Prediction of peak oxygen consumption using cardiorespiratory parameters from warm-up and submaximal stage of treadmill cardiopulmonary exercise test

Maciej Rosol¹*, Monika Petelczyc², Jakub S. Gąsior³, Marcel Młyńczak¹

¹ Warsaw University of Technology, Faculty of Mechatronics, Institute of Metrology and Biomedical Engineering, Boboli 8 Street, Warsaw, Poland
² Warsaw University of Technology, Faculty of Physics, Warsaw, Poland
³ Medical University of Warsaw, Department of Pediatric Cardiology and General Pediatrics, Warsaw, Poland
*maciej.rosol.dokt@pw.edu.pl

Abstract

This study investigates the quality of peak oxygen consumption (VO₂peak) prediction based on cardiac and respiratory parameters calculated from warmup and submaximal stages of treadmill cardiopulmonary exercise test (CPET) using machine learning (ML) techniques and assesses the importance of respiratory parameters for the prediction outcome. The database consists of the following parameters: heart rate (HR), respiratory rate (RespRate), pulmonary ventilation (VE), oxygen consumption (VO₂) and carbon dioxide production (VCO₂) obtained from 369 treadmill CPETs. Combinations of features calculated based on the HR, VE and RespRate time-series from different stages of CPET were used to create 11 datasets for VO₂peak prediction. Thirteen ML algorithms were employed, and model performances were evaluated using cross-validation with mean absolute percentage error (MAPE), R² score, mean absolute error (MAE), and root mean squared error (RMSE) calculated after each iteration of the validation. The results demonstrated that incorporating respiratory-based features improves the prediction of VO₂peak. The best results in terms of R² score (0.47) and RMSE (5.78) were obtained for the dataset which included both cardiac- and respiratory-based features from CPET up to 85% of age-predicted HRmax, while the best results in terms of MAPE (10.5%) and MAE (4.63) were obtained for the dataset containing cardiorespiratory features from the last 30 seconds of warmup. The study showed the potential of using ML models based on cardiorespiratory features from submaximal tests for prediction of VO₂peak and highlights the importance of the monitoring of respiratory signals, enabling to include respiratory parameters into the analysis. Presented approach offers a feasible alternative to direct VO₂peak measurement, especially when specialized equipment is limited or unavailable.

Keywords: cardiorespiratory fitness, peak oxygen consumption, cardiorespiratory signals, submaximal exercise test, respiratory rate, machine learning, time-series modeling

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
1. Introduction

Peak oxygen consumption (VO₂peak) obtained through cardiopulmonary exercise test (CPET) is the gold standard measure of cardiorespiratory fitness (1). It is a reliable predictor of cardiac events, as well as lung cancer and liver transplantation survival and risk of postoperative complications (2–6). Moreover, VO₂peak is a predictor of sport performance (7–9) and planetary mission task performance during spaceflight (10). Although CPET is the most reliable form of test, it is costly, requires specialized personnel and advanced equipment (11).

While conducting CPET, heart rate (HR) data are usually obtained through electrocardiography (ECG), while respiratory rate (RespRate) and pulmonary ventilation (VE) are gathered using tight-fitting masks. Nevertheless, this data can be acquired with relative ease, using heart rate monitors or smartwatches in case of HR, and impedance pneumography (IP) in case of RespRate and VE (12,13). Moreover, CPET is physically demanding as assumes the participants’ exhaustion and thus it is contraindicated for patients with acute myocardial infarction, unstable angina, uncontrolled arrhythmia causing symptoms or hemodynamic compromise, uncontrolled asthma, and other pathological conditions (11). Maximal cardiopulmonary exercise test might also interfere with an athletes training program (14).

