Deep Learning for Metabolic Rate Estimation from Biosignals: A Comparative Study of Architectures and Signal Selection

Sarvenaz Babakhani[1]
sarvenaz.babakhani@ki.uni-stuttgart.de
David Remy[2]
david.remy@iams.uni-stuttgart.de
Alina Roitberg[3]
roitberg@uni-hildesheim.de

[1] Institute for Artificial Intelligence
University of Stuttgart
Stuttgart, Germany

[2] Institute for Adaptive Mechanical Systems
University of Stuttgart
Stuttgart, Germany

[3] Intelligent Assistive Systems Lab
University of Hildesheim
Hildesheim, Germany

Abstract

Energy expenditure estimation aims to infer human metabolic rate from physiological signals such as heart rate, respiration, or accelerometer, and has been studied primarily with classical regression methods. The few existing deep learning approaches rarely disentangle the role of neural architecture from that of signal choice. In this work, we systematically evaluate both aspects. We compare classical baselines with newer neural architectures across single signals, signal pairs, and grouped sensor inputs for diverse physical activities. Our results show that minute ventilation is the most predictive individual signal, with a transformer model achieving the lowest root mean square error (RMSE) of 0.87 W/kg across all activities. Paired and grouped signals, such as those from the Hexoskin smart shirt (5 signals), offer good alternatives for faster models like CNN and ResNet with attention. Per-activity evaluation revealed mixed outcomes: notably better outcomes in low-intensity activities (RMSE down to 0.29 W/kg; NRMSE = 0.04), while higher-intensity tasks showed larger RMSE but more comparable normalized errors. Finally, subject-level analysis highlights strong inter-individual variability, motivating the need for adaptive modeling strategies. Our code and models will be publicly available at this GitHub repository.

1 Introduction and Related Work

Wearable assistive devices are promising for improving mobility, optimizing body energy expenditure, and enhancing the quality of life for older adults and individuals with mobility impairments [8]. Designing such systems is challenging due to the complexity of the human neuromuscular system. To address this, human- and body-in-the-loop optimization methods

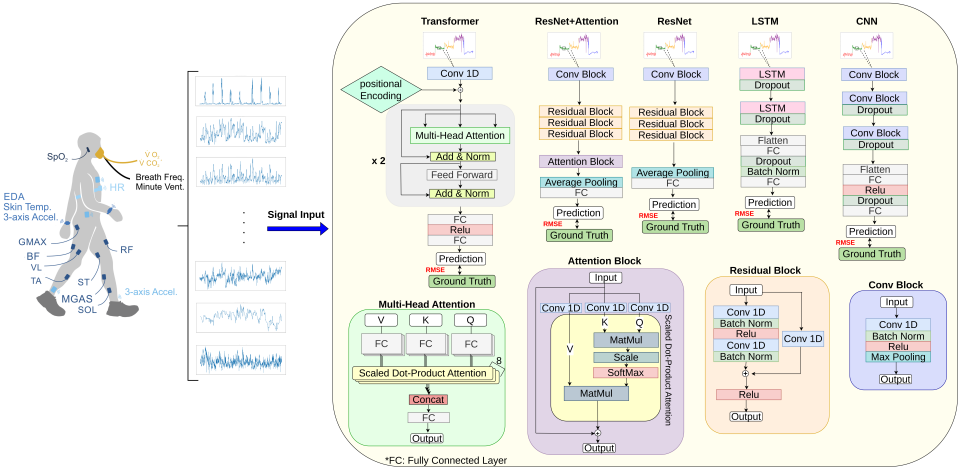

Figure 1: Multimodal physiological signal processing pipeline for EE. Wearable sensors placed across the body collect multimodal signals. These signals are processed and fed as input into multiple neural network architectures. (Image of sensor placement on the body is adapted from Ingraham *et al*. K.A.Ingraham).

adapt device parameters in real time based on user feedback, thereby reducing reliance on complete biomechanical models and avoiding manual tuning in clinical settings [5, 11, 22].

