Oxynet: a collective intelligence that detects ventilatory thresholds in cardiopulmonary exercise tests

A. **Zignoli**[1,2,3,*], A. **Fornasiero**[2,4], P. **Rota**[5], V. **Muollo**[6], L. **Peyré-Tartaruga**[7], D.A. **Low**[8], F.Y. **Fontana**[9], D. **Besson**[10], M. **Pühringer**[11], S. **Ring-Dimitriou**[11] and L. **Mourot**[12,13]

1. Department of Industrial Engineering, University of Trento, Trento, Italy
2. CeRiSM Research Centre, University of Verona, Trento, Italy
3. ProM Facility, Trentino Sviluppo, Trento, Italy
4. Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy
5. Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
6. Department of Medicine, Clinical and Experimental Biomedical Sciences, University of Verona, Verona, Italy
7. Exercise Research Laboratory, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
8. Research Institute of Sport and Exercise Sciences, Liverpool John Moores University, Liverpool, U.K.
9. Team Novo Nordisk professional cycling team, Atlanta, USA
10. INSERM, CIC 1432, Module Plurithématique, Plateforme d'Investigation Technologique, Dijon, France; CHU Dijon-Bourgogne, Centre d'Investigation Clinique, Module Plurithématique, Plateforme d'Investigation Technologique, Dijon, France
11. Department of Sport and Exercise Science, University of Salzburg, Salzburg, Austria
12. EA3920 Prognostic Factors and Regulatory Factors of Cardiac and Vascular Pathologies, Exercise Performance Health Innovation (EPHI) platform, University of Bourgogne Franche-Comté, Besançon, France
13. National Research Tomsk Polytechnic University, Tomsk, Russia

*Corresponding author
e-mail to: andrea.zignoli@unitn.it
Mail to: Via Fortunato Zeni 8, 38068, Rovereto, Trento, Italy

Keywords: automatic methods; artificial intelligence; machine learning; deep learning;

Abstract

The problem of the automatic determination of the first and second ventilatory thresholds (VT1 and VT2) from cardiopulmonary exercise test (CPET) still leads to controversy. The reliability of the gold standard methodology (i.e. expert visual inspection) feeds into the debate and several authors call for more objective automatic methods to be used in the clinical practice. In this study, we present a framework based on a collaborative approach, where a web-application was used to crowd-source a large number (1245) of CPET data of individuals with different aerobic fitness. The resulting database was used to train and test a machine learning (i.e. a convolutional neural network) algorithm. This automatic classifier is currently implemented in another web-application and was used to detect the ventilatory thresholds in CPET. A total of 206 CPET were used to evaluate the accuracy of the estimations against the expert opinions. The neural network was able to detect the ventilatory thresholds with an average mean absolute error of 178 (198) mlO_2/min (11.1%, r=0.97) and 144 (149) mlO_2/min (6.1%, r=0.99), for VT1 and VT2 respectively. The performance of the neural network in detecting VT1 deteriorated in case of individuals with poor aerobic fitness. Our results suggest the potential for a collective intelligence system to outperform isolated experts in ventilatory thresholds detection. However, the inclusion of a larger number of VT1 examples certified

by a community of experts will be likely needed before the abilities of this collective intelligence can be translated into the clinical use of CPET.

Introduction

Cardiopulmonary exercising test (CPET) is used to evaluate an individual's metabolic response to an increasing exercise intensity/workload that relies mainly on the provision of energy through aerobic metabolism. CPET therefore provides a global assessment of the systems involved in oxygen (O_2) transport and utilisation and carbon dioxide (CO_2) removal throughout the exercise intensity spectrum (Wasserman et al., 2005). These capacities are related to an individual's fitness level and they are affected by any pathological condition causing circulatory, respiratory and/or metabolic impairments. CPET applications include aerobic capacity/level and athletic assessment for performance purposes, medical diagnosis and prognosis of several chronic diseases and the individual prescription of adequate/symptom-limited exercise intensities for optimal physical training/rehabilitation (Balady et al., 2010).

During CPET, the engagement of the cardiovascular and respiratory systems increases progressively. To estimate the body metabolic activity during exercise, a metabolic cart is used to measure gas exchange (i.e., in terms of volume and concentration of inspired and expired gases, O_2 and CO_2, by means of amperometric and infrared gas-receptors) and ventilation (by means of a spiroceptor/flowmeter) at the airways. During the test, the following variables are usually considered: O_2 consumption ($\dot{V}O_2$), CO_2 production ($\dot{V}CO_2$), heart rate (HR), ventilation ($\dot{V}E$), tidal volume ($\dot{V}_T$), end-tidal partial pressure of O_2 ($PetO_2$) and CO_2 ($PetCO_2$), as well as other derived variables reflecting ventilatory efficiency (i.e. ventilatory equivalents for O_2 and CO_2, $\dot{V}E/\dot{V}O_2$ and $\dot{V}E/\dot{V}CO_2$ or respiratory exchange ratio ($RER=\dot{V}CO_2/\dot{V}O_2$)). Specific patterns of variations in cardiorespiratory variables identify important indexes of exercise capacity and parameters for exercise prescription, such as the first (VT1) and second ventilatory thresholds (VT2), which identify the boundaries between moderate- and heavy- and between heavy- and very-heavy-intensity exercise domains, respectively (Jones et al., 2010; Meyer et al., 2005). Additionally, abnormal patterns of responses in cardiorespiratory variables are connected to specific pathophysiological states which can be detected via CPET (Meyer et al., 2005).

The accepted gold standard methodology to evaluate ventilatory thresholds involves a visual detection of breakpoints in the cardio-ventilatory variables conducted by multiple reviewers (Prud'Homme et al., 1984). Usually, VT1 is defined as: 1) the first disproportionate increase in $\dot{V}E$; with 2) a concomitant increase in $\dot{V}E/\dot{V}O_2$ and no increase in $\dot{V}E/\dot{V}CO_2$ and with 3) a concurrent increase in $PetO_2$ with no consequent fall in $PetCO_2$. VT2 is defined as: 1) the second disproportionate increase in $\dot{V}E$; 2) the first systematic increase in $\dot{V}E/\dot{V}CO_2$ and 3) the first systematic decrease in $PetCO_2$. However, other methods considering other variables (e.g. respiratory frequency (M. Jones & Doust, 1998) or HR variability (Merati et al., 2004)) and other relationships between variables (e.g., excess $\dot{V}CO_2$ (Gaskill et al., 2001): $(\dot{V}CO_2)^2/\ \dot{V}O_2 - \dot{V}CO_2$ *vs* $\dot{V}O_2$ or $\dot{V}O_2$/HR (Gaskill et al., 2001)) have been proposed.

