Received: 21 September 2021 | Revised: 12 February 2022 | Accepted: 8 March 2022

DOI: 10.1002/ail2.65

LETTER

WILEY

Deep learning to predict power output from respiratory inductive plethysmography data

Erik Johannes B. L. G Husom[1] | **Pierre Bernabé**[2] | **Sagar Sen**[1]

[1]Sustainable Communication Technologies, SINTEF Digital, Oslo, Norway

[2]Department of Validation Intelligence for Autonomous Software Systems, Simula Research Laboratory, Oslo, Norway

Correspondence
Erik Johannes B. L. G Husom, Software and Service Innovation, SINTEF Digital, Oslo, Norway.
Email: erik.husom@sintef.no

Funding information
European Union's Horizon 2020 Research and Innovation programme, Grant/Award Number: 958357

Abstract

Power output is one of the most accurate methods for measuring exercise intensity during outdoor endurance sports, since it records the actual effect of the work performed by the muscles over time. However, power meters are expensive and are limited to activity forms where it is possible to embed sensors in the propulsion system such as in cycling. We investigate using breathing to estimate power output during exercise, in order to create a portable method for tracking physical effort that is universally applicable in many activity forms. Breathing can be quantified through respiratory inductive plethysmography (RIP), which entails recording the movement of the rib cage and abdomen caused by breathing, and it enables us to have a portable, non-invasive device for measuring breathing. RIP signals, heart rate and power output were recorded during a N-of-1 study of a person performing a set of workouts on a stationary bike. The recorded data were used to build predictive models through deep learning algorithms. A convolutional neural network (CNN) trained on features derived from RIP signals and heart rate obtained a mean absolute percentage error (MAPE) of 0.20 (ie, 20% average error). The model showed promising capability of estimating correct power levels and reactivity to changes in power output, but the accuracy is significantly lower than that of cycling power meters.

KEYWORDS

breathing, deep learning, machine learning, power estimation, respiratory inductive plethysmography

1 | INTRODUCTION

The modern practice of quantified training monitoring started in the 1930s with recordings of speed and heart rate, which now are two of the most common guiding measures of exercise intensity.[1] One of the challenges with heart rate as a measure of physical effort is *cardiovascular drift*, which means that the heart rate increases during prolonged exercise, even though the workload remains constant.[2] Speed as an effort gauge is also problematic, since it fails to give a fair representation of the intensity when we have external factors such as varying terrain, surface and weather. Power output is another exercise metric, which expresses the actual work performed by the muscles over time with the unit

watts. It may be regarded as one of the most direct methods of measuring physical effort during exercise.[3] Power meters are commonly used in cycling, and they usually consist of a set of electronic sensors called strain gauges. The deformation of these is proportional to the torque created by the cyclist.[4] However, it requires an embedded sensor propulsion system, which is infeasible in many common endurance sports such as running, rowing, cross-country skiing and swimming. Moreover, power meters are in general quite expensive.[5]

Is there a non-invasive and inexpensive alternative to measure power for any physical activity? This is the overarching question that intrigues us. We turn to an ubiquitous but largely under-explored variable in the context of general activity tracking in fitness and sports: *Breathing*. Breathing supplies the body with oxygen necessary to produce energy, and is thus fundamentally connected to any form of physical activity. It is typically studied using exercise spirometry where we measure the flow of air through a tube held in the mouth.[6,7] Exercise spirometry is however invasive and not easily portable. An alternative approach is to use a non-invasive, portable method of recording breathing called *respiratory inductive plethysmopgrahy* (RIP).[8] This method measures the expansion and contraction of the chest and abdomen, using straps or wires worn around the upper body. Our motivation to use breathing variables stems from previous work where it is shown to be a possible metric for activity tracking,[9] and the review by Gastinger et al[10] concludes that devices such as RIP sensors increase the potential of ventilation as a measure of energy expenditure. The objective of our research is to estimate physical effort from a person's breathing, or more specifically: Estimate the power output in watts during exercise based on RIP signals. We choose power output as our estimation target due to its direct relationship to the actual physical effort. Cycling is an activity where it is trivial to measure the power output using a power meter.[5,11] We also investigate whether heart rate as an additional input variable can improve the performance of our predictive models. In order to perform power estimation, we use machine learning. In this paper, we are dealing with time series data from RIP signals and a power meter, and because of this we focus on two types of neural networks that can handle data with local dependencies across multiple samples: Convolutional neural networks (CNNs)[12] and recurrent neural networks.[13] In addition, we use a simple fully-connected dense neural network (DNN) as a baseline model for comparison.

