Check for updates

ARTICLE   OPEN

Temporal convolutional networks predict dynamic oxygen uptake response from wearable sensors across exercise intensities

Robert Amelard[1,2,4 ✉], Eric T. Hedge [2,3,4] and Richard L. Hughson [2,3]

Oxygen consumption ($\dot{V}O_2$) provides established clinical and physiological indicators of cardiorespiratory function and exercise capacity. However, $\dot{V}O_2$ monitoring is largely limited to specialized laboratory settings, making its widespread monitoring elusive. Here we investigate temporal prediction of $\dot{V}O_2$ from wearable sensors during cycle ergometer exercise using a temporal convolutional network (TCN). Cardiorespiratory signals were acquired from a smart shirt with integrated textile sensors alongside ground-truth $\dot{V}O_2$ from a metabolic system on 22 young healthy adults. Participants performed one ramp-incremental and three pseudorandom binary sequence exercise protocols to assess a range of $\dot{V}O_2$ dynamics. A TCN model was developed using causal convolutions across an effective history length to model the time-dependent nature of $\dot{V}O_2$. Optimal history length was determined through minimum validation loss across hyperparameter values. The best performing model encoded 218 s history length (TCN-VO2 A), with 187, 97, and 76 s yielding <3% deviation from the optimal validation loss. TCN-VO2 A showed strong prediction accuracy (mean, 95% CI) across all exercise intensities (-22 ml min^{-1}, [-262, 218]), spanning transitions from low–moderate (-23 ml min^{-1}, [-250, 204]), low–high (14 ml min^{-1}, [-252, 280]), ventilatory threshold–high (-49 ml min^{-1}, [-274, 176]), and maximal (-32 ml min^{-1}, [-261, 197]) exercise. Second-by-second classification of physical activity across 16,090 s of predicted $\dot{V}O_2$ was able to discern between vigorous, moderate, and light activity with high accuracy (94.1%). This system enables quantitative aerobic activity monitoring in non-laboratory settings, when combined with tidal volume and heart rate reserve calibration, across a range of exercise intensities using wearable sensors for monitoring exercise prescription adherence and personal fitness.

INTRODUCTION

Cardiorespiratory fitness is an established risk factor for cardiovascular disease and all-cause mortality[1] and is an important determinant for endurance exercise performance[2]. Cardiorespiratory fitness is conventionally assessed by measuring the rate of oxygen consumption ($\dot{V}O_2$) and its dynamic response to exercise. Biomarkers such as peak oxygen uptake ($\dot{V}O_{2peak}$) and the rate of adaptation to changes in exercise intensity provide important information about the integrative responses of the pulmonary, cardiovascular, and muscular systems[3], which can provide insights into different disease states[4]. Accordingly, $\dot{V}O_2$ monitoring has become a crucial objective measure for advanced clinical therapies[5].

$\dot{V}O_{2peak}$ is often considered the gold standard metric of cardiorespiratory function. In heart failure, $\dot{V}O_{2peak}$ is a strong predictor of 1-year mortality[5] and is clinically used to select patients for advanced therapies. In cases where maximal exercise is infeasible, the dynamic response to sub-maximal exercise also provides important indicators of health[6–10] and fitness status[11–14]. Despite its established importance, monitoring $\dot{V}O_2$ in non-laboratory settings remains challenging. Direct measurement of $\dot{V}O_2$ requires a metabolic cart and trained technician, which limits its applicability to laboratory assessment. Heart rate (HR) has traditionally been used as an inexpensive and non-intrusive proxy to $\dot{V}O_2$ response to activity and estimate energy expenditure under the assumption that HR varies linearly with $\dot{V}O_2$[15,16];

however, dynamic $\dot{V}O_2$ and HR responses do not always have a direct correspondence with each other, such as following prior exercise[17]. Thus, ambulatory physiological monitoring using wearable sensors may provide early detection of sub-clinical biomarkers of disease and enable more widespread assessment of cardiorespiratory function[18].

Recent advances in wearable technologies and artificial intelligence have led to new developments in non-intrusive cardiorespiratory monitoring. These approaches are generally modeled as regression problems, where a machine learning model learns a transformation function between physiological inputs from wearable sensors and $\dot{V}O_2$ measured using a gas analyzer system. Earlier work primarily used a combination of HR and activity-related inputs to predict $\dot{V}O_2$ during various forms of exercise[19,20]. Sensors embedded in textile fabrics have enabled $\dot{V}O_2$ prediction in low-to-moderate-intensity exercise during activities of daily living[21,22]. Recent studies have leveraged the time-dependent nature of $\dot{V}O_2$ by modeling the regression as a time-series, or sequential, prediction, where previous physiological states were used to predict $\dot{V}O_2$[23] and ventilatory threshold (VT)[24] during stationary cycling on a cycle ergometer. Although sequential prediction has been shown to model the temporal nature of oxygen uptake well, further investigations are needed to assess model efficiency and performance across a range of exercise intensities, different days, and demographics.

[1]KITE-Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada. [2]Schlegel-UW Research Institute for Aging, Waterloo, ON, Canada. [3]University of Waterloo, Waterloo, ON, Canada. [4]These authors contributed equally: Robert Amelard, Eric T. Hedge. ✉email: robert.amelard@uhn.ca

In this paper, we propose and evaluate a sequential deep learning model based on temporal convolutional networks (TCNs)[25] for predicting $\dot{V}O_2$ from physiological inputs derived from smart textiles and a cycle ergometer. The model used causal convolutions to incorporate only past and present physiological response, and the system architecture was designed to provide a tunable effective history length or receptive fields. We assessed the effect of receptive field and model complexity on prediction accuracy to investigate the temporal relationship between physiological inputs and $\dot{V}O_2$ response to provide guidance on optimal model design and assessment across a range of different exercise intensities.
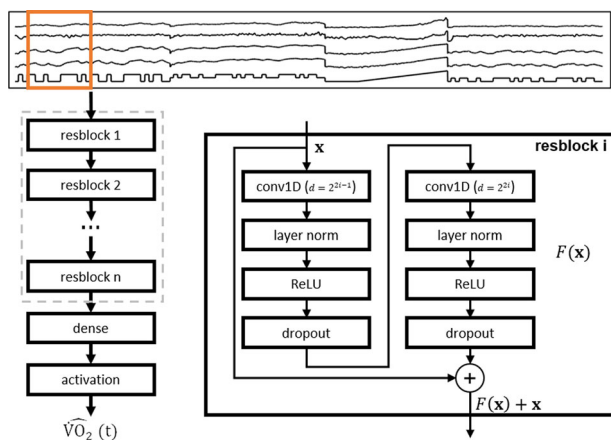
RESULTS