Despite the fact that it is extensively and widely used in exercise physiology practices, visual inspection is time-consuming and is affected by intra- and inter-evaluator variability (Gladden et al., 1985). From our own experience, visual detection of the ventilatory thresholds can take several minutes to be properly determined. Furthermore, agreement between evaluators can range from 195 mlO_2/min (expert) to 790 mlO_2/min (novice) regarding VT1 inspection (Dolezal et al., 2017). Similar levels (100 mlO_2/min for VT1 and 130 mlO_2/min for VT2) between experts have been reported by others (Myers et al., 2010; Santos & Giannella-Neto, 2004). To aid experts in providing an objective and time-efficient determination of ventilatory thresholds, a considerable number of different

computational algorithms have been proposed (Ekkekakis et al., 2008; Santos & Giannella-Neto, 2004). The currently available algorithms suffer from three main limitations: 1) they need pre-processing operations, 2) they are highly sensitive to the signal-to-noise ratio and 3) they cannot be refined even if new data is provided. To date, the general consensus is that the best methodology does not consist in using a unique algorithm, but rather combining different algorithms (Ekkekakis et al., 2008; Gaskill et al., 2001) or both visual and automatic methods (Pühringer et al., 2020).

The problem of determining the ventilatory thresholds from CPET data is characterised by ambiguity, contradiction and complexity (Hopker et al., 2011; Meyer et al., 2005). To tame this problem, we decided to act collaboratively, knowing that this hyper-connected world offers great opportunities (e.g. cloud computing, crowdsourcing and computer assisted techniques) to improve the effectiveness of cross-expert communication and collaboration. Therefore, instead of seeking the answer that eliminates all the issues, we recognized that this is an ongoing process, and further actions will always be needed. The aim of this paper is to present the *Oxynet* project, and therefore the potential of a collaborative approach to solve the problem of ventilatory thresholds detection in CPET.

Methods

The structure of the project

In general, artificial intelligence in CPET interpretation is defined as the use of algorithms and software to approximate human cognition in the analysis of CPET data (Hearn et al., 2018; Myers et al., 2014; Zignoli, Fornasiero, Stella, et al., 2019). Deep learning technologies, such as deep neural networks, can be used for CPET data interpretation and, most importantly, their performance can be improved if new data is provided for supervised training (Zignoli, Fornasiero, Bertolazzi, et al., 2019). Particularly, ventilatory thresholds detection in CPET can be considered as a time-series classification problem, whereby an algorithm can be used to classify the domain of exercise intensity and then detect the time that corresponds with a change in the domain. One of the biggest limitations to the development of new artificial intelligence algorithms for CPET interpretation and analysis is data availability. Therefore, we created: 1) a crowdsourcing web-application (oxynet.promfacility.eu), to collect data and then train and test new artificial intelligence algorithms and 2) a web application (oxynetresearch.promfacility.eu) that automatically detects the ventilatory thresholds in new uploaded CPET files. Uploaded data are stored on a server (i.e. a dedicated virtual machine running Ubuntu 18.04) and they are not accessible from the web. The second application has been built to process the CPET data currently available on the server automatically. This application works side-by-side with the web-based application. The application has been developed in Python (ver. 3.8) and the neural network was implemented by means of Tensorflow (ver. 2.2.0). Finally, a website (oxynet.net) with a single landing-page has been created to redirect the users to the web applications and for advertising the project. New collaborators can join the project by sending an email to oxynetcpetinterpreter@gmail.com (see Documentation on oxynet.net). The neural network is scheduled to automatically train daily, using the data available on the server.

The datasets

A total of 1245 CPET files from both healthy subjects and patients and with different incremental protocols were crowd-sourced by a number of contributors (**Tab. 1**). Individual CPET were divided in subgroups by age (young (age<40), adult (40<age<60) and old (age>60)), fitness level (low, medium and high levels based on maximal oxygen consumption relative to body mass (American College of Sports Medicine (Indianapolis & Pescatello, 2014)) and gender (M/F) (**Tab.2**). Informed consents have been collected by the participants before every test. All the tests have been

conducted in accordance with the Helsinki Declaration. CPET files were uploaded in a raw format (breath-by-breath basis) or alternatively with data averaged every 5 or 10 sec. This is because different laboratories use different metabolic carts with different exporting options. Irrespective of the time-basis, data were filtered with a rolling average (20 sec time-window) and interpolated at 1 sec (Robergs et al., 2010). Contributors also provided the values of the ventilatory thresholds for every CPET on a different file. Contributors were all experts, and detected the first and second ventilatory thresholds using the gold-standard methodology (Prud'Homme et al., 1984). In this case, with 'experts' we mean 'people working with cardiorespiratory data and the detection of the ventilatory thresholds on an almost daily basis as part of their research/work activities". We also want to point out that the different contributors have particular expertise in detecting the ventilatory thresholds in the specific populations they tested (e.g. sedentary older adults vs. highly trained athletes). This has ensured that the data included in the manuscript (i.e. the dataset used to train the neural network algorithm) represent the highest standards of quality, and the accuracy obtained in VT detection by the algorithm is the highest attainable at the current stage of the project. At every time-sample, exercise intensity was labelled as follows: below VT1; "moderate", between VT1-VT2; "heavy" and above VT2; "very-heavy". Intensity categorical data were converted to binary vectors with one-hot encoding. The maximal and minimal values of the time-series data were used to standardise the data between −1 and 1, to facilitate the convergence of the optimiser used during the training phase.

	References	Characteristics of the participants	Test type	Metabolic cart	n
1	Zignoli et al. 2019 (Zignoli, Fornasiero, Stella, et al., 2019)	Healthy active and trained males and females	Running, XC skiing, cycling; both ramp and graded tests	Quark PFT Ergo CPET, Cosmed, Italy	253
2	Fornasiero et al. (Fornasiero et al., 2019; Fornasiero, Savoldelli, Fruet, et al., 2018; Fornasiero, Savoldelli, Skafidas, et al., 2018)	Healthy active and trained males and females	Running, XC skiing, cycling; both ramp and graded tests	Quark PFT Ergo CPET, Cosmed, Italy	128
3	Kleinnibbelink et al. (Kleinnibbelink et al., n.d.)	Healthy active males and females	Running ramp test (speed and grade increments)	Oxycon Pro, Carefusion, Germany	21
4	Muollo et al. 2019 (Muollo et al., 2019)	Sedentary males and females, aged 50–80, BMI >27 kg/m^2	Cycling ramp test (15 W/min and 10 W/min for males and females, respectively)	Quark PFT, Cosmed, Italy	81
5	F.Y.F. laboratory data	Healthy active males and females	Cycling ramp test (15 W/min and 10 W/min for males and females, respectively)	MetaMax 3R-B2, Cortex, Biophysics, Germany	74
6	Da Rosa et al., 2019 (da Rosa et al., 2019)	Endurance male runners, aged 18-40	Running ramp test [(0.3 km/h)/25s]	K5, Cosmed, Italy	41
7	Masiero, 2019 (Masiero, 2019)	Distance male runners, aged 20-50.	Running ramp test [(1 km/h)/min]	K5, Cosmed, Italy	16
8	Lanferdini et al., 2020 (Lanferdini et al., n.d.)	Distance male runners, aged 22-48.	Running ramp test [(1 km/h)/min]	K5, Cosmed, Italy	19
9	L.M. laboratory data	Coronary male and female patients, aged 50-83, BMI 21-30	Cycling ramp test (10 W/min)	k4b2, Cosmed, Italy	16
10	L.M. laboratory data	Chronic heart failure female and male patients; aged 50-75; BMI 22-30	Cycling ramp test (10 W/min)	k4b2, Cosmed, Italy	178
11	Pühringer et al. (Pühringer et al., 2020)	Healthy untrained males and females, aged 50-60	Graded test on a cycle ergometer	Master Screen CPX, Jaeger, CareFusion, Germany	100
12	L.M. laboratory data	Coronary and heart failure male	Cycling ramp test	SensorMedics Corporation,	318

		patients, aged 29-75, BMI 22-30	(10 W/min)	Yorba Linda, California	
Tot.					1245

Table 1: Characteristics of the database used to train and test the neural network.