Actually, thanks to: a) the growing development of aforementioned measurement devices, b) availability of simply field-based tests such as incremental shuttle walk test (15,16) and c) new statistical prediction models and equations, clinicians and/or researchers are able to estimate VO₂peak, and/or VO₂max, based on selected parameters without performing maximal CPET (17–22). Unfortunately, estimated VO₂peak using, e.g., only 6-min Walk Test distance demonstrated poor agreement with measured VO₂peak from a CPET (23). Addition of other data such as demographic, anthropometric, and functional characteristics improved the accuracy of VO₂peak estimate based on walking tests at least in elderly patients with stable coronary artery disease (model with all variables explained 73% of VO₂peak variance) (24). Thus, estimation of peak oxygen consumption based on combination of demographic factors and cardiac parameters obtained during submaximal (or even not) physical effort is possible, however, may be biased.

Reliable and accurate estimation of VO₂peak without performing maximal CPET may require more input physiological data to perform more sophisticated analyses. Thus, the development of new prediction models or equations, which will be able to accurately estimate VO₂peak, and/or VO₂max, and will not relies on performing maximal CPET, is still ongoing (18,25). In recent years with the growth of the popularity of machine learning tools (ML) incorporated during the data analysis phase, those techniques were also utilized for the prediction of VO₂ kinetics and VO₂max (26,27). ML models were also used by Szijarto et al. for prediction of VO₂peak based on the anthropometric data and 2D echocardiography (2DE) (28). This approach was more accurate than a model based on anthropometric factors, however, it required performing a 2DE examination with sophisticated equipment and a trained physician. Importantly, not only the model or prediction algorithm might be important in terms of the prediction accuracy, but also the features used for the training. There are existing studies utilizing respiratory rate and ML for prediction of oxygen uptake dynamics during CPET (29–31). However, to the best of our knowledge, there have been no previous studies utilizing
features from cardiorespiratory time-series obtained from submaximal CPET, for the prediction of \( \text{VO}_{2\text{peak}} \) using ML models.

The aim of this paper was hence to investigate the quality of \( \text{VO}_{2\text{peak}} \) prediction by models based on cardiac and respiratory features obtained from different stages of CPET. Additionally, we assessed the importance of respiratory-based features included in the models for \( \text{VO}_{2\text{peak}} \) prediction.

2. Materials and Methods

2.1. Data and study population

The database of cardiorespiratory time-series acquired during treadmill maximal cardiopulmonary exercise tests presented by Mongin et al. was used (32,33). The database comprises 992 recordings from experiments undertaken among amateur and professional athletes in the Exercise Physiology and Human Performance Lab of the University of Málaga between 2008 and 2018 with two types of protocols: a continuous increase of treadmill speed and a graded approach. The test itself was preceded by a warmup at 5 km/h. The study was conducted according to the principles of the Declaration of Helsinki, the study protocol was approved by the Research Ethics Committee of the University of Málaga, written informed consent was obtained from the participants and all the data were analyzed anonymously.

During each test, the following cardiorespiratory time-series were acquired: heart rate (HR), respiratory rate (RespRate), pulmonary ventilation (VE), oxygen uptake (\( \text{VO}_{2} \)) and carbon dioxide production (\( \text{VCO}_{2} \)). Data were acquired on a breath-to-breath basis. HR was monitored via a 12-lead ECG (Mortara Instrument, Inc., USA), while respiratory signals were obtained using the CPX MedGraphics gas analyzer system (Medical Graphics Corporation, USA) (32).

Participants between 18 and 40 years old were chosen for the analysis reducing the sample size to 692. Tests only with ramp speed increments were selected in order to obtain more consistent conditions along the study population. In result, 485 recordings have left. Next, subjects who were determined as outliers based on the 1.5 interquartile range method in terms of weight, height, and \( \text{VO}_{2\text{peak}} \), with respect to the given sex, were excluded from the study, limiting to 462 recordings. Furthermore, the obtained data was visually evaluated in order to discard measurements during which there were visible artefacts in HR acquisition (e.g., sudden drop of over 30 bpm or lack of continuity of HR time-series during CPET probably due to electrode detachment); ultimately 369 recordings became background for the analysis. The final recordings belong to 327 unique subjects (42 subjects had more than one test) including 275 men and 52 women. The demographic summary of the final group is presented in Table 1.

