A Deep Learning based Architecture for Energy Expenditure Estimation from Multiple Sensors

Group 5B: Sarthak Bisht, Soumith Udatha, Zeda Xu, Zhenyu Lu

Abstract

This paper investigates the use of deep learning techniques for predicting energy expenditure in wearable health technologies. Specifically, we propose a deep learning based trimodal architecture that supports multiple basic deep learning models. Our results demonstrate that our LSTM-based models with separate encoders outperforms the Ingraham baseline, when using a comparable number of sensors, ~20% and 13% when using just Heart Rate sensor and accelerometer respectively.To validate our approach, we test our model on the Empatica EmbracePlus watch for walking and running, and report results that align with the expected trends. Our findings suggest that deep learning techniques hold promise for more accurate and efficient prediction of energy expenditure in wearable health technologies.

Introduction

Energy expenditure measurements are widely used for athletes to achieve their optimal performance during their resistance training. As the public's fitness awareness increases in recent years, Energy expenditure measurements have become a need for everyday fitness goals in our daily life. Traditional precise methods of measuring EE include double label water and indirect calorimetry [1]. Other activity based assessments, like combination between physical activity and questionnaires [2], overestimate EE to a large extent. Other than expense, the aforementioned traditional methods of energy estimation are not suitable for pervasive monitoring of physical activities due to the following limitations -

- Lack of portability and practicality of room calorimeters
- Impracticality for free-living conditions of indirect mask-based calorimeters
- Lack of temporal & activity-specific granularity of measurements in doubly labeled water

This drives the search for more convenient and inexpensive wearable sensors that can match the accuracy of the traditional methods while being practical for mass-market deployment. The early attempts in this space started by attaching a single accelerometer to the user's body and running its readings through a single linear regression model. Crouter et al. [3] improved upon this by using different regression models for different activities types, namely - sedentary, ambulatory, or lifestyle. Subsequent literature focused upon improving activity recognition [4] and leveraging multiple accelerometers and other sensors such as heart-rate monitor, skin and near-body temperature sensors, breath-rate sensor, photoplethysmography, galvanic skin response, etc.

Cvetković et al. [5] concluded that for light activities, EE estimation should only utilize acceleration data as these are accompanied by normal heart rate and temperature readings. Only moderate and vigorous activities benefit from additional sensor information such as HR and near body temperature, especially if the activity has a large range of possible EE values. The

contribution of skin temperature and galvanic skin response does not improve EE estimates probably due to its high correlation with heart rate. Álvarez-García et al. [6] confirmed this conclusion in his recent survey that accelerometer and heart rate are the most important sensor modalities for EE and suggested using scaledHR technique of Munguia-Tapia [7].

The existing research tends to use simple models (such as linear regression or shallow neural networks) for predicting the energy expenditure. Slade et al. [8] demonstrate that neural networks can outperform simple linear regression, and this inspires us to employ more complex deep learning algorithms. In addition, previous work collects their data from experiment subjects, and uses their own data alone for analysis and modeling. However, we have yet to see the effort of merging heterogeneous datasets together.

In this study, we investigate whether utilizing a more complicated deep learning model and a combined dataset can achieve better energy expenditure estimation performance, and furthermore, we hypothesized that heart rate and accelerometer data all play an important role on EE estimation. To test our hypothesis, we propose an architecture that takes three sensor inputs, heart rate, accelerator data on the wrist and EDA to perform EE estimation and such an architecture can support multiple basic deep learning models. Leave-One-Out cross subject approach is used to train and validate the architecture proposed and we also generalize the model to qualitatively test on data collected from Empatica Watch worn by human subjects.