We validate our approach to predict power using an N-of-1 study where the author (Husom), an active adult male of age 25, performed 21 workouts on a stationary bike (Concept2 BikeErg) to obtain real-time power and two RIP sensors worn around the ribcage and abdomen to obtain breathing data. We use a trial-and-error process based on visual inspection and mean squared error to derive promising neural network configurations to evaluate and compare prediction of power from breathing. After selecting optimal architectures, we investigate combinations of input features in the breathing signals from raw data to features such as slope and gradient of the breathing pattern and the heart rate. Finally, we test the performance of our neural networks on unseen breathing and heart rate data from the same subject and use the mean absolute percentage error and visual inspection to evaluate model performance.

The rest of the paper is organized as follows. In Section 2, we discuss related work. We present our approach to predict power from breathing in Section 3. We validate our approach using an N-of-1 study to predict power output during cycling and discuss our results in Section 4. We present our conclusions in Section 5.

2 | RELATED WORK

Gastinger et al[10] have done a thorough review of the case for using ventilation as a representative measure of energy expenditure. The study states that while there exists a range of portable devices that estimate energy expenditure, such as pedometers, accelerometers and heart rate monitors, these show significant variability when it comes to accuracy of their estimations. The review states that sensors based on RIP estimate energy expenditure reasonably well. This is our starting point to ask whether RIP data can be used to predict other metrics in energy expenditure such as power in watts.

Indirect estimation of power output during exercise has been previously attempted for cycling. Costa et al[14] studied in 2017 the validity of a device called PowerCal, which estimates the power output during cycling based on heart rate. The PowerCal device uses only heart rate to calculate the power output, based on a secret embedded algorithm. The details of the algorithm are not revealed in the paper. A cycle ergometer has been used to measure the actual power output during a set of workouts, and compared with the power estimation of the PowerCal. The mean power output over each kilometre during the workouts was from 5.8% to 23.4% larger on the ergometer compared to the PowerCal. The researchers concluded that the power estimation based on heart rate calculated by the PowerCal device was not reliable and under-estimated the actual power output. This result further strengthens our need to explore other physiological variables such as breathing to estimate power.

Hilmkil et al[15] applied deep learning on data from cyclists. They used a long short-term memory (LSTM) neural network to train a model for predicting heart rate response of cyclists during workouts. A wide range of input variables was used: Speed (km/h), distance (km), power (watt), cadence (pedal strokes/minute), power/weight (watt/kg) and heart rate 30 seconds prior to the current time steps (beats per minute). The study concluded that the heart rate predicted by the trained model was very close to the true heart rate, and that deep learning algorithms are promising methods for applications on sports performance data.

The mean deviation of commercial cycling power meters has been reported to be $-0.9 \pm 3.2\%$ (mean $\pm$ SD) by Maier et al[16] 11% of the 54 power meters they tested had deviated more than $\pm 5\%$.

We have previously investigated how minute ventilation in litres per minute can be estimated from ribcage RIP signals using a machine learning approach called DeepVentilation.[17] Deep learning presents the possibility to take into account individual differences such as age, gender weight, height, variation in the placement of sensors across several different people. It offers the possibility to improve prediction and specificity over time with more availability of data. This is something standard time series prediction models such as ARIMA[18] cannot address very well. Our work aims to use a similar approach to predict power in watts from RIP signals from both the ribcage and abdomen. To the best of our knowledge, no research has attempted to estimate the power output from RIP signals, which is the contribution of this paper.

3 | METHODS

The following section describes the methodology followed in this paper, with the goal of estimating power output from RIP signals. In Section 3.1 we present our methods for data acquisition. In Section 3.2 we discuss data preparation, where we extract robust features from the breathing signals. Section 3.3 presents our approach to building our predictive models.