Experimental setup

Figure 1 shows the data flow through the TCN network architecture. The four cardiorespiratory biosignals derived from the smart shirt (HR, HR reserve, breathing frequency, and minute ventilation ($\dot{V}_E$)) and the work rate (WR) profile were used as inputs into a chain of residual blocks, followed by a dense layer and linear activation to predict $\dot{V}O_2$ at each time point. Results of the TCN networks were compared against a stacked long short-term memory (LSTM) network[26] and random forest (RF)[27] prediction models. The stacked LSTM model was trained using the originally proposed features[23], as well as adding HR reserve and $\dot{V}_E$ to the feature set, with a sequence length of 140 s approximating 70 breaths at low intensity exercise. The RF model was built using the optimal number of trees according to the validation loss (30 trees; see Supplementary Fig. 1). $\dot{V}O_2$ data were converted to ml min^{-1} kg^{-1} to compute metabolic equivalent of task (METs) for quantifying physical activity levels (METs = $\dot{V}O_2/3.5$[28]). METs were classified as light (<3.0 METs), moderate (3.0–5.9 METs), or vigorous (≥6.0 METs) intensity exercise according to established guidelines[28].

Hyperparameters

Figure 2 shows the effect on validation loss of receptive field (by modifying kernel size k and dilation depth d) and the number of filters in the causal convolutional layers across a combination of hyperparameter values (see Table 1). Trend lines were generated using exponential weighting moving average ($a = 0.5$) for visualization purposes. Validation loss decreased to convergence

with larger receptive fields, apart from the two-filter models that were insufficient for learning the prediction function, as evidenced by a flat and highly variable loss curve over all receptive fields. Performance increase was marginal beyond eight filters. The optimal hyperparameters that produced the smallest hold-out validation loss were 24 filters and a receptive field of 218 s using a kernel size of 8 s and 5 dilations (2 residual blocks). However, a smaller receptive field and/or a smaller set of model parameters may be preferable to a marginal loss increase. Smaller receptive fields allow for reduced "cold start" time, and fewer parameters result in decreased computational load. Thus, we assessed the accuracy of models that were within 5% of the best (minimum) validation loss below (see Supplementary Table 1 for full list of models with validation losses).

Network performance

We assessed the performance of models exhibiting validation loss within 5% of the minimum validation loss, as well as models with no history (TCN-VO2 NH) and using only HR as input (TCN-VO2 A (HR)), and compared the results to existing $\dot{V}O_2$ prediction methods (Table 2). There were eight models within the 5% performance threshold, of which we reported four with varied receptive fields and parameter sets. The best performing model according to hold-out validation loss (TCN-VO2 A) had a 218 s receptive field and 19,921 parameters. TCN-VO2 B had marginally higher validation loss with 187 s receptive field, indicating minimal performance gain for additional history beyond 187 s. TCN-VO2 C and D had near identical validation loss (3% increase over model A) but required a receptive field of only 76 and 91 s, respectively, thus providing a reduced cold start period for initial prediction.
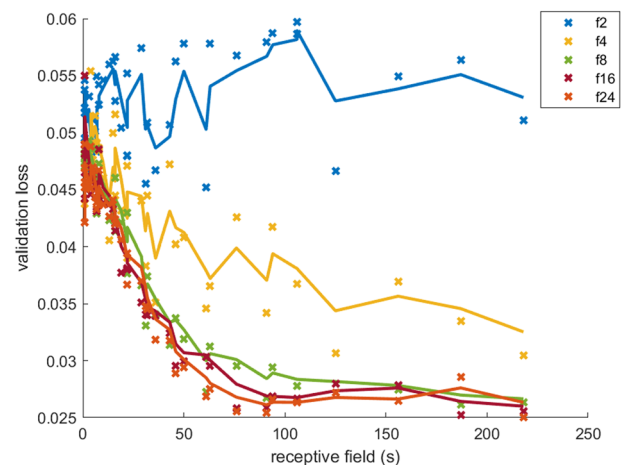
Fig. 2 Model validation performance across different receptive fields (determined by kernel size and dilation depth) grouped by the number of filters in the convolutional layers (f#). Exponential weighted moving average fits were plotted for visualization purposes. The optimal model hyperparameters were 24 filters, 8 kernel sizes, and 5 dilations. Two filters per layer were insufficient for learning the transformation function.

Fig. 1 Convolutional neural network architecture for temporal prediction of $\dot{V}O_2$. Temporal features (heart rate, heart rate reserve, breathing frequency, minute ventilation, and work rate) are processed through a series of residual blocks with causal convolutions and dilations for feature extraction, followed by a fully connected layer and linear activation to predict $\dot{V}O_2$ at time t.

Table 1. Hyperparameter values for investigating the optimal temporal convolutional network configuration.

Hyperparameter	Values
Filters	2, 4, 8, 16, 24
Kernel size	1, 2, 3, 4, 5, 6, 7, 8
Dilation depth	1, 2, 3, 4, 5

**Table 2.** Error rates and characteristics of TCN and comparison models.

Method	Filters	Receptive field (s)	Validation loss	# params	L-M	L-H	VT-H	MAX	Combined	$\dot{V}O_{2peak}$
Random forest[21]				298,106	20 ± 134	144 ± 187	54 ± 169	15 ± 111	62 ± 164	134 ± 151
SLSTM[23]				21,589	93 ± 133	52 ± 145	-58 ± 148	-16 ± 178	21 ± 160	-48 ± 243
SLSTM + {HRR, $\dot{V}_E$}				21,845	52 ± 132	7 ± 163	-72 ± 152	-37 ± 163	-10 ± 159	-17 ± 205
TCN-VO2 A	24	218	0.02506	19,921	-23 ± 116	14 ± 136	-49 ± 115	-32 ± 117	-22 ± 122	18 ± 182
TCN-VO2 B	16	187	0.02525	8081	-7 ± 109	30 ± 138	-25 ± 115	-2 ± 113	-1 ± 120	62 ± 162
TCN-VO2 C	16	76	0.02585	5393	-21 ± 113	16 ± 159	-57 ± 136	3 ± 123	-17 ± 137	93 ± 214
TCN-VO2 D	16	91	0.02586	6241	-24 ± 115	31 ± 142	-54 ± 123	9 ± 117	-11 ± 128	80 ± 165
TCN-VO2 NH	24	1	0.04895	361	10 ± 127	133 ± 179	69 ± 129	75 ± 123	72 ± 149	223 ± 196
TCN-VO2 A(HR)	24	218	0.10733	19,057	197 ± 286	133 ± 373	40 ± 412	145 ± 445	127 ± 379	-220 ± 445