Method

There are three datasets available that could be used to build the EE estimation model. The Ingraham dataset [9] is collected from 10 participants doing 5 different activities for around 160 minutes. The Slade dataset [8] is similarly structured, capturing 24 participants doing 9 different activities for around 50 minutes. The JSI dataset [6], on the other hand, has labeled the different scenarios for all different activities. It collected 10 participants doing 19 different activities for around 180 minutes. The Slade dataset and the JSI dataset have Energy Expenditure data in the units of METs, and for the Ingraham dataset, the energy is measured using VO2 and VCO2. Using the Weir Formula, we obtain the energy expenditure in MET:

$$EE_{Ingraham} = [1440 \times (3.9\,VO_2 + 1.1\,VCO_2) \times 0.06] / [24 \times 70]$$

where VO2 and VCO2 is in mL/s, and EE_Ingraham is Energy Expenditure in MET.

Sensor types are vastly different across the three datasets (Table A3). The only shared sensor among all three datasets is heart rate but three sensor signals, which are Heart Rate, EDA, Accelerometer data on the right Wrist, are found across JSI and Ingraham and Empatica Watch. However, we decide to drop the Slade dataset in order to have at least three different raw sensor signals that we can process.

We adopt a Leave-One-Out Cross-Subject evaluation approach to train and validate our model. For the two dataset we use, there are 9 subjects in JSI and 10 subjects in Ingraham. To reduce the number of trials, we combine two datasets together in a way that one subject in JSI is

randomly paired with another subject in Ingraham, forming the new subject for the combined dataset. In such a way, the combined dataset has 9 new subjects in total.

In order to transform the raw data into data format expected by the model, some necessary data preprocessing steps are applied on the raw data. Firstly, all raw signals need to downsample to 1 Hz via interpolation or averaging. We categorize the activity labels in two datasets and reassign them in the same format - walking, running and other. For our human subject experiment, the physiological signals of the participants are captured by the Empatica watch. The Empatica watch only provides the heart rate per minute and we upsample the aggregated heart rate to 1 Hz by having all the per-second values the same for each minute. The sliding time window size is set to 4 with 1 frame overlapped in between two windows, which is good for real time monitoring as suggested by Álvarez-García et al. [6] Furthermore, we implement a subject-wise Min-Max normalization option to tackle individual differences.

Since there are three signal types available, we design a tri-modal architecture (Figure 1), that can perform EE estimation and activity classification. Overall, the model will have three loss terms and we use the most commonly used SimCLRv2 [10] as our contrastive loss function:

$$Loss_{Final} = \alpha * MSE\,(EE_{pred}, EE_{GT}) + \beta * SimCLR(Feature_{fused}) + \gamma * CrossEntroy(Label_{pred}, Label_{GT})$$

Where α, β, and γ are weight factors that determine the relative importance of each loss term, EE_{pred} is predicted EE, EE_{GT} is ground truth EE, $Feature_{fused}$ is the fused feature, $Label_{pred}$ is predicted activity label, $Label_{GT}$ is ground truth label

The encoder module is a neutral network that can extract important features from input data and the encoder choice is flexible, ranging from linear layer, multi-layer perceptron, LSTM and transformer. Through experiments, the LSTM encoder is proved to be slightly better, and therefore is deployed for our final model.