3.1 | Data acquisition

The N-of-1 data set was collected from an active adult male of age 25, who performed 21 workouts on a stationary bike. Four different variables were recorded during these workouts:

- RIP from rib cage, unit: millivolt (mV). Sampled at 10 Hz.
- RIP from abdomen, unit: millivolt (mV). Sampled at 10 Hz.
- Heart rate, unit: beats per minute (bpm). Sampled at 1 Hz.
- Power output, unit: Watt (W). Sampled at 1 Hz.

The RIP data were collected using a product called SweetZpot Flow (www.sweetzpot.com), shown in Figure 1A. During data collection the subject wore two of these sensors. One was placed such that the belt was lying directly below the

(A) SweetZpot Flow RIP sensor

(B) Wahoo Tickr heart rate monitor

(C) Concept2 BikeErg stationary bike

FIGURE 1 Sensors used for data acquisition

TABLE 1 The three main categories of workouts performed during data collection

Workout category	Description	Example
Steady-state effort	Cycling at a constant intensity.	45 minutes easy cycling at 180 W.
High-intensity intervals	Alternating between working at a high and low intensity.	4 × 4 minute at 350 W, with a break of 2 minutes.
Ramp structure	Increasing the intensity step-wise.	Starting at 100 W, increase by 100 W every 5. min up to 500 W.

TABLE 2 Features engineered from the raw RIP and heart rate signals

Feature name	Unit
Range of RIP rib cage signal	mV
Range of RIP abdomen signal	mV
Respiratory rate calculated from the RIP rib cage signal	breaths/min
Respiratory rate calculated from the RIP abdomen signal	breaths/min
Gradient of RIP from rib cage	mV/s
Gradient of RIP from abdomen	mV/s
Sine of the slope of RIP from rib cage	-
Cosine of the slope of RIP from rib cage	-
Sine of the slope of RIP from abdomen	-
Cosine of the slope of RIP from abdomen	-
Sine of the slope of heart rate	-
Cosine of the slope of heart rate	-

pectoral muscles, while the other was placed over the navel, with the belt around the waist. Movements of the chest or abdomen will give increasing and decreasing strain on the belt, and this force is measured by a strain-gauge in the sensor itself. The heart rate data were collected using a Wahoo Tickr heart rate monitor, shown in Figure 1B. It is fastened to a chest strap that the subject wears around the chest below the pectoral muscles, just below the upper RIP sensor described above. Power values were recorded from a stationary bike called Concept2 BikeErg, shown in Figure 1C. The bike pedals drive a chain connected to a flywheel, which is used to calculate the work done by the cyclist, measured in watts. Data from all sensors were sent to a computer using the Bluetooth 4.0 protocol.

The workouts performed during data collection can be divided into three categories: steady-state effort, high-intensity intervals and ramp structure. Table 1 contains the descriptions of the three different categories, and includes an example workout for each of them. The main idea behind these workouts was to imitate typical workout routines with a mixture of intensities and workout structures, in order to make a robust data set with a wide power distribution. One of the aims of this project was to create a model that can accurately estimate power output for arbitrary workout intensities, and that is why a mixture of workouts was chosen instead of following a specific protocol. Data were also collected when the subject was *sedentary*, in which case the power values were manually set to zero for all time steps.

3.2 | Data preparation

The raw data need pre-processing before it is used to create predictive models in order to extract the most pertinent information from it. We derive a set of features on the input data (breathing and heart rate) as presented in Table 2. These features were designed based on several hypotheses about what features might be used to estimate the power output. The range of the RIP values is a feature that can express the depth of our breath, and is calculated by taking the difference between the maximum and minimum value of the RIP signal in the last few seconds of the current time step. The respiratory rate (RR) was found by finding the peaks of the signal, and looking at the distance between the peaks.