Errors are reported as mean ± SD [ml min^{-1}].
SLSTM stacked long short-term memory.
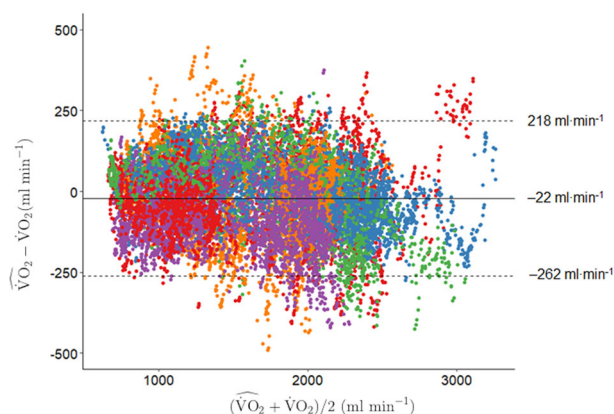
Fig. 3 Repeated-measures Bland–Altman analysis of the predicted oxygen uptake ($\dot{V}O_2$) using TCN-VO2 A and directly measured $\dot{V}O_2$ with all exercise conditions combined. Dotted horizontal lines represent the 95% limits of agreement and the solid line represents the prediction bias. Each color represents data from a unique participant in the test set.

Repeated-measures Bland–Altman analysis showed no systematic error between predicted and true $\dot{V}O_2$ across all TCN-VO2, LSTM, and RF models and no proportional error (Fig. 3). The limits of agreement (LoA) remained relatively constant across TCN-VO2 models during low-to-moderate (L-M), VT-to-high (VT-H), and MAX data but increased with decreased model complexity in low-to-high (L-H) and combined data. Model bias on all combined data was <25 ml min^{-1} in the best performing TCN-VO2 models and 72 ml min^{-1} for 1 s receptive field (model NH), and the equality line fell within the confidence interval of the mean difference. TCN-VO2 models exhibited smaller LoA compared to LSTM in all exercise protocols, indicating a stronger overall fit and smaller error variance. Similarly, mean error of $\dot{V}O_{2peak}$ was smallest in TCN-VO2 A (18 ml min^{-1}), strongly outperforming models with no history (RF: 134 ml min^{-1}, TCN-VO2 NH: 22 ml min^{-1}) and HR only (-220 ml min^{-1}), as well as exhibiting smaller LoA than LSTM models. Furthermore, TCN-VO2 models required substantially fewer network parameters to achieve comparable bias and lower LoA, requiring 1.1×, 4.1× and 60.5× fewer parameters in the TCN-VO2 A, TCN-VO2 C, and TCN-VO2 NH respectfully, compared to the stacked LSTM model. TCN-VO2 A(HR) exhibited high error bias and variance across protocols, showing that HR is not a robust estimator of $\dot{V}O_2$.

TCN-VO2 with no history (NH), comprising 1 s receptive field, was compared against RF, which performs point-wise predictions without input from previous states. TCN-VO2 NH and RF performed comparably, although the TCN-VO2 NH parameter set contained 825.8× fewer parameters than RF (quantified as the number of split nodes in the forest). Both TCN-VO2 NH and RF exhibited larger bias in L-H data compared to the other protocols due to overestimating $\dot{V}O_2$ during low WR transients (see Fig. 4 inset). $\dot{V}O_2$ prediction during exercise with relatively small WR transitions was not as severely affected. Top performing TCN models were the only models to correctly predict the lowest $\dot{V}O_2$ values during the off-transients, making them good candidates for assessing kinetics.

Quantifying physical activity levels

Predicted $\dot{V}O_2$ were used in accordance with globally established guidelines on exercise prescription for health[28] to classify second-by-second activity levels. These guidelines are often used in practice, alongside other biomarkers, for optimizing cardiovascular health during rehabilitation exercise prescription. Figure 5 shows the MET data derived from the true (measured) and predicted $\dot{V}O_2$ data, as well as the confusion matrix. All test data (4 protocols/participant across the 5-participant test set) were classified on a second-by-second basis and visually concatenated into a single plot spanning a total of 16,090 s of predicted metabolic activity across 20 individual exercise trials. Overall, 15,147 s (94.1%) of the 1 Hz data were correctly classified into appropriate physical activity categories. The data spanned across light (2.7%), moderate (33.8%), and vigorous (63.5%) intensity exercise. During moderate-intensity exercise (5445 s), 4882 s (89.7%) were correctly classified, while 153 s (2.8%) and 410 s (7.5%) were classified as light and vigorous activity, respectively. During vigorous-intensity exercise (10,216 s), 10,004 s (97.9%) were correctly classified, while 212 s (2.1%) were classified as moderate activity. Less than 3% of the data were in the "light" category (429 s), and of those data, 261 s (60.8%) were correctly classified and 168 s (39.2%) were classified as moderate activity. In all, 90% of these data were within 0.5 METs of the threshold.