<table>
<thead>
<tr>
<th>Age [years]</th>
<th>Height [cm]</th>
<th>Weight [kg]</th>
<th>BMI</th>
<th>( \text{VO}_{2\text{peak}} ) [ml/min/kg]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>27.3 ± 5.8</td>
<td>177.4 ± 6.3</td>
<td>76.6 ± 8.3</td>
<td>24.3 ± 2.2</td>
</tr>
<tr>
<td>Woman</td>
<td>26.9 ± 6.3</td>
<td>165.2 ± 6.1</td>
<td>62.2 ± 8.2</td>
<td>22.8 ± 2.3</td>
</tr>
</tbody>
</table>
2.2. Modeling

Based on the aforementioned dataset, we decided to investigate the quality of \( \text{VO}_2\text{peak} \) prediction from different stages of CPET based on cardiac and respiratory parameters, and to assess the importance of respiratory-based features included in the modeling of \( \text{VO}_2\text{peak} \). For this purpose, we utilized recorded time-series of HR, RespRate, and VE. \( \text{VO}_2\text{peak} \) was determined as the maximal value of the signal obtained after a 15-breath \( \text{VO}_2 \) moving average window according to the recommendation presented by Robergs et al. (34).

As features for ML models, basic statistics such as mean, standard deviation, maximal and minimal value, median, 25\(^{th}\) and 75\(^{th}\) quantile, skewness, kurtosis, coefficient from linear regression, impulse and shape factors were calculated for HR, RespRate, and VE, for a given stage of the maximal CPET. On this basis, 11 datasets were created based on different combinations of parameters and CPET stages, as presented in Table 2. Our research is focused on the submaximal stage from the cardiopulmonary exercise test, which equals 85% of the maximal measured and age-predicted HR\(_{\text{max}}\) as a threshold. Studied value of HR termination is commonly used in submaximal testing (35–37). We also used both actual HR\(_{\text{max}}\) obtained during the treadmill cardiopulmonary exercise test, and age-predicted HR\(_{\text{max}}\) (220-age) in order to provide insights about the utility of the prediction of \( \text{VO}_2\text{peak} \) in submaximal tests without prior knowledge about the value of HR\(_{\text{max}}\) for a given subject. The example plot of the signals, alongside the threshold for all the stages of the CPET for which the features were calculated, is presented in Figure 2.

The 10-fold cross-validation (CV) was used to assess the accuracy of the prediction. In each iteration, standardization of the non-categorical features based on the mean and standard deviation from the training dataset was performed. The only feature that was not standardized was participants’ sex: -1 was assigned to male, and 1 to female subjects. Different ML algorithms were utilized: Linear, Lasso and Ridge Regression, Random Forest, XGBoost, Multilayer perceptron, Epsilon-Support Vector Regression, Bayesian Ridge Regression, Bayesian Automatic Relevance Determination (ARD) Regression, Gaussian Process Regression, Gradient Boosting for Regression, Huber Regression and Theil-Sen Estimator (38–40). The hyperparameter tuning was performed for each algorithm using the grid-search technique. In each iteration of the validation, metrics like mean absolute percentage error (MAPE), \( R^2 \) score, mean absolute error (MAE) and root mean squared error (RMSE) were calculated. The best model for each dataset was determined based one the lowest MAPE score (which was chosen arbitrarily) obtained from the cross-validation. For the best model, Lin concordance correlation coefficient was calculated and results were visualized as the dependency between predicted and actual values of \( \text{VO}_2\text{peak} \) and as Bland-Altman plot.

Metrics obtained from all datasets were pairwise compared using the Wilcoxon signed-rank test. The significance level was set to 0.05. For the calculations, Python 3.9.13 was used. The whole modeling pipeline is presented in Figure 2.