TABLE 3 Metrics for evaluating model performance

Metric	Definition
Mean squared error (MSE)	$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$
Coefficient of determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
Mean absolute percentage error (MAPE)	$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \widehat{y}_i}{y_i}\right\rvert$

Note: The variable $\widehat{y}$ represents the predicted output values and y is the actual output values. The index i runs from 1 and up to n, the number of samples.

The gradient $\widehat{f}$ and the slope θ of the RIP signal f can express how fast we draw our breath, and these were calculated as follows:

$$\widehat{f}_t = \frac{f_{t+1} - f_{t-1}}{2d}, \tag{1}$$

$$\theta_t = \arctan\left(\frac{f_t - f_{t-1}}{d}\right). \tag{2}$$

Here t represents the index of the current time step, and d is the distance between each data sample of f. While the gradient and the slope both give information on the rate of change in the signal, the slope (in radians) is a representation that is bounded between 0 and 2π. Furthermore, we choose to represent the slope θ as a pair of values: $\sin(\theta)$ and $\cos(\theta)$. These two features give a unique representation of the slope, and have the added benefit that the cosine-encoding gives the same values for upwards and downwards slopes with the same absolute angle. We experimented with the same engineered features for both RIP and heart rate, but have only included the slope features of heart rate in Table 2 and in our results, as these were the only ones who affected the model performance significantly.

The data were divided into overlapping sequences of a certain size, which we denote the *history size*, or s. Each sequence was matched to the power value recorded at the last time step of the sequence. In practice, this would mean that the last few seconds of recorded input data will be used to estimate the current power output. Consecutive sequences overlapped with $s-1$ samples (ie, all but one sample), in order to get the maximum number of input sequences from the data set. We also scaled all input features to contain only values between 0 and 1.

3.3 | Building predictive models

We use deep learning, and more specifically neural networks, to create predictive models for power output estimation. In this paper, we deal with time series data, and several deep learning methods have the advantage of being designed to process input data where there are relationships between individual observations. Deep learning methods also have the advantage of improving with increased amounts of data[19] from people with different heights, genders, weights, ages, positioning of the sensors on the body, and physical health conditions.

Convolutional neural networks (CNNs)[12] is one of the neural network types we use for creating our models. CNNs are originally designed for image processing, but they can be adapted to work for time series data by applying convolution filters only along one axis of the input, namely the temporal axis. We also use long short-term memory networks (LSTMs),[20] which is a type of recurrent neural network (RNN). LSTMs are designed specifically to handle data sequentially, and is therefore a good fit for processing time series data. The network can "remember" information from previous steps, which can be used to identify temporally dependent features in an input sequence. Furthermore, we test our approach with a fully-connected or dense neural network (DNN). The DNN models will serve as a baseline for comparison against the other types of neural networks.

When building our predictive models, we divided the data into three categories: Training, validation and test data. Out of the total 21 workouts in the data set, 4 were separated out as a test data set, on which the final models would be evaluated. After having partitioned the test set, the remaining workouts were divided into a training and a validation set, with a split of respectively 70% and 30%. The training set was used to do the actual model fitting, while the

validation set was used to tune the architecture of the neural networks and all hyper-parameters defining the models. The training, validation and test data sets had a near equal distribution of the three different workout types. The training set was further split during the training, where 25% of the data were used to evaluate the model after each epoch. This provides information on the performance in relation to number of epochs, and lets us know when the model starts over-fitting to the training data. When the evaluation after each epoch starts to give deteriorating results, training is stopped.

The metrics we use for evaluating model performance are presented in Table 3. The MSE is one of the most common evaluation criteria for regression problems. It is used during model training because it decreases as the difference between predictions and the ground truth becomes smaller, and enables the algorithm to adjust the model parameters. Additionally, we use R^2 score, which expresses the ratio between the variance explained by the model and the total variance, since it can be interpreted independently of the scale of the input variables. A score of 1 means a perfect fit. We also use MAPE, which is a fraction that gives the average error in the model's predictions. It gives the most intuitive measure of a model's overall performance, but it suffers among other things from the limitation that it is heavily influenced by the error when the true values y are small. It is therefore important to compare multiple metrics when evaluating models.

