



Machine learning predicts peak oxygen uptake and peak power output for customizing cardiopulmonary exercise testing using non-exercise features

Charlotte Wenzel¹ · Thomas Liebig² · Adrian Swoboda³ · Rika Smolareck¹ · Marit L. Schlagheck¹ · David Walzik¹ · Andreas Groll⁴ · Richie P. Goulding⁵ · Philipp Zimmer¹

Received: 28 March 2024 / Accepted: 22 June 2024 / Published online: 3 July 2024
© The Author(s) 2024

Abstract

Purpose Cardiopulmonary exercise testing (CPET) is considered the gold standard for assessing cardiorespiratory fitness. To ensure consistent performance of each test, it is necessary to adapt the power increase of the test protocol to the physical characteristics of each individual. This study aimed to use machine learning models to determine individualized ramp protocols based on non-exercise features. We hypothesized that machine learning models will predict peak oxygen uptake ($\dot{V}O_{2\text{peak}}$) and peak power output (PPO) more accurately than conventional multiple linear regression (MLR).

Methods The cross-sectional study was conducted with 274 (♀168, ♂106) participants who performed CPET on a cycle ergometer. Machine learning models and multiple linear regression were used to predict $\dot{V}O_{2\text{peak}}$ and PPO using non-exercise features. The accuracy of the models was compared using criteria such as root mean square error (RMSE). Shapley additive explanation (SHAP) was applied to determine the feature importance.

Results The most accurate machine learning model was the random forest (RMSE: 6.52 ml/kg/min [95% CI 5.21–8.17]) for $\dot{V}O_{2\text{peak}}$ prediction and the gradient boosting regression (RMSE: 43watts [95% CI 35–52]) for PPO prediction. Compared to the MLR, the machine learning models reduced the RMSE by up to 28% and 22% for prediction of $\dot{V}O_{2\text{peak}}$ and PPO, respectively. Furthermore, SHAP ranked body composition data such as skeletal muscle mass and extracellular water as the most impactful features.

Conclusion Machine learning models predict $\dot{V}O_{2\text{peak}}$ and PPO more accurately than MLR and can be used to individualize CPET protocols. Features that provide information about the participant's body composition contribute most to the improvement of these predictions.

Trial registration number DRKS00031401 (6 March 2023, retrospectively registered).

Keywords Cardiopulmonary exercise testing · Machine learning · Peak oxygen uptake · Peak power output · Prediction

Abbreviations

BIA Bioelectrical impedance analysis
BMI Body-mass-index

CPET Cardiopulmonary exercise testing
ECW Extracellular water
FFM Fat free mass
FM Fat mass

Communicated by Guido Ferretti.

✉ Philipp Zimmer
philipp.zimmer@tu-dortmund.de

¹ Institute for Sport and Sport Science, Performance and Health (Sports Medicine), TU Dortmund University, Dortmund, Germany

² Institute for Computer Science, Department of Artificial Intelligence, TU Dortmund University, Dortmund, Germany

³ Institute for Training Optimization for Sport and Health, iQ Athletik, Frankfurt am Main, Germany

⁴ Department of Statistics, Statistical Methods for Big Data, TU Dortmund University, Dortmund, Germany

⁵ Faculty of Behavioral and Movement Sciences, Department of Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, Amsterdam, The Netherlands

HR	Heart rate
HR _{max}	Maximal heart rate
MET	Metabolic equivalent task
PAL	Physical activity level
PPO	Peak power output
PSQI	Pittsburgh Sleep Quality Index
RER	Respiratory exchange ratio
RMSE	Root mean square error
SD	Standard deviation
SHAP	Shapley additive explanation
SMM	Skeletal muscle mass
TBW	Total body water
TTE	Time to exhaustion
$\dot{V}O_{2max}$	Maximal oxygen uptake
$\dot{V}O_{2peak}$	Peak oxygen uptake
W	Watts
WSD	Wasserstein distance

Introduction

Cardiopulmonary exercise testing (CPET) on a cycle ergometer is widely applied in endurance sports as well as in clinical settings. It provides a comprehensive insight into integrated cardiopulmonary function within a single laboratory session. In particular, maximal oxygen uptake ($\dot{V}O_{2max}$) reflects the integrated capacity of the cardiopulmonary and neuromuscular systems to take up, transport, and utilize oxygen during exercise (Poole and Jones 2017). It thus represents the greatest attainable rate of aerobic adenosine triphosphate generation and is a marker of exercise capacity (Bassett and Howley 2000). In addition to $\dot{V}O_{2max}$, peak power output (PPO) is typically used as a measure of exercise capacity. It is quantified in external units of power output, thus requiring less specialist equipment. Both outcomes are strongly predictive of all-cause mortality and the risk of developing chronic diseases (Ross et al. 2016). They can be applied to manage exercise training by determining training intensity and to validate and monitor the success of training interventions (Myers 2005).

To measure valid and interpretable CPET values on a cycle ergometer, the appropriate rate of increase in power and associated test duration are relevant factors. A too-rapid increase and a short test duration may lead to hyperventilation, lack of determinability of the gas exchange threshold (Glaab and Taube 2022), and premature end of the test due to the occurrence of task failure prior to the attainment of $\dot{V}O_{2max}$ (Hill et al. 2002). Too slow increase in power and a long test duration could result in insufficient $\dot{V}O_2$ drive to reach $\dot{V}O_{2max}$. This may lead to test termination due to factors not typically associated with reaching the tolerance limit during severe exercise (e.g. peripheral muscle fatigue, accumulation of metabolic products associated with fatigue,

etc.) (Vanhatalo et al. 2010; Burnley et al. 2012). Consequently, an inappropriate increase in power can lead to early or delayed termination of CPET and failure to accurately determine $\dot{V}O_{2max}$. Due to the challenge of attaining $\dot{V}O_{2max}$, peak oxygen uptake ($\dot{V}O_{2peak}$) is alternatively employed as an indicator of physical performance, representing the highest $\dot{V}O_2$ value determined during the CPET.

To reach peak performance values, the test protocol should be precisely adjusted to the participant achieving voluntary exhaustion within the recommended 8 to 12 min (Buchfuhrer et al. 1983; American College of Sports Medicine 2021). Standardization of test duration across participants would ensure the comparability of results between different test facilities or clinical environments and lead to optimized processes. To enable valid completion of each test, it is necessary to adapt the power increase to the physical characteristics of every individual. It would be advantageous if these characteristics are non-exercise features that are convenient to collect in daily practice before conducting a CPET and are expected to have an impact on $\dot{V}O_{2peak}$ and PPO. Furthermore, appropriate prediction models are important for customizing the protocol. In addition to $\dot{V}O_{2peak}$, which is commonly used as an outcome parameter (Myers et al. 2001; da Silva et al. 2012; Cunha et al. 2015), PPO depicts a further outcome parameter that can be collected with less equipment.