A critical component of such body-in-the-loop optimization systems is the accurate estimation of energy expenditure (EE). One precise but intrusive way for measuring EE is indirect calorimetry, which requires the measurement of oxygen consumption ($\dot{V}_{O_2}$) and carbon dioxide production ($\dot{V}_{CO_2}$) via a metabolic mask; thus limiting long-term use. As an alternative, portable wearable sensors can provide physiological signals, such as heart rate, respiration, or accelerometry, which can be combined to estimate energy expenditure. To achieve this estimation, classical ML methods were used largely in earlier studies, often with limited sensor modalities and task types. For instance, [4, 14, 16, 17, 21] applied linear regression, SVMs, or Gaussian Process Regression with a limited number of signals like accelerometry and heart rate. Some of them reported errors as low as 0.31–0.66 W/kg, but the number of activities was restricted. Other studies, such as Cvetković *et al*. [3] attempted richer sensor sets, but still optimized only for narrow activity ranges. Ingraham *et al*. [8], evaluated the importance of physiological signals for EE estimation and released an accompanying public dataset. They utilized linear regression models and indicated that using minute ventilation alone achieved an RMSE of 1.24 W/kg. While their dataset has since enabled broader evaluation, their analysis was limited to ML approaches. In contrast, our work leverages the same dataset to benchmark both ML and Deep Learning methods across a wide range of signals and activities.

Deep learning studies have achieved stronger performance but under narrower conditions. In vision-based approaches for estimate metabolic, neural architectures (CNNs and Transformers) [9, 15, 19] are a common choice, but they provide only coarse estimates. In contrast, wearable-sensor research has seen less widespread use of deep learning, though also here several studies exist with promises of better estimation through such architectures. For

example, a Deep Multi-Branch Two-Stage Regression Network (DMTRN) was introduced by Ni *et al*. [18] that utilized ECG and IMU data, and achieved an RMSE of 0.71 kcal/min. Other studies, such as those by Lopes *et al*. [13], Lee and Lee [12], and Yuan *et al*. [25] employed CNNs, LSTMs, or hybrid CNN–LSTM models on signals such as IMU, EMG, and motion velocity, but activity diversity in their studies was focused on walking-based tasks, which reduces generalizability. In parallel, Kim and Seong [10] introduced a personalized EE estimation method that combines a modified MET formulation with a heart rate–driven Deep Q-Network, achieving improved per-subject accuracy but without demonstrating cross-subject generalization. Other research has explored specific scenarios, further limiting broader applicability. For instance, [20] estimated EE during assisted and loaded walking, reaching RMSE values as low as 0.40 W/kg across novel subjects and conditions, while [7] evaluated model performance separately for each activity, without assessing generalization across all activities. Additionally, [24] introduced a spatial-temporal fusion network with hybrid attention mechanisms, using multi-sensor data (sEMG, IMU, and HR). Results indicated strong performance with an RMSE of 0.342 kcal/min in individual scenarios and a cross-subject RMSE of 0.646 kcal/min. However, the study did not address the average performance across all subjects or scenarios, making it difficult to assess overall generalization. Our study extends this line of work by considering a broader sensor set, diverse activities, and a wider variety of architectures beyond a single model class.

In summary, most prior work on EE estimation has relied on classical machine learning techniques, while only a few recent studies have explored deep learning. However, these deep learning approaches rarely disentangle the role of neural architecture from that of signal selection, leaving open the question of architectural and signal choices. In this study, we expanded on recent research by comparing various models, including linear regression, CNN, ResNet, ResNet+Attention, LSTM, and Transformer models, across multiple input configurations (single, paired, and grouped signals) to check both overall activity and activity-specific performance. We studied the ability of the models to generalize and investigated the impact of transitions between activities. Additionally, we examined inter-individual variability in signal effectiveness and model performance by studying a per-subject evaluation. By systematically comparing classical and deep learning methods on the Ingraham *et al*. [8] dataset, we establish a new state of the art on this benchmark, achieving substantially lower error rates than prior work.