4 | RESULTS AND DISCUSSION

We validate our method by addressing three research questions, presented below:

RQ1. *What neural network architectures are promising for estimating power output from breathing?*

We use a trial-and-error process to address RQ1. The process of choosing a fitting architecture for neural networks can be very challenging, primarily because the possible combinations and configurations are infinite. While there are some frameworks such as AutoML[21] to automate this process, we found it to be easier to control and supervise the process by manually crafting the network architectures. We started with networks with few layers and nodes/units, and expanded them until we reached architectures that yielded promising results. When testing different configurations, we compute the MSE on our validation set, which was not directly used in training. We monitor how the validation MSE behaved as a function of different numbers of layers and nodes, and how it is affected by different choices of activation functions and filter sizes (the latter only applicable for CNNs). Our trial-and-error process is also guided by a bird's eye visual inspection of how well the prediction of power matches the ground truth power for steady rate cycling and for rare cases where there are sudden increase in power. The reactivity of the model to changes in power application was an important factor in selecting optimal architectures along with its performance during steady-state cycling. The optimal configurations of the networks are also affected by the size of the input sequence, and because of this we also tested all configurations with different history sizes. Our search for effective neural network architectures yielded the following promising configurations for the three different network types:

- DNN: three hidden layers with 256, 128 and 64 nodes, respectively. Rectified Linear Unit (ReLU) activation in each layer.
- CNN: four conv. layers with 64 filters and ReLU activation, followed by 1 hidden layer with 32 nodes and ReLU activation.
- LSTM: one LSTM layer with 100 hidden units and ReLU activation.

RQ2. *What features of breathing data are useful for estimating power output from breathing?*

In Table 4, we have defined 11 different sets of input variables/features, together with the MSE and R^2 score for our three network architectures. We want to compare these in order to answer research question RQ2. Feature set 11 only contains heart rate as an input feature, and this can be used as a baseline to indicate the usefulness of RIP signals compared with heart rate as a predictor for power output. We have used bold typeface to highlight the best performing models for each type of neural network. In all of the models in Table 4, a history size of 100 time steps, that is, 10 seconds, was used. While we focused on engineering features that we can link to an intuitive understanding of how our breathing works, these are not necessarily the best features for a machine learning algorithm. The trade-off between

TABLE 4 Overview over model performance for various feature sets when tested on the validation data set

	Features							Performance					
	RIP rib cage and abdomen					Heart rate		DNN		CNN		LSTM	
Feat. Set	Raw	Range	RR	Gradient	Slope	Raw	Slope	MSE	R^2	MSE	R^2	MSE	R^2
1	x							0.021	−0.26	0.017	0.00	0.019	0.00
2	x					x		0.008	0.50	0.011	0.35	0.011	0.34
3		x				x		0.007	0.60	0.009	0.49	**0.006**	**0.64**
4			x			x		0.010	0.38	0.017	0.00	0.010	0.37
5				x	x	x		0.008	0.51	0.007	0.62	0.009	0.47
6				x	x	x	x	0.007	0.60	**0.006**	**0.66**	0.008	0.55
7		x	x	x	x			0.018	−0.08	0.017	0.00	0.016	0.06
8		x	x	x	x	x		0.011	0.35	0.017	0.00	0.007	0.56
9				x	x			0.026	0.28	0.010	0.43	0.016	0.07
10					x			0.016	0.07	0.009	0.45	0.014	0.20
11						x		**0.006**	**0.64**	0.007	0.60	0.007	0.61

Bold indicates best performing models for each type of neural network.

predictive and descriptive accuracy is explained in detail by Murdoch et al.[22] The predictive accuracy is defined as how well a model performs in terms of the error metrics used to evaluate it. This is the performance of the model, or how well the model is able to estimate a desired function or mapping. The descriptive accuracy is how well a method of interpretation is able to describe the relationships that the model has learned. In this article, we focus on what Murdoch et al[22] define as *model-based interpretability*, which is creating predictive models that enable understanding of the relationships learned by the machine learning algorithm. In practice, this means that we prioritize simpler models, which are easier to interpret, rather than complex models with a large number of input features. Simpler models will also require less computing power and storage, which is an important factor if it should be used in a portable, resource-constrained devices.