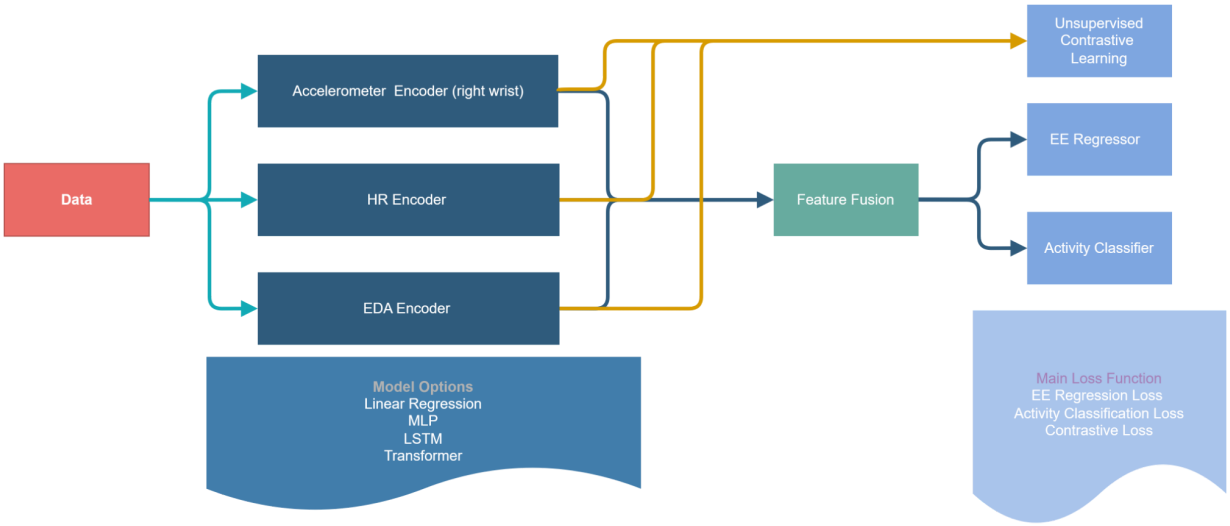

Figure 1. Full Model Pipeline. The backbone of the model is composed of three encoders, one for each sensor signal. The three processed signals are fed into three encoders respectively to generate their embedded features. While the three features are fused, unsupervised contrastive learning is applied on the features to force them to interact with each other and find the shared information across three modalities. The fused features are also fed into a linear regression layer to output the predicted energy consumption and, in parallel, a classification layer to classify the activity.

While our full model pipeline takes signals from all three sensors, the pipeline also supports solely feeding either single modal data into the model in order to study the impact of heart rate and accelerometer data on energy expenditure (Figure 2).

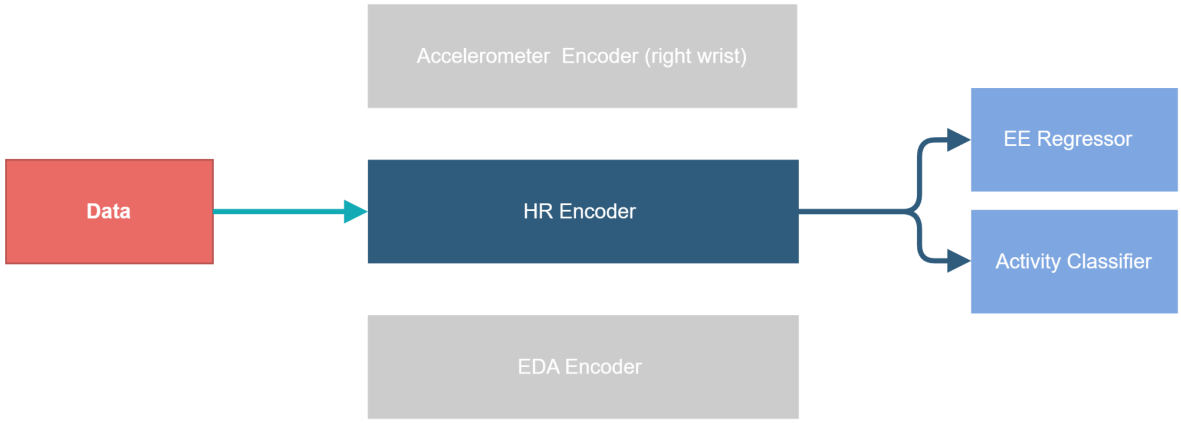

Figure 2. Model with HR input only. The model takes only one modality feature and passes the data into the corresponding encoder and then directly feeds into the last layers. Contrastive loss is not implemented in this model. The accelerometer-data-only model has the exact same structure except it only takes the accelerometer data.