Previous work investigating protocol adaptations for CPET included small (Saengsuwan et al. 2017) or homogeneous (Myers et al. 1994; Cunha et al. 2015) populations and usually used conventional linear predictive models (Myers et al. 1994, 2001; da Silva et al. 2012; Saengsuwan et al. 2017). This may lead to a possible overestimation of the explanatory power of the models. Indeed, the American College of Sports Medicine recommends formulas for predicting $\dot{V}O_{2max}$ based on multiple linear regression (American College of Sports Medicine 2021). Thus, potential non-linear relationships between the features and the outcome parameter cannot be identified. Meanwhile, several machine learning models, like decision tree (Song and Lu 2015), random forest (Breiman 2001), *k*-nearest-neighbor (Sreevalsan-Nair 2020), and gradient boosting regression (Friedman 2001), are able to capture non-linear patterns. They can incorporate many features, and deal with heterogeneous data and the associated outliers (Friedman 2001; Singh et al. 2016). Due to this background, this work aims (1) to compare two different non-exercise feature sets to predict $\dot{V}O_{2peak}$ and PPO using four machine learning models and one linear model and (2) to identify the most impactful features to adapt the power increase to the physical conditions of each participant on a cycle ergometer. We hypothesize that machine learning models provide more accurate predictions of $\dot{V}O_{2peak}$ and PPO than the conventional multiple linear regression technique.

Materials and methods

Participants

The cross-sectional study included $n = 274$ (♀168, ♂106) participants who were at least 18 years old and physically able to perform CPET. Participants were asked not to engage in vigorous exercise or drink alcohol or coffee for 24 h and not to eat for 2 h prior to the measurement. Participants were excluded from the study if they had a pacemaker, an acute infection, or an orthopedic injury. Female subjects were also excluded if they were pregnant.

Procedures

Participants' demographic and anthropometric data were collected. Subsequently, body composition was assessed by bioelectrical impedance analysis (BIA) (SECA mBCA 525), and handgrip strength measurement (Jamar hand dynamometer hydraulic) was conducted. Afterward, participants answered questionnaires on physical activity level (PAL) (Godin and Shephard 1985; Armstrong and Bull 2006) and sleep quality (Buysee et al. 1989). The CPET was performed in an upright position on a cycle ergometer (CORTEX Bike M). The test began with a one-minute resting measurement and a 2-min period of baseline cycling at 25 watts (W) for females and 50 W for males. Subsequently, we increased the workload by 15, 20 or 25 W/min for females and 20, 25 or 30 W/min for males, depending on PAL category and body-mass-index (BMI) (supplements Table A). The power increased continuously until participants were no longer able to maintain a frequency above 60 revolutions per minute. This was followed by a recovery phase of 3 min at 25 W for females or 50 W for males.

Respiratory gas exchange and ventilation were measured continuously on a breath-by-breath basis via spirometry (CORTEX METAMAX® 3B). Heart rate (HR) was monitored permanently via a bluetooth chest strap (Polar H10). The highest 15-s average determined by the software (CORTEX MetaSoft® Studio) during the CPET was regarded as peak $\dot{V}O_2$ ($\dot{V}O_{2peak}$), respiratory exchange ratio (RER) and maximal HR (HR_{max}). The highest power output achieved during the CPET prior to exhaustion was considered the PPO. Time to exhaustion (TTE) was defined by the time of exercise test minus baseline and recovery periods. To ensure that most participants had achieved their full capacity, they had to reach a $\dot{V}O_2$ plateau with 150 ml/min difference between the last two 30-s intervals or meet two of the following three criteria at the time of $\dot{V}O_{2peak}$: (1) a RER ≥ 1.1 , (2) a rating of perceived

exertion > 17 on the 6–20 scale and (3) a HR_{max} within 10 beats/min of the age-predicted HR_{max} (Tanaka et al. 2001).

Feature selection

The selection of non-exercise features for the prediction of $\dot{V}O_{2peak}$ and PPO was based on possible associations with physical performance (Wier et al. 2006; Schembre and Riebe 2011; Booth et al. 2012; Antunes et al. 2017; Saengsuwan et al. 2017; Przednowek et al. 2018; Langer et al. 2020; American College of Sports Medicine 2021; Shen et al. 2022). The small feature set comprised 15 features that are convenient to assess in practice and have a low time requirement. These include anthropometric and demographic data, as well as self-perceived health status and activity level. The big feature set was extended by body composition variables, handgrip strength, and questionnaires (Godin and Shephard 1985; Buysee et al. 1989; Armstrong and Bull 2006), and comprised 41 features. All features are listed in Table 1 and explained in more detail in the supplements (Table B).

Applied machine learning algorithms

Multiple linear regression (Jobson 1991) as a conventional model and four different supervised machine learning models were used to predict the $\dot{V}O_{2peak}$ and PPO, each utilizing the two feature sets. Table 2 lists the multiple linear regression, and the machine learning models, their description, and the reasons for selection.

Table 1 Small and big feature set

Feature set	Features
Small feature set	Age (years), sex, smoking, smoking behavior (years), chronic diseases, allergies, supplements, medications, PAL, weight (kg), height (m), BMI, waist circumference (cm), hip circumference (cm), waist-hip-ratio
Big feature set	Age (years), sex, smoking, smoking behavior (years), chronic diseases, allergies, supplements, medications, PAL, weight (kg), height (m), BMI, waist circumference (cm), hip circumference (cm), waist-hip-ratio, FM (%), FM (kg), FFM (%), FFM (kg), SMM (%), SMM (kg), SMM torso (kg), SMM torso (%), SMM legs (kg), SMM legs (%), TBW (%), TBW (L), ECW (%), ECW (L), ECW/TBW (%), total energy expenditure (kcal/day), resting energy expenditure (kcal/day), phase angle (°), handgrip strength (kg), handgrip strength (%), MET, PAL category, Godin score, Godin category, PSQI score, PSQI category

PAL physical activity level, BMI body-mass-index, FM fat mass, FFM fat free mass, SMM skeletal muscle mass, TBW total body water, ECW extracellular water, MET metabolic equivalent task, PSQI Pittsburgh Sleep Quality Index

Table 2 Description of the supervised machine learning models and justification of the selection