DISCUSSION

This work has shown that the complex dynamic $\dot{V}O_2$ response to changes in exercise intensity can be accurately predicted using sequential deep learning models across a range of low-, moderate-, and high-intensity exercise, as well as maximal aerobic exercise.
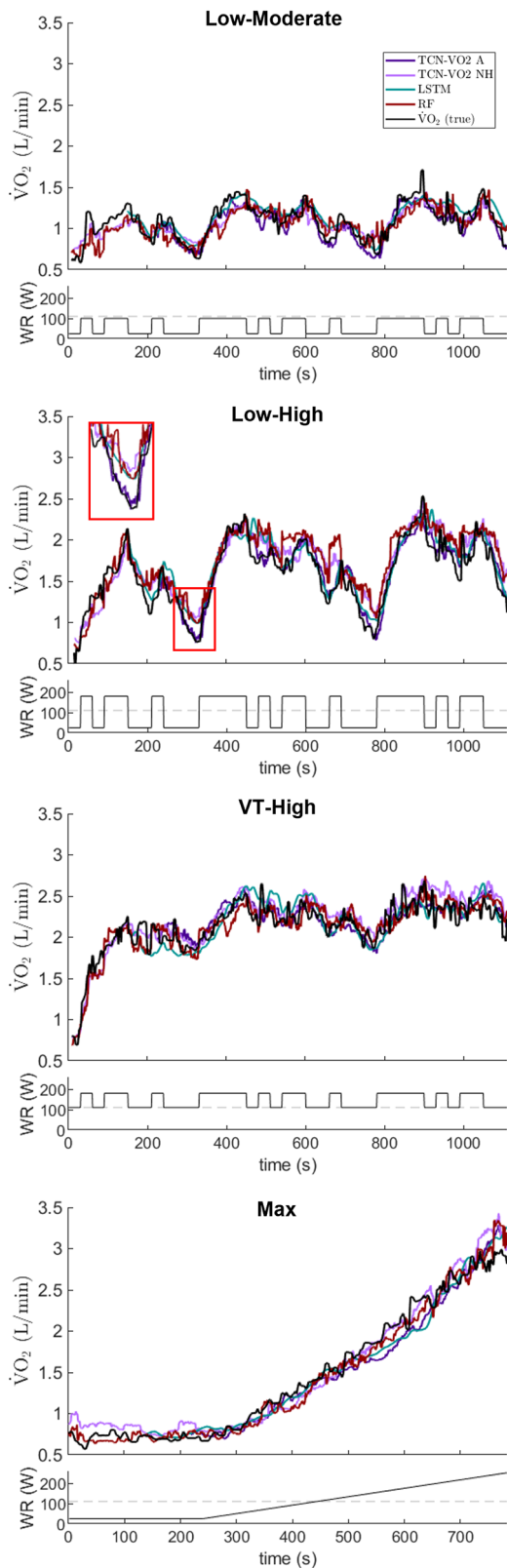
Fig. 4 Results for a representative participant across the four exercise protocols. Prediction curves are shown for VO2-TCN (proposed) with the lowest validation loss (VO2-TCN A) and 1 s receptive field (VO2-TCN NH), random forest (RF)[21], and stacked long short-term memory network (LSTM)[23]. Dashed gray line represents the participant's ventilatory threshold.

The baseline model TCN-VO2 NH (with 1 s receptive field) and point-wise RF regression[21] were both blind to previous states, using only the current time point for prediction. In both cases, we observed a large and similar bias in the L-H data, albeit being two fundamentally different model types. The source of error was largely due to overestimation during recovery in the off-transients. The overestimation is likely a result of the point-wise predictor's blindness to the previous state, as the models were unable to adequately learn the relationships between model inputs and outputs during on- and off-transients. This is an important consideration, as differences in HR dynamics have been observed between on- and off-transients, with the slower off-transients being amplified following high-intensity exercise[17,29]. HR kinetics are also reported to be slower than $\dot{V}O_2$ kinetics during recovery[17]. Similarly, the ventilation response has been observed to be slower during recovery compared to exercise onset[30], with the rate of change in ventilation being markedly slowed when recovering from higher intensity exercise[29]. Furthermore, the ventilatory response is slower than the $\dot{V}O_2$ response[31,32]. Accordingly, without knowing the previous history of the system, these point-wise models are naive to status of the system leading to erroneous $\dot{V}O_2$ predictions. Thus, it appears that the temporal models were able to more accurately learn the relationships between HR and ventilation and $\dot{V}O_2$ during exercise onset as well as recovery.

The best performing models, when ranked by validation loss, were consistently those with receptive fields of 218, 187, 76, and 91 s. We observed marginal difference between the test errors of these models, although there were some differences in predicting $\dot{V}O_2$ minima at the end of off-transients, which did not significantly affect the global loss function. Considering that the standard $\dot{V}O_2$ time constant is typically 20–30 s in healthy populations[33], these receptive fields range from approximately 3 to 11 time constants, which suggests that the best models tend to use a receptive field that includes most, if not all, of the transient phase for a sustained step change in WR to achieve a new constant $\dot{V}O_2$. Receptive fields of 76 and 91 s appear consistent with the protocol's longest off-transient (90 s). The longest on-transient is 120 s. Thus, further investigations are needed to determine the effects of receptive field during different exercise protocols.

Both classes of sequential deep learning models (TCN and LSTM) exhibited strong predictive power. TCN architectures have become popular alternatives to recurrent neural network instances largely due to their data parallelism, flexible receptive field size, and stable gradients[25]. In this work, top performing TCN models were much smaller than the stacked LSTM models, yielding more computationally efficient prediction networks.

The $\dot{V}O_2$ predictions were converted to METs, which is an established metric for quantifying physical activity levels[28]. Quantifying activity levels is helpful for exercise prescription in cardiovascular disease management[34]. However, traditional patient recall may be affected by recall and/or social desirability bias[35], providing uncertainty in an important biomarker for cardiovascular health. This work showed strong accuracy across a range of MET categories, which provides supporting evidence for quantitative and objective at-home activity monitoring using wearable sensors. The errors in the light activity category were mainly during category transition, not during sustained activity. In these data, the majority of light data was close to threshold. It would be expected that incorporating data from resting of very low-intensity exercise would lead to a more representative error profile.