2 Methods

2.1 Dataset

In this study, we use the public dataset provided by Ingraham *et al*. Please refer to [8] for more information on data collection and processing. In total, sixteen signals were gathered with wearable sensors from 10 different subjects, performing six types of physical activities. The signals are provided in Table 1, with additional details available in the supplementary material. The ground truth energy expenditure was computed using the Brockway equation [2] and normalized based on the subject's body weight.

Grouping Signals: In addition to evaluating models on 1) individual signals, we considered 2) all possible signal pairs as well as 3) physiologically motivated signal groups proposed by [8]. Group memberships are listed in Table 1, with each signal annotated by its group label in parentheses (e.g., G for Global signals). *Global signals*, such as minute ventilation

and heart rate, reflect whole-body physiological state. *Local signals*, such as ankle and wrist acceleration, capture activity in specific body segments. The *Local+Global* setting combines both groups, incorporating all 16 signals. *Hexoskin signals* refer to those measured using the Hexoskin smart shirt [1], like breath frequency and chest acceleration.

2.2 Neural Architectures

We design and analyze different neural network-based models for estimating human metabolic rate from wearable sensor signals, aiming to disentangle the impact of neural architecture from that of signal choice by comparing models with distinct inductive biases. We consider six representative models: Linear Regression, CNN, LSTM, ResNet, ResNet+Attention, and Transformer. All deep learning models in this study operate directly on temporal signal inputs. An overview of the task and implemented approaches is given in Figure 1.

Linear Regression: This simple and interpretable model serves as an important baseline, as it is widely used in prior work on EE estimation, and is the key approach used in the benchmark we build on [8]. We implemented both single and multiple linear regression variants.

CNN: This model is designed to capture local temporal patterns in the input signals by applying one-dimensional convolutions through time. The model consists of three 1D convolutional blocks followed by fully connected layers. The convolutional output is flattened and passed through two fully connected layers. Finally, the output layer has a linear activation function to match the prediction with the target dimensions. (The training time is 550.98 s).

LSTM: We implemented a stacked LSTM-based regression network to leverage both short-term and long-term memory to monitor changes in the input signals over time. The model consists of two sequential Long Short-Term Memory (LSTM) layers. The final LSTM output is flattened and passed through a fully connected layer, followed by batch normalization and dropout. (The training time is 264.06 s).

ResNet: We build on the popular ResNet architecture [6] and adapt it for 1D time-series input. The main idea is to utilize residual (skip) connections to allow the network to pass information from earlier layers directly to later layers. The model architecture begins with a 1D convolutional block. Next, there are three residual blocks with increasing output dimensions. Global average pooling is applied after the last residual block, followed by a linear layer mapping to the output size. (The training time is 227.01 s).

ResNet+Attention: We extend the ResNet architecture with a self-attention block after the residual layers to capture longer-range dependencies (see Figure 1). This block uses a self-attention mechanism over the temporal dimension. In this block, there are three separate 1×1 convolutions to produce query, key, and value, which represent the input. The attention score is calculated with the related equation for $Attention(Q,K,V)$ in [23] After re-weighting the values, a residual connection adds the attention output back to the input and preserves the original features. (The training time is 590.99 s).

Transformer: To model complex temporal dependencies that extend beyond local patterns, we implemented a multi-head attention model as a Transformer-based architecture inspired by the original Transformer encoder framework [23]. This model combines 1D convolutional feature projection, positional encoding, and multi-head self-attention. A stack of two Transformer encoder layers is applied, each consisting of multi-head self-attention (with eight heads), feedforward network, residual connections, and layer normalization. Finally, the output is passed through a small feedforward network to match the output dimension of the EE prediction. (The training time is 4877.41 s).