Models	Description	Reason
Multiple linear regression	Multiple linear regression is used to model a continuous variable and make predictions. The independent features and the dependent feature must be linearly related to fit a straight line to the data set (Ray 2019).	Multiple linear regression is the most commonly used statistical technique especially in exercise physiology (Jobson 1991; Akay and Abut 2015). It provides insights into the relationship between independent and dependent features. Multiple linear regression is a simple and comprehensible method (Ray 2019).
Decision tree	In decision tree, actions are executed by if–then conditions by using a single attribute for splitting. The data are split in the nodes and the decisions are in the leaves (Breiman et al. 1984; Ray 2019).	Decision tree is easy to implement. It can handle categorical and quantitative values and is easy to interpret and visualize (Ray 2019).
Random forest	The random forest contains several decision trees that are applied in parallel to different subsamples of the data set. For each split, a random subset of the available features is used for splitting. The result or final value is determined by majority decisions or averages (Breiman 2001).	Random forest is an effective tool for prediction and is relatively robust against outliers and noise. Random forest reduced the over-fitting problem and usually achieves a good bias-variance tradeoff (Singh et al. 2016).
<i>k</i> -Nearest-neighbors	The <i>k</i> -nearest-neighbor is a non-parametric algorithm. <i>k</i> -Nearest-neighbor approximates the association between independent variables and the continuous outcome by averaging the values of the <i>k</i> -nearest neighbors (Sreevalsan-Nair 2020).	<i>k</i> -Nearest-neighbor is a simple technique, non-parametric and quick to implement. In addition, the scheme is very flexible (Ray 2019).
Gradient boosting regression	Gradient boosting regression creates a sophisticated model based on a combination of multiple weak individual models, which are mostly decision trees (Sarker 2021; Li et al. 2022).	Boosting is one of the most successful techniques introduced to solve complex problems (Li et al. 2022). Gradient boosting regression is a meaningful and robust model that is suitable for unclean data (Friedman 2001).

Statistical analysis

All statistical analyses were conducted with Python 3.9 (Van Rossum and Drake 2009) (supplements Table C). First, participant characteristics were presented by descriptive statistics [mean $\pm$ standard deviation (SD)]. Two missing values of the Pittsburgh Sleep Quality Index (PSQI) Score were replaced by the single imputation method (Glas 2010). Since the outcome parameters are continuous, the machine learning models were trained for regression tasks. All variables were standardized using the z transformation. The final study population was divided into a training (80%) and a validation (20%) set. On the training data, fivefold cross-validation (Refaeilzadeh et al. 2016) was used for hyperparameter tuning using Bayes search (Lindauer et al. 2019) and to train the final models. The performance of the final models was evaluated on the validation set. The evaluation was based on quality criteria including the mean of the reset standardized root mean square error (RMSE), the R squared (R^2), the Wasserstein distance (WSD), and the respective 95% confidence intervals obtained from the validation set using 1000 replicates. The Shapley additive explanation (SHAP) (Nohara et al. 2022) was used to determine the feature importance. This involved assessing the relevance of features using SHAP values to identify the relative contribution of the feature to $\dot{V}O_{2\text{peak}}$ and PPO prediction. In a further step, the entire procedure described above was performed again for $\dot{V}O_{2\text{peak}}$ and PPO separately for females and males to investigate sex-specific differences. Accordingly, the feature sex was removed from this part of the analysis.

Results

Characteristics and CPET values of participants included in the predictions

In total, $n = 274$ potential participants attended the study. Finally, $n = 258$ (♀101, ♂157) adults were included in the analysis for the prediction of $\dot{V}O_{2\text{peak}}$. Of the $n = 16$ excluded participants, $n = 14$ did not reach a $\dot{V}O_2$ plateau or at least two out of three exhaustion criteria and $n = 2$ were excluded due to missing BIA and CPET values. In the analysis for predicting PPO, $n = 272$ (♀106, ♂166) participants were included. Only $n = 2$ were excluded due to missing CPET and BIA values. The participants' characteristics and CPET values are shown in Table 3.

Model comparison for predicting $\dot{V}O_{2\text{peak}}$ and PPO

Figure 1 illustrates the performance of the multiple linear regression and the applied machine learning models concerning the mean of the three quality criteria: RMSE, R^2 and

Table 3 Participants' characteristics and CPET values for the predictions (mean $\pm$ SD)

Characteristics and CPET values	Prediction of $\dot{V}O_{2peak}$			Prediction of PPO		
	Females ($n=101$)	Males ($n=157$)	All ($n=258$)	Females ($n=106$)	Males ($n=166$)	All ($n=272$)
Age (years)	27.40 $\pm$ 9.96	27.17 $\pm$ 8.46	27.26 $\pm$ 9.06	27.29 $\pm$ 9.74	27.52 $\pm$ 8.97	27.43 $\pm$ 9.26
Height (m)	1.69 $\pm$ 0.06	1.81 $\pm$ 0.07	1.77 $\pm$ 0.09	1.69 $\pm$ 0.06	1.81 $\pm$ 0.07	1.76 $\pm$ 0.09
Weight (kg)	65.97 $\pm$ 9.60	79.97 $\pm$ 10.97	74.49 $\pm$ 12.48	65.83 $\pm$ 9.52	79.96 $\pm$ 10.84	74.45 $\pm$ 12.42
BMI	23.09 $\pm$ 3.24	24.32 $\pm$ 2.98	23.84 $\pm$ 3.14	23.03 $\pm$ 3.20	24.39 $\pm$ 2.99	23.86 $\pm$ 3.14
Waist-Hip-Ratio	0.76 $\pm$ 0.05	0.83 $\pm$ 0.05	0.80 $\pm$ 0.06	0.76 $\pm$ 0.05	0.83 $\pm$ 0.05	0.80 $\pm$ 0.06
Rel. FM (%)	27.74 $\pm$ 7.60	17.02 $\pm$ 6.73	21.22 $\pm$ 8.80	27.78 $\pm$ 7.53	17.04 $\pm$ 6.83	21.23 $\pm$ 8.83
Rel. FFM (%)	72.29 $\pm$ 7.61	82.98 $\pm$ 6.73	78.79 $\pm$ 8.80	72.25 $\pm$ 7.54	82.96 $\pm$ 6.83	78.78 $\pm$ 8.82
PPO (W)	214 $\pm$ 44	322 $\pm$ 57	280 $\pm$ 74	212 $\pm$ 44	320 $\pm$ 60	278 $\pm$ 75
$\dot{V}O_{2peak}$ (L/min)	2.53 $\pm$ 0.51	3.87 $\pm$ 0.65	3.34 $\pm$ 0.88	2.52 $\pm$ 0.51	3.84 $\pm$ 0.66	3.33 $\pm$ 0.89
$\dot{V}O_{2peak}$ (ml/kg/min)	38.89 $\pm$ 8.05	48.89 $\pm$ 8.34	44.97 $\pm$ 9.56	38.71 $\pm$ 7.93	48.60 $\pm$ 8.56	44.74 $\pm$ 9.61
TTE (min:s)	09:41 $\pm$ 1:57	10:14 $\pm$ 1:49	10:01 $\pm$ 1:53	09:37 $\pm$ 1:56	10:07 $\pm$ 1:55	09:55 $\pm$ 1:56