The primary limitations impacting the widespread generalizability of these results stem from the dataset's constrained demographic (young healthy adults) and structured exercise protocol. We selected the three-stage pseudorandom binary
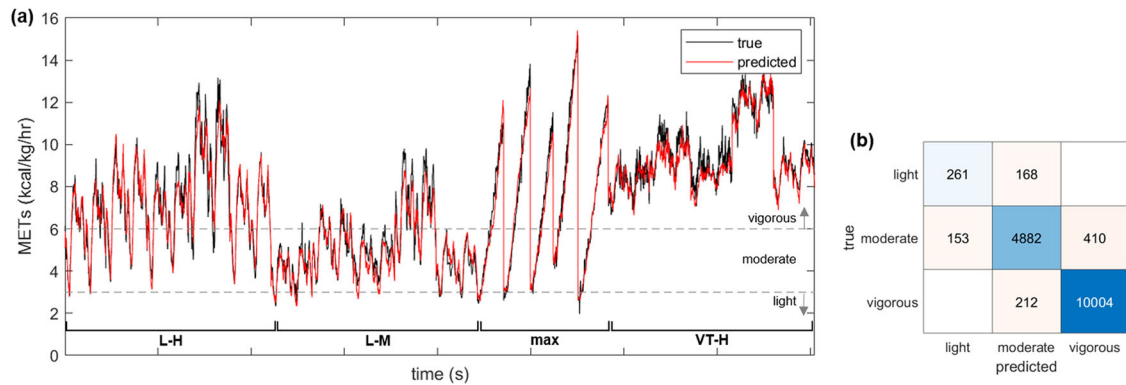
Fig. 5 Temporal prediction of physical activity in the test set. a Prediction of metabolic equivalent of task (METs) over all test data (four protocols per participant across the five-participant test set) concatenated into one plot and visually grouped by protocol for quantifying physical activity levels according to global guidelines. **b** Confusion matrix of physical activity classification (light: <3.0; moderate: 3.0–5.9; vigorous: ≥6.0).
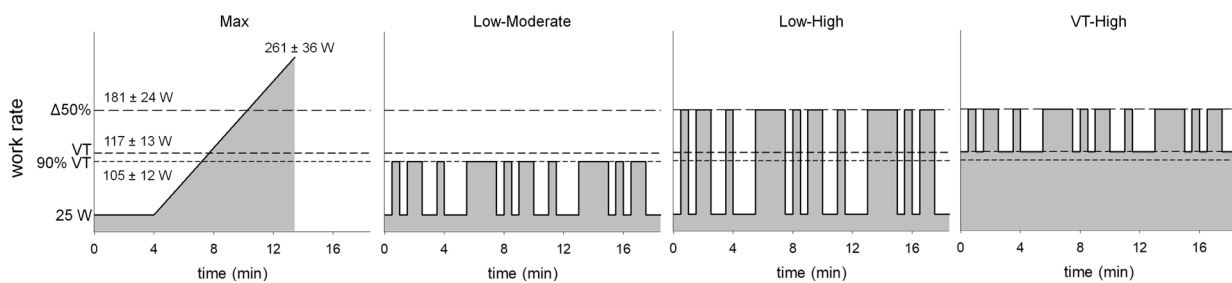
Fig. 6 Schematic showing the incremental ramp and three pseudorandom binary sequence (PRBS) cycling tests. PRBS protocols were designed using work rates at 90% ventilatory threshold (VT), VT, and the midpoint between VT and $\dot{V}O_{2peak}$ (Δ50%) using participant-specific VT and $\dot{V}O_{2peak}$ determined from the ramp test.

sequence (PRBS) protocol because it challenged the dynamic response of aerobic metabolism across a wide range of intensities that could occur in real-life situations. PRBS exercise provides an opportunity to directly quantify an index of physical fitness[36]; however, PRBS cycling exercise is a controlled laboratory protocol. Accordingly, our optimization results may not be directly applicable to unstructured exercise or different exercise protocols (i.e., constant load exercise). Further investigations in different exercise situations with a more diverse participant sample, including cardiovascular-related diseases, are needed to assess its generalizability to different kinds of physical activity and out-of-sample populations. Additionally, the exercise in this study was performed in a temperature-controlled environment. Further investigations are needed to evaluate model performance in specialized environments that may alter cardiovascular response to exercise (e.g., hypoxia, heat stress, etc.).

To conclude, using causal convolutions in a temporal deep learning model, the effect of receptive field (i.e., effective history) on $\dot{V}O_2$ prediction was assessed by performing a grid search across hyperparameter values. The best performing models, according to validation loss, comprised receptive fields of 218, 187, 97, and 76 s. Results showed low prediction error across a wide range of exercise intensities with drastically reduced parameter sets compared to existing methods. Using HR as the only input feature into the same model architecture yielded substantially larger errors, reinforcing that HR alone is insufficient for predicting $\dot{V}O_2$, thus necessitating more complex approaches. Using the temporal prediction outputs, physical activity levels were quantified to provide a breakdown of time spent in light, moderate, and vigorous activity according to global health guidelines. These results suggest that cardiorespiratory function

may be assessed in non-laboratory settings, when combined with tidal volume calibration and exact determination of HR reserve, across a wide range of activity levels using wearable sensors and smart textiles.

METHODS
Data collection and preprocessing
Twenty-two young healthy adults (13 males, 9 females; age: 26 ± 5 years; height: 1.71 ± 0.08 m; mass: 70 ± 11 kg; $\dot{V}O_{2peak}$: 42 ± 6 ml min^{-1} kg^{-1}) with no known musculoskeletal, respiratory, cardiovascular, or metabolic conditions volunteered to participate in the study. The study was approved by a University of Waterloo Research Ethics committee (ORE #32164) and conducted in accordance with the Declaration of Helsinki. All participants signed an informed consent before participating.

Participants visited the laboratory on four separate occasions to perform a ramp-incremental exercise test and three different PRBS exercise tests[37]. Each exercise session was separated by at least 48 h, and participants were instructed to arrive for testing at least 2 h postprandial and abstain from alcohol, caffeine and vigorous exercise in the 24 h preceding each test. All exercise tests were performed in an environmentally controlled laboratory on an electronically braked cycle ergometer (Lode Excalibur Sport, Lode B. V., Groningen, Netherlands). Participants were instructed to maintain cadence at 60 revolutions per minute for all exercise tests.