Finally, we conduct experiments to validate our model by having participants perform activities while wearing the Empatica watch. The experiment participants are three of the team members. Each participant is asked to conduct three activities, resting, running and walking, each for 6 minutes. We choose 2 mile/hour, 6 mile/hour for walking and running conditions respectively and for the resting condition, the participants are free to do anything as long as they are sitting. The walking speed and running speed are kept consistent during the experiment. Walking and running experiments are conducted roughly during the same time: the participants first start walking, take 1 min gap and then continue with running. Resting conditions are tested at a different time when participants don't do any high intensity movement before the condition starts.

Results

We evaluate different versions of our model with LOOCV (Table 1) and show the full model with Min-Max Normalization achieves the best performance. Overall, both heart rate and accelerometer data are important features that contribute to the model performance, since solely using either one is able to estimate the EE, though less accurate. The model using heart rate data

only can achieve 1.45 of RMSE with 0.64 R2 and the model using Accelerometer data on the wrist can achieve 1.73 of RMSE with 0.5 R2.

Table 1. Summary of Average RMSE and R2 for all cross-subject evaluation trials. The best model evaluated on validation data from combined dataset is the full model pipeline with Min-Max Normalization and the full model denotes all three modalities (accelerometer data, Heart Rate, EDA) were taken as input. Single modality model results for acceleration and heart rate are all proved to be useful with considerable R2 score but the RMSE were undermined.

Model	Full + Min-Max	Full	ACC-only	HR-only	HR-only + Min-Max
RMSE(MET)	**1.24**	1.3	1.73	1.45	1.37
R2	**0.74**	0.72	0.5	0.64	0.68

Based on the full result for each cross-subject trial (Table A2),the best version among all 9 trials is the model trained for subject 2 with the lowest validation error(RMSE = 0.88, R^2 = 0.84) and therefore this version is the model used for our human subject experiment.

We feed the data collected from Empatica Watch and the estimated output from the model is basically as expected (Table 2). Since we don't have ground truth, there is no way that we can further evaluate the validity of the model but we do compare our estimation result with the average EE calculated from Ingraham.

Table 2. Average Energy Expenditure for three conditions. The average energy consumption for walking, running and resting are 3.29, 5.47, 1.37 MET respectively (averaged over three participants). In general, The energy consumption for running conditions is the highest, and matches the average value from Ingraham. The EE for walking is slightly higher than average for walking from Ingraham.

	Walking	Running	Resting
Average Energy Expenditure(MET)	3.29	5.47	1.37
Average EE from Ingraham(MET)	2.45	5.71	NA

We visualize the energy expenditure across time for one of our participants (Figure 3). Since no filtering has been applied to the raw signal, the noisy signal could possibly explain why the output EE is not very stable especially for walking conditions and therefore, the walking estimated EE (3.29) is higher than EE calculated from Ingraham dataset(2.45).