SD standard deviation, CPET cardiopulmonary exercise testing, BMI body-mass-index, Rel. FM relative fat mass, Rel. FFM relative fat free mass, PPO peak power output, $\dot{V}O_{2max}$ maximal oxygen consumption, TTE time to exhaustion

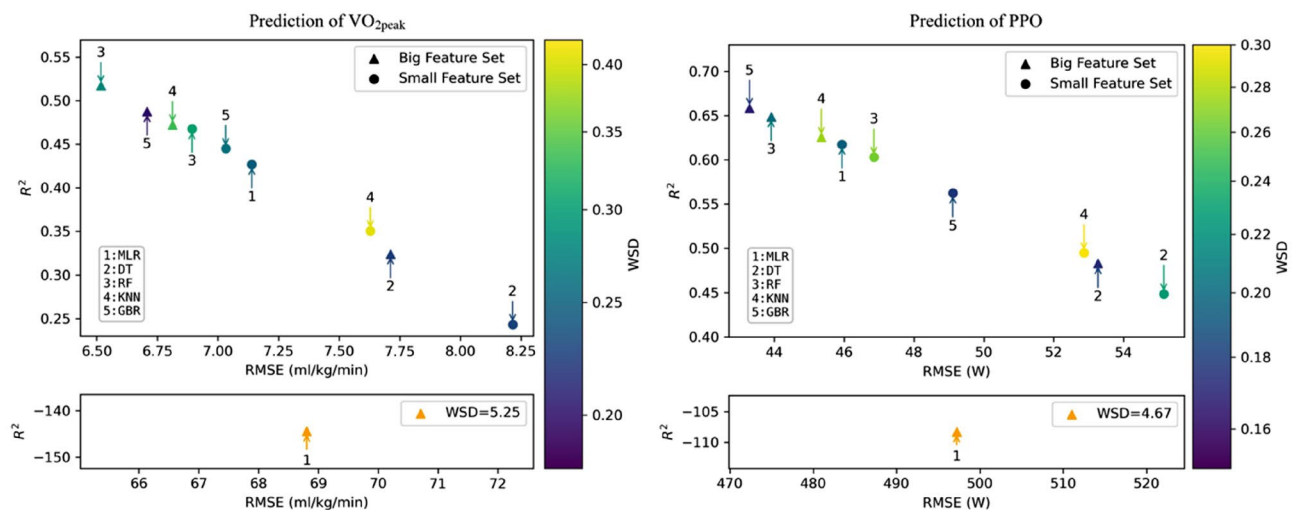


Fig. 1 Quality criteria of the models for the prediction of $\dot{V}O_{2peak}$ and PPO. RMSE root mean squared error, R^2 R squared, WSD Wasserstein distance, MLR multiple linear regression, DT: decision tree, RF Random forest, KNN k-nearest-neighbor, GBR gradient boosting regression

WSD. The RMSE is a standard statistical parameter and is used to evaluate model performance. The units of RMSE in this work are ml/kg/min for the prediction of $\dot{V}O_{2peak}$ and W for the prediction of PPO. The R^2 represents the proportion of the variance of the outcome parameter that is explained by the features of the model. The WSD measures differences between probability distributions. The smaller the RMSE and the WSD and the larger the R^2 the more accurate the results of the models. The multiple linear regression with the big feature set has an additional coordinate system, as all quality criteria differ significantly from those of the other models. The mean values of the quality criteria and the 95% confidence intervals can be found in the supplements (Table D–F).

Machine learning models such as random forest and gradient boosting regression with the big feature set have a low RMSE and WSD as well as a high R^2 and perform better overall than models with the small feature set. Figure 2 shows the performance of the sex-separated models for the prediction of $\dot{V}O_{2peak}$ and PPO using the mean of quality criteria.

In the prediction of $\dot{V}O_{2peak}$ for females, the applied machine learning models consistently outperformed the multiple linear regression models. In all sex-separated models, the machine learning models with the big feature set perform better overall than the models with the small set.