<table>
<thead>
<tr>
<th></th>
<th>27.3 ± 5.9</th>
<th>175.5 ± 7.6</th>
<th>74.5 ± 9.7</th>
<th>24.1 ± 2.3</th>
<th>46.3 ± 8.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Characteristics of all datasets with an indication of features belonging to individual datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>subjects’ demography</th>
<th>HR features from the last 30 s of warmup</th>
<th>RespRate and VE features from the last 30 s of warmup</th>
<th>HR features from CPET up to 85% of HR(_{\text{max}})</th>
<th>RespRate and VE features from CPET up to 85% of HR(_{\text{max}})</th>
<th>HR features from CPET up to 85% of age-predicted HR(_{\text{max}})</th>
<th>RespRate and VE features from CPET up to 85% of age-predicted HR(_{\text{max}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D5</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D6</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D7</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D8</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>D9</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>D10</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>D11</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>
Figure 1. Typical representation of the time-series for participants with selected fragments used in the analysis. Part A presents the linearly increasing treadmill speed, part B heart rate fluctuations, part C respiratory rate and part D pulmonary ventilation kinetics. The segment between the blue and orange dashed lines is the last 30 seconds of warmup. The segment between the orange and green lines corresponds to the section of CPET up to 85% of the age-predicted HR$_{\text{max}}$. Finally, the segment between the orange and red lines corresponds to the increasing workload in CPET up to 85% of the measured HR$_{\text{max}}$, which is marked with a red circle on the HR plot.

Figure 2. Modeling pipeline applied for each dataset and algorithm.

2.3. Explainable AI

In order to investigate the importance of the individual features used for ML modeling, explainable artificial intelligence (XAI) tools were applied. For this purpose, the Dalex Python package was used (41). During each iteration of the cross-validation, Shapley values and model-level variable importance based on drop-out loss values were calculated on the test set. After the whole cross-validation, all Shapley values for each sample and feature, as well as mean variable importance values were visualized. For the variable importance, model_parts function of dalex.Explainer class was used. 30 permutation rounds were performed on each variable with MAE as a loss function and no data sampling (argument $N$ was equal to None) due to the small number of samples.

3. Results

The metrics obtained for the best algorithm in terms of the lowest MAPE from the cross-validation for each dataset are presented in Table 3 alongside the model names. The violin-plots of the obtained metrics for each dataset were visualized in Figure 3. The p-values from the Wilcoxon signed-rank test from a pairwise comparison of the metrics are presented in Figure 3.

The lowest MAPE and MAE - 10.51% and 4.63, respectively - were obtained for dataset D11 (demographic data along with cardiac and respiratory features from the last 30 seconds of warmup and CPET up to 85% of age-predicted HR$_{\text{max}}$), while the lowest RMSE and highest R$^2$ score (5.78 and 0.47, respectively) were obtained for D9 (demographic data along with cardiac and respiratory features from CPET up to 85% of age-predicted HR$_{\text{max}}$). The worst prediction of VO$_{\text{2peak}}$ in terms of all metrics was achieved by using the D1 (demographic data) dataset. Results obtained for D11 were statistically significantly better in terms of all metrics than results for all the rest of the datasets excluding D9 as presented in Figure 3. Regarding R$^2$ score and RMSE metrics, datasets that included respiratory-based features from the part of CPET (irrespective of HR$_{\text{max}}$ determination, whether measured or estimated) showed statistically
significant superiority over datasets lacking features based on VE and respiratory rate during
the corresponding period as presented in Figure 3. Similarly, for MAPE and MAE, datasets
containing respiratory-based features calculated up to 85% of age-predicted HR$_{max}$
demonstrated significantly better metrics than datasets without such features.

