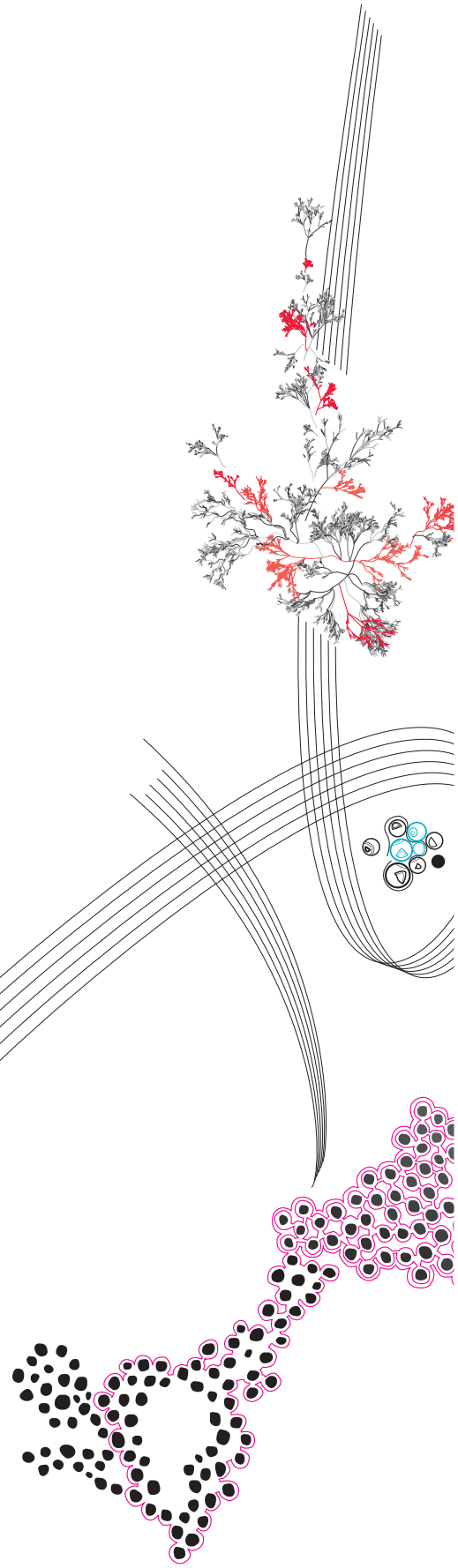


MSc Computer Science
Final Project

Estimating $\dot{V}O_2$ using peripheral and central measurements of wearable sensor data

Gies den Broeder

Supervisor: Richie Goulding, Dees Postma, Alexia Briassouli

January, 2025

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

Oxygen uptake ($\dot{V}O_2$) and its kinetics are important indicators of fitness and health, yet their measurement requires expensive, specialized equipment. This study aimed to validate previously achieved results in predicting $\dot{V}O_2$ and its kinetics using wearable sensors, as well as evaluating whether peripheral measurements could improve predictive accuracy. Data were collected from 14 participants performing three different pseudorandom binary sequence (PRBS) exercise tests at varying intensities. The model achieved high prediction accuracy for $\dot{V}O_2$ with a mean and best repeated-measures correlation of 0.928 and 0.943, respectively. Slightly lower than prior results, likely due to sample heterogeneity. However, predicting $\dot{V}O_2$ kinetics using the Mean Normalized Gain (MNG) method was less successful, showing no meaningful correlation. Incorporating muscle oxygenation (via near-infrared spectroscopy) and body composition data marginally improved prediction accuracy. These findings illustrate the significant potential of wearable sensors for estimating $\dot{V}O_2$ and consequently accessible fitness and health monitoring, though methodological improvements and larger datasets are still required.

Keywords: oxygen uptake prediction, machine learning, wearable sensors, body composition, near-infrared spectroscopy

Chapter 1

Introduction

Oxygen uptake ($\dot{V}O_2$) kinetics and $\dot{V}O_2$max are key indicators of the body's ability to meet changing oxygen demands during physical activity [1]. These metrics integrate responses from the pulmonary, cardiovascular, and muscular systems, reflecting fitness and overall health [1, 2]. Therefore, these metrics are indicative of both exercise performance [3, 4, 5] as well as disease prognosis [6, 7, 8, 9]. As such, being able to easily obtain this measurement during daily activities does not only have the potential to detect changes in physical fitness, but may also identify disease states before symptoms appear [10, 11].

However, obtaining these measurements typically requires specialized, expensive equipment, limiting their availability to clinical and research settings [1]. To enable measurements during daily activities, methods that enable cost-effective and user-friendly estimation of $\dot{V}O_2$ and its kinetics are required. Wearable sensors provide a promising alternative to the conventional systems used to measure $\dot{V}O_2$ data. They are minimally intrusive, can be used outside of laboratory settings and have the ability to acquire large amounts of physiologically relevant data [12]. These data may include measures such as heart rate, air intake or muscle oxygen saturation, which contain relationships with $\dot{V}O_2$. These relationships can be determined by supervised machine learning techniques, presenting an opportunity to measure fitness and health more accessibly and conveniently [13].

Particularly measuring $\dot{V}O_2$ kinetics during (pseudo)random and daily activities has been a significant hurdle, as traditional kinetics analysis is not applicable to activities with varying work rate. However, a promising alternative method has been proposed in the form of Mean Normalized Gain (MNG) [14], though this method has not been used or validated by any other research group. Subsequently, Amelard et al. has shown the potential of a Temporal Convolutional Network (TCN) in predicting $\dot{V}O_2$ using central measurements captured by wearable sensors [13]. Recently, Hedge et al. have achieved encouraging results in predicting both $\dot{V}O_2$ and its kinetics [12], setting the stage for nonintrusive and continuous $\dot{V}O_2$ monitoring and health assessments.

$\dot{V}O_2$ is an integration of the respiratory, cardiovascular and muscular systems. However, the muscles are ultimately using most of the extracted oxygen [15]. Therefore, utilizing techniques such as near-infrared spectroscopy (NIRS) and bioelectric impedance analysis (BIA), through which oxygen utilization of the muscles and body composition statistics can non-invasively be measured [16, 17], respectively, may help improve $\dot{V}O_2$ prediction results. Consequently, the effect on the $\dot{V}O_2$ prediction accuracy of the addition of these measurements will be examined in this work.

Chapter 2

Scientific Background

Oxygen uptake serves as an important indicator of the metabolic demands placed upon the body [1, 18]. The relationship between the power output and $\dot{V}O_2$ has traditionally been characterized as a dynamic first-order linear and symmetric system, where the ratio between power output and $\dot{V}O_2$, the gain of the $\dot{V}O_2$ response, has been shown to be ∼10 $mL\,min^{-1}\,W^{-1}$ [1, 18]. A characteristic of such a system is that the kinetics and gain of the system should be invariant to the amplitude of the work rates [1, 18]. However, the body's response to exercise varies substantially based on work rate, challenging the assumption of dynamic linearity and symmetry of the $\dot{V}O_2$ system [1, 18]. Therefore, exercise is categorized into 4 intensity domains; moderate, heavy, severe and extreme, each bounded by physiological thresholds and distinguished by different physiological responses.

In the moderate domain, O_2 delivery and utilization are balanced, allowing $\dot{V}O_2$ to reach a steady state while the exercise stimulus remains constant [1, 18]. The Lactate Threshold (LT) (or Gas Exchange Threshold (GET)), marks the transition to the heavy domain. At work rates above LT, the slow component emerges, which reflects a slow increase in $\dot{V}O_2$ until steady state is reached. The slow component disrupts both the on- and off-transient symmetry as well as increases the gain of the $\dot{V}O_2$ response [1, 18, 19]. The upper bound of the heavy domain is the Critical Power (CP), which denotes the maximum power output that can be sustained for prolonged periods of time, as all energy demands should be able to be facilitated through aerobic means [20]. The severe domain exists above CP, and within this domain the slow component will force $\dot{V}O_2$ (and lactate) to increase until either maximal $\dot{V}O_2$ ($\dot{V}O_2max$) is reached or the exercise is ceased [1, 18, 21]. Finally, work rates above $\dot{V}O_2max$ are of extreme intensity and $\dot{V}O_2$ will not reach $\dot{V}O_2max$ before exercise has to be ceased.

Pulmonary $\dot{V}O_2$ is primarily driven by the O_2 delivery and uptake occurring in the muscles, as they are the primary consumers of oxygen during exercise [22, 15, 23]. The O_2 utilization can be measured non-invasively using Near-infrared spectroscopy (NIRS). By emitting near-infrared (NIR) light into tissues and measuring transmitted and absorbed light intensities, NIRS devices can measure concentrations of oxygenated hemoglobin (O_2Hb), deoxygenated hemoglobin (HHb) and total hemoglobin (tHB) as the sum of O_2Hb and HHb [16, 24]. Additionally, they can calculate the tissue saturation index (TSI), which represents the percentage of oxygenated hemoglobin relative to total hemoglobin. NIRS has a limitation to extend to broader populations due to its susceptibility to signal interference from adipose tissue [16]. Despite this, NIRS' ability to measure local oxygen dynamics allows it to bridge the gap between O_2 delivery and utilization.