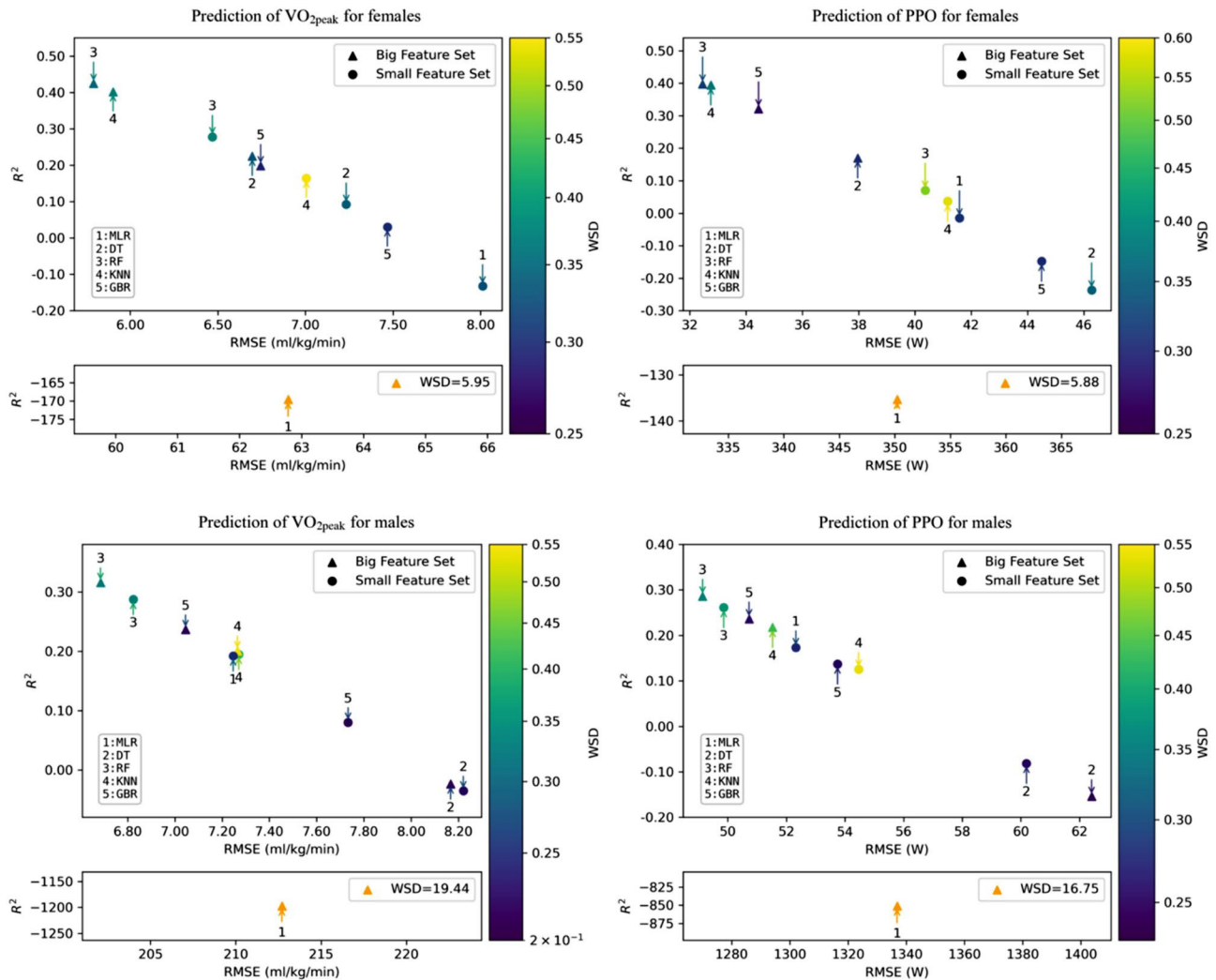


Fig. 2 Quality criteria of the models for the prediction of $\dot{V}O_{2peak}$ and PPO for females and males. *RMSE* root mean squared error, R^2 *R* squared, *WSD* Wasserstein distance, *MLR* multiple linear regression,

DT decision tree, *RF* random forest, *KNN* *k*-nearest-neighbor, *GBR* gradient boosting regression

Feature importance

The ten most impactful features for predicting $\dot{V}O_{2peak}$ and PPO were selected for each model by SHAP. SHAP values are assigned to each feature for prediction. The prediction result is the sum of the contributions of each feature. The *x*-axis represents the impact of each feature on the prediction for each participant represented by a dot and the *y*-axis shows the feature in descending order of overall importance. The color of the gradient denotes the magnitude of the original value for that feature. Since random forest and multiple linear regression performed best among the machine learning models with the small feature set (Fig. 1) and random forest with the big feature set consistently performed best among the sex-separated models, the SHAP values of these models are shown in Fig. 3. The supplements also contain all

figures of the SHAP values of the most accurate model with the big and the small feature set (Figure A–F).

Discussion

We hypothesized that machine learning models will provide more accurate predictions of $\dot{V}O_{2peak}$ and PPO than the conventional multiple linear regression technique. Moreover, we aimed (1) to compare two non-exercise feature sets to predict $\dot{V}O_{2peak}$ and PPO using four machine learning models as well as multiple linear regression and (2) to identify the most impactful features. The results confirm that machine learning models provide more precise results in comparison to multiple linear regression. Our analysis further indicates that machine learning

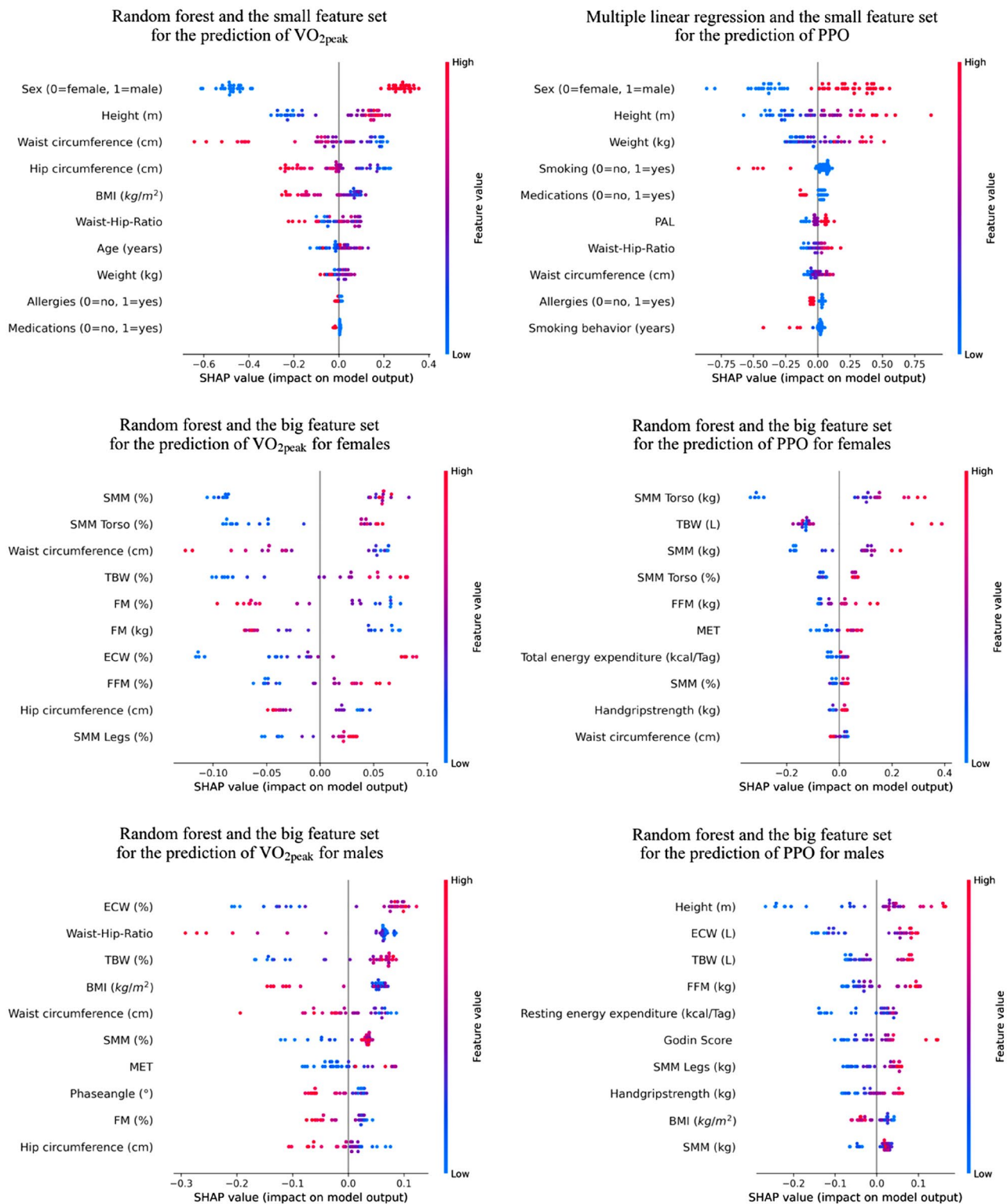


Fig. 3 Importance of the features by SHAP with the random forest and multiple linear regression with the small feature set as well as the random forest with the big feature set for the prediction of $\dot{V}O_{2peak}$ and PPO for females and males. *SHAP* Shapley additive explanation,