On the first visit, 5 min of seated resting data were collected to determine each participant's resting HR. After the resting period, participants performed a ramp-incremental exercise test to exhaustion (25 W baseline for 4 min followed by a 25 W min^{-1} ramp) to determine each participant's VT[38], $\dot{V}O_{2peak}$, and the WRs for the PRBS exercise tests (Fig. 6). The test was terminated when the cadence dropped below 55 revolutions per minute despite strong verbal encouragement. $\dot{V}O_{2peak}$ was defined as the highest $\dot{V}O_2$ computed from a 20 s moving average during the exercise test. $\dot{V}O_2$ at VT was estimated by visual inspection using standard ventilatory and gas exchange indices, and their ratios, as previously described[38]. WRs at 90% VT, VT, and the

midpoint between VT and $\dot{V}O_{2peak}$ (referred to as Δ50%) were estimated by left-shifting the $\dot{V}O_2$ response by each individual's mean response time to align the $\dot{V}O_2$ and WR profiles. Mean response time was determined by fitting a double-linear model to the ramp-incremental data and finding the point of intersection between the forward extrapolation of the average $\dot{V}O_2$ during the 25 W baseline cycling in the 2 min prior to ramp onset and the backward extrapolation of the linear portion of the ramp $\dot{V}O_2$ response below VT[39].

In visits 2–4, participants performed one of the three different PRBS exercise tests in a randomized order. WRs systematically alternated in the three PRBS exercise test between 25 W and 90% of the VT (L-M), 25 W and Δ50% (L-H), or VT and Δ50% (VT-H). The time series for the changes in WR for PRBS protocols were generated by a digital shift register with an adder module feedback[37,40–42]. This process pseudo-randomized the changes in WR and ensured that there would be sufficient $\dot{V}O_2$ signal amplitude while performing non-constant load exercise. A single PRBS was composed of 15 units, each of 30 s in duration, totaling 7.5 min. Each complete PRBS testing session consisted of a 3.5 min warm-up (the last 3.5 min of the 7.5 min PRBS) and then two full repetitions of the PRBS for a total of 18.5 min of continuous cycling per session.

A portable metabolic system (MetaMax 3B-R2, CORTEX Biophysik, Leipzig, Germany) was used to measure gas exchange during all exercise tests. Participants breathed through a mask (7450 SeriesV2 Mask, Hans Rudolph Inc., Shawnee, KS, USA), and inspired and expired flow were measured using a bi-directional turbine. The turbine was calibrated before each testing session using a 3 L syringe. Oxygen and carbon dioxide gas concentrations were continuously sampled at the mouth and were analyzed using a chemical fuel cell and nondispersive infrared sensor, respectively. Precision-analyzed gas mixtures were used to calibrate the oxygen and carbon dioxide gas concentrations. $\dot{V}O_2$ and carbon dioxide output were calculated using the standard breath-by-breath algorithms[3]. $\dot{V}O_2$ data were filtered using a sliding 5-breath median filter to correct spurious outlier breaths.

Participants wore a HR monitor (Polar H7, Polar Electro Oy, Kempele, Finland) that wirelessly communicated with the portable metabolic system, such that HR data were logged synchronously with the gas exchange data. Participants were also fitted with a wearable integrated sensors shirt that was sized to each participant based on the manufacturer's guidelines (Hexoskin, Carre Technologies, Montreal, Canada). The shirt contained a textile electrocardiogram to measure HR and thoracic and abdominal respiration bands to obtain estimates of breathing frequency and minute ventilation ($\dot{V}_E$) via respiratory inductance plethysmography. Estimates of $\dot{V}_E$ provided by the smart shirt were calibrated by linear regression to the known $\dot{V}_E$ measured throughout each protocol with the bi-directional turbine. The estimates of HR, $\dot{V}_E$, and breathing frequency from the smart shirt have been previously validated[43]. Data recorded by the metabolic system and smart shirt were time-aligned by cross-correlating the two different HR signals. After processing, all data were interpolated to 1 Hz to ensure signal synchronization and a constant sampling rate. The final set of physiological features were: WR (W), $\dot{V}_E$ (L min^{-1}), breathing frequency (breath min^{-1}), HR (bpm), and HR reserve (%). HR reserve at time t was calculated as $(HR_t - HR_{rest})/(HR_{max} - HR_{rest})$, where HR_{rest} and HR_{max} were determined during 5 min rested seated baseline and the peak HR during ramp-incremental test, respectively.

Network architecture

$\dot{V}O_2$ kinetics have a systematic, albeit complex, temporal response to exercise[44]. Exponential models for quantifying $\dot{V}O_2$ kinetics during controlled exercise protocols have demonstrated that the intensity of the exercise being performed strongly influences the dynamic $\dot{V}O_2$ response[36,45,46]. Thus, as an alternative to conventional convolutional neural networks that use a symmetric kernel about the current time (or space) location, we developed a sequential convolutional model using causal convolutions[25]. Specifically, given a sequence of time-series inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T \in \mathbb{R}^n$ extracted from wearable sensors and a cycle ergometer, the goal was to predict the $\dot{V}O_2$ at time t, denoted $\dot{V}O_{2,t}$. Specifically, given a prediction model $M : \mathcal{X}^T \to \mathcal{Y}$, $\dot{V}O_{2,t}$ is predicted using historical temporal features $\mathbf{x}_i \in \mathbb{R}^n$ only up to the current time point:

$$\widehat{\dot{V}O_{2,t}} = M(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-w}) \tag{1}$$

where w is the effective history.

Our model incorporated a temporal feature extraction network and a regression network. The feature extraction network was implemented as a TCN with a tunable receptive field through multiple sequential layers with kernel dilation for multi-scale aggregation of the input data[25,47,48]. The TCN was implemented as a sequence of stacked residual blocks comprised of repetitions of dilated causal one-dimensional convolution, layer normalization, rectified linear activation, and dropout (see Fig. 1). The input into the residual block $\mathbf{x}$ was added to the residual function $F(\mathbf{x})$ through identity mapping (or 1×1 convolution in the first residual block when the number of channels did not match the residual function shape) to encourage learning of the residual modifications of the input data, which has been shown to improve the performance of deep networks[49]. Each residual block is composed of successive pairs of dilations, rather than applying the same dilation twice as in the original TCN description[25], for better control of the receptive field. In models with an odd number of dilations, the first three were grouped into a single residual block, inspired by ResNet, which learns a function over two or three layers[49]. Thus, the (causal) receptive field was determined by the kernel size k and the number of exponential dilations N:

$$RF = 1 + (k-1)(2^N - 1) \tag{2}$$

A splice function was used to extract the features at the last known time point t for input into the regression network. This network was defined simply as a fully connected dense layer and linear activation to predict $\dot{V}O_{2,t}$.