Table 3. Mean and standard deviation of metrics from cross-validation for each dataset for the
model which resulted in the lowest MAPE for the given dataset. The highest metric values were
highlighted.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAPE [%]</th>
<th>$R^2$</th>
<th>MAE [ml/min/kg]</th>
<th>RMSE [ml/min/kg]</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>12.52 ± 2.11</td>
<td>0.26 ± 0.09</td>
<td>5.50 ± 0.80</td>
<td>6.84 ± 0.81</td>
</tr>
<tr>
<td>D2</td>
<td>11.95 ± 1.84</td>
<td>0.31 ± 0.07</td>
<td>5.24 ± 0.84</td>
<td>6.61 ± 0.71</td>
</tr>
<tr>
<td>D3</td>
<td>11.63 ± 1.84</td>
<td>0.34 ± 0.05</td>
<td>5.13 ± 0.75</td>
<td>6.45 ± 0.65</td>
</tr>
<tr>
<td>D4</td>
<td>11.51 ± 1.72</td>
<td>0.36 ± 0.06</td>
<td>5.07 ± 0.68</td>
<td>6.36 ± 0.64</td>
</tr>
<tr>
<td>D5</td>
<td>10.86 ± 1.23</td>
<td>0.44 ± 0.06</td>
<td>4.78 ± 0.52</td>
<td>5.95 ± 0.51</td>
</tr>
<tr>
<td>D6</td>
<td>11.67 ± 1.72</td>
<td>0.34 ± 0.07</td>
<td>5.15 ± 0.69</td>
<td>6.46 ± 0.64</td>
</tr>
<tr>
<td>D7</td>
<td>11.10 ± 1.26</td>
<td>0.42 ± 0.08</td>
<td>4.90 ± 0.53</td>
<td>6.07 ± 0.50</td>
</tr>
<tr>
<td>D8</td>
<td>11.36 ± 1.49</td>
<td>0.38 ± 0.06</td>
<td>4.99 ± 0.61</td>
<td>6.29 ± 0.59</td>
</tr>
<tr>
<td>D9</td>
<td>10.54 ± 1.20</td>
<td>0.47 ± 0.06</td>
<td>4.64 ± 0.49</td>
<td>5.78 ± 0.50</td>
</tr>
<tr>
<td>D10</td>
<td>11.50 ± 1.49</td>
<td>0.36 ± 0.07</td>
<td>5.06 ± 0.62</td>
<td>6.37 ± 0.59</td>
</tr>
<tr>
<td>D11</td>
<td><strong>10.51 ± 1.24</strong></td>
<td>0.47 ± 0.07</td>
<td><strong>4.63 ± 0.52</strong></td>
<td>5.78 ± 0.52</td>
</tr>
</tbody>
</table>

Figure 3. Violin-plots of the calculated metrics for each dataset with the visualization of the
metrics obtained in each iteration of the 10-fold cross-validation. Black dots represent metrics
obtained from datasets without respiratory-based features, while red dots represent these that
include such features.
Figure 4. The p-values from Wilcoxon signed-rank test from pairwise comparison of the metrics obtained from different datasets. P-values smaller than 0.05 are marked with a black background.

The measured values of VO_{2peak} and values predicted for the dataset that obtained the lowest MAPE score (D11) were visualized in Figure 5. The Lin concordance correlation coefficient between predicted and measured VO_{2peak} values was 0.66. The Bland-Altman plot for this dataset is presented in Figure 6.

Figure 5. The plot of measured and predicted VO_{2peak} values for dataset D11. The solid black line represents the function where predicted value is equal to the measured one.

Figure 6. Bland-Altman plot of measured (gold standard from CPET) and predicted VO_{2peak} values based on the results for dataset D11.

As the smallest mean MAPE was obtained for D11, Shapley values and feature importance were visualized for this dataset in Figures 7 and 8, respectively. The discussion of the XAI results can be found in the next section.

Figure 7. Shapley values obtained for dataset D11. Feature names are explained in the Appendix 1.

Figure 8. Variable importance for dataset D11. Feature names are explained in the Appendix 1.