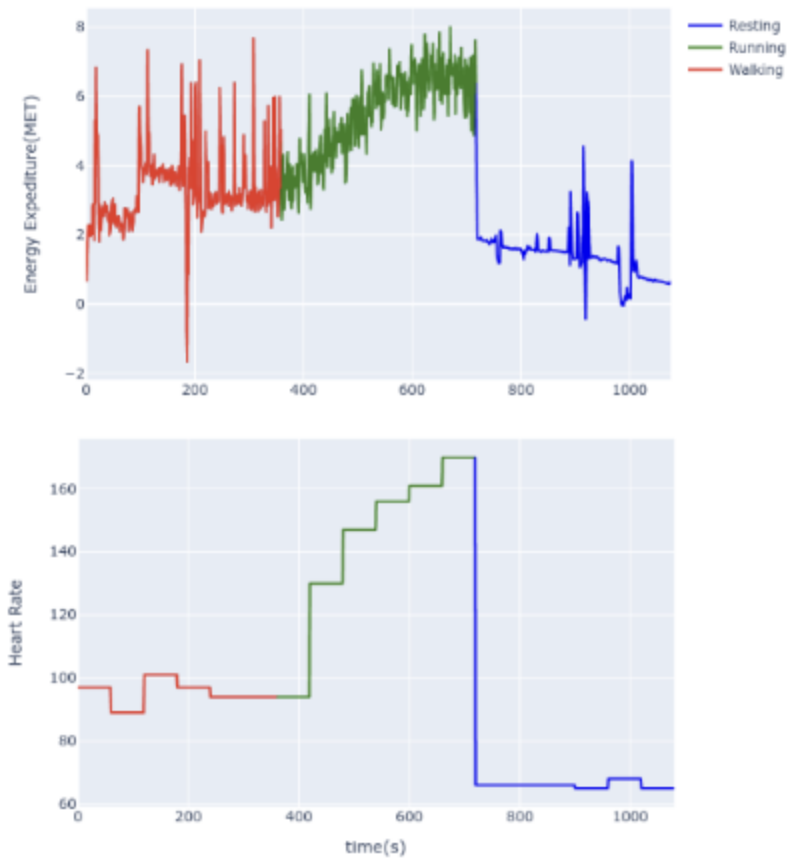

Figure 3. Visualization of Energy Expenditure vs Heart rate across three conditions for one participant. The energy expenditure is consistent with the trend of heart rate change, and the energy expenditure keeps increasing as the participant starts running. However, for the walking period, the Energy estimation is not quite stable and there is even one negative value, which could be possibly explained by the random hand movement during the walking.

Discussion

The goal of this project is to explore whether deep learning approaches can be used for energy expenditure estimation from wearable sensors. We design an architecture that can support multiple simple deep learning algorithms and has shown that our architecture can successfully estimate energy consumption with a considerably small error.

Additionally, we also show that both heart rate and accelerometer data play an important role on EE estimation (Figure 4). Heart rate, as a biomarker that doesn't fluctuate heavily in a short time, is good at capturing the global level of energy expenditure, but it might have a hard time to do a fine estimation locally. On the other hand, accelerometer data is more sensitive to body movement and can capture the local fluctuation better(indicator of motion intensity). It might not work very well solely in terms of EE estimation but it can make up for a deficiency from information that can be obtained from heart rate.

Figure 4. Visualization of Linear regression Estimation with acceleration(Top)/heart rate(Bottom) only. With heart rate as the only input signal, the estimation of EE(red) is still able to match the trend of ground truth(blue). However, if the model only takes acceleration on the wrist only, the estimation performance is not as good as heart rate. As the dataset also included a reference estimation from Bodymedia (green) [1], another device designed for EE estimation, we also plot it for reference.

We compare our model performance with Ingraham et al's approach [9] which simply uses a Linear Regression algorithm on the whole input signal (Figure 5). With minute ventilation as the only predictor, Ingraham records a RMSE of 1.07 MET, and with all eight modalities used to train the model, the RMSE score can achieve 0.89 MET. In this sense, we still do not exceed their reported performance. However, minute ventilation is a biomarker that is hard to collect unobtrusively in our daily life, and is not used as the input to our model. Our model outperforms their result if using either heart rate or accelerometer data at the right wrist only.