The input to the neural network contained 7 time-series, i.e.: $\dot{V}O_2$, $\dot{V}CO_2$, $\dot{V}E/\dot{V}O_2$, $\dot{V}E/\dot{V}CO_2$, $\dot{V}E$, $PetO_2$ and $PetCO_2$. The output of the neural network contained the probability of falling in the "moderate", "heavy" or "very-heavy" exercise intensity domains. The entire dataset was randomly divided into a training set (~80% of the data, i.e. 996 tests) and the test set (~20% of the data, i.e. 249 tests). VT1 was set when the value of the second output neuron became greater than the value of the first output neuron, and VT2 was set when the value of the third output neuron became greater than the value of the second output neuron (**Fig. 1**).

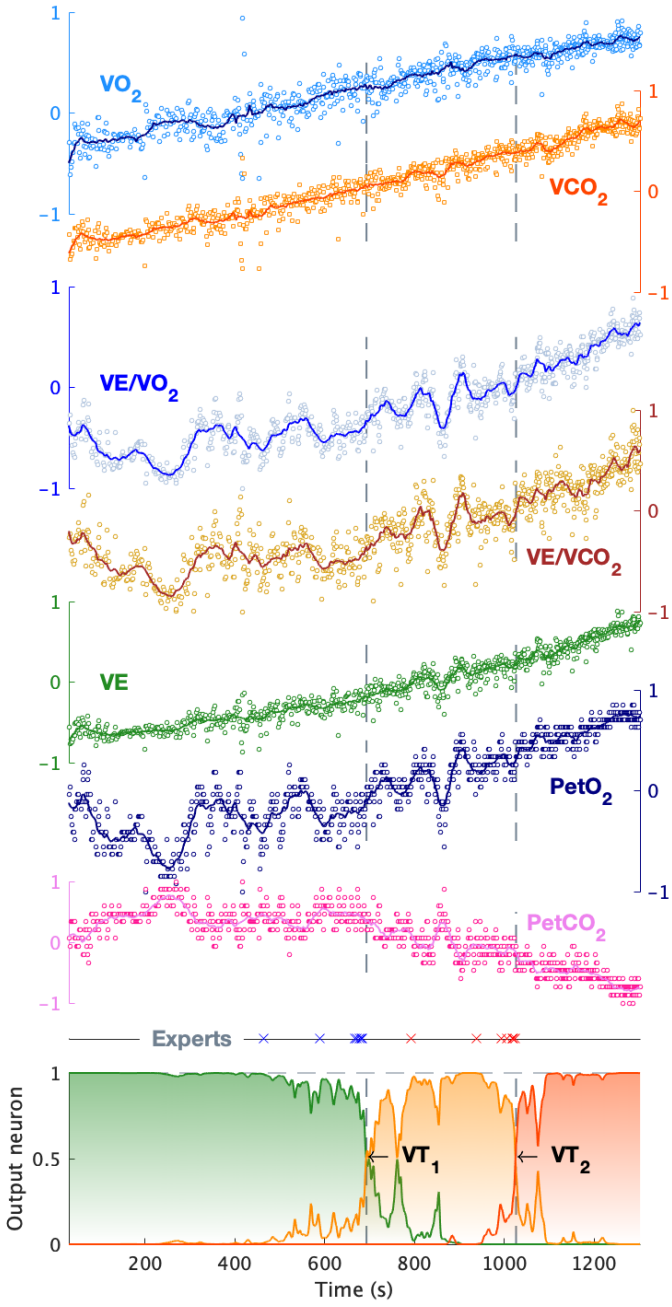


Figure 1: Input and output of the convolutional neural network for a representative subject (data from the current study). From top to bottom: normalised values of O_2 consumption ($\dot{V}O_2$), CO_2 production ($\dot{V}CO_2$), ventilatory equivalents for O_2 and CO_2 (i.e. $\dot{V}E/\dot{V}O_2$ and $\dot{V}E/\dot{V}CO_2$) and minute ventilation ($\dot{V}E$), end-tidal partial pressure of O_2 ($PetO_2$) and CO_2 ($PetCO_2$). The outputs of the three output neurons are provided in green, yellow and red, and represent the probability to be in the "moderate", "heavy" and "very-heavy" exercise intensity domain. First and second ventilatory thresholds (i.e. VT1 and VT2) are detected when the probabilities move from an intensity domain the other. On an additional axis, the VT1 and VT2 detected by 7 couples of experts are also reported (expert evaluations kindly provided by Zignoli et al.).

Statistics

The methodology presented in (Hanneman, 2008) was used to express the agreement between the visual inspection analysis and the ventilatory thresholds estimated by the neural network. 1) We computed model residuals and we checked if they approximated a Normal distribution with the Kolmogorov-Smirnov test, 2) a scatter diagram was created and the correlation coefficient calculated (Pearson's or Spearman's *r* if the residuals were or were not normally distributed, respectively), confidence interval (CI) of the regression coefficient and explained variance R^2 have also been computed, 3) the method of the Bland-Altman plot was applied to compute the bias (with LA) and to detect any trend in the magnitude of the error, 4) the mean absolute and percentage error were calculated (i.e. MAE and MAE%), 5) the standardised difference of the means (Cohen's d) was used to evaluate the magnitude of the differences between the methods (pooled SD were used). A multiple linear regression model with categorical covariates was used to test whether the MAE% was associated with individual age, gender or fitness level (Tab. 2). After fitting the linear regression model, an ANOVA was used to test the significance of the single categorical variable.

	Complete dataset (n=1245)																	
Gender	F (n=251)									M (n=994)								
Aerobic fitness	Low (n=94)			Medium (n=4)			High (n=154)			Low (n=228)			Medium (n=326)			High (n=440)		
Age	Young	Adult	Old	Young	Adult	Old	Young	Adult	Old	Young	Adult	Old	Young	Adult	Old	Young	Adult	Old
Sample size	3	24	66	2	1	1	151	1	2	16	76	136	10	302	14	370	37	33

Table 2: Sample size of the different subgroups of the complete dataset.