Main experiments were conducted using PyTorch 2.5.1 with CUDA 12.4 on a single NVIDIA

Signal	Lin-Reg [8]	Lin-Reg (ours)	CNN	LSTM	ResNet	ResNet+Att	Transformer
Waist Acceleration (L,H)	-	2.41	2.22	2.04	2.30	3.04	**1.89**
Chest Acceleration (L,H)	-	2.35	2.01	2.02	2.09	2.01	**1.92**
Left Ankle Acceleration (L)	-	2.33	**1.84**	2.02	2.01	1.89	1.85
Right Ankle Acceleration (L)	-	2.33	**1.83**	1.96	1.92	1.87	1.83
Left Wrist Acceleration (L)	-	2.70	2.17	2.16	2.19	**2.08**	2.09
Left Wrist Electrodermal (G)	2.93	3.19	2.60	2.36	2.53	2.83	**2.11**
Left Wrist Temperature (G)	-	3.11	2.55	2.81	2.73	2.54	**2.52**
Right Wrist Acceleration (L)	-	2.73	2.16	2.22	2.25	2.25	**2.07**
Right Wrist Electrodermal (G)	-	3.01	2.37	2.46	2.59	2.88	**2.24**
Right Wrist Temperature (G)	-	3.13	**2.53**	2.74	2.85	2.56	2.56
EMG Magnitude Left (L)	-	2.86	2.40	2.48	2.58	**2.37**	2.40
EMG Magnitude Right (L)	-	2.83	**2.40**	2.55	2.50	2.43	2.42
Heart Rate (G,H)	-	2.29	**1.81**	2.08	1.97	1.84	1.95
SpO_2 (G)	-	2.81	2.34	2.51	2.48	**2.33**	2.34
Breath Frequency (G,H)	-	2.89	2.46	2.67	2.57	**2.35**	2.43
Minute Ventilation (G,H)	1.24	1.30	1.00	1.03	1.03	0.97	**0.87**
Global Signals	1.25	1.34	**0.97**	1.08	1.16	1.17	1.18
Global Signals W/O MinVent	-	2.35	1.82	**1.77**	1.96	1.81	2.17
Local Signals	-	1.99	1.98	1.73	1.84	2.69	**1.54**
Local+Global Signals	1.28	1.27	**0.93**	1.13	1.34	1.21	1.27
Local+Global W/O MinVent	-	1.88	**1.60**	1.79	1.95	1.88	1.58
Hexoskin Signals	1.24	1.28	**0.92**	0.98	1.12	1.10	1.07
$\dot{V}_{O_2}$ (part of ground truth)	0.93	0.91	0.62	0.56	0.58	0.74	**0.40**

Table 1: RMSE (W/kg) of Models using physiological input signals (individual and grouped). Signal grouping's initials: (L) local, (G) Global, and (H) Hexoskin. The first column reports baseline (linear regression) results from [8](where available), while the remaining columns present our reproduced linear regression and deep learning models.

GeForce RTX 4090 GPU with 24 GB memory; additional hyperparameters and training configurations are provided in the supplementary material.

3 Experiments and Result

We follow the preprocessing and evaluation protocol of [8] and evaluate the architectures using leave-one-subject-out cross-validation. In each fold, one subject is used for testing, while the remaining subjects are used for training, with 15% of the training data held out for validation. This process is repeated for all 10 subjects. For each test subject, root mean square error (RMSE) is computed as the average error across all predicted time steps, and the final RMSE is obtained by averaging the results across all folds. When analyzing performance for each activity, RMSE is normalized by the average EE of that activity to obtain the normalized RMSE (NRMSE).
A manual grid search over hyperparameters is performed using minute ventilation as the input signal, and the selected values are then fixed for training on all other signals.