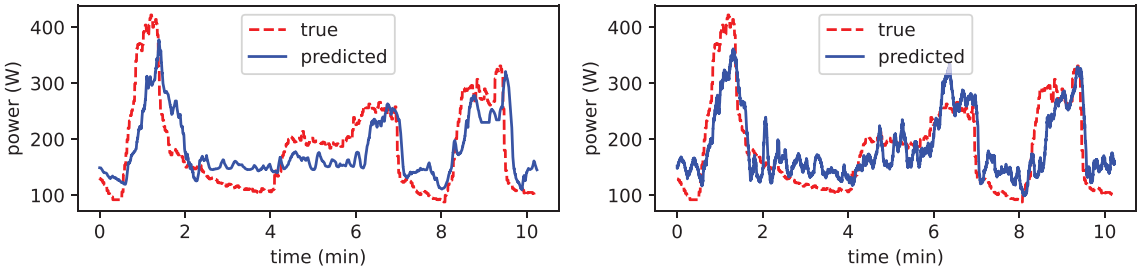
The error metrics in Table 4 let us see which feature sets work best with each type of neural network. The metrics are calculated from the validation data set, which we also used to tune hyper-parameters during the training process. The DNN obtains the highest performance when trained solely on heart rate, while the CNN and LSTM performs best on a combination of heart rate and RIP signals (feature set 6 and 3, respectively). Our analysis of the feature sets indicates that the range, the gradient and the slope are the most useful features of the RIP signals, and that these give a significant improvement over using only raw RIP data. We hypothesize that these features are most important because they express the depth of breath (range) and how fast the subject inhales and exhales (slope and gradient), since the exercise intensity will affect how deep and quickly a person inhales/exhales. The choice of high-level features gives us the opportunity to observe which characteristics contribute to higher performance, but the "black-box" nature of neural networks makes it challenging to interpret why the various architectures have different performance on the respective feature sets. Approaches in explainable AI for time series data such as LIME[23] could be used to investigate the influence of certain input features and output. However, this is outside the scope of this study. The history size, that is, the length of the input sequences, was also analysed based on the model performance on the validation data set. For the CNN and the LSTM network, a history size of 100 gave optimal results (for feature sets 6 and 3 respectively). The DNN was tested with feature set 3, which was one of the best performing DNN model that included RIP signals, and in this case a history size of 120 gave the best result.

RQ3. *What is the effectiveness of estimating power output from unseen breathing data?*

We evaluated our models on the unseen test set, and the results are shown in Table 5. We evaluated the three best performing models for each type of neural network, in addition to the best model that was trained only in RIP signals (CNN on feature set 10).

TABLE 5 Performance metrics for final models when evaluated on the test data set

Network	Feature set	MSE	R^2	MAPE
DNN	11	0.006	0.43	0.22
CNN	6	0.004	0.56	0.20
LSTM	3	0.006	0.35	0.22
CNN	10	0.005	0.50	0.24



(A) Results using DNN architecture on feature set 11. **(B)** Results using CNN architecture on feature set 6.

FIGURE 2 Prediction of power output on an interval workout from the test set. The red stapled curve shows the true values, and the blue solid curve shows the predictions of each respective model

If we look at the error metrics in Table 5, the CNN model trained on feature set 6 gives the best performance, with an MAPE of 0.20. The difference in performance of the five models in Table 5 is however quite small, with the highest MAPE being 0.24 for the CNN trained on feature set 10. The random initialization of weights in neural networks affects the performance to some degree. Our assessment of the results is that the differences in performance are not significant enough to judge one of the network architectures or feature set to be clearly superior to the others, at least not exclusively based on error metrics.

In Figure 2, we show predicted vs true values across a 10 minutes excerpt from a workout for two of our models. The DNN model trained solely on heart rate, in Figure 2A, has less noise in its predictions compared with the model with the lowest MAPE, which is the CNN model trained on feature set 6, in Figure 2B. The CNN model is however more reactive to sharp changes decreases in power output, exemplified at the 1.75 and 9.75 minute mark. This could be explained by breathing being more reactive to decreases in intensity compared with heart rate, which may lag behind. This feature contributed by breathing can help address the problem with cardiovascular drift by making the power estimation more reactive to changes in intensity. Furthermore, less data for high wattage (very high-intensity workout) can explain why both models fail to predict the highest peak at above 400 W.