Statistical analyses were conducted for both VT1 and VT2 in terms of time and $\dot{V}O_2$. Statistical significance for the *p*-value was set to 0.05. The following criteria were adopted to interpret the magnitude of the correlation *r* between the estimates: < 0.1 trivial, 0.1-0.3 small, 0.3-0.5 moderate, 0.5-0.7 large, 0.7-0.9 very large, and 0.9-1.0 almost perfect. Threshold values for the description of Cohen's *d* were: <0.2, trivial; 0.2–0.6, small; 0.6-1.2, moderate. Bias and precision estimates of ± 270 mlO_2/min (110 s) and ± 120 mlO_2/min (52 s), respectively, were established *a priori* as the maximum parameters that would indicate acceptable agreement between methods and precision of the difference in VT1 (Keir et al., 2014; Zignoli, Fornasiero, Stella, et al., 2019). Similarly, bias and precision of ±171 mlO_2/min (71 s) and ±120 mlO_2/min (47 s), respectively, were established for VT2 (Keir et al., 2014; Zignoli, Fornasiero, Stella, et al., 2019). All analyses were conducted with GraphPad software (ver. 6).

The neural network

The neural network adopted in this study was constituted by two one-dimension (1D) convolutional layers (filters=64 and 32 respectively, kernel size = 2 and RELU activation), a batch normalization layer, a drop-out layer, a max-pooling 1D layer (pool size=2), a dense layer of 8 neurons, a dense layer of 3 (output) neurons with *softmax* activation. The output neurons represented the probability of falling in the "moderate", "high" or "very-heavy" domain, respectively. An *adam* optimiser with *categorical cross-entropy* loss function was used to maximise the accuracy of the estimations. The training dataset entries were shuffled, and the whole dataset was crossed in epochs with 120 of batch size and 40 sec of data. Therefore, the input of the neural network had shape 120x40x5. The weights and biases of the neural network were initialised as random signed values. The learning rate was set to 0.001 and then it was progressively reduced by a factor of 10 every 20 epochs. We implemented the dropout and the class-imbalance restoration technique to prevent overfitting. The hyper-parameters of the neural network were set by trial and error by looking the evolution of the loss and accuracy metrics at different training epochs. A comparison between training and validation metrics was used to determine the number of epochs. The number of training epochs was increased if the training loss was lower than the validation loss and the validation loss trend suggested further improvements where possible.

Results

The ventilatory thresholds values obtained with the neural network were strongly associated with the VT values estimated with the visual inspection analysis (**Tab. 3 and Fig. 2**, Additional Material upper panels). The ability of the model to generalise on new samples is described by the confidence interval (CI) of the correlation coefficients, suggesting that when the neural network will be used on other tests in similar conditions, the strong association will be conserved (applicable in usual circumstances with multiple individuals). Residuals (i.e. differences between the estimates form the neural network and the visual methodology) were normally distributed for both VT1 and VT2 in terms of time (i.e. TVT1 and TVT2) ($p<0.001$) and in terms of $\dot{V}O_2$ (i.e. VO2VT1 and VO2VT2) ($p<0.001$), indicating that the data could be subjected to a parametric analysis. Mean absolute and percentage error values suggested that the differences in the mean ventilatory thresholds values estimated with the neural network and the visual method were trivial ($d<0.2$). The comparisons between the *a priori* established biases suggested that these differences were not clinically meaningful for the great majority of the samples. However, the precision of the neural network might not be sufficient in some isolated cases. The mean absolute errors (MAE) of the estimates (in absolute and relative (%) values) and the values of the standardised difference of the means are reported in **Tab. 3**.

	Regression		Bland-Altman		Mean absolute error		Standardised difference
Var.	**r [CI]**	**R^2**	**Bias (LA)**	**r [CI]**	**MAE (SD)**	**MAE% (SD)**	**Cohen's *d***
TVT1	0.9 [0.87-0.93] **	81%	-14 (223)	0.47 [0.36-0.57] **	77 (85)	21.3 (37)	0.045
TVT2	0.97 [0.96-0.98] **	94%	-28 (157)	0.21 [0.08-0.34] *	61 (60)	9.2 (10.7)	0.03
VO2VT1	0.97 [0.96-0.97] **	94%	-42 (516)	0.1 [-0.04-0.23]	178 (98)	11.1 (18)	0.06
VO2VT2	0.99 [0.98-0.99] **	98%	-75 (378)	-0.09 [-0.22-0.05]	144 (149)	6.1 (7.1)	0.05

Table 3: Results of the statistical analysis: comparisons between time (T) and oxygen consumption ($\dot{V}O_2$) values in correspondence of the first and second ventilatory thresholds (i.e. VT1 and VT2). **p<0.001 *p<0.05

The Bland-Altman plot (**Fig. 2**, Additional Material lower panels) revealed significant moderate and poor correlations between estimation errors and absolute values of VT1 and VO2VT1, respectively (**Tab. 3**). The multiple linear regression model and subsequent ANOVA revealed a

significant association between 1) the MAE% in VT1 estimation and fitness level (p=0.015) and age (p=0.014); and 2) the MAE% in VO2VT1 estimation and fitness level (p=0.006). There was no significant association between 1) the MAE% in VT1 estimation and gender (p=0.483); 2) the MAE% in VO2VT1 estimation and gender (p=0.231) and age (p=0.187); 3) the MAE% in VT2 estimation and gender (p=0.687) and fitness level (p=0.439) and age (p=0.095); 4) the MAE% in VO2VT2 estimation and gender (p=0.349) and fitness level (p=0.162) and age (p=0.194). These results suggest that the accuracy in the estimation of VT1 and VO2VT1 deteriorates for individuals with poor aerobic fitness.

Discussion

We set out to build: 1) a web-based application that facilitates the collection of large CPET datasets and 2) a web-based application that automatically detects ventilatory thresholds in new CPET files. The same approach has been successfully applied in many other fields of medicine, e.g.: LungNet (Mukherjee et al., 2020) and NiftyNet (Gibson et al., 2018) and additionally, in the previous few years, other projects based on crowd-sourced labelled data such as ImageNet (Deng et al., 2009) and WordNet (Miller, 1995), provided millions of annotated samples to the community of data scientists and model developers. It is certainly not beyond the realm of possibility that a similar framework can be replicated, in due proportions, with the *Oxynet* project. In fact, in the last few years, the number of extensive CPET databases used in the scientific literature is surprisingly increasing (e.g. while Beaver et al. (Beaver et al., 1986) in 1986 adopted only 10 CPET to validate their *v-slope* method, Myers et al. (Myers et al., 2010) in 2010 presented a reliability study where a total of 1679 CPET was used and Vainshelboim et al. (Vainshelboim et al., 2017) in 2017 compared different methods to detect the VT with a total of 328 CPET). There are nevertheless key differences between the dataset presented here and the other existing datasets: 1) the participation to this project and hence to this dataset is open to all the experts in the field of CPET, 2) data (including experts' annotations) come from decentralised and delocalized research centres and 3) this dataset is available to researchers who wants to develop new algorithms for CPET interpretation (individual data will not be accessible, but *features* or normalised data can be provided to the interested researchers).