BMI body-mass-index, *PAL* physical activity level, *SMM* skeletal muscle mass, *TBW* total body water, *FM* fat mass, *FFM* fat free mass, *ECW* extracellular water, *MET* metabolic equivalent task

models with comprehensive features make more accurate predictions than models containing only anthropometric and demographic data. In particular, features that include information about the participant's body composition seem to have a relevant impact on the prediction of $\dot{V}O_{2\text{peak}}$ and PPO. These results may be helpful in developing new standards for performing CPETs and improving prediction models for $\dot{V}O_{2\text{peak}}$ and PPO.

In addition to the four machine learning models decision tree, random forest, *k*-nearest-neighbor and gradient boosting regression, we used multiple linear regression as a conventional technique that is commonly used in exercise physiology to predict $\dot{V}O_{2\text{peak}}$ and PPO (Myers et al. 1994, 2001; da Silva et al. 2012; Akay and Abut 2015; Saengsuwan et al. 2017). Previous literature has already shown that intelligent machine learning models can predict $\dot{V}O_{2\text{peak}}$ more accurately than existing multiple linear regression-based prediction models (Akay and Abut 2015; Liu et al. 2022). These results are confirmed by our work, which, in contrast to previous literature, compared machine learning and multiple linear regression models based on the same population and the same conditions. Moreover, these indicate that some predictor variables showed non-linear relationships with $\dot{V}O_{2\text{peak}}$ and PPO. The applied machine learning models can effectively analyze and capture these non-linear relations, explaining their greater performance over the traditional multiple linear regression technique.

In this study, the random forest proved to be the most robust prediction model, as it possessed the lowest RMSE, the highest R^2 , and typically displayed a lower WSD for almost all predictions. The random forest estimated $\dot{V}O_{2\text{peak}}$ considering both sexes with a mean error of 6.52 ml/kg/min, a variance explanation of approximately 52% and a difference in probability distributions of 0.28. The gradient boosting regression model performed slightly better than the random forest in predicting PPO when both sexes were considered. It predicted PPO with a mean error of 43 W, a variance explanation of about 66% and a difference in the probability distributions of 0.18. In particular, the multiple linear regression had an unusually high RMSE and WSD as well as a negative R^2 if many features were included in the prediction. The results were outside the interpretable range and are due to the fact that the multiple linear regression cannot handle a large number of predictor variables that exhibit multicollinearity (Jobson 1991). In contrast to multiple linear regression, random forest handles outliers and avoids overfitting by capturing underlying patterns rather than overlearning the training data (Singh et al. 2016). The gradient boosting regression is considered a robust method that can also deal with very heterogeneous data (Friedman 2001). In addition to these advantages, machine learning models are able to recognize linear relationships between variables. Therefore, the effectiveness of conventional linear

methods in predicting $\dot{V}O_{2\text{peak}}$ and PPO should be critically reconsidered.

Beside the correct selection of suitable prediction models, the identification of relevant features is crucial. This allows practical recommendations regarding the parameters that should be recorded before conducting a CPET. Previous literature aimed at individualizing test protocols has used features that are usually assessed prior CPET, such as questionnaires on PAL, sex, age, BMI or resting HR (Myers et al. 2001; da Silva et al. 2012; Cunha et al. 2015; Saengsuwan et al. 2017). In our study, we divided the features into two sets to determine if collecting only anthropometric and demographic data before conducting a CPET is sufficient to adapt a ramp protocol to the participant's characteristics.

Previously, data from the National Health and Nutrition Examination Survey (Liu et al. 2022) have been used to develop machine learning models for the prediction of $\dot{V}O_{2\text{max}}$ with non-exercise features. In line with our results, the authors concluded that models with a comprehensive feature set performed significantly better than previous methods using a limited number of predictors and mainly linear models. However, the work was limited by the fact that some predictor variables cannot be readily implemented in other healthcare settings. Furthermore, existing studies have often used submaximal features to predict maximal physical performance, which are less practical and more time-consuming (Evans et al. 2015; Kokkinos et al. 2018; Abut et al. 2019; Ashfaq et al. 2022). In our work, only non-exercise features that are convenient to collect in various environments were included.

To elaborate which features contribute most to the prediction of $\dot{V}O_{2\text{peak}}$ and PPO, the results of the SHAP analysis were considered. As in previous studies (Myers et al. 2001; da Silva et al. 2012), our findings indicate that sex has a significant influence on the prediction when the small feature set is applied. To determine what accounts for the difference between males and females, we fitted the models to the sex-separated data and examined the big feature set in the following.

Body composition variables proved to be the most influential features of $\dot{V}O_{2\text{peak}}$ and PPO. Especially SMM, ECW, and TBW seem to be important predictors. The SHAP analysis showed that high body composition values associated with high SMM led to increased $\dot{V}O_{2\text{peak}}$ and PPO. This can be explained by the fact that muscle fibers consume oxygen and fiber cross-sectional areas increase linearly with PPO (Appelman et al. 2024). Furthermore, there exists a linear relationship between the power output and the $\dot{V}O_2$ increment rate.

Subjective features such as health-related questionnaires, as well as demographic data appeared to be less relevant, particularly for the prediction of $\dot{V}O_{2\text{peak}}$ for females and PPO for both sexes. The SHAP values for the prediction of

$\dot{V}O_{2\text{peak}}$ and PPO in males attribute a relevant significance to the waist–hip ratio and height. This indicates that anthropometric data should continue to be used for the predictions and should not be completely excluded.

The SHAP analysis showed that it may be beneficial to determine the body composition of the participant before conducting a CPET to adapt a ramp protocol on the cycle ergometer to the characteristics of the participant. With the prediction of PPO, an adaptation of the protocol can be implemented quickly. To effectively utilize the predicted $\dot{V}O_{2\text{peak}}$ from this work, the $\dot{V}O_2$ and power output relationship can be considered, which is approximately 10 ml/W/min. The mean response time of $\dot{V}O_2$ for ramp protocols is about 40 s (Caen et al. 2020). Using these two variables, the rate of power increase required to reach the predicted $\dot{V}O_{2\text{peak}}$ in a given time can be calculated.