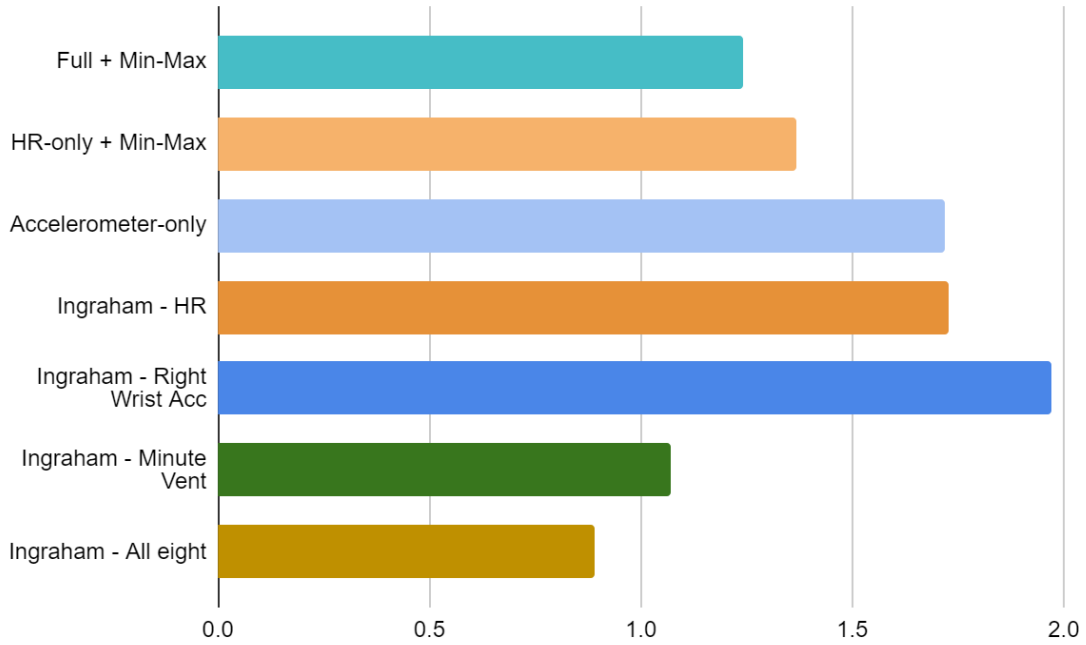

Figure 5. Comparison between our results and Ingraham's. Although our result didn't achieve a lower RMSE than 0.89, which is the lowest estimation error documented by Ingraham. However, if we only use accelerometer data solely, our model can achieve a RMSE of 1.72 lower than 1.97(Ingraham). If heart rate is the only input signal, our model can achieve a RMSE of 1.37 lower than 1.73(Ingraham).

We evaluated the impact of subject-specific normalization on the final energy expenditure by applying the Min-Max normalization. Overall, after the normalization, we get better performance no matter for the full model or the HR-only model. However, the normalization only successfully improves the performance on Ingraham dataset and actually slightly undermines the performance on JSI dataset. The overall improved performance is caused by the much larger data volume of Ingraham compared with JSI. For future work, we can further explore what is the correct scenario to use the min-max normalization approach.

Contrastive learning has been proven to be useful in computer vision and NLP domain, while it is less studied in time series domain. For our model, the incorporation of contrastive learning doesn't necessarily improve the upper bound of validation performance but it can help stabilize the result produced. More importantly, contrastive learning makes personalized fine tuning possible as it is unsupervised and no ground truth is needed. Fine tuning with contrastive learning needs further experiments as it is not producing stable estimation and is therefore not taken into account for the result presented in this report because in the real setting, we can not tell which epoch gives the best performance.

Unlike some previous models [3], which EE estimation is built on top of the activity recognition model, our model is designed to perform energy expenditure estimation and simple activity classification in a parallel way, such that we can build a soft relationship between EE estimation

and activity classification. Instead of a cascaded model, of which the performance of activity recognition will directly impact the ee estimation performance, treating activity classification as an auxiliary loss relaxes the dependence of correct classification of activity while also allowing the activity classification loss to help regularize the EE estimation. However, due to a design fault of our activity classification head, our model's activity classification performance isn't not ideal, but it is a potential direction to work on.

Our architecture supports the basic transformer encoder, which has been proved to be useful in many domains. The initial attempt isn't ideal as the transformer model suffers from big overfitting issues. One further step to improve on the design is that we can manually increase the dimension size of the signal which most deep learning algorithms favor. Relying on expertise from related domains, we can extract multiple features from raw signals. For instance, from raw BVP signals, we can extract high-level features more than just heart rate, but also HRV, crest time, systolic peak information, etc.

Altogether, this work proposes a preliminary architecture for energy expenditure estimation and shows its usefulness when processing data collected from Empatica Watch. However, many novel designs still need to be further developed and tested in the future.