With the rapid increase of data availability, several authors call for a more efficient use of CPET and ventilatory thresholds in clinical practice (Mezzani, 2017), therefore calling for more sophisticated data mining techniques. The solution offered by the *Oxynet* project is based on a collective intelligence, which is used to integrate the opinions and expertise of different centres and laboratories and improve their collaboration with a network. Specifically, a machine learning technology was used to compute a weighted estimation of the ventilatory thresholds, therefore aiming at the best possible solution for each individual. The technology refers to deep neural networks, already widely implemented in the process of CPET data (Hearn et al., 2018; Myers et al., 2014; Zignoli, Fornasiero, Stella, et al., 2019). There are a number of alternative methods for time series classification, such as: time warping, k-nearest neighbour and support vector machine (SVM). However, all these methods require some kind of feature engineering to be conducted before the actual classification is performed. Convolutional neural networks are able to extract features and automatically create informative representations of time series automatically in a single step. Unlike the conventional regression algorithms currently available, the accuracy of the neural networks improves with increased data availability. It is true that the absolute errors reported in the present manuscript are compatible with those reported by Zignoli et al. (Zignoli, Fornasiero, Stella, et al., 2019) (who adopted a recurrent neural network trained with 228 CPET to detect ventilatory thresholds in trained individuals). However, the heterogeneous nature of the CPET files adopted in this work come from different laboratories and cover a wider spectrum of aerobic fitness, testing protocols and exercise modalities. It seems, therefore, more generalizable.

The use of artificial intelligence techniques in automatic ventilatory thresholds detection raises a number of ethical questions, particularly regarding how we report standards for automatic methods *vs* experts. The proposed approach works as a surrogate and as a support to human ability in detecting ventilatory thresholds, while providing a means to reduce variability of the estimates across different experts and centres. In addition, a neural network can perform the inference of the ventilatory thresholds in a few milliseconds for a single CPET (tested on a Mac Book Pro, Intel core 7, 2.8 GHz), while the visual inspection methodology often requires several minutes. Importantly, the neural network output is not the result of the application of a single algorithm; it can be seen as an average of the different opinions of the contributors.

Concerning the state-of-the-art automatic methodologies, our neural network does not require pre-processing operations. This means that once the file is exported from the metabolic cart, the neural network can be directly used for inference (at the moment this is of course true for the metabolic cart models listed in **Tab. 1**). Algorithms for ventilatory detection that do not rely on a machine learning technology are progressively using more and more complex equations to detect the pattern of changes in the cardiorespiratory variables. A common argumentation against the implementation of machine learning models in exercise physiology is that these models do not have a physiological meaning (Zignoli, Fornasiero, Bertolazzi, et al., 2019). However, even the high-degree polynomial regression models currently available in the literature do not underlie any physiological mechanism. The practice of using fixed polynomial equations to detect deflection/breaking points in the cardiorespiratory variables is only justifiable if researchers and clinicians were able to implement those equations easily in their own daily routine. However, high-degree polynomial equations are hard to be handle by non-mathematics experts and require a number of pre-processing operations (e.g. selecting lower and upper limits of the ventilatory thresholds) that might have influence on the final outcome. Today, the complexity of applying machine learning algorithms is compatible to that required for high-degree polynomial equations. Conversely, the web-application we developed, can directly provide estimates for the ventilatory thresholds without any pre-processing effort. Whilst we do not suggest a whole-scale rejection of linear/polynomial regression methods, we strongly sustain the development of new algorithms based on collaborative approaches and collective intelligences.

This study has potential limitations. In terms of $\dot{V}O_2$, the differences between ventilatory thresholds estimated by the neural network and the visual inspection analysis (11.1% and 6.1% for VO2VT1 and VO2VT2, respectively) were slightly greater than the reliability of the visual inspection methodology (9.5% and 4.7% for VO2VT1 and VO2VT2, respectively) (Gladden et al., 1985; Zignoli, Fornasiero, Stella, et al., 2019), lower than the "natural variation" (19% for VO_2VT1) in cardiorespiratory fitness (Rose et al., 2018) and lower than the typical variations observed following endurance training periods (Jones & Carter, 2000). Here, it is important to highlight the potentially low reliability of the visual inspection methodology in individuals with poor aerobic fitness, who constituted a large portion (~25%) of the entire sample of this study. Of particular concern is the association of the error in the estimations with the individual fitness level (VT1 and VO2VT1) and age (only VT1) highlighted by the multiple linear regression model and subsequent ANOVA analyses. However, we think that this is mostly due to the low reliability of the visual inspection methodology (Gladden et al., 1985; Prud'Homme et al., 1984) (which is subjective and depends on experience and can negatively affect the training of the neural network) rather than to a limitation of the collective intelligence approach presented here. However, to the best of our knowledge, poorly is known about the association between the reliability of the visual inspection methodology and the individual fitness level. The fact that the ventilatory thresholds might be hard to be detected is common (Beaver et al., 1986) and, for our own experience, this is particularly true for individuals with poor aerobic fitness and for both visual and automatic methods. For example, the second ventilatory threshold can be only detected by linear-regression methods if there is a 15% increase in the $\dot{V}E$ vs $\dot{V}CO_2$ slope (Beaver et al., 1986). An important contribution to this project could come

from a reliability study, where a large number of CPET are determined by different groups of experts for data validation. An "idiosyncratic" noise is likely associated with each individual evaluation but taking the average over a large number of evaluations will likely get us closer to the ground-truth. This process of course will take time to take place, but it will eventually cancel the effects of the noise (Yi et al., 2012). At the moment, the single inaccurate opinion of one expert might have a negative influence on what the neural network eventually learns. It would be impossible to evaluate the impact of an inaccurate evaluation on the whole algorithm performance, but this impact will be largely mitigated if more trusted collaborators will join the project in the future. Indeed, if a large number of "verified" VT1 examples on the same dataset of CPET will be provided, the output of the neural network output could be optimised to be valuable in many clinical situations, particularly when no clear breakpoints in the ventilatory variables can be detected (Myers et al., 2010). However, more specific research is needed to determine whether the estimations' reliability can be potentially significant in prognostic applications, e.g., when the CPET is adopted for a preoperative assessment (Vainshelboim et al., 2017). For example, the U.S. Food & Drug Administration proposed a regulatory framework (Food and Drug Administration, 2019) to deliver safe and effective artificial intelligence-based software as medical devices. The algorithms developed within the *Oxynet* project must therefore be further developed (especially in terms of reliability and accuracy in VT1 estimation) before they can be adopted in clinical practice.

Furthermore, 43 CPET data files (i.e. 17% of our validation dataset) were not used to compute errors in the estimations. They were excluded from the validation dataset because one or both the ventilatory thresholds could not be visually detected. Other studies reported similar or even higher proportions of "indeterminate" cases following the adoption of automatic methods for ventilatory threshold detection (Ekkekakis et al., 2008). It is worth mentioning that the neural network does not report an "indeterminate" case because it always finds a solution to the problem. This might be due to the fact that that the neural network is exceptionally able to detect changing patterns in the cardio-respiratory variables, but we do not have the data that can support this notion. Regardless, the neural network provided an estimation for the ventilatory threshold in every test, which might not be always a positive characteristic.