4. Discussion

Considering the features calculated from HR, VE, and RespRate time-series (attainable without the specialized equipment used in CPET), it is possible to predict VO_{2peak} from a submaximal test relying on age-predicted HR_{max}, achieving a mean absolute percentage error of 10.51% (for D11), using Bayesian ARD regression method. The addition of respiratory-based parameters resulted in an improvement of prediction compared to datasets based solely on the corresponding stage of the treadmill cardiopulmonary exercise test in 4 out of 5 cases in terms of R^2 score and RMSE, and 2 out of 5 cases in terms of MAPE and MAE. When limiting treadmill cardiopulmonary exercise test to 85% of age-based HR_{max}, the inclusion of features based on VE and RespRate improved the prediction in terms of all the specified metrics. The fact that the best results were achieved for the dataset considering 85% aged-based HR_{max} and parameters obtained from easily accessible time-series indicates the possibility of using the presented method in clinical practice to determine VO_{2peak} without the prior knowledge of the actual HR_{max} value and the necessity to perform a maximal treadmill cardiopulmonary exercise test.
Obtaining VO\textsubscript{2peak} from maximal CPET might be costly, time-consuming and in some cases impossible or contraindicated to carried out due to observed cardiac or pulmonary dysfunction, musculoskeletal diseases, or strict training programs. Therefore, there is a growing interest in the prediction of VO\textsubscript{2peak} and/or VO\textsubscript{2max} from submaximal tests (14,42–48). Our study focused on investigating ML algorithms to predict VO\textsubscript{2peak} with the set of features, which could be obtained using simpler techniques than commonly used spirometry, and the significance of incorporating respiration into the prediction process. The presented results are similar or superior compared to some other presented VO\textsubscript{2peak} prediction methods like WFI VO\textsubscript{2peak} prediction equation, deep-learning model based on 2DE, or regression models from PACER 20-m shuttle run (19,28,49–52). However, in the existing literature, there are also techniques, which managed to obtain better performance like regression models based on submaximal exercise test protocol using a total body recumbent stepper (53–55). Nonetheless, in those studies more heterogeneous groups of patients were present in terms of age or health status (patients after heart failure or individuals with low to moderate risk of cardiovascular diseases). Further improvement of the prediction of VO\textsubscript{2peak} might be achieved by increasing sample size, and inclusion of other parameters based on the raw signals (especially ECG) like HRV and parameters from information and causal domain (56–59).

Another notable aspect of the study was the utilization of XAI tools, specifically Shapley values and model-level variable importance, to obtain insights into the feature importance for prediction. For most datasets (including D9 and D11, which produced the best results) Bayesian ARD Regression model was used, which has an ability to automatically determine the relevance of each feature, effectively pruning irrelevant or redundant information, while accentuating the impactful variables (60). In our analysis, we found that the top five most influential features were consistent between Shapley values and variable importance. The most impactful feature of the prediction was the maximal value of VE during the test, up to 85% of age-predicted \text{HR}_{\text{max}}. Additionally, subjects' weight and sex influenced the prediction results, with higher VO\textsubscript{2peak} observed in lighter individuals and males compared to females. Notably, 13 out of the 20 features with the highest Shapley values and 10 out of the 15 features with the highest variable importance score were related to respiratory signals. Those findings seem to be in line with results presented in other studies, where the importance of respiratory signals in the context of oxygen consumption was presented (31,61,62). The presented configuration offers the benefit of avoiding monitoring \text{O}_2 consumption and \text{CO}_2 production through laboratory device, instead allowing for the application of less sophisticated respiratory monitoring techniques, such as IP. Simultaneous acquisition of both ECG and IP can be performed using e.g., Pneumonitor device, which is a recently developed device, designed for research in the fields of physiology and sports medicine (12,13,63). Thus, all cardiorespiratory features under current study could be obtained using Pneumonitor without any additional equipment.