References

[1] Liden CB, Wolowicz M, Stivoric J. Benefits of the SenseWearTM armband over other physical activity and energy expenditure measurement techniques. White Papers Body Media 2001; 1:1–14.

[2] Andre D, Wolf DL. Recent advances in free-living physical activity monitoring: a review. Diabetes Sci Technol 2007; 1(5):760–767

[3] S. E. Crouter, J. R. Churilla, and D. R. Bassett, Jr., "Estimating energy expenditure using accelerometers," Eur. J. Appl. Physiol., vol. 98, no. 6, pp. 601–612, 2006

[4] D. Kim and H. C. Kim, "Estimation of activity energy expenditure based on activity classification using multi-site triaxial accelerometry," Electron. Lett., vol. 44, no. 4, pp. 266–267, 2008.

[5] Cvetković, Božidara, Radoje Milić, and Mitja Luštrek. "Estimating energy expenditure with multiple models using different wearable sensors." IEEE journal of biomedical and health informatics 20, no. 4 (2015): 1081-1087.

[6] Álvarez-García, Juan A., Božidara Cvetković, and Mitja Luštrek. "A Survey on Energy Expenditure Estimation Using Wearable Devices." ACM Computing Surveys (CSUR) 53, no. 5 (2020): 1-35.

[7] Emmanuel Munguia Tapia. 2008. Using machine learning for real-time activity recognition and estimation of energy expenditure. Ph.D. Dissertation. Massachusetts Institute of Technology.

[8] Slade, Patrick, Rachel Troutman, Mykel J. Kochenderfer, Steven H. Collins, and Scott L. Delp. "Rapid energy expenditure estimation for ankle assisted and inclined loaded walking." (2019).

[9] Ingraham, Kimberly A., et al. "Evaluating Physiological Signal Salience for Estimating Metabolic Energy Cost from Wearable Sensors." *Journal of Applied Physiology*, vol. 126, no. 3, 2019, pp. 717–729, doi:10.1152/japplphysiol.00714.2018.

[10] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, *33*, 22243-22255.

Appendix

Table A1. Average Energy Expenditure for each activity. This table summarizes the average Energy Expenditure (in MET) across participants for each activity in three datasets. Overall, the average EE values are comparable across datasets, but the values in the JSI dataset are noticeably higher than those in the Ingraham dataset for walking and running, indicating the need for further investigation if both datasets are used for model training.

Ingraham		**Slade**		**JSI**	
Activity	Avg. EE (MET)	Activity	Avg. EE (MET)	Activity	Avg. EE (MET)
Cycling	3.985	C01	1.550	Additional postures 1	1.988
Incline	4.044	C02	3.057	Additional postures 2	2.710
Running	5.713	C03	4.032	Basic postures 1	1.282
Stairs	4.250	C04	7.445	Basic postures 2	2.097
Walking	2.452	C05	8.217	Cycling	6.117
		C06	6.358	Kitchen chores	2.130
		C07	7.682	Lying exercising	2.292
		C08	4.374	Office work	1.278
		C09	5.710	Resting	1.190
				Running	7.798
				Scrubbing the floor	2.646
				Shoveling snow	3.388
				Walking	4.256