Another potential limitation is that we adopted the crowdsourcing strategy to collect data for this study. This strategy aims at engaging a large number of experts to find the best possible solution for everyone. A disadvantage of this approach is that achieving a shared understanding and commitment is a time-consuming process. In fact, different contributors might have slightly different approaches to the ventilatory thresholds identification and might export the data in different formats.

In light of the aforementioned strengths and limitations, we can consider *Oxynet* the first working example of a collective intelligence created to automatically process a CPET. The constant increase of the internet usage worldwide is evident, and CPET data availability will likely increase in parallel (Reeves et al., 2018). These trends suggest that initiatives like *Oxynet* will find fertile ground where to grow, and that collaborative efforts will be needed to develop the new generation of automatic CPET interpreters.

Conclusion

Our findings suggest that a convolutional neural network can be used to estimate ventilatory thresholds from CPET with appropriate accuracy, especially in individuals with medium to high aerobic fitness levels. Likely, the reliability of the method proposed here is affected by the reliability of the gold standard methodology, especially when the first ventilatory threshold is concerned. Therefore, we suggest being cautious when using this neural network instead of the visual inspection methods to determine essential fitness parameters from CPET in clinical settings. This algorithm

should be rather used in parallel with the visual inspection methodology carried out by two independent experts. However, the potential of a collaborative approach based on a collective intelligence to tame the problem of ventilatory thresholds detection in CPET is clearly acknowledged in the current study.

Additional material

Please visit oxynet.net and read the documentation (follow the link to Overleaf).

Competing interest statement

The authors declare no competing interests.

Bibliography

American College of Sports Medicine (Indianapolis, Ind.), & Pescatello, L. S. (2014). *ACSM's guidelines for exercise testing and prescription*. Wolters Kluwer, Lippincott Williams & Wilkins.

Balady, G. J., Arena, R., Sietsema, K., Myers, J., Coke, L., Fletcher, G. F., Forman, D., Franklin, B., Guazzi, M., & Gulati, M. (2010). Clinician's guide to cardiopulmonary exercise testing in adults: A scientific statement from the American Heart Association. *Circulation*, *122*(2), 191–225.

Beaver, W. L., Wasserman, K., & Whipp, B. J. (1986). A new method for detecting anaerobic threshold by gas exchange. *Journal of Applied Physiology*, *60*(6), 2020–2027. https://doi.org/10.1152/jappl.1986.60.6.2020

da Rosa, R. G., Oliveira, H. B., Gomeñuka, N. A., Masiero, M. P. B., da Silva, E. S., Zanardi, A. P. J., de Carvalho, A. R., Schons, P., & Peyré-Tartaruga, L. A. (2019). Landing-Takeoff Asymmetries Applied to Running Mechanics: A New Perspective for Performance. *Frontiers in Physiology*, *10*, 415. https://doi.org/10.3389/fphys.2019.00415

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dolezal, B. A., Storer, T. W., Neufeld, E. V., Smooke, S., Tseng, C.-H., & Cooper, C. B. (2017). A Systematic Method to Detect the Metabolic Threshold from Gas Exchange during Incremental Exercise. *Journal of Sports Science & Medicine*, *16*(3), 396–406.

Ekkekakis, P., Lind, E., Hall, E. E., & Petruzzello, S. J. (2008). Do regression-based computer algorithms for determining the ventilatory threshold agree? *Journal of Sports Sciences*, *26*(9), 967–976. https://doi.org/10.1080/02640410801910269

Food and Drug Administration. (2019). *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)-discussion paper*.

Fornasiero, A., Savoldelli, A., Fruet, D., Boccia, G., Pellegrini, B., & Schena, F. (2018). Physiological intensity profile, exercise load and performance predictors of a 65-km mountain ultra-marathon. *Journal of Sports Sciences*, *36*(11), 1287–1295. https://doi.org/10.1080/02640414.2017.1374707

Fornasiero, A., Savoldelli, A., Skafidas, S., Stella, F., Bortolan, L., Boccia, G., Zignoli, A., Schena, F., Mourot, L., & Pellegrini, B. (2018). Delayed parasympathetic reactivation and sympathetic withdrawal following maximal cardiopulmonary exercise testing (CPET) in hypoxia. *European Journal of Applied Physiology*, *118*(10), 2189–2201. https://doi.org/10.1007/s00421-018-3945-5

Fornasiero, A., Skafidas, S., Stella, F., Zignoli, A., Savoldelli, A., Rakobowchuk, M., Pellegrini, B., Schena, F., & Mourot, L. (2019). Cardiac Autonomic and Physiological Responses to Moderate- intensity Exercise in Hypoxia. *International Journal of Sports Medicine*, *40*(14), 886–896. https://doi.org/10.1055/a-1015-0647

Gaskill, S. E., Ruby, B. C., Walker, A. J., Sanchez, O. A., Serfass, R. C., & Leon, A. S. (2001). Validity and reliability of combining three methods to determine ventilatory threshold: *Medicine and Science in Sports and Exercise*, *33*(11), 1841–1848. https://doi.org/10.1097/00005768-200111000-00007

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D. C., Ourselin, S., Cardoso, M. J., & Vercauteren, T. (2018). NiftyNet: A deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, *158*, 113–122. https://doi.org/10.1016/j.cmpb.2018.01.025

Gladden, L. B., Yates, J. W., Stremel, R. W., & Stamford, B. A. (1985). Gas exchange and lactate anaerobic thresholds: Inter- and intraevaluator agreement. *Journal of Applied Physiology*, *58*(6), 2082–2089. https://doi.org/10.1152/jappl.1985.58.6.2082

Hanneman, S. K. (2008). Design, Analysis, and Interpretation of Method-Comparison Studies: *AACN Advanced Critical Care*, *19*(2), 223–234. https://doi.org/10.1097/01.AACN.0000318125.41512.a3

Hearn, J., Ross, H. J., Mueller, B., Fan, C.-P., Crowdy, E., Duhamel, J., Walker, M., Alba, A. C., & Manlhiot, C. (2018). Neural Networks for Prognostication of Patients With Heart Failure: Improving Performance Through the Incorporation of Breath-by-Breath Data From Cardiopulmonary Exercise Testing. *Circulation: Heart Failure*, *11*(8), e005193.

Hopker, J. G., Jobson, S. A., & Pandit, J. J. (2011). Controversies in the physiological basis of the 'anaerobic threshold' and their implications for clinical cardiopulmonary exercise testing: Controversies in the 'anaerobic threshold.' *Anaesthesia*, *66*(2), 111–123. https://doi.org/10.1111/j.1365-2044.2010.06604.x

Jones, A. M., & Carter, H. (2000). The effect of endurance training on parameters of aerobic fitness. *Sports Medicine*, *29*(6), 373–386.