The interplay between pulmonary and muscular oxygen dynamics becomes evident in the concept of the O_2 deficit, which is the difference between energy demand and aerobically facilitated energy. It arises due to the vascular transit time and mitochondrial oxidative limitations, leading to an initial reliance on anaerobic pathways [22]. Anaerobic pathways can facilitate energy much faster than aerobic pathways [25], however anaerobic resources are finite and the produced metabolites are associated with fatigue [26]. $\dot{V}O_2$ kinetics describe the speed at which the $\dot{V}O_2$ adjusts to a new stimulus and are thus inversely related to the O_2 deficit. It is not by chance that faster $\dot{V}O_2$ kinetics are associated with better exercise performance and disease is associated with slower $\dot{V}O_2$ kinetics [1].

The speed of the $\dot{V}O_2$ kinetics is generally defined in terms of the time constant τ [27]. τ denotes the time it takes for the $\dot{V}O_2$ to reach 63% of the new steady state and after 4 τ's the new steady state is reached in full [26, 1]. Theoretically, the τ is an underlying system parameter that defines the $\dot{V}O_2$ response. However, due to the non-linearities of the $\dot{V}O_2$ response, kinetics are slower at higher exercise intensities [28]. The τ is typically derived via non-linear least squares regression of a single exponential function on the $\dot{V}O_2$ of a step-increase in work rate. Often times multiple repetitions of such a protocol are performed to increase the signal-to-noise ratio, as $\dot{V}O_2$ data is noisy and the regression is sensitive to noise [29, 27]. Furthermore, the regression inherently assumes system parameters, which combined with the other limitations make τ unsuitable for kinetics analysis outside of specific lab protocols.

Frequency-domain analysis offers a promising alternative to conventional time-domain methods by transforming the $\dot{V}O_2$ signal into the frequency domain through Fourier transformation. This yields harmonics, each representing a specific frequency, the amplitude of which indicates the contribution of that frequency to the overall response [30, 31, 32]. The base frequency has a frequency of the inversion of the duration of the signal. The other harmonics are integer multiples of the base frequency and typically have a lower amplitude [14]. The harmonics can be used to reconstruct the original signal with arbitrary precision. However, frequency decomposition can only be used on signals with sufficient variation, making it suitable for daily activities. The most common lab protocol for this is the pseudorandom binary sequence (PRBS), where the work rate is pseudo-randomly switched between 2 work rates as it allows for the detection of many different frequencies from a single exercise protocol [32, 30, 31], though pseudorandom ternary sequencues (PRTS) have also been used [33, 34, 35].

One method using frequency-domain analysis has been proposed by Beltrame et al., named Mean Normalized Gain (MNG) [14]. MNG attempts to quantify the kinetics by comparing the system gain of frequencies up to a specified cut-off frequency with the base frequency. The cut-off frequency helps filter out noise in the $\dot{V}O_2$ signal, as the noise is almost invariably contained within the higher frequency components [32, 30, 31]. MNG has demonstrated strong correlations with τ measured during cycling, walking and activities of daily living, at least within the moderate domain [14, 36, 33, 35]. The inherent noise-filtering capabilities, the absence of system parameter assumptions and the applicability to a wider range of exercise protocols make MNG a promising alternative for τ. However, it should be noted that MNG has only been used and validated by studies within a single research group. Broader validation of MNG is essential to establish reliability as a valid analytical tool.

The estimation of non-maximal $\dot{V}O_2$ during activities is a remarkably novel field of research with less than a handful of research papers. The end goal of this field is to be able to accurately predict $\dot{V}O_2$ as well as its kinetics in real-time during activities of daily living (ADL). The first papers have been performed by Beltrame et al. aiming to estimate $\dot{V}O_2$ during PRTS walking sequences and controlled ADL [35, 34]. A Random Forest machine learning algorithm was trained on heart rate, breathing and walking derived features as measured by a smart shirt and a hip accelerometer. Promising results were achieved with Pearson correlation of 0.87 and 0.69 during ADL and PRTS, respectively.

To enable real-time predictions it is important that the chosen machine learning model architecture is both sufficiently powerful to achieve high accuracy, but also small enough to be used by embedded systems that are typically constrained in battery capacity, computational power and storage [37, 38]. Therefore, Amelard et al. compared the size and performance of different machine learning models, including the previously used Random Forest method, a stacked Long Short Term Memory (LSTM) network and different sets of parameters for a Temporal Convolutional Network (TCN) [13]. The TCN had the best performance, most likely attributable to the receptive field it boasts, which allows it to use data of previous time points for its next prediction, while also being the smallest model in terms of parameters.

The TCN architecture with the best performance compared to its size was used by Hedge et al. to estimate $\dot{V}O_2$ and its kinetics during medium and heavy intensity cycling PRBS sequences [12]. For this, heart rate and breathing derived features were used as measured by the same smart shirt as used by Beltrame et al. in addition to the resistance of the cycle ergometer. Repeated-measures correlation between predicted and measured $\dot{V}O_2$ was exceptionally high at r=0.974, and repeated-measures correlation between predicted and measured MNG was present with r=0.68 [12]. These results provide high confidence that accurate predictions of $\dot{V}O_2$ and its kinetics are indeed feasible, though more research is needed to validate whether such accuracy is consistently possible.

Therefore, the goals of this research are threefold. First, we aim to validate Hedge et al.'s $\dot{V}O_2$ prediction results by using a separately collected dataset, with the same TCN architecture and input features. Second, we aim to validate the use and estimation of the MNG method, proposed by Beltrame et al. to analyze the $\dot{V}O_2$ kinetics. Finally, up until now only central measurements of the cardiovascular and pulmonary systems have been used to estimate $\dot{V}O_2$. However, the muscular system is ultimately the largest consumer of O_2 within the body. Therefore, we hypothesize that adding NIRS and body composition data as model inputs can improve $\dot{V}O_2$ prediction results.

Chapter 3

Methods

3.1 Participants

Data were collected from 14 young healthy adults (8 men, 6 women) who were recruited through local sports associations or the personal network of the researchers. Participant characteristics are displayed in table 3.1. All participants performed multiple exercise tests including one ramp-incremental test and different pseudorandom binary sequence (PRBS) exercise tests. Collected data was used to train a TCN model to predict the oxygen uptake of the participants during cycling. All participants signed an informed consent form before participating in the study, and were made aware of their right to withdraw without prejudice. The research was approved by the Natural Sciences & Engineering Sciences (NES) ethics committee, request number 240302.

Table 3.1: Participant Anthropometry

(n=14)	mean ± sd	min − max
Age (years)	25.8 ± 3.2	22.3 − 31.6
Height (cm)	180 ± 9	161 − 194
Weight (kg)	74.2 ± 9.5	58.8 − 93.2
Peak $\dot{V}O_2$ ($mL\,min^{-1}\,kg^{-1}$)	46.9 ± 7.1	33.1 − 60.9

3.2 Data Collection

Each participant was expected to perform a ramp-incremental test, followed by 3 PRBS sequences of different intensities performed in random order, totaling 4 visits to the lab. All visits were performed on a cycle ergometer (Lode Excalibur Sport, Lode, Groningen, Netherlands).

Before the first visit, height was measured and before all visits, body composition data such as weight, muscle mass and bodyfat % were collected using a bioelectric impedance scale (Sacoma Original©).

$\dot{V}O_2$ data was collected through use of a metabolic cart (Quark CPET, Cosmed, Rome, Italy) as ground truth for the machine learning model.

Central measurements such as Heart Rate (HR), minute ventilation ($\dot{V}E$) and breathing frequency (BF) were collected using an unreleased smart strap from the company Tymewear. The smart strap is worn similarly to a typical heart rate strap and estimates $\dot{V}E$ and BF by measuring the chest expansion during breathing. Before collecting data, a calibration would be performed that included breathing in and out maximally 3 times.

Peripheral measurements were collected using a NIRS device called the Portamon (Artinis, Netherlands) which measures TSI and O_2Hb, HHb and tHb using 3 different frequencies to examine different depths of the muscle. Additionally it provides a measure called the TSI Fit Factor, which represents the signal quality. According to Artinis, if it is above 99, the signal should be representative.

The greater trochanter and the lateral condyle of the right femur were located using palpation. The halfway point between these landmarks was used to align the top left corner of the Portamon, such that the entire Portamon is anterior to the line connecting the two landmarks. The Portamon was then kept in place by an elastic strap closed by velcro and subsequently covered with a blackout cloth to ensure no light would pollute the Portamon signal.