Table A2. RMSE and R2 for all cross-subject evaluation trials

Subject ID	1	2	3	4	5	6	7	8	9	Average
FULL model										
RMSE(MET)	1.43	**1.08**	1.1	1.36	1.38	1.8	1.27	1.16	1.1	1.3
R2	0.74	0.75	0.69	0.64	0.73	0.59	0.75	**0.79**	0.76	0.72
FULL_model_min_max										
RMSE(MET)	1.25	**0.88**	1.05	1.4	1.28	1.67	1.27	1.34	1	1.24
R2	0.81	**0.84**	0.72	0.61	0.77	0.65	0.75	0.71	0.8	0.74
Accelmeter_only										
RMSE(MET)	1.87	**1.56**	1.51	1.72	1.73	2.11	1.71	1.73	1.61	1.73
R2	0.57	0.5	0.43	0.42	**0.58**	0.44	0.56	0.53	0.48	0.5
HR_only										
RMSE(MET)	1.53	**0.82**	1.22	1.64	1.66	2.1	1.64	1.33	1.1	1.45
R2	0.71	**0.86**	0.62	0.47	0.61	0.44	0.59	0.72	0.75	0.64
HR_only_min_max										
RMSE(MET)	1.37	**0.82**	1.15	1.75	1.38	2.03	1.34	1.47	1.01	1.37
R2	0.77	**0.86**	0.67	0.4	0.73	0.47	0.73	0.66	0.79	0.68

Table A3. Dataset summary.A table summarizing the sensor types that the three datasets contain, with each entry in the table has the format *#features [name of features] (device).*

		Ingraham	**JSI**	**slade**	**Empatica(our device)**
Num of subject		10	10	24(1 to 26, skipping 14 and 18)	0

conditions		cc	13(Resting , basic postures, office work, lying exercise, etc)	standing , two walking, two running, two stairs, and two biking	0
time		1	1	1	1
Ground truth					
EE		0	1 (COSMED EE in MET)	1	0
VO2 (mL/s)		1	0	0	0
VCO2 (mL/s)		1	0	0	0
Cal (Estimated calories per hour)		0	1(BodyMedia)	0	0
Estimated EE in MET		0	1(BodyMedia)	0	0
IMU related sensors					
ACC	Chest	9 [Acceleration, Angular Velocity,Magnetic Field] (APDM Accel)	3(Shimmer)	0	0
	Waist	9 [Acceleration, Angular Velocity,Magnetic Field] (APDM Accel)	0	0	0
	Thigh	0	3(Shimmer)	3(IMU)	0
	Left wrist	3 (Empatica Accel)	0	0	0
	Right wrist	3 (Empatica Accel)	3(Shimmer)	0	3

	Left Ankle	9 [Acceleration, Angular Velocity,Magnetic Field] (APDM Accel)	0	0	0
	Right Ankle/Shank	9 [Acceleration, Angular Velocity,Magnetic Field] (APDM Accel)	3(Shimmer)	3(IMU)	0
	Left Foot	9 [Acceleration, Angular Velocity,Magnetic Field] (APDM Accel)	0	0	0
	Right Foot	9 [Acceleration, Angular Velocity,Magnetic Field] (APDM Accel)	0	0	0
	trousers pocket	0	3 (Android)	0	0
Gyroscope		0	0	3(IMU)	3
Muscular System					
EMG	Left	8	0	0	0
	Right	8	0	0	0
EDA /GSR	Left wrist	1 (Empatica Physiological)	1(BodyMedia)	0	1
	Right wrist	1 (Empatica Physiological)	0	0	0
Skin Temperature	Left wrist	1 (Empatica Physiological)	0	0	0
	Right wrist	1 (Empatica	0	0	1

		Physiological)			
	NBT Near Body temperature	0	1(BodyMedia)	0	0
Metabolics System					
Respiratory Exchange Ratio (RER)		1	0	0	0
Breath Frequency (1/min)/ BR		1	1(ZephyrG)	0	0
Minute Ventiltation (L/min) / Respiratory Rate		1	0	0	1
Oxygen Saturation SpO2 (%)		1	0	0	0
Heart Rate (1/min) / Pulse rate / PRV /HRV/ BVP		1 [Heart Rate]	1(ZephyrG)	1	3 [Pulse rate, prv, BVP]
ST		0	1(ZephyrG,BodyMedia)	0	0
R-R interval		0	1(ZephyrG)	0	0
No dataset for training					
movement intensity		0	0	0	1
systolic peaks		0	0	0	1
steps		0	0	0	1
sleep detection		0	0	0	1
wearing detection		0	0	0	1
** Format: #features [name of features] (device)*					