Jones, A. M., Vanhatalo, A., Burnley, M., Morton, R. H., & Poole, D. C. (2010). Critical Power: Implications for Determination of V˙O2max and Exercise Tolerance: *Medicine & Science in Sports & Exercise*, *42*(10), 1876–1890. https://doi.org/10.1249/MSS.0b013e3181d9cf7f

Keir, D. A., Murias, J. M., Paterson, D. H., & Kowalchuk, J. M. (2014). Breath-by-breath pulmonary O_2 uptake kinetics: Effect of data processing on confidence in estimating model parameters: Breath-by-breath pulmonary O_2 uptake kinetics. *Experimental Physiology*, *99*(11), 1511–1522. https://doi.org/10.1113/expphysiol.2014.080812

Kleinnibbelink, G., Stens, N., Fornasiero, A., Speretta, G., Van Dijk, A., Low, D. A., Oxborough, D., & Thijssen, D. (n.d.). *The acute and chronic effects of high-intensity exercise in hypoxia on blood pressure and post-exercise hypotension*.

Lanferdini, F. J., Silva, E. S., Machado, E., Fischer, G., & Peyré-Tartaruga, L. A. (n.d.). *Physiological predictors of maximal incremental running performance*.

M. Jones, A., & Doust, J. H. (1998). Assessment of the lactate and ventilatory thresholds by breathing frequency in runners. *Journal of Sports Sciences*, *16*(7), 667–675. https://doi.org/10.1080/026404198366470

Masiero, M. P. B. (2019). *Physiomechanics of interval and continuous running at velocity associated with VO2max* [Thesis in Master Science in Human Movement Sciences, Universidade Federal do Rio Grande do Sul]. http://hdl.handle.net/10183/202184

Merati, G., Rampichini, S., Ce, E., Sangiovanni, M., Castiglioni, P., Di Rienzo, M., & Veicsteinas, A. (2004). *Ventilatory threshold detection: A new method based on heart rate variability*. 221–224. https://doi.org/10.1109/CIC.2004.1442912

Meyer, T., Lucía, A., Earnest, C. P., & Kindermann, W. (2005). A Conceptual Framework for Performance Diagnosis and Training Prescription from Submaximal Gas Exchange Parameters—Theory and Application. *International Journal of Sports Medicine*, *26*, S38–S48. https://doi.org/10.1055/s-2004-830514

Mezzani, A. (2017). Cardiopulmonary Exercise Testing: Basics of Methodology and Measurements. *Annals of the American Thoracic Society*, *14*(Supplement_1), S3–S11. https://doi.org/10.1513/AnnalsATS.201612-997FR

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mukherjee, P., Zhou, M., Lee, E., Schicht, A., Balagurunathan, Y., Napel, S., Gillies, R., Wong, S., Thieme, A., Leung, A., & Gevaert, O. (2020). A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets. *Nature Machine Intelligence*, *2*(5), 274–282. https://doi.org/10.1038/s42256-020-0173-6

Muollo, V., Rossi, A., Milanese, C., Masciocchi, E., Taylor, M., Zamboni, M., Rosa, R., Schena, F., & Pellegrini, B. (2019). The effects of exercise and diet program in overweight people – Nordic walking versus walking. *Clinical Interventions in Aging*, *Volume 14*, 1555–1565. https://doi.org/10.2147/CIA.S217570

Myers, J., de Souza, C. R., Borghi-Silva, A., Guazzi, M., Chase, P., Bensimhon, D., Peberdy, M. A., Ashley, E., West, E., Cahalin, L. P., Forman, D., & Arena, R. (2014). A neural network approach to predicting outcomes in heart failure using cardiopulmonary exercise testing. *International Journal of Cardiology*, *171*(2), 265–269. https://doi.org/10.1016/j.ijcard.2013.12.031

Myers, J., Goldsmith, R. L., Keteyian, S. J., Brawner, C. A., Brazil, D. A., Aldred, H., Ehrman, J. K., & Burkhoff, D. (2010). The Ventilatory Anaerobic Threshold in Heart Failure: A Multicenter Evaluation of Reliability. *Journal of Cardiac Failure*, *16*(1), 76–83. https://doi.org/10.1016/j.cardfail.2009.08.009

Prud'Homme, D., Bouchard, C., Leblance, C., Landry, F., Lortie, G., & Boulay, M. R. (1984). Reliability of assessments of ventilatory thresholds. *Journal of Sports Sciences*, *2*(1), 13–24. https://doi.org/10.1080/02640418408729692

Pühringer, M., Ring-Dimitriou, S., Stöggl, T., Iglseder, B., & Paulweber, B. (2020). Comparison of visual, automatic and semiautomatic methods to determine ventilatory indices in 50 to 60 years old adults. *Journal of Sports Sciences*, *38*(6), 692–702. https://doi.org/10.1080/02640414.2020.1725993

Reeves, T., Bates, S., Sharp, T., Richardson, K., Bali, S., Plumb, J., Anderson, H., Prentis, J., Swart, M., & Levett, D. (2018). Cardiopulmonary exercise testing (CPET) in the United Kingdom—A national survey of the structure, conduct, interpretation and funding. *Perioperative Medicine*, *7*(1), 1–8.

Robergs, R. A., Dwyer, D., & Astorino, T. (2010). Recommendations for improved data processing from expired gas analysis indirect calorimetry. *Sports Medicine*, *40*(2), 95–111.

Rose, G. A., Davies, R. G., Davison, G. W., Adams, R. A., Williams, I. M., Lewis, M. H., Appadurai, I. R., & Bailey, D. M. (2018). The cardiopulmonary exercise test grey zone; optimising fitness stratification by application of critical difference. *British Journal of Anaesthesia*, *120*(6), 1187–1194. https://doi.org/10.1016/j.bja.2018.02.062

Santos, E. L., & Giannella-Neto, A. (2004). Comparison of computerized methods for detecting the ventilatory thresholds. *European Journal of Applied Physiology*, *93*(3), 315–324. https://doi.org/10.1007/s00421-004-1166-6

Vainshelboim, B., Rao, S., Chan, K., Lima, R. M., Ashley, E. A., & Myers, J. (2017). A comparison of methods for determining the ventilatory threshold: Implications for surgical risk stratification. *Canadian Journal of Anesthesia/Journal Canadien d'anesthésie*, *64*(6), 634–642. https://doi.org/10.1007/s12630-017-0862-8