Training and hyperparameter optimization

The data across all 22 participants were split into train (40 protocols), test (20 protocols), and validation (19 protocols) datasets. Each participant's whole data were only comprised within one of train, test, or validation, with no participant data spread to encourage generalizability. Data of specific sequence lengths (according to the receptive field) were extracted from whole exercise protocols through a sliding window method, where the first sequence of sequence length T was extracted from the onset of exercise to time T, and subsequent sequences were extracted until the end of the exercise was reached. This yielded a dataset of size $\mathbb{R}^{N \times T \times F}$, where N is the number of sequences, T is the sequence length, and F is the number of feature signals. The sequence length was determined by the network's receptive field. Each feature, except WR, was standardized to zero mean and unit variance according to the training data statistics. WR was normalized to [0, 1] due to its non-normal distribution.

The hyperparameters of importance are the number of convolutional filters and the receptive field, which is defined by kernel size, dilation rate, and network depth (Eq. (2)). This has practical implication in that having a very long receptive field results in a delayed initial prediction (or "cold start") of that length of time. Hyperparameter search (number of filters, kernel size, dilation depth) was performed on the hold-out validation set. Table 1 lists the hyperparameter values that were searched. A grid search was performed on a distributed computing cluster using 108 cores and 36 NVIDIA T4 Turing GPUs across 9 compute nodes.

The network was trained using the Adam optimizer, 0.2 dropout rate, 32 minibatch size, and a learning rate of 0.0005 over 100 epochs. For each hyperparameter combination, the epoch with the lowest validation loss was saved. The network hyperparameters producing the lowest mean squared error in the validation set was chosen as the final network architecture.

Data analysis

Signals were analyzed in MATLAB (2020b, MathWorks, Portola Valley, CA, USA). Statistical analyses were conducted in R (version 3.5.1). The agreement between the predicted and directly measured $\dot{V}O_2$ were assessed using repeated-measures Bland–Altman analysis, which accounts for the within-participant variance of the repeated-measures data[50]. $\dot{V}O_{2peak}$ agreement was assessed using the standard Bland–Altman.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Data are available by contacting the corresponding author upon reasonable request.

CODE AVAILABILITY
The code used to generate the results are available by contacting the corresponding author upon reasonable request.