4.1 | Limitations and threats to validity

One of the limitations of this project is the small sample size of 21 workouts (totalling approximately 10 hours of recorded data of the data set). In the field of deep learning, large amounts of diverse data are fundamental for the algorithms to generalize well.[19] Another limitation of the study is that there has been no measurement of the computational cost of model inference. If a model is to be used for real-time estimations during cycling, the inference must necessarily happen on a small, portable device such as a smartphone, which puts constraints on computational power and power usage.

Threats to external validity: The main threat to external validity of our project is that we have used a single subject for data acquisition. The global pandemic of COVID-19 put limits on our ability to arrange data collection from multiple subjects because of infection control considerations. While the N-of-1 method has reduced the total workload by eliminating the need for administrating experiments with multiple subjects, it has also placed limits on our investigation into RIP signals as predictors for physical effort. We are unable to discuss whether our chosen methods, especially the engineered features, will give similar results when applied to data from other subjects. One of our outlooks towards

an N-of-1 study has been to verify if power prediction can be personalized to an individual who can have several personal variations in usage of the sensor such as body posture, personal anthropometry, placement of sensor on the body, and usage environment (room temperature, altitude, etc.). Another threat to external validity is that the data collection took place indoors on a stationary bike, and differs from cycling in an outdoor environment, which introduces factors such as weather, gear shifts and different movement patterns due to bike handling.

Threats to internal validity: Regarding threats to internal validity, it seems beyond doubt that breathing is related to the exercise intensity during a workout, which is also the case for heart rate. However, a possible problem is the cause-and-effect relationship between our input and output variables. Our models estimate the power output of the current time step based on the last few seconds of input data, because we aimed at producing models that can provide real-time power estimation during a workout. In reality, when a person increases the power output during a workout, the change in breathing will happen as a response to the increase in power output because of increased need for oxygen. Experiments using target values of earlier time steps did not improve the model performance, and therefore we stayed with our initial idea of real-time estimation. Nevertheless, this problem deserves further investigation, since both breathing and heart rate undoubtedly are variables that respond to exercise intensity, and not the other way around. Furthermore, we have not considered physiological effects that may appear when exercising for a longer period of time (several hours), and whether fatigue might change the relationship between breathing, heart rate and effort.

5 | CONCLUSION

We have studied how we can use deep learning to create personalized models to estimate power output during exercise from breathing data. Respiratory inductive plethysmography (RIP) signals, heart rate and power output were recorded while a subject was performing a set of workouts on a stationary bike, and these data were used to train neural networks to estimate power output during cycling. The models were trained using input features based on either RIP signals, heart rate, or a combination of the two.

We investigated the performance of three different types of neural network: Dense neural networks (DNN), convolutional neural networks (CNN) and long short-term memory (LSTM) networks. Our research into the effectiveness of estimating power output from breathing has shown that a CNN trained on features derived from both RIP signals and heart rate gives a mean absolute percentage error (MAPE) of 0.20 when evaluated on our test data set.

In comparison, a DNN trained on raw heart rate data resulted in an MAPE of 0.24. These error rates are significantly higher than that of cycling power meters, which has been found to deviate with $-0.9 \pm 3.2\%$.[16] We conclude that although our results are promising, further research is needed to reach acceptable accuracy when predicting power output from breathing and heart rate. A large and balanced data set, especially for higher intensity, in addition to research into a wider set of input features may contribute to improved performance when using deep learning models to predict power.

A continuation of this research is to gather a diverse data set from subjects of various ages, genders, heights, weights and physical fitness levels. Applying our methods to such data might give insights to whether generalized, non-personalized models are able to give accurate power output estimation. Potentially, the prediction accuracy may be greater for personalized models, but an option is to use personal data to "tune" a pre-trained generalized model by using transfer learning,[24] and avoid the need to train personalized models from scratch. This research can also be extended to other sports, for example, indoor rowing or other activities that take place on exercise ergometers, where it is trivial to record power output.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 958357 (InterQ).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://github.com/ejhusom/respiratory-inductive-plethysmography-for-power-prediction-data.