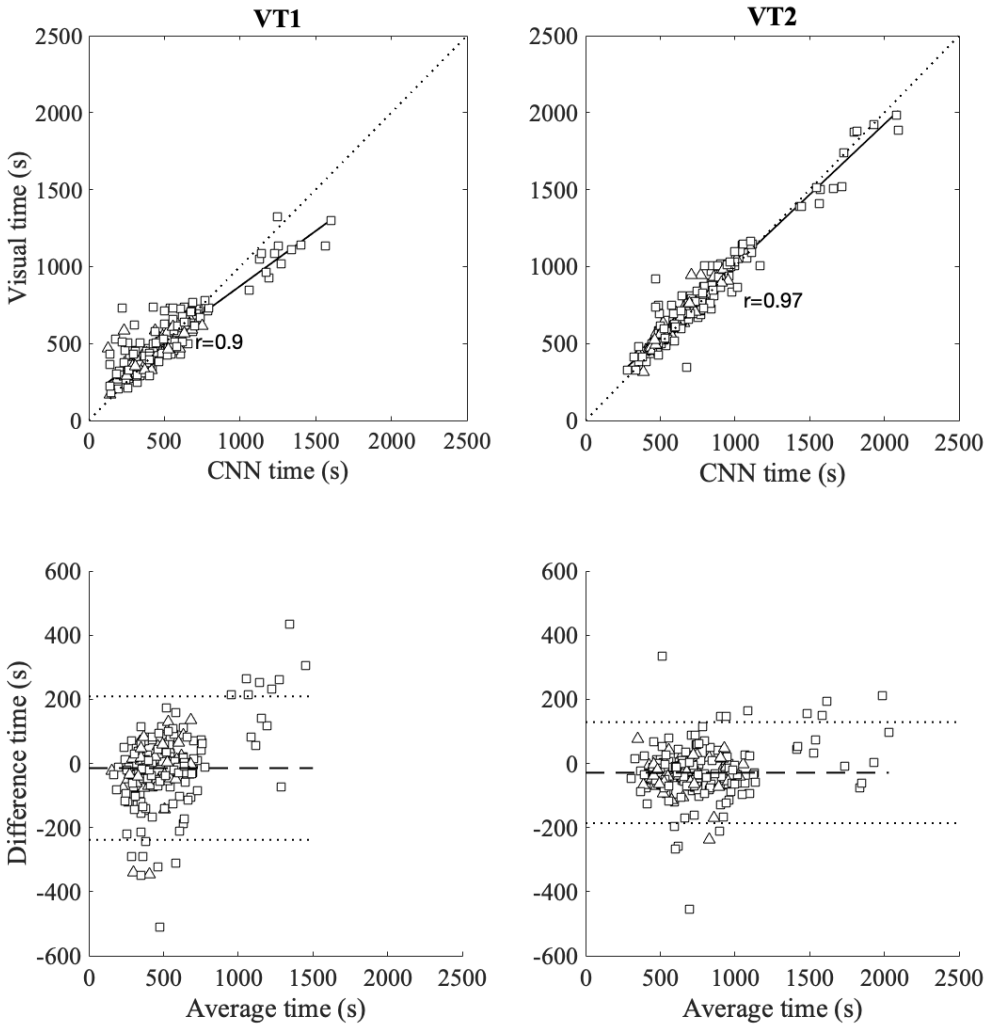
Wasserman, K., Hansen, J. E., Sue, D. Y., Stringer, W. W., & Whipp, B. J. (2005). *Principles of exercise testing and interpretation: Including pathophysiology and clinical applications*. Lippincott Williams & Wilkins Philadelphia.

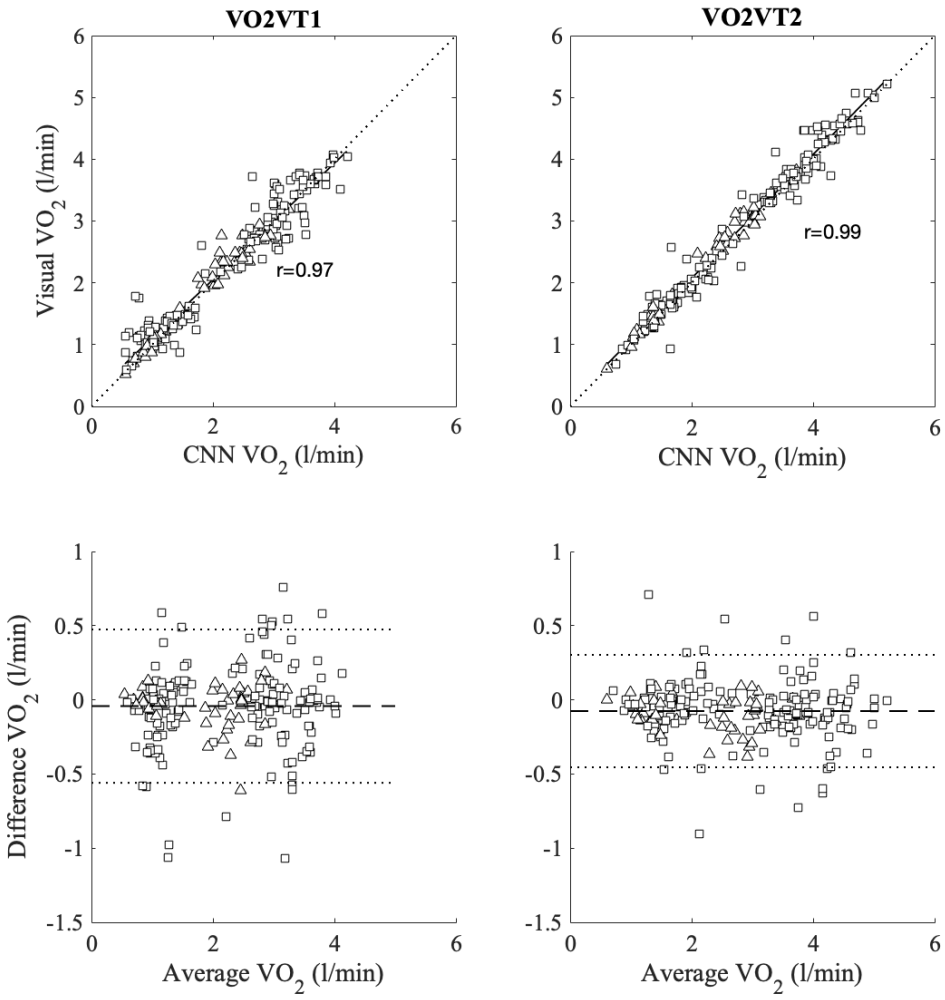
Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The Wisdom of the Crowd in Combinatorial Problems. *Cognitive Science*, *36*(3), 452–470. https://doi.org/10.1111/j.1551-6709.2011.01223.x

Zignoli, A., Fornasiero, A., Bertolazzi, E., Pellegrini, B., Schena, F., Biral, F., & Laursen, P. B. (2019). State-of-the art concepts and future directions in modelling oxygen consumption and lactate concentration in cycling exercise. *Sport Sciences for Health*.

Zignoli, A., Fornasiero, A., Stella, F., Pellegrini, B., Schena, F., Biral, F., & Laursen, P. B. (2019). Expert-level classification of ventilatory thresholds from cardiopulmonary exercising test data with recurrent neural networks. *European Journal of Sport Science*, 1–9.

Additional material: Scatter diagrams (upper graphs) and Bland-Altman plots (lower graphs) for the first (VT1) and second (VT2) ventilatory threshold estimations from visual inspection analysis and the convolutional neural network (CNN). Individual results are classified between men (squares) and women (triangles).




VO2VT1
VO2VT2
Visual VO2 (l/min)
CNN VO2 (l/min)
r=0.97
r=0.99
Difference VO2 (l/min)
Average VO2 (l/min)