D. Forecasting the outcomes of sports events: A review. *Eur. J. Sport Sci.* **21**, 944–957 (2021).
35. Yin, Y. et al. Sensor fusion of GNSS and IMU data for robust localization via smoothed error state Kalman filter. *Sensors* **23**, 3676 (2023).
36. 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. *J. Strength Cond. Res.* 10–1519 (2022).
37. di Prampero, P. E. & Osgnach, C. Metabolic power in team sports-part 1: An update. *Int. J. Sports Med.* **39**, 581–587 (2018).
38. Osgnach, C. & di Prampero, P. E. Metabolic power in team sports-part 2: Aerobic and anaerobic energy yields. *Int. J. Sports Med.* **39**, 588–595 (2018).



39. Simão, R. et al. Comparison between nonlinear and linear periodized resistance training: Hypertrophic and strength effects. *J. Strength Cond. Res.* **26**, 1389–1395 (2012).
40. Aquino, R. L. et al. Periodization training focused on technical-tactical ability in young soccer players positively affects biochemical markers and game performance. *J. Strength Cond. Res.* **30**, 2723–2732 (2016).
41. 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. Act. Rev.* **11** (2023).
42. 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).
43. Saribekyan, G. & Yarovoy, N. Football prediction model based on the teams' Elo ratings and scoring indicators. *Res. Square*[SPACE]<https://doi.org/10.21203/rs.3.rs-3861295/v1> (2024).
44. Berrar, D., Lopes, P. & Dubitzky, W. Incorporating domain knowledge in machine learning for soccer outcome prediction. *Mach. Learn.* **108**, 97–126 (2019).
45. Antequera, D. R. et al. Asymmetries in football: The pass-goal paradox. *Symmetry* **12**, 1052 (2020).
46. Sarkar, S. Paradox of crosses in association football (soccer)—a game-theoretic explanation. *J. Quant. Anal. Sports* **14**, 25–36 (2018).
47. 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).
48. 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).
49. Mattsson, C. M. Silicon valley exercise analytics case study—Swedish swimming. <https://svexa.com/case-studies/swedish-swimming/> (2020). Last access: March 9, 2024.
50. McMillan, K., Simpkin, A., Moore, B. & Newell, J. Predicting and individualising training load using historical GPS data in elite soccer. In *Proceedings of the [Sports Tomorrow Congress, Analytics in Sports Tomorrow 2020]* (Barça Innovation Hub, Barcelona, 2020).

## 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 Śledź and Dariusz Grzegorzółka, for constantly searching for a sports advantage on the pitch using the potential of science.

## Author contributions

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.

## 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.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to M.N. or M.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024