A limiting factor of this work is that the PPO is influenced by the choice of power increase (Poole and Jones 2017). This reduces the reliability of the models for predicting PPO as they are based on the ramp protocols performed in this study. Consequently, we included $\dot{V}O_{2\text{peak}}$ as an outcome parameter since it can be achieved despite different power output slopes (Iannetta et al. 2020). Moreover, the generalizability of the prediction models is limited to the investigated population, comprising mainly healthy young European adults who were physically able to perform a CPET.

The results can help to adjust power increase in a ramp protocol to achieve volitional exhaustion within a certain duration. This facilitates the comparison of CPETs between different test settings, clinical environments, and studies. In addition, the results can be used to evaluate the effectiveness of an intervention to increase PPO or $\dot{V}O_{2\text{peak}}$. This involves adapting the power increase in a CPET before and after the intervention using the machine learning prediction models. The results could be used to assess an individual's exercise tolerance by using the machine learning prediction models to determine when an individual's measured $\dot{V}O_{2\text{peak}}$ is significantly different from the predicted values. In future analysis, the machine learning approach can be extended to predictions for clinical populations by adding disease-specific features. In addition, the population can be extended to a wider age range, different body mass classes and lower fitness levels.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00421-024-05543-x>.

Author contributions PZ and DW conceptualized the study. PZ, DW and MS contributed to the study design. RS and DW helped with the implementation. CW was the principal investigator. TL, AG, RG, AS and CW contributed to data analysis and interpretation. CW drafted the manuscript, and all authors contributed to revisions and approved the final manuscript. PZ is responsible for the overall content.

Funding Open Access funding enabled and organized by Projekt DEAL. No funds, grants, or other support was received.

Data availability statement Data are available on reasonable request.

Declarations

Conflict of interest No conflicts of interest, financial or otherwise, are declared by the authors.

Ethical approval Written informed consent of the participants was required. The study was reviewed and approved by the ethics committee of the Leibniz Institute at TU Dortmund University (No. 209 and date of approval: November 2021).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

References

- Abut F, Akay MF, George J (2019) A robust ensemble feature selector based on rank aggregation for developing new $VO_{2\text{max}}$ prediction models using support vector machines. *Turk J Electr Eng Comput Sci* 27:3648–3664. <https://doi.org/10.3906/elk-1808-138>
- Akay F, Abut F (2015) Machine learning and statistical methods for the prediction of maximal oxygen uptake: recent advances. *MDER*. <https://doi.org/10.2147/MDER.S57281>
- Antunes BM, Campos EZ, Parmezani SS et al (2017) Sleep quality and duration are associated with performance in maximal incremental test. *Physiol Behav* 177:252–256. <https://doi.org/10.1016/j.physbeh.2017.05.014>
- Appelman B, Charlton BT, Goulding RP et al (2024) Muscle abnormalities worsen after post-exertional malaise in long COVID. *Nat Commun* 15:17. <https://doi.org/10.1038/s41467-023-44432-3>
- Armstrong T, Bull F (2006) Development of the world health organization Global Physical Activity Questionnaire (GPAQ). *J Public Health* 14:66–70. <https://doi.org/10.1007/s10389-006-0024-x>
- Ashfaq A, Cronin N, Müller P (2022) Recent advances in machine learning for maximal oxygen uptake ($VO_{2\text{max}}$) prediction: a review. *Inform Med Unlocked* 28:100863. <https://doi.org/10.1016/j.imu.2022.100863>
- Bassett DR, Howley ET (2000) Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Med Sci Sports Exerc* 32:70–84. <https://doi.org/10.1097/00005768-20001000-00012>
- Booth FW, Roberts CK, Laye MJ (2012) Lack of exercise is a major cause of chronic diseases. *Compr Physiol* 2:1143–1211. <https://doi.org/10.1002/cphy.c110025>
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>

- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Taylor & Francis. <https://doi.org/10.1201/9781315139470>
- Buchfuhrer MJ, Hansen JE, Robinson TE et al (1983) Optimizing the exercise protocol for cardiopulmonary assessment. *J Appl Physiol* 55:1558–1564. <https://doi.org/10.1152/jappl.1983.55.5.1558>
- Burnley M, Vanhatalo A, Jones AM (2012) Distinct profiles of neuromuscular fatigue during muscle contractions below and above the critical torque in humans. *J Appl Physiol* 113:215–223. <https://doi.org/10.1152/japplphysiol.00022.2012>
- Buysee DJ, Reynolds CF, Monk TH et al (1989) Pittsburgh Sleep Quality Index. *J Clin Psychol Med Settings*. <https://doi.org/10.1037/t05178-000>
- Caen K, Boone J, Bourgois JG et al (2020) Translating ramp $\dot{V}O_2$ into constant power output: a novel strategy that minds the gap. *Med Sci Sports Exerc* 52:2020–2028. <https://doi.org/10.1249/MSS.0000000000002328>
- American College of Sports Medicine (ed) (2021) ACSMs guidelines for exercise testing and prescription, 11th edn. Wolters Kluwer, Philadelphia
- Cunha F, Midgley A, Montenegro R et al (2015) Utility of a non-exercise $\dot{V}O_{2\max}$ prediction model for designing ramp test protocols. *Int J Sports Med* 36:796–802. <https://doi.org/10.1055/s-0034-1395590>
- da Silva SC, Monteiro WD, Cunha FA et al (2012) Determination of best criteria to determine final and initial speeds within ramp exercise testing protocols. *Pulm Med* 2012:1–10. <https://doi.org/10.1155/2012/542402>
- Evans HJL, Ferrar KE, Smith AE et al (2015) A systematic review of methods to predict maximal oxygen uptake from submaximal, open circuit spirometry in healthy adults. *J Sci Med Sport* 18:183–188. <https://doi.org/10.1016/j.jsams.2014.03.006>
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29:1189–1232
- Glaab T, Taube C (2022) Practical guide to cardiopulmonary exercise testing in adults. *Respir Res* 23:9. <https://doi.org/10.1186/s12931-021-01895-6>
- Glas CAW (2010) Missing data. In: Peterson P, Baker E, McGaw B (eds) International encyclopedia of education, 3rd edn. Elsevier, Oxford, pp 283–288
- Godin G, Shepard R (1985) A simple method to assess exercise behavior in the community. *Can J Appl Sport Sci* 10(3):141–146. <https://doi.org/10.1097/00005768-199706001-00009>
- Hill DW, Poole DC, Smith JC (2002) The relationship between power and the time to achieve $\dot{V}O_{2\max}$. *Med Sci Sports Exerc* 34:709
- Iannetta D, De Almeida AR, Ingram CP et al (2020) Evaluating the suitability of supra- $\dot{P}O_{\text{peak}}$ verification trials after ramp-incremental exercise to confirm the attainment of maximum $\dot{O}_2$ uptake. *Am J Physiol Regul Integr Comp Physiol* 319:R315–R322. <https://doi.org/10.1152/ajpregu.00126.2020>
- Jobson JD (1991) Multiple linear regression. In: Jobson JD (ed) Applied multivariate data analysis: regression and experimental design. Springer, New York, pp 219–398
- Kokkinos P, Kaminsky LA, Arena R et al (2018) A new generalized cycle ergometry equation for predicting maximal oxygen uptake: the Fitness Registry and the Importance of Exercise National Database (FRIEND). *Eur J Prev Cardiol* 25:1077–1082. <https://doi.org/10.1177/2047487318772667>
- Langer RD, da Costa KG, Bortolotti H et al (2020) Phase angle is associated with cardiorespiratory fitness and body composition in children aged between 9 and 11 years. *Physiol Behav* 215:112772. <https://doi.org/10.1016/j.physbeh.2019.112772>
- Li K, Yao S, Zhang Z et al (2022) Efficient gradient boosting for prognostic biomarker discovery. *Bioinformatics* 38:1631–1638. <https://doi.org/10.1093/bioinformatics/btab869>
- Lindauer M, Feurer M, Eggenberger K, et al (2019) Towards assessing the impact of bayesian optimization's own hyperparameters. arXiv. <https://doi.org/10.48550/arXiv.1908.06674>
- Liu Y, Herrin J, Huang C et al (2022) Non-exercise machine learning models for maximal oxygen uptake prediction in national population surveys. medRxiv. <https://doi.org/10.1101/2022.09.30.22280471>
- Myers J (2005) Applications of cardiopulmonary exercise testing in the management of cardiovascular and pulmonary disease. *Int J Sports Med* 26:S49–S55. <https://doi.org/10.1055/s-2004-830515>
- Myers J, Do D, Herbert W et al (1994) A nomogram to predict exercise capacity from a specific activity questionnaire and clinical data. *Am J Cardiol* 73:591–596. [https://doi.org/10.1016/0002-9149\(94\)90340-9](https://doi.org/10.1016/0002-9149(94)90340-9)
- Myers J, Bader D, Madhavan R, Froelicher V (2001) Validation of a specific activity questionnaire to estimate exercise tolerance in patients referred for exercise testing. *Am Heart J* 142:1041–1046. <https://doi.org/10.1067/mhj.2001.118740>
- Nohara Y, Matsumoto K, Soejima H, Nakashima N (2022) Explanation of machine learning models using Shapley additive explanation and application for real data in hospital. *Comput Methods Programs Biomed* 214:106584. <https://doi.org/10.1016/j.cmpb.2021.106584>
- Poole DC, Jones AM (2017) Measurement of the maximum oxygen uptake $\dot{V}O_{2\max}$: $\dot{V}O_{2\text{peak}}$ is no longer acceptable. *J Appl Physiol* 122:997–1002. <https://doi.org/10.1152/japplphysiol.01063.2016>
- Przednowek K, Barabasz Z, Zadarko-Domaradzka M et al (2018) Predictive modeling of $\dot{V}O_{2\max}$ based on 20 m shuttle run test for young healthy people. *Appl Sci* 8:2213. <https://doi.org/10.3390/app8112213>
- Ray S (2019) A quick review of machine learning algorithms. In: 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon), pp 35–39
- Refaeilzadeh P, Tang L, Liu H (2016) Cross-validation. In: Liu L, Özsu MT (eds) Encyclopedia of database systems. Springer, New York, pp 1–7
- Ross R, Blair SN, Arena R et al (2016) Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association. *Circulation*. <https://doi.org/10.1161/CIR.0000000000000461>
- Saengsuwan J, Nef T, Hunt KJ (2017) A method for predicting peak work rate for cycle ergometer and treadmill ramp tests. *Clin Physiol Funct Imaging* 37:610–614. <https://doi.org/10.1111/cpf.12344>
- Sarker IH (2021) Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci* 2:160. <https://doi.org/10.1007/s42979-021-00592-x>
- Schembre SM, Riebe DA (2011) Non-exercise estimation of $\dot{V}O_{2\max}$ using the International Physical Activity Questionnaire. *Meas Phys Educ Exerc Sci* 15:168–181. <https://doi.org/10.1080/1091367X.2011.568369>
- Shen T, Liu D, Lin Z et al (2022) A machine learning model to predict cardiovascular events during exercise evaluation in patients with coronary heart disease. *JCM* 11:6061. <https://doi.org/10.3390/jcm11206061>
- Singh A, Thakur N, Sharma A (2016) A review of supervised machine learning algorithms. In: 2016 3rd international conference on computing for sustainable global development (INDIACom), pp 1310–1315
- Song Y, Lu Y (2015) Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry* 27:130–135. <https://doi.org/10.11919/j.jissn.1002-0829.215044>
- Sreevalsan-Nair J (2020) K-nearest neighbors. In: Daya Sagar BS, Cheng Q, McKinley J, Agterberg F (eds) Encyclopedia of

- mathematical geosciences. Springer International Publishing, Cham, pp 1–3
- Tanaka H, Monahan KD, Seals DR (2001) Age-predicted maximal heart rate revisited. *J Am Coll Cardiol* 37:153–156. [https://doi.org/10.1016/S0735-1097\(00\)01054-8](https://doi.org/10.1016/S0735-1097(00)01054-8)
- Van Rossum G, Drake FL (2009) Python 3 reference manual. CreateSpace, Scotts Valley
- Vanhatalo A, Fulford J, DiMenna FJ, Jones AM (2010) Influence of hyperoxia on muscle metabolic responses and the power–duration relationship during severe-intensity exercise in humans: a ³¹P magnetic resonance spectroscopy study. *Exp Physiol* 95:528–540. <https://doi.org/10.1113/expphysiol.2009.050500>
- Wier LT, Jackson AS, Ayers GW, Arenare B (2006) Nonexercise models for estimating $\dot{V}O_{2\max}$ with waist girth, percent fat, or BMI. *Med Sci Sports Exerc* 38:555. <https://doi.org/10.1249/01.mss.0000193561.64152>

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