The signal quality would be observed while the participant would stand and perform an isometric squat. It was expected that the TSI would be between 50 and 80 during standing. O_2Hb and TSI should drop, and HHb should increase during the isometric squat as oxygen is continuously being used and the isometric contraction prevents much blood from flowing into the quadriceps. If the TSI Fit Factor was consistently above 99.9 during these movements, the signal was deemed sufficient.

Subsequently, the blackout cloth was removed and a bandage was wrapped around the Portamon to keep it into place, followed by a blackout wrap. Finally, the same standing and isometric squat would be performed to verify the signal quality was still sufficient. If it was not, the Portamon would be moved slightly distally or anteriorly and the same steps would be performed until the signal was of sufficient quality. If it was, the exercise test could begin.

3.3 Exercise Tests

All exercise tests started with 2 minutes of rest whilst baseline measurements were recorded, meaning no cycling was done during this time. The ramp-incremental test subsequently consisted of 4 minutes of baseline pedaling at a low power output of 25 W, followed by a continuous increase in power output until participants could no longer continue (20-35 $\mathrm{W\,min^{-1}}$, depending upon participant's body composition and self-reported physical activity levels). Throughout the test, participants were required to keep a cadence of 70-90 revolutions per minute. When the cadence dropped below 60 revolutions per minute, this marked the end of the test.

The subsequent 3 exercise tests follow the 2 minute rest with a 3.5 minute warmup, followed by 2 repetitions of a 16 minute long PRBS sequence. Each PRBS sequence uses different work rates to elicit different physiological responses from the participant. The work rates depend on the participants' incremental ramp test results. The work rates that were used are 25 Watt, 90% GET, GET and Δ30%. Δ30% refers to the work rate equivalent of GET + 30% of the difference between GET and the $\dot{V}O_2$max. Hedge et al. used Δ50% instead [12], however, Δ30% was chosen because it is highly likely that Δ30% puts participants into heavy intensity [39], whereas Δ50% could also be of severe intensity for some participants.

The **Easy** protocol had participants pseudorandomly switch between 25 Watt and 90% GET. This will keep participants within moderate intensity for the duration of the protocol.
The **Medium** protocol had participants switch between 25 Watt and Δ30%. This ensures

participants switch between moderate and heavy intensity training.
The **Hard** protocol had participants switch between GET and Δ30%. This had participants exercising within heavy intensity for the duration of the protocol.
An example of the Easy protocol is shown in figure 1. The work rate profile is consistent with the Medium and Hard protocols, but actual work rates are different as described above.

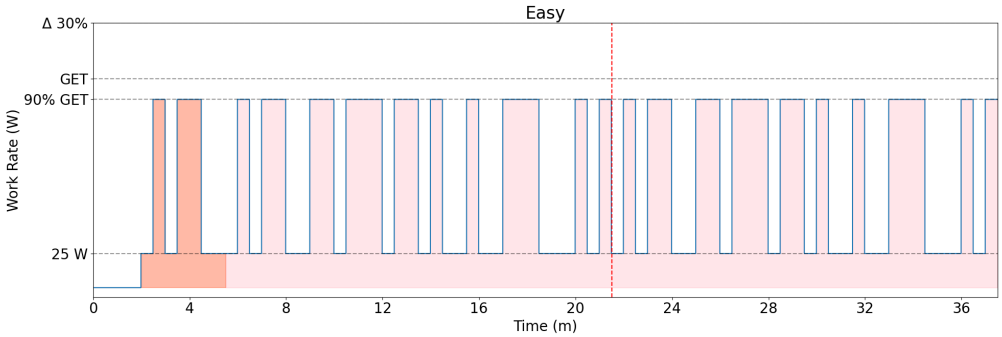


FIGURE 1: PRBS sequence showing the Easy protocol. The work rates used during the other protocols are also shown on the y-axis. The dark shaded area indicates the warm-up period that starts after the 2 minute rest and the light-shaded area indicates the exercise protocol. The vertical dotted line indicates the transition from the first to the second repetition.

3.4 Machine Learning

In this work, the same model architecture was used as by the most recent paper by Hedge et al. [12]. This model is based on the Temporal Convolutional Neural Network (TCN) architecture described by Amelard et al. [13] which was used to estimate the $\dot{V}O_2$ of the participants. This type of model architecture features causal convolutions and dilations to extract features from a preceding time window called the receptive field. This particular network uses a kernel size of 7 data points and the 4 layers of the network each dilate the kernel width by a factor of 2, meaning that the time between the examined data points is doubled each layer, ultimately yielding a receptive field of 187 seconds. Each layer contains 16 filters.

Due to a relatively small and non-homogeneous sample of 14 participants, k-fold cross validation was expected to introduce significant bias into the results as it is impossible to cut the sample into representative slices. Therefore, leave-one-out cross validation (LOOCV) was used. With LOOCV, a model will be trained on every participant except for one and subsequently tested on the one participant it was not trained on. This was repeated for every participant, after which results of all (14) models were combined.

Each model was trained for 30 epochs using mean squared error as loss function and the version with the lowest training loss was selected. Because neural networks are non-deterministic, at least 5 sets of models are trained for any set of input features to be able to show the consistency of the results.

3.4.1 Validation

With the goal of validating the results of Hedge et al. [12], the same input features were used to train the model as were used by them to be able to compare results. These parameters include heart rate (HR), heart rate reserve (HRR), minute ventilation ($\dot{V}E$), breathing frequency (BF) and work rate (WR). If any of these parameters were missing or invalid, the visit was excluded from training and testing data.

HRR is a variable derived using resting and maximal HR through the formula $(HR - HR_{rest})/(HR_{peak} - HR_{rest})$. HR_{rest} was taken as the minimum of the average HR of a participant during the 2 minutes rest before each visit. HR_{peak} was taken as the peak HR during the Ramp protocol. If HR data from the Ramp was missing, maximum HR from the Hard protocol was multiplied with the average ratio between maximum Ramp HR and maximum Hard HR of the participants where both were available. If HR data from the Hard protocol was also missing (n=1), HR_{peak} was selected to be 220 - age.

Hedge et al. applied a 5 breath median filter over the $\dot{V}O_2$ data in an effort to provide a less noisy ground truth for the model, which was also done for this work [12]. In addition to this, outlier rejection was applied to all the breath by breath parameters, which include $\dot{V}O_2$, $\dot{V}E$ and BF. A 5 breath window was rolled through the data and every data point subsequent to the window that was more than 3 standard deviations from the mean was excluded from the final result. Outliers were still included within the rolling window. This method of rejecting outliers is not perfect, but it does not interfere with the ultimate ideal of real time predictions as it can be applied to a stream of data in real time.

3.4.2 Extension

The final goal of this research was to examine whether body composition and NIRS data can improve the model performance compared to only using the previously described input data. For this, first data from these 2 were separately be added as model inputs to examine the difference with the original input features. Finally, both were added simultaneously, to determine whether there was an additive effect.

Body Composition

For the extension of body composition features, previous work has shown that gender, skeletal muscle mass (SMM) and body fat (BF) have among the strongest predictive values of body composition features for predicting peak power output and $\dot{V}O_2$max [40]. As such, it can be assumed that these features also influence submaximal $\dot{V}O_2$ substantially. Given that the gender differences of the oxygen transport systems and fat (free) weight have the strongest effect on the $\dot{V}O_2$max differences between men and women [41], and representative features are already included for both ($\dot{V}E$, SMM), including gender seemed redundant. Therefore only muscle mass and body fat were included as features.

NIRS

For the extension of NIRS features, typically O_2Hb and HHb are used from the channel that goes deepest into the muscle tissue. However, blood volume changes often occur in the collected data and the effect on O_2Hb is largest, followed by HHb. Additionally, not every channel always had representative data, meaning sometimes a certain channel was not usable. While relatively infrequent, this would have required more visits to be excluded from the training/testing data due to quality reasons.

Therefore, it was chosen that only TSI was to be used as a feature, as it was most consistent across participants, leading to the least amount of excluded visits while still representing the muscle $\dot{V}O_2$. If TSI Fit Factor was too low or substantial blood volume changes occurred in TSI, a visit or repetition was still excluded from the training/testing data. If only 1 repetition was of insufficient quality, the other repetition was still used to train and test the model.

3.4.3 Data preprocessing

Before being used as model inputs, data from different sensors were time-aligned. Breath by breath data were interpolated to 1 Hz. NIRS data was captured at 10 Hz and downsampled to 1 Hz. Each feature, except WR, was standardized to zero mean and unit variance across all data. WR was normalized to [0, 1] across all data due to its non-normal distribution.

3.5 Data Analysis

3.5.1 Quality

Some visits were excluded from analysis due to missing features, insufficient data quality or were not recorded at all. The only exception being P4's Hard visit where the second repetition of $\dot{V}O_2$ data was not recorded due to technical issues. Wearable sensor data was intact and under the (slightly incorrect) assumption that $\dot{V}O_2$ data should be the same for both repetitions, $\dot{V}O_2$ data from the first repetition was used for the second repetition as well.