3.1 Single and Grouped signals Comparison

We began by examining model performance across single and grouped signals in Table 1. Minute ventilation emerged as the most reliable predictor of EE, consistently outperforming other modalities. Among models, the Transformer-based approach achieved the lowest over-

all error (RMSE of 0.87 W/kg for minute ventilation), and was also best for several other signals, such as waist and chest acceleration. If we prefer an alternative for minute ventilation (challenging to measure, see Sec. 3.3), heart rate is a viable alternative due to its ease of measurement and the second lowest RMSE (1.81 W/kg) among other signals. Beyond the transformer, other architectures also showed strengths: the ResNet+Attention model outperformed on four signals, whereas the CNN attained the lowest error on five signals and was particularly effective on grouped inputs. Recognition quality was high for Hexoskin signals (RMSE of 0.92 W/kg), similar to the Transformer's best result on minute ventilation.

3.2 Pair Combination of Signals

In the next step, we considered pairwise combinations of physiological signals. While individual signals alone may carry strong predictive power, combining their sources can reveal useful synergistic effects. As expected, the combination of minute ventilation and other signals consistently outperformed all other combinations. Among these combinations, we selected the best ones and visualized them in Figure 2. The first row showed the result when

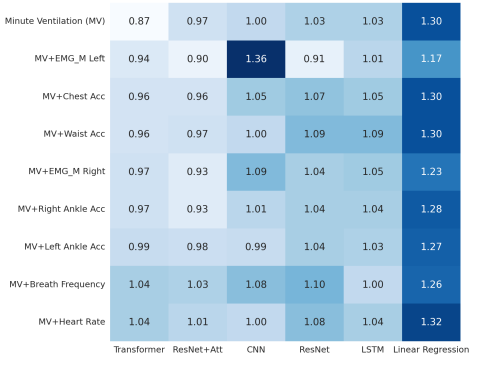

Figure 2: Heatmap of RMSE values using Minute Ventilation (MV) alone and in combination with secondary signals (rows). The columns correspond to different prediction models. Lower RMSE values (lighter colors) indicate better predictive performance.

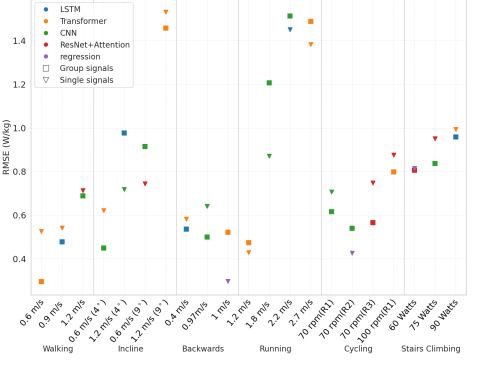

Figure 3: Model performance across different activities and conditions. The x-axis shows six activities with variations in speed or resistance. Model types are distinguished by color, while input type (single and grouped) is indicated by marker shape.

minute ventilation is the only input, and the other rows illustrated the results when additional signals were combined with minute ventilation. Both the Transformer and ResNet+Attention yielded the best overall results. Notably, adding EMG magnitude (left) further boosted the performance of both ResNet+Attention and ResNet. Beyond accuracy, both networks benefited from faster training times compared to the Transformer (see Sec. 2.2). When compared to the results in Table 1, pairing signals enabled ResNet+Attention and ResNet to surpass the CNN with Hexoskin inputs (RMSE 0.92 W/kg), underscoring the added value of EMG signals in combination with minute ventilation.

3.3 Alternatives to Minute Ventilation

While minute ventilation is the strongest predictor in our study, its measurement is technically demanding, costly, and often uncomfortable, as it typically requires the use of a mask. This motivated the search for practical alternatives.

In Figure 4, we compared five candidate physiological signals beyond minute ventilation. The right side of the figure shows their individual performance, while the left side highlights the best-performing pairwise combinations. Pairing signals resulted in lower RMSE, indicating that when minute ventilation is removed, using other signals in pairs