ORCID

Erik Johannes B. L. G Husom https://orcid.org/0000-0002-9325-1604

REFERENCES

1. Foster C, Rodriguez-Marroyo JA, De Koning JJ. Monitoring training loads: the past, the present, and the future. *Int J Sports Physiol Perform*. 2017;12(s2):S2-2-S2-8.
2. Coyle EF, Gonzalez-Alonso J. Cardiovascular drift during prolonged exercise: new perspectives. *Exerc Sport Sci Rev*. 2001;29(2):88-92.
3. Jeukendrup A, Diemen AV. Heart rate monitoring during training and competition in cyclists. *J Sports Sci*. 1998;16(sup1):91-99. doi:10.1080/026404198366722
4. Passfield L, Hopker JG, Jobson S, Friel D, Zabala M. Knowledge is power: issues of measuring training and performance in cycling. *J Sports Sci*. 2017;35(14):1426-1434.
5. Novak AR, Dascombe BJ. Agreement of power measures between Garmin vector and SRM cycle power meters. *Meas Phys Educ Exerc Sci*. 2016;20(3):167-172.
6. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *Eur Respir J*. 2005;26(2):319-338.
7. Askanazi J, Silverberg P, Foster R, Hyman A, Milic-Emili J, Kinney J. Effects of respiratory apparatus on breathing pattern. *J Appl Physiol*. 1980;48(4):577-580.
8. Carry PY, Baconnier P, Eberhard A, Cotte P, Benchetrit G. Evaluation of respiratory inductive plethysmography: accuracy for analysis of respiratory waveforms. *Chest*. 1997;111(4):910-915.
9. Gastinger S, Sorel A, Nicolas G, Gratas-Delamarche A, Prioux J. A comparison between ventilation and heart rate as indicator of oxygen uptake during different intensities of exercise. *J Sports Sci Med*. 2010;9(1):110-118.
10. Gastinger S, Donnelly A, Dumond R, Prioux J. A review of the evidence for the use of ventilation as a surrogate measure of energy expenditure. *J Parenter Enter Nutr*. 2014;38(8):926-938.
11. Turner KJ, Rice AJ. Physiological responses on the concept II BikeErg and concept II RowErg in well-trained male rowers. *Int J Sports Sci Coach*. 2021;16(3):741-748.
12. LeCun Y. Generalization and network design strategies. Connectionism in Perspective. 1989;19;143-155.
13. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533-536.
14. Costa VP, Guglielmo LG, Paton CD. Validity and reliability of the PowerCal device for estimating power output during cycling time trials. *J Strength Cond Res*. 2017;31(1):227-232.
15. Hilmkil A, Ivarsson O, Johansson M, Kuylenstierna D, Erp vT. Towards Machine Learning on data from Professional Cyclists. https://arxiv.org/abs/1808.00198. 2018.
16. Maier T, Schmid L, Müller B, Steiner T, Wehrlin JP. Accuracy of cycling power meters against a mathematical model of treadmill cycling. *Int J Sports Med*. 2017;38(06):456-461.
17. Sen S, Bernabé P, Husom EJB. DeepVentilation: Learning to Predict Physical Effort from Breathing. 2020.
18. Liu C, Hoi SC, Zhao P, Sun J. Online ARIMA Algorithms for Time Series Prediction. 2016.
19. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intell Syst*. 2009;24(2):8-12.
20. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780.
21. He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. *Knowl-Based Syst*. 2021;212:106622.
22. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci*. 2019;116(44):22071-22080.
23. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. 2016;1135-1144.
24. Bozinovski S. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*. 2020;44(3):291-302.

How to cite this article: Husom EJBLG, Bernabé P, Sen S. Deep learning to predict power output from respiratory inductive plethysmography data. *Applied AI Letters*. 2022;e65. doi:10.1002/ail2.65