Considering 14 participants were recruited, a total of 56 visits should have taken place. Instead, in total 58 visits were recorded, of which 50 have sufficient data quality to remain included for the validation step. A further 9 visits are completely removed due to NIRS data quality issues, and 2 visits have their second repetition excluded for the same reason. In chapter 6, figure 6.1 displays a concrete view of which visits and repetitions were in- and excluded.

3.5.2 Kinetics

In order to evaluate the $\dot{V}O_2$ kinetics, MNG was calculated from the measured and predicted $\dot{V}O_2$ data. First the $\dot{V}O_2$ data of both repetitions of a PRBS sequence were ensemble averaged to yield a single 16 minute response for each visit. Then, the WR and $\dot{V}O_2$ were both converted to the frequency domain and system gain was calculated by dividing $\dot{V}O_2$ harmonics by WR harmonics. Finally, MNG was calculated by taking the mean of the harmonics up to a cut-off frequency of 0.01Hz with regards to the base harmonic [14]. In cases where 1 of the 2 repetitions were excluded due to signal quality, the ensemble averaging step was skipped and MNG was calculated from the single repetition.

3.5.3 Statistical

To evaluate both the $\dot{V}O_2$ and MNG prediction results of the model, repeated-measures correlation and repeated-measures Bland-Altman analysis were used. Both account for the within-participant variance of the observations over time. Repeated-measures correlation uses analysis of covariance (ANCOVA) to adjust for differences between the measurements and predictions within each person [42]. By fitting parallel regression lines with a common slope but varying intercepts for each individual, repeated-measures correlation evaluates

the overall within-person association. It yields an r-value that describes the strength of the relationship between the predictions and measurements that is bounded by -1 to 1, where 1 means the predictions are perfectly correlated.

Unlike standard Bland-Altman analysis, which assumes independent observations, repeated-measures Bland-Altman adjusts for the clustering of repeated measurements within participants. It separates variance into between-individual differences and within-individual variability using analysis of variance (ANOVA). This yields more accurate bias and limits of agreement (LoA), reflecting the variability at both individual and group levels [43].

All data analysis and processing has been performed in python. For repeated-measures correlation, package pingouin was used. For repeated-measures Bland-Altman analysis, code was used from Wade et al. [44]. Statistical significance is set at p=0.05.

Chapter 4

Results

4.1 Ramp-incremental exercise responses

In table 4.1, the results of the ramp-incremental test are shown. Included are peak WR and $\dot{V}O_2$, GET $\dot{V}O_2$ and all work rates used during the different PRBS sequences.

TABLE 4.1: Mean, standard deviation and range of peak and GET $\dot{V}O_2$ and corresponding work rates based on the ramp-incremental test results for all 14 participants.

(n=14)	mean ± sd	min − max
$\dot{V}O_2$ Peak ($L\,min^{-1}$)	3.47 ± 0.62	2.44 − 4.28
$\dot{V}O_2$ GET ($L\,min^{-1}$)	2.22 ± 0.37	1.63 − 2.74
WR Peak (W)	348 ± 62	250 − 471
WR 90% GET (W)	138 ± 27	107 − 184
WR GET (W)	153 ± 30	119 − 204
WR Δ30% (W)	194 ± 37	147 − 256

4.2 Oxygen uptake

Tables 4.2 and 4.3 present the mean, standard deviation and range of the $\dot{V}O_2$ prediction and model training test losses for the different input feature sets, respectively. During the rest of this work, a **set of models** will refer to a set of 14 models, one for each participant, that together represent one data point for the results. Such that the described analysis methods are to be performed on a single set of models and thus the results of this single set will have a single repeated-measures correlation, Bland-Altman bias and limits of agreement, etc... In tables 4.2 and 4.3 the $n = x$ denotes that x sets of models were included in the results for that input feature set. Furthermore, in table 4.2, a *best* column is included, which denotes the results for the set of models that includes only the best performing model for each participant.

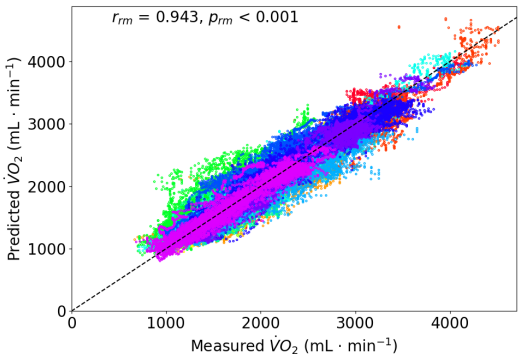
Validation

Mean repeated-measures correlation is high with an r-value of 0.928. Systemic bias is low with a mean of -14 $mL\,min^{-1}$, but the mean LoA are wide with a rage of -512 to 485 $mL\,min^{-1}$. Results for the best set of models are better with an r-value of 0.943 and LoA

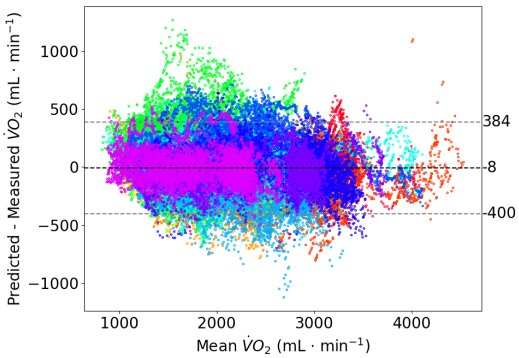
of -401 to 385 mL min^{-1}.

Participant test losses are relatively consistent, though P8 is a large outlier with a mean test loss of 0.4. P1, P3, P11, P12 and P15 show relatively high mean test losses around 0.2 - 0.3.

FIGURE 2: $\dot{V}O_2$ prediction results for best performing set of models using Validation input features. Each participant is represented by a different color.



A Repeated-measures correlation between measured and predicted $\dot{V}O_2$. Black dashed line represents the line of identity (i.e., y = x).



B Repeated-measures Bland–Altman plot. The black dashed line represents the bias, and the grey dashed lines represent the 95% limits of agreement.

Extension

NIRS input features achieve 0.932 and 0.943, Body Composition input features achieve 0.930 and 0.944 and Both input features achieve 0.931 and 0.947 mean and best repeated-measures correlation, respectively. All different extension input features show a slight increase in both mean and best repeated-measures correlation compared to the validation input features, except for NIRS, which shares the same best correlation as the Validation input features. Additionally, highest correlation is attained using Both input features.

Mean LoA are similar between all input feature sets, though best LoA are widest for the Body Composition input features and smallest for Both input features. Both input features attained the lowest bias for the best set of models, but the standard deviation of the mean bias is largest of any set of input features, though differences are small.

Mean participant test losses seem lower for the NIRS input feature set compared to the Validation input features, with exception of P15, which seems to be an outlier for the NIRS data. Mean test losses are less consistent for the Body Composition input features, as their standard deviation are generally higher than for the Validation and NIRS input feature sets. Additionally, P15 is an even bigger outlier for the Body Composition features. Both input feature set seems to have slightly more consistent test losses than the Body Composition results, though relative outliers seem to be present more often.

TABLE 4.2: $\dot{V}O_2$ results across different input features. r_{rm} is the repeated-measures correlation. The bias, LLoA, and ULoA values are the mean bias, lower limit of 95% agreement, and upper limit of 95% agreement in mL min^{-1}, respectively, for the repeated-measures Bland-Altman analysis. The 'best' columns refer to the results of the set of models that includes only the best performing model for each participant.

	Validation (n=8)			**NIRS (n=5)**			**Body Comp (n=5)**			**Both (n=7)**		
	mean± sd	min − max	best	mean± sd	min − max	best	mean± sd	min − max	best	mean± sd	min − max	best
r_{rm}	0.928±0.007	0.918−0.935	0.943	0.932±0.006	0.923−0.940	0.943	0.930±0.007	0.919−0.940	0.944	0.931±0.006	0.920−0.940	0.947
bias	-14 ± 14	-35 − 2	-8	-18 ± 15	-38 − 3	-28	12 ± 22	-25 − 41	-11	4 ± 26	-30 − 37	4
LLoA	-512 ± 24	-550 − -480	-401	-495 ± 35	-545 − -446	-417	-539 ± 36	-582 − -489	-451	-510 ± 41	-602 − -472	-377
ULoA	485 ± 40	426 − 554	385	458 ± 15	437 − 482	360	563 ± 39	507 − 612	429	519 ± 52	435 − 575	384

TABLE 4.3: Model test losses when trained on a participant according to LOOCV across different input features.

	Validation (n=8)		**NIRS (n=5)**		**Body Comp (n=5)**		**Both (n=7)**	
	mean ± sd	min − max	mean ± sd	min − max	mean ± sd	min − max	mean ± sd	min − max
P1	0.21 ± 0.11	0.07 − 0.42	0.09 ± 0.02	0.07 − 0.11	0.15 ± 0.09	0.05 − 0.29	0.12 ± 0.07	0.06 − 0.29
P3	0.20 ± 0.07	0.11 − 0.35	0.13 ± 0.04	0.08 − 0.20	0.14 ± 0.04	0.10 − 0.20	0.13 ± 0.05	0.09 − 0.24
P4	0.13 ± 0.01	0.11 − 0.15	0.13 ± 0.01	0.12 − 0.14	0.26 ± 0.03	0.20 − 0.30	0.22 ± 0.04	0.16 − 0.30
P5	0.06 ± 0.01	0.05 − 0.08	0.09 ± 0.01	0.07 − 0.11	0.10 ± 0.02	0.08 − 0.13	0.10 ± 0.03	0.07 − 0.15
P6	0.08 ± 0.01	0.06 − 0.11	0.08 ± 0.02	0.07 − 0.13	0.14 ± 0.15	0.06 − 0.45	0.14 ± 0.06	0.07 − 0.28
P7	0.12 ± 0.04	0.06 − 0.19	0.06 ± 0.02	0.03 − 0.09	0.11 ± 0.06	0.05 − 0.22	0.09 ± 0.05	0.03 − 0.16
P8	0.40 ± 0.06	0.26 − 0.46	0.31 ± 0.10	0.17 − 0.48	0.27 ± 0.04	0.21 − 0.31	0.28 ± 0.06	0.20 − 0.42
P9	0.12 ± 0.04	0.08 − 0.21	0.13 ± 0.01	0.11 − 0.15	0.17 ± 0.08	0.09 − 0.30	0.15 ± 0.07	0.06 − 0.25
P10	0.09 ± 0.01	0.08 − 0.11	0.10 ± 0.02	0.08 − 0.13	0.13 ± 0.02	0.11 − 0.16	0.16 ± 0.04	0.09 − 0.20
P11	0.24 ± 0.02	0.20 − 0.26	0.17 ± 0.03	0.13 − 0.21	0.20 ± 0.07	0.12 − 0.30	0.38 ± 0.27	0.11 − 0.88
P12	0.28 ± 0.10	0.14 − 0.43	0.17 ± 0.07	0.10 − 0.31	0.11 ± 0.07	0.05 − 0.23	0.10 ± 0.02	0.07 − 0.13
P13	0.11 ± 0.01	0.09 − 0.13	0.15 ± 0.07	0.10 − 0.29	0.13 ± 0.01	0.12 − 0.15	0.23 ± 0.02	0.20 − 0.27
P14	0.06 ± 0.01	0.05 − 0.07	0.09 ± 0.02	0.07 − 0.11	0.14 ± 0.05	0.09 − 0.22	0.12 ± 0.05	0.05 − 0.20
P15	0.19 ± 0.11	0.04 − 0.38	0.47 ± 0.21	0.20 − 0.80	0.89 ± 0.24	0.50 − 1.24	0.47 ± 0.35	0.09 − 1.02

4.3 Oxygen Uptake Kinetics

In table 4.4 the results are shown for the comparison between measured and predicted kinetics. Repeated-measures correlation is on average 0.07, with an average p-value of 0.59, showing that there is no meaningful correlation at all. Maximum correlation is 0.449, with a statistically significant p-value of 0.017. Bias is relatively low at 5%, but 95% limits of agreement are substantial with mean values of -59 to 69%.

TABLE 4.4: MNG results for the Validation input feature set. r_{rm} and p_{rm} are repeated-measures correlation and p-value, respectively. $r_{rm_{Hard}}$ and $p_{rm_{Hard}}$ are these same measurements but with the Hard condition excluded. The bias, LLoA, and ULoA values are the mean bias, lower limit of 95% agreement, and upper limit of 95% agreement, respectively, for the repeated-measures Bland-Altman analysis. The 'best' columns refer to the results of the set of models that includes only the best performing model in terms of training loss for each participant.

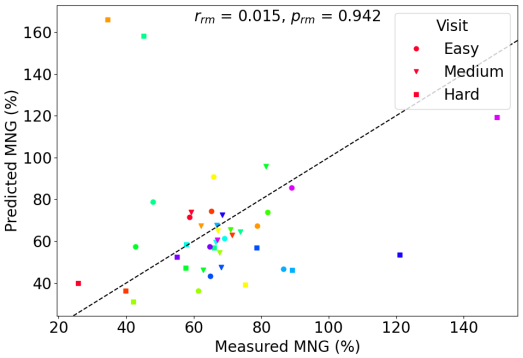
(n=8)	mean ± sd	min – max	best
r_{rm}	0.070 ± 0.170	-0.161 – 0.449	0.015
p_{rm}	0.586 ± 0.269	0.017 – 0.913	0.942
$r_{rm_{Hard}}$	0.292 ± 0.096	0.146 – 0.386	0.311
$p_{rm_{Hard}}$	0.34 ± 0.18	0.17 – 0.62	0.28
bias	5 ± 3	-1 – 9	1
LLoA	-59 ± 16	-91 – -33	-65
ULoA	69 ± 11	47 – 88	68

Figure 3 displays the repeated-measures correlation and boxplot of the measured and predicted MNG of the best set of models using the Validation input feature set. It can be seen that there is indeed no correlation. However, all of the largest outliers are of the Hard protocol. When excluding the Hard protocol from the correlation, mean correlation is indeed higher, though not enough to make it statistically significant.

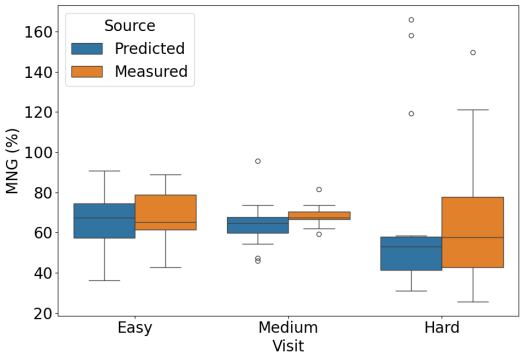
The boxplot shows that the ranges do not match up well at all, although a case could be made that the Easy and Medium visits' results are not poor. The range for the Hard visit is particularly wide. It is expected to see kinetics slow as exercise intensity increases [1]. However, at most there exists some semblance towards lower MNG as exercise intensity increases, though the spread of the data is too large to state conclusively.

In chapter 6, figure 6.3 shows the results for the extension input features, though there is no significant difference between any of the input feature sets. Very few sets of models have statistically significant repeated-measures correlation and the correlations do not go beyond 0.45. When the Hard protocol is excluded, some sets of models achieved correlation between 0.6 and 0.7. However, the sets of models with the highest MNG prediction accuracy do not correspond necessarily with the sets of models with the highest $\dot{V}O_2$ prediction accuracy.

FIGURE 3: MNG prediction results for lowest test loss set of models using Validation input features.



A Repeated-measures correlation between measured and predicted MNG. Black dashed line represents the line of identity (i.e., y = x). Each participant is represented by a different color, and each type of visit is represented by a different shape.



B Boxplot of the measured and predicted MNG results across the different PRBS protocols. Orange and blue filled boxes represent the measured and predicted kinetics, respectively. Solid lines within the boxes represent medians and open circles represent outliers.

Chapter 5

Discussion

This study aimed to validate and extend the predictive capabilities of a TCN model for estimating $\dot{V}O_2$ and its kinetics using wearable sensor data. It was found that high accuracy can consistently be achieved, with a mean repeated-measures correlation of 0.928 and a best repeated-measures correlation of 0.943 using Validation input features. However, prediction of $\dot{V}O_2$ kinetics using the MNG method was less successful, showing no meaningful correlation. Adding muscle oxygenation (NIRS) and body composition data was input features improved the $\dot{V}O_2$ prediction results marginally, but did not change the kinetics estimation.

5.1 Oxygen uptake

5.1.1 Validation

Hedge et al. achieved an overall repeated-measures correlation of 0.974, a bias of -17 $\mathrm{mL\,min^{-1}}$and 95% LoA of -289 to 254 $\mathrm{mL\,min^{-1}}$ [12]. Given that Hedge et al. only report single results, without mention of how many times models have been trained, it is likely the results are simply the congregated best results for all participants. Under those circumstances, the correlation of 0.943, bias of -8 and 95% LoA of -401 to 385 $\mathrm{mL\,min^{-1}}$ achieved in this study fell short of the results reported by Hedge et al.

The Tymewear strap that was used to measure $\dot{V}E$ only requires a simple calibration of 3 maximal in- and exhales. This allows it to be used as a stand-alone measuring tool, but comes at the cost of accuracy. When comparing the strap's $\dot{V}E$ measurements compare to the CPET's $\dot{V}E$ measurements, the signals are temporally very similar. However, the ratio between them can vary between 1 and $\tilde{2}.3$ in the collected sample, making it an inconsistent measuring tool. Hedge et al. calibrated their Hexoskin vest with $\dot{V}E$ measured by the CPET to improve their $\dot{V}E$ measurements making it likely that their $\dot{V}E$ values, and consequently their prediction results, are more accurate [12].

A major difference in methodology was the choice to use $\Delta 30\%$ instead of $\Delta 50\%$ for the high work rate of the Hard protocol in this work. It remains true that it is substantially more likely to equalize the responses to exercise between different people, though it significantly reduces the absolute work rate difference with the GET work rates. Given that the $\dot{V}O_2$ signal is relatively noisy at higher work rates, which in combination with the small absolute difference in $\dot{V}O_2$ makes the ultimate signal not always clear. This could have had an effect on the ground truth $\dot{V}O_2$ values, even with the applied outlier rejection.

Furthermore, the participant sample is substantially different. In addition to the mean person being taller and fitter compared to theirs, the standard deviation of work rate or anthropometric measurement is larger by a factor ranging from 1.4 to 2.5, indicating that the sample is significantly less homogeneous. Homogeneity of the characteristics can have a significant effect on the prediction results as LOOCV is quite prone to outliers.

Finally, it is typical to control participants' exercise and consumption of food, alcohol and caffeine prior to each visit, as these can influence the test results and physiological measurements. However, due to strong time constraints on the data collection period, these things were not controlled for as they typically would be. This can result in more inconsistent measurements between and within participants. An example of this is caffeine, as it will elevate heart rate and can improve exercise performance [45]. Consequently, this may have added further heterogeneity to the sample.

5.1.2 Extension

A marginal improvement in $\dot{V}O_2$ prediction accuracy was observed when adding either or both NIRS and body composition features compared to the Validation input features. Despite the included data being physiologically relevant, the improvement is very small and likely statistically insignificant. It could be that the variance within the participant sample is limiting the improvement in prediction accuracy.

Looking at test loss, results from NIRS input features are consistently slightly better than those from the Validation input features, with an exception of P15, indicating that there seems to be some improvement. However, for both Body Composition and Both input features, there seems to be an increased variance in the observed test losses. Mean minimum test loss is only different for the body composition input feature set, where it is higher.

The body composition features are different from the NIRS features in that they add 2 features, rather than 1, as well as being the only non-temporal features. Perhaps the model can not consistently learn the most suitable relationships for a larger amount of input features. Additionally, the TCN architecture is designed to use temporal features, and so using non-temporal features may yield inconsistent or undesirable results.

5.2 Oxygen Uptake Kinetics

Hedge et al. reported a repeated-measures correlation between predicted and measured MNG to be 0.68, with a bias of 1.12% and 95% LoA of -16.56 to 18.86%. These results are substantially better than what is achieved in this work, given that generally no correlation was observed and the highest found correlation was 0.449 or 0.694 when excluding the Hard visits. The progressive slowing of kinetics as exercise intensity increases observed in their results, was not conclusively found in this work. The slight reduction in $\dot{V}O_2$ prediction accuracy compared to Hedge et al.'s work is unlikely to force such a strong reduction in MNG prediction accuracy.

Furthermore, the *measured* MNG values range from about 20 to 150%, whereas all previous papers using the MNG method have not reported MNG values above 100% [14, 33, 34, 36, 35, 12]. Higher MNG indicates faster kinetics and the average participant in this sample is quite fit. However, the MNG outliers are primarily from data collected during the Hard protocol, which should display slower kinetics.

Previous work found significant correlation between MNG and τ within the moderate intensity domain, but not within the heavy intensity domain [36]. It was argued that this was due to increased breath-by-breath noise during the heavy intensity cycling. The noise would impact the calculations of τ, but not MNG, as the inherent noise-filtering capabilities make it more robust to noise [36, 14]. However, the large range and outliers of measured MNG during the Hard visits contradict their findings.

The largest methodological differences between Hedge et al.'s work and this work with regard to MNG are the PRBS sequence that is twice as long in this work, as well as the usage of $\Delta 30\%$ instead of $\Delta 50\%$. However, if MNG were an ideal method of analyzing $\dot{V}O_2$ kinetics, these methodological differences should not yield such different results.

Figures 4 and 5 display on the left the averaged $\dot{V}O_2$ and the reconstructed signal using the cut-off frequency, and on the right the $\dot{V}O_2$ and WR harmonics from which MNG is calculated for P13 and P15's Hard protocols, respectively. These are the largest outliers with regard to MNG. It can be seen that the $\dot{V}O_2$ signal is still very noisy, despite being averaged to increase the signal to noise ratio. The reconstructed $\dot{V}O_2$ signal is substantially less noisy, but also filters out significant parts of the actual signal, smoothing out some of the valleys and peaks.

In the right figures, a higher amplitude of the $\dot{V}O_2$ and WR harmonics indicate that there is a larger proportion of the original signal present within that frequency. Generally, it is both expected and observed that the higher frequency $\dot{V}O_2$ harmonics drop off in amplitude and are considered to be just noise. For most participant visits this happens soon past the used cut-off frequency of 0.01 Hz, indicating that it is a relatively accurate cut-off, though definitely not perfect.

More importantly, and the reason why calculated MNG is so high for the visits displayed in the figures, is that the harmonics before the cut-off frequency do not display a clear trend downwards. To calculate MNG, first the gain of each harmonic is calculated by dividing the $\dot{V}O_2$ by the WR amplitude. Then, the first gain harmonic is divided by the mean of the 2nd harmonic until the last harmonic before the cut-off frequency. Given that no MNG values below 100% have been reported before and that MNG's unit is a percentage, it makes intuitive sense that the first harmonic should be one of the highest in amplitude. Evidently, this is not reliably the case.

Furthermore, dividing the $\dot{V}O_2$ by the WR harmonics makes sense from a physiological perspective, as the relationship between $\dot{V}O_2$ and WR is a well known characteristic known as the gain of the $\dot{V}O_2$ response. Clearly, by dividing the two, they try to capture whether the $\dot{V}O_2$ increases relatively faster or slower than the WR does. However, the different frequency harmonics of the $\dot{V}O_2$ signal already correspond to slower and faster oscillations within the $\dot{V}O_2$ signal, making the division by the WR harmonics a questionable choice.

5.3 Limitations

Due to data quality and data collection time restrictions, there are some visits with unrepresentative data that had to be excluded. Due to this, there exists some inherent bias within the sample as some visits are included twice and some are not included at all. Prediction results are generally worse for the Hard visit compared to the Easy and Medium visits. Consequently, for participants that had their Hard visit excluded, test loss is generally lower and participants who had more than 1 of a particular visit included may have different test loss because of it. Furthermore, the model is trained on an inconsistent amount of data per participant, which could also introduce bias.

Looking at the PRBS sequence in figure 1 it can be seen that both repetitions have the exact same work rate profile. However, the warmup ends at the lower work rate, whereas the first repetition ends at the higher work rate. Therefore, from a kinetics point of view, the start of the first and second repetition are not the same. It only impacts roughly the first 30 seconds of each 16 minute long sequence. Additionally, for the kinetics analysis, the 2 repetitions are averaged, so the visits should have a similar profile regardless. However, because some of the NIRS repetitions have been excluded, it is not possible to average the predicted repetitions in those cases. Thus, there is a small discrepancy between the $\dot{V}O_2$ profile of some specific visits when using NIRS input features.

It is important to note that despite using outlier rejection and the 5 breath median filter, not all visits' $\dot{V}O_2$ ground truth is perfectly representative of actual $\dot{V}O_2$. Large outliers exist and are still reflected after the median filter. This inherently decreases model accuracy by tainting the ground truth. While this is expected to also have been the case for Hedge et al., there is a chance that the longer and different PRBS sequence, the choice to use $\Delta 30\%$ or a different participant sample may have exacerbated the limitation.

5.4 Conclusion

This work demonstrated the potential of wearable sensors combined with machine learning to predict $\dot{V}O_2$, paving the way for more accessible fitness and health monitoring tools. High prediction accuracy was achieved with repeated-measures correlation reaching up to 0.947 by utilizing the TCN model architecture.

However, predictions for $\dot{V}O_2$ kinetics, evaluated through the MNG method, showed no meaningful correlation, challenging previous statements about the robustness of the method [36, 12]. Furthermore, inclusion of additional peripheral measurements, such as muscle oxygenation from NIRS and body composition data from BIA, marginally improved $\dot{V}O_2$ prediction results, but did not change kinetics estimation.

Several factors have potentially played a role in the slightly poorer results compared to previous work [12]. Particularly sample heterogeneity and data quality issues have likely contributed significantly, though differences in study design may have exacerbated the discrepancy in prediction results. Future research should focus on gathering larger and more diverse datasets, improving data (pre)processing techniques and exploring more robust methods for analysing $\dot{V}O_2$ kinetics.

Bibliography

[1] David C. Poole and Andrew M. Jones. Oxygen uptake kinetics. *Comprehensive Physiology*, 2(2):933–996, 4 2012.

[2] T. Beltrame and R. L. Hughson. Linear and non-linear contributions to oxygen transport and utilization during moderate random exercise in humans. *Experimental Physiology*, 102(5):563–577, 5 2017.

[3] S M Phillips, H J Green, M J Macdonald, R L Hughson, and S Ri. Progressive effect of endurance training on a vo l 2 kl netics at the onset of submaximal exercise. Technical report, 1995.

[4] Stephen R. Norris and Stewart R. Petersen. Effects of endurance training on transient oxygen uptake responses in cyclists. *Journal of Sports Sciences*, 16(8):733–738, 11 1998.

[5] Scott K Powers, Stephen Dodd, and Ralph E Beadle. Oxygen uptake kinetics in trained athletes differing in VO2ma x. Technical report, 1985.

[6] Audrey Borghi-Silva, Thomas Beltrame, Michel Silva Reis, Luciana Maria Malosá Sampaio, Aparecida Maria Catai, Ross Arena, and Dirceu Costa. Relationship between oxygen consumption kinetics and BODE index in COPD patients. *International Journal of COPD*, 7:711–718, 10 2012.

[7] Judith G Regensteiner, Timothy A Bauer, Jane E B Reusch, Suzanne L Brandenburg, Jeffrey M Sippel, Andria M Vogelsong, Susan Smith, Eugene E Wolfel, Robert H Eckel, and William R Hiatt. Abnormal oxygen uptake kinetic responses in women with type II diabetes mellitus. Technical report, 1994.

[8] Bruna Varanda Pessoa, Thomas Beltrame, Valéria A. Pires Di Lorenzo, Aparecida M. Catai, Audrey Borghi-Silva, and Mauricio Jamami. COPD patients' oxygen uptake and heart rate on-kinetics at cycle-ergometer: Correlation with their predictors of severity. *Brazilian Journal of Physical Therapy*, 17(2):152–162, 4 2013.

[9] Marco Guazzi, Ross Arena, Martin Halle, Massimo F. Piepoli, Jonathan Myers, and Carl J. Lavie. 2016 focused update: Clinical recommendations for cardiopulmonary exercise testing data assessment in specific patient populations. *European Heart Journal*, 39(14):1144–1161, 4 2018.

[10] Joshua Rudner, Carol McDougall, Vivek Sailam, Monika Smith, and Alfred Sacchetti. Interrogation of Patient Smartphone Activity Tracker to Assist Arrhythmia Management. *Annals of Emergency Medicine*, 68(3):292–294, 9 2016.

[11] Toru Nakamura, Ken Kiyono, Herwig Wendt, Patrice Abry, and Yoshiharu Yamamoto. Multiscale Analysis of Intensive Longitudinal Biomedical Signals and Its Clinical Applications, 2 2016.

[12] Eric T. Hedge, Robert Amelard, and Richard L. Hughson. Prediction of oxygen uptake kinetics during heavy-intensity cycling exercise by machine learning analysis. *Journal of applied physiology (Bethesda, Md. : 1985)*, 134(6):1530–1536, 6 2023.

[13] Robert Amelard, Eric T. Hedge, and Richard L. Hughson. Temporal convolutional networks predict dynamic oxygen uptake response from wearable sensors across exercise intensities. *npj Digital Medicine*, 4(1), 12 2021.

[14] Thomas Beltrame and Richard L. Hughson. Mean normalized gain: A new method for the assessment of the aerobic system temporal dynamics during randomly varying exercise in humans. *Frontiers in Physiology*, 8(JUL), 7 2017.

[15] Andrew M. Jones and David C. Poole. Oxygen uptake dynamics: From muscle to mouth - An introduction to the symposium. In *Medicine and Science in Sports and Exercise*, volume 37, pages 1542–1550, 9 2005.

[16] Stephane Perrey, Valentina Quaresima, and Marco Ferrari. Muscle Oximetry in Sports Science: An Updated Systematic Review, 4 2024.

[17] R. F. Kushner. Bioelectrical impedance analysis: A review of principles and applications, 1992.

[18] Harry B. Rossiter. Exercise: Kinetic considerations for gas exchange. *Comprehensive Physiology*, 1(1):203–244, 1 2011.

[19] Thomas J Barstow and Paul A Mole. Linear and nonlinear characteristics of oxygen uptake kinetics during heavy exercise. Technical report, 1991.

[20] David W Hill. The Critical Power Concept A Review. Technical Report 4, 1993.

[21] N. C. Spurway, B. Ekblom, T. D. Noakes, and P. D. Wagner. What limits VO 2max?: A symposium held at the BASES Conference, 6 September 2010. *Journal of Sports Sciences*, 30(6):517–531, 3 2012.

[22] Brian J. Whipp, Susan A. Ward, and Harry B. Rossiter. Pulmonary O2 uptake during exercise: Conflating muscular and cardiovascular responses. In *Medicine and Science in Sports and Exercise*, volume 37, pages 1574–1585, 9 2005.

[23] Jan Boone, Thomas J Barstow, Bert Celie, Fabrice Prieur, and Jan Bourgois. The interrelationship between muscle oxygenation, muscle activation and pulmonary VO2 to incremental ramp exercise: Influence of aerobic fitness. Technical report, 2016.

[24] Darren S Delorey, John M Kowalchuk, Donald H Paterson, and D H Paterson. Relationship between pulmonary O 2 uptake kinetics and muscle deoxygenation during moderate-intensity exercise. *J Appl Physiol*, 95:113–120, 2003.

[25] Nicola Lai, Marco Camesasca, Gerald M. Saidel, Ranjan K. Dash, and Marco E. Cabrera. Linking pulmonary oxygen uptake, muscle oxygen utilization and cellular metabolism during exercise. In *Annals of Biomedical Engineering*, volume 35, pages 956–969, 6 2007.

[26] Mark Burnley and Andrew M. Jones. Oxygen uptake kinetics as a determinant of sports performance. *European Journal of Sport Science*, 7(2):63–79, 6 2007.

[27] Norman Lamarra, Brian J Whipp, Susan A Ward, and Karlman Wasserman. Effect of interbreath fluctuations on characterizing exercise gas exchange kinetics. Technical report, 1987.

[28] C. J. Brittain, H. B. Rossiter, J. M. Kowalchuk, and B. J. Whipp. Effect of prior metabolic rate on the kinetics of oxygen uptake during moderate-intensity exercise. *European Journal of Applied Physiology*, 86(2):125–134, 2001.

[29] T Beltrame, M Gois, U Hoffmann, J Koschate, and R Hughson. Relationship between maximal aerobic power with aerobic fitness. 2020.

[30] Uwe Hoffmann, Dieter Efflfeld, Hans-Georg Wunderlich, and Jiirgen Stegemann. Dynamic linearity of I O2 responses during aerobic exercise. Technical report, 1992.

[31] Uwe Hoffmann, Dieter Egfeld, Dieter Leyk, Hans-Georg Wunderlich, and Jiirgen Stegemann. Prediction of individual oxygen uptake on-step transients from frequency responses. Technical report, 1994.

[32] R L Hughson, D A Winter, A E Patla, G D Swanson, and L A Cuervo. Investigation of VO2 kinetics in humans with pseudorandom binary sequence work rate changes. Technical report, 1990.

[33] X Thomas Beltrame and Richard L Hughson. Aerobic system analysis based on oxygen uptake and hip acceleration during random over-ground walking activities. *Am J Physiol Regul Integr Comp Physiol*, 312:93–100, 2017.

[34] Thomas Beltrame, Robert Amelard, Alexander Wong, and Richard L Hughson. Extracting aerobic system dynamics during unsupervised activities of daily living using wearable sensor machine learning models. *J Appl Physiol*, 124:473–481, 2018.

[35] T. Beltrame, R. Amelard, A. Wong, and R. L. Hughson. Prediction of oxygen uptake dynamics by machine learning analysis of wearable sensors during activities of daily living. *Scientific Reports*, 7, 4 2017.

[36] Eric T Hedge and Richard L Hughson. Frequency domain analysis to extract dynamic response characteristics for oxygen uptake during transitions to moderate-and heavy-intensity exercises. *J Appl Physiol*, 129:1422–1430, 2020.

[37] Eleftherios Batzolis, Eleni Vrochidou, and George A. Papakostas. Machine Learning in Embedded Systems: Limitations, Solutions and Future Challenges. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference, CCWC 2023*, pages 345–350. Institute of Electrical and Electronics Engineers Inc., 2023.

[38] Sérgio Branco, André G. Ferreira, and Jorge Cabral. Machine learning in resource-scarce embedded systems, FPGAs, and end-devices: A survey, 11 2019.

[39] Hester Van Der Vaart, Scott R. Murgatroyd, Harry B. Rossiter, Carey Chen, Richard Casaburi, and Janos Porszasz. Selecting constant work rates for endurance testing in COPD: The role of the power-duration relationship. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 11(3):267–276, 2014.

[40] Charlotte Wenzel, Thomas Liebig, Adrian Swoboda, Rika Smolareck, Marit L. Schlagheck, David Walzik, Andreas Groll, Richie P. Goulding, and Philipp Zimmer. Machine learning predicts peak oxygen uptake and peak power output for customizing cardiopulmonary exercise testing using non-exercise features. *European Journal of Applied Physiology*, 11 2024.

[41] Kirk Cureton, Phillip Bishop, Patricia Hutchinson, Hillary Newland, Susan Vickery, and Linda Zwiren. Sex difference in maximal oxygen uptake Effect of' equating haemoglobin concentration. Technical report, 1986.

[42] Jonathan Z. Bakdash and Laura R. Marusich. Repeated measures correlation. *Frontiers in Psychology*, 8(MAR), 4 2017.

[43] J. Martin Bland and Douglas G. Altman. Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics*, 17(4):571–582, 7 2007.

[44] Logan Wade, Laurie Needham, Murray Evans, Polly McGuigan, Steffi Colyer, Darren Cosker, and James Bilzon. Examination of 2D frontal and sagittal markerless motion capture: Implications for markerless applications. *PLoS ONE*, 18(11 NOVEMBER), 11 2023.

[45] Todd A Astorino and Daniel W Roberson. Efficacy of acute caffeine ingestion for short-term high-intensity exercise performance: a systematic review. Technical report, 2010.

Chapter 6

Appendix

6.1 Data Collection

TABLE 6.1: Viability of collected data by participant and visit. VAL columns refer to data necessary for validation step. N1 and N2 refer to NIRS data for the first and second repetition of the sequence, respectively. Red: missing or invalid data, Lightgreen: valid and sufficient, Darkgreen: 2 sufficient recordings

	Ramp			**Easy**			**Medium**			**Hard**		
Participant	**VAL**	**N1**	**N2**	**VAL**	**N1**	**N2**	**VAL**	**N1**	**N2**	**VAL**	**N1**	**N2**
P1												
P3												
P4												
P5												
P6												
P7												
P8												
P9												
P10												
P11												
P12												
P13												
P14												
P15												

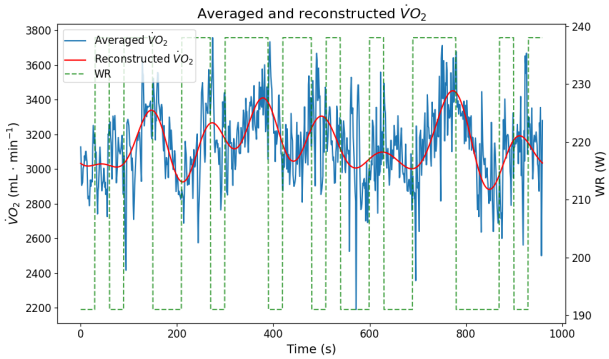
TABLE 6.2: $\dot{V}O_2$ results with P15 excluded. Displaying results across different input features. r_{rm} is the repeated-measures correlation. The bias, LLoA, and ULoA values are the mean bias, lower limit of agreement, and upper limit of agreement, respectively, for the repeated-measures Bland-Altman analysis. The 'best' columns refer to the results of the set of models that includes only the best performing model for each participant.

	Validation (n=8)			**NIRS (n=5)**			**Body Comp (n=5)**			**Both (n=7)**		
	mean± sd	min − max	best	mean± sd	min − max	best	mean± sd	min − max	best	mean± sd	min − max	best
r_{rm}	0.932±0.004	0.924−0.938	0.942	0.938±0.003	0.932−0.940	0.944	0.937±0.004	0.931−0.942	0.950	0.936±0.004	0.930−0.940	0.947
bias	-23 ± 8	-36 − -6	-9	7 ± 16	-18 − 29	-11	-30 ± 28	-80 − 2	-43	-22 ± 25	-72 − 12	-6
LLoA	-519 ± 30	-566 − -469	-412	-432 ± 27	-470 − -389	-388	-508 ± 30	-558 − -472	-428	-506 ± 37	-587 − -471	-388
ULoA	474 ± 35	433 − 553	395	446 ± 14	434 − 472	366	448 ± 59	347 − 524	343	462 ± 30	430 − 519	377

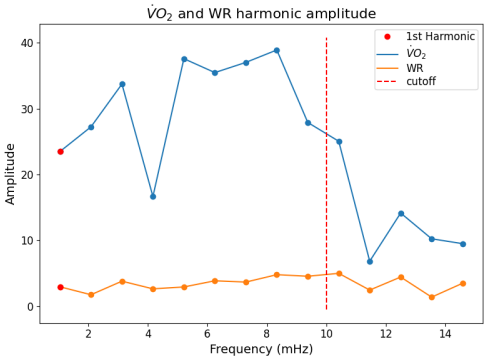
TABLE 6.3: MNG results, across different input feature sets. r_{rm} and p_{rm} are repeated-measures correlation and p-value, respectively. $r_{rm_{Hard}}$ and $p_{rm_{Hard}}$ are these same measurements but with the Hard condition excluded. The bias, LLoA, and ULoA values are the mean bias, lower limit of 95% agreement, and upper limit of 95% agreement, respectively, for the repeated-measures Bland-Altman analysis.

	NIRS (n=5)			**Body Comp (n=5)**			**Both (n=7)**		
	mean ± sd	min − max	best	mean ± sd	min − max	best	mean ± sd	min − max	best
r_{rm}	0.142 ± 0.256	-0.317 − 0.391	-0.060	-0.282 ± 0.150	-0.484 − -0.046	-0.093	0.097 ± 0.186	-0.246 − 0.322	-0.024
p_{rm}	0.279 ± 0.229	0.065 − 0.590	0.786	0.261 ± 0.299	0.009 − 0.815	0.636	0.473 ± 0.307	0.134 − 0.965	0.912
$r_{rm_{Hard}}$	0.432 ± 0.067	0.350 − 0.526	0.312	0.343 ± 0.291	-0.046 − 0.694	0.627	0.318 ± 0.216	0.065 − 0.662	0.351
$p_{rm_{Hard}}$	0.22 ± 0.08	0.12 − 0.32	0.38	0.38 ± 0.39	0.01 − 0.88	0.02	0.45 ± 0.31	0.04 − 0.86	0.32
bias	10 ± 3	7 − 15	10	-10 ± 28	-66 − 6	4	7 ± 2	5 − 10	7
LLoA	-47 ± 4	-53 − -40	-50	-238 ± 366	-970 − -44	-45	-44 ± 6	-51 − -33	-45
ULoA	67 ± 8	58 − 82	69	218 ± 310	56 − 838	53	57 ± 5	48 − 64	59

FIGURE 4: Averaged $\dot{V}O_2$ and frequency domain harmonics of P13's Hard protocol, MNG=121.1%

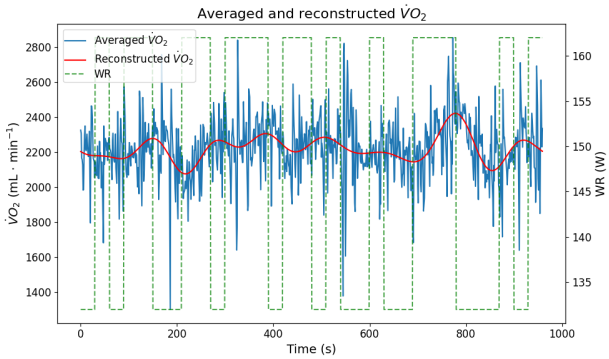


A $\dot{V}O_2$ of averaged repetitions and the same signal reconstructed through Fourier using the 0.01 Hz cut-off frequency.

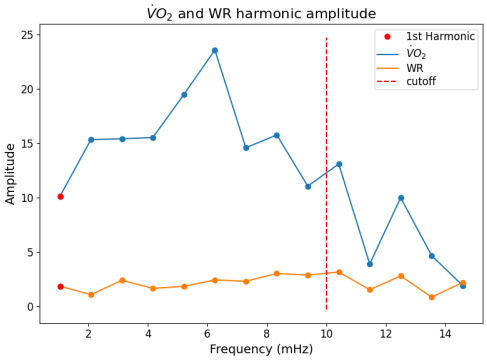


B Amplitude of harmonics of deconstructed $\dot{V}O_2$ and WR signal. Vertical red dotted line indicates the cut-off frequency. First harmonic is indicated in red.

FIGURE 5: Averaged $\dot{V}O_2$ and frequency domain harmonics of P15's Hard protocol, MNG=149.8%



A $\dot{V}O_2$ of averaged repetitions and the same signal reconstructed through Fourier using the 0.01 Hz cut-off frequency.



B Amplitude of harmonics of deconstructed $\dot{V}O_2$ and WR signal. Vertical red dotted line indicates the cut-off frequency. First harmonic is indicated in red.