REFERENCES
1. Ross, R. et al. Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association. *Circulation* **134**, e653–e699 (2016).
2. Joyner, M. J. & Coyle, E. F. Endurance exercise performance: the physiology of champions. *J. Physiol.* **586**, 35–44 (2008).
3. Wasserman, K., Hansen, J. E., Sue, D. Y., Casaburi, R. & Whipp, B. J. *Principles of Exercise Testing and Interpretation* 3rd edn (Lippincott Williams & Wilkins, Philadelphia, 1999).
4. Guazzi, M. et al. Clinical recommendations for cardiopulmonary exercise testing data assessment in specific patient populations. *Eur. Heart J.* **33**, 2917–2927 (2012).
5. Mancini, D. M. et al. Value of peak exercise oxygen consumption for optimal timing of cardiac transplantation in ambulatory patients with heart failure. *Circulation* **83**, 778–786 (1991).
6. Rocca, H. P. B.-L. et al. Prognostic significance of oxygen uptake kinetics during low level exercise in patients with heart failure. *Am. J. Cardiol.* **84**, 741–744 (1999).
7. Alexander, N. B., Dengel, D. R., Olson, R. J. & Krajewski, K. M. Oxygen-uptake (VO2) kinetics and functional mobility performance in impaired older adults. *J. Gerontol. Ser. A* **58**, M734–M739 (2003).
8. Schalcher, C. et al. Prolonged oxygen uptake kinetics during low-intensity exercise are related to poor prognosis in patients with mild-to-moderate congestive heart failure. *Chest* **124**, 580–586 (2003).
9. Borghi-Silva, A. et al. Relationship between oxygen consumption kinetics and BODE Index in COPD patients. *Int. J. Chron. Obstr. Pulm. Dis.* **7**, 711–718 (2012).
10. Malhotra, R., Bakken, K., D'Elia, E. & Lewis, G. D. Cardiopulmonary exercise testing in heart failure. *JACC Heart Fail.* **4**, 607–616 (2016).
11. Hickson, R. C., Bomze, H. A. & Hollozy, J. O. Faster adjustment of o_2 uptake to the energy requirement of exercise in the trained state. *J. Appl. Physiol.* **44**, 877–881 (1978).
12. Hagberg, J. M., Hickson, R. C., Ehsani, A. A. & Holloszy, J. O. Faster adjustment to and recovery from submaximal exercise in the trained state. *J. Appl. Physiol.* **48**, 218–224 (1980).
13. Powers, S. K., Dodd, S. & Beadle, R. E. Oxygen uptake kinetics in trained athletes differing in VO$_{2max}$. *Eur. J. Appl. Physiol. Occup. Physiol.* **54**, 306–308 (1985).
14. Chilibeck, P. D., Paterson, D. H., Petrella, R. J. & Cunningham, D. A. The influence of age and cardiorespiratory fitness on kinetics of oxygen uptake. *Can. J. Appl. Physiol.* **21**, 185–196 (1996).
15. Swain, D. P., Leutholtz, B. C., King, M. E., Haas, L. A. & Branch, D. J. Relationship between % heart rate reserve and % VO$_2$ reserve in treadmill exercise. *Med. Sci. Sports Exerc.* **30**, 318–321 (1998).
16. Strath, S. J. et al. Evaluation of heart rate as a method for assessing moderate intensity physical activity. *Med. Sci. Sports Exerc.* **32**, S465–S470 (2000).
17. Bearden, S. E. & Moffatt, R. J. VO$_2$ and heart rate kinetics in cycling: transitions from an elevated baseline. *J. Appl. Physiol.* **90**, 2081–2087 (2001).
18. Sana, F. et al. Wearable devices for ambulatory cardiac monitoring: JACC state-of-the-art review. *J. Am. Coll. Cardiol.* **75**, 1582–1592 (2020).
19. Beltrame, T. et al. Estimating oxygen uptake and energy expenditure during treadmill walking by neural network analysis of easy-to-obtain inputs. *J. Appl. Physiol.* **121**, 1226–1233 (2016).
20. Altini, M., Penders, J. & Amft, O. Estimating oxygen uptake during nonsteady-state activities and transitions using wearable sensors. *IEEE J. Biomed. Health Inform.* **20**, 469–475 (2016).
21. Beltrame, T., Amelard, R., Wong, A. & Hughson, R. L. Prediction of oxygen uptake dynamics by machine learning analysis of wearable sensors during activities of daily living. *Sci. Rep.* **7**, 45738 (2017).
22. Beltrame, T., Amelard, R., Wong, A. & Hughson, R. L. Extracting aerobic system dynamics during unsupervised activities of daily living using wearable sensor machine learning models. *J. Appl. Physiol.* **124**, 473–481 (2018).
23. Zignoli, A. et al. Estimating an individual's oxygen uptake during cycling exercise with a recurrent neural network trained from easy-to-obtain inputs: a pilot study. *PLoS ONE* **15**, e0229466 (2020).
24. Miura, K. et al. Feasibility of the deep learning method for estimating the ventilatory threshold with electrocardiography data. *npj Digital Med.* **3**, 141 (2020).
25. Bai, S., Kolter, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. Preprint at https://arxiv.org/abs/1803.01271 (2018).
26. Zignoli, A. et al. Estimating an individual's oxygen uptake during cycling exercise with a recurrent neural network trained from easy-to-obtain inputs: a pilot study. *PLoS ONE* **15**, e0229466 (2020).
27. Beltrame, T., Amelard, R., Wong, A. & Hughson, R. L. Prediction of oxygen uptake dynamics by machine learning analysis of wearable sensors during activities of daily living. *Sci. Rep.* **7**, 45738 (2017).
28. World Health Organization. *Global Recommendations on Physical Activity for Health* (WHO, 2010).
29. Linnarsson, D. Dynamics of pulmonary gas exchange and heart rate changes at start and end of exercise. *Acta Physiol. Scand. Suppl.* **415**, 1–68 (1974).
30. Hughson, R. L. Coupling of ventilation and gas exchange during transitions in work rate by humans. *Respir. Physiol.* **101**, 87–98 (1995).
31. Bell, C., Kowalchuk, J. M., Paterson, D. H., Scheuermann, B. W. & Cunningham, D. A. The effects of caffeine on the kinetics of O$_2$ uptake, CO$_2$ production and expiratory ventilation in humans during the on-transient of moderate and heavy intensity exercise. *Exp. Physiol.* **84**, 761–774 (1999).
32. Hughson, R. L. & Morrissey, M. Delayed kinetics of respiratory gas exchange in the transition from prior exercise. *J. Appl. Physiol.* **52**, 921–929 (1982).
33. Poole, D. C. & Jones, A. M. Oxygen uptake kinetics. *Compr. Physiol.* **2**, 933–996 (2011).
34. Wahid, A. et al. Quantifying the association between physical activity and cardiovascular disease and diabetes: a systematic review and meta-analysis. *J. Am. Heart Assoc.* **5**, e002495 (2016).
35. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *J. Multidiscip. Healthc.* **9**, 211–217 (2016).
36. Hedge, E. T. & Hughson, R. L. 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. Hughson, R. L., Winter, D. A., Patla, A. E., Swanson, G. D. & Cuervo, L. A. Investigation of V̇O$_2$ kinetics in humans withpseudorandom binary sequence work rate change. *J. Appl. Physiol.* **68**, 796–801 (1990).
38. Beaver, W. L., Wasserman, K. & Whipp, B. J. A new method for detecting anaerobic threshold by gas exchange. *J. Appl. Physiol.* **60**, 2020–2027 (1986).
39. Keir, D. A., Paterson, D. H., Kowalchuk, J. M. & Murias, J. M. Using ramp-incremental V̇O$_2$ responses for constant-intensity exercise selection. *Appl. Physiol. Nutr. Metab.* **43**, 882–892 (2018).
40. Beltrame, T. & Hughson, R. L. Mean normalized gain: a new method for the assessment of the aerobic system temporal dynamics during randomly varying exercise in humans. *Front. Physiol.* **8**, 504 (2017).
41. Beltrame, T. & Hughson, R. L. Linear and non-linear contributions to oxygen transport and utilization during moderate random exercise in humans. *Exp. Physiol.* **102**, 563–577 (2017).
42. Bennett, F. M., Reischl, P., Grodins, F. S., Yamashiro, S. M. & Fordyce, W. E. Dynamics of ventilatory response to exercise in humans. *J. Appl. Physiol.* **51**, 194–203 (1981).
43. Villar, R., Beltrame, T. & Hughson, R. L. Validation of the hexoskin wearable vest during lying, sitting, standing, and walking activities. *Appl. Physiol. Nutr. Metab.* **40**, 1019–1024 (2015).
44. Whipp, B. J., Ward, S. A. & Rossiter, H. B. Pulmonary O$_2$ uptake during exercise: conflating muscular and cardiovascular responses. *Med. Sci. Sports Exerc.* **37**, 1574–1585 (2005).
45. Koppo, K., Bouckaert, J. & Jones, A. M. Effects of training status and exercise intensity on phase II VO$_2$ kinetics. *Med. Sci. Sports Exerc.* **36**, 225–232 (2004).
46. McNarry, M. A., Kingsley, M. I. C. & Lewis, M. J. Influence of exercise intensity on pulmonary oxygen uptake kinetics in young and late middle-aged adults. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **303**, R791–R798 (2012).
47. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. In *Proc. International Conference on Learning Representations* (ICLR, 2016).
48. van den Oord, A. et al. WaveNet: a generative model for raw audio. Preprint at https://arxiv.org/abs/1609.03499 (2016).
49. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
50. Bland, J. M. & Altman, D. G. Agreement between methods of measurement with multiple observations per individual. *J. Biopharmaceut. Stat.* **17**, 571–582 (2007).

AUTHOR CONTRIBUTIONS

R.A., E.T.H., and R.L.H. designed the study. E.T.H. collected the data. R.A. developed the machine learning code and related analyses. E.T.H. conducted the statistical analyses. R.A. and E.T.H. independently validated the results. R.A. and E.T.H. wrote the first draft of the manuscript. All authors revised and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-021-00531-3.

Correspondence and requests for materials should be addressed to Robert Amelard.

Reprints and permission information is available at http://www.nature.com/reprints

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