There are several limitations of the study. First of all, the raw ECG/RR-intervals signals and raw respiratory curves were unavailable, and thus more sophisticated parameters and parameters from information and causal domains, which could provide additional insights into the predictive models could have not been calculated. Moreover, the sample size in this study was limited, as only 369 recordings from the initial database of 992 CPET recordings were used for analysis after applying exclusion criteria based on outlier detection methods and visual inspection of the signals. Furthermore, the dataset was imbalanced in terms of patients’ sex as
there were 275 men and 52 women. A larger and more balanced dataset could prove beneficial for ML model training. There was also lack of information about the amount of sport activity undertaken by the participants, which might introduce inconsistency in the study population. Additionally, one approach of age-predicted HR\textsubscript{max} calculation and one threshold of HR\textsubscript{max} were introduced. Some of these limitations could be overcome by the usage of the Pneumonitor device, which allows for the simultaneous acquisition of raw ECG and IP signals (63). Thanks to this, the pulmonary activity (including RespRate and VE) can be monitored without the usage of sophisticated apparatus for gas analysis and tight-fitting masks may stress some groups of patients (e.g., children). Future studies may explore the optimal percentage of HR\textsubscript{max} as well as other than treadmill forms of cardiopulmonary exercise tests in order to determine the optimal settings for the prediction of VO\textsubscript{2peak} for clinical practice.

This study expands the discussion on predicting cardiorespiratory fitness by highlighting the important role of submaximal testing and incorporating respiratory signals in the prediction process. The presented analysis indicates that the inclusion of respiratory parameters might improve the quality of the VO\textsubscript{2peak} prediction. The use of a submaximal test based on age-predicted HR\textsubscript{max} and the utilization of cardiological and respiratory parameters that can be obtained without specialized CPET equipment is an advantage of the presented approach and facilitates its potential application in clinical practice.

5. Supporting information

Feature names presented in Figures 7 and 8 are explained in the S1 Appendix.

6. Author Contributions

**Conceptualization:** Maciej Rosol, Monika Petelczyc

**Data Curation:** Maciej Rosol, Monika Petelczyc

**Formal Analysis:** Maciej Rosol

**Methodology:** Maciej Rosol, Monika Petelczyc, Jakub S. Gąsior, Marcel Młyńczak

**Software:** Maciej Rosol

**Visualization:** Maciej Rosol

**Writing – Original Draft Preparation:** Maciej Rosol

**Writing – Review & Editing:** Maciej Rosol, Monika Petelczyc, Jakub S. Gąsior, Marcel Młyńczak

7. Acknowledgments

Research was founded by POB Biotechnology and biomedical engineering of Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme.
8. References


Figure 1
Dataset

10-fold cross-validation

Test set

Train set

Data standardization

Data standardization

Standardized test set

Standardized train set

Model training

Metrics calculation

Final metrics calculation and results visualization

Figure 2
Figure 4
Figure 6
Figure 7
Figure 8

Variable Importance

VO2_peak

VE_RUN_85_HRMAX_AGE_maximum

Weight

Sex

RespRate_RUN_85_HRMAX_AGE_std

RespRate_RUN_85_HRMAX_AGE_Ir_alpha

VE_RUN_85_HRMAX_AGE_mean

VE_WARMUP_impuls_factor

VE_RUN_85_HRMAX_AGE_median

RespRate_WARMUP_Ir_alpha

VE_RUN_85_HRMAX_AGE_q75

HR_WARMUP_kurtosis

RespRate_RUN_85_HRMAX_AGE_q25

HR_WARMUP_skewness

RespRate_RUN_85_HRMAX_AGE_q75

HR_RUN_85_HRMAX_AGE_max_min

4.5  5  5.5  6  6.5  7  7.5

drop-out loss

+2.687  +1.836  +1.106  +0.17  +0.093  +0.09  +0.061  +0.05  +0.039  +0.036  +0.028  +0.021  +0.015  +0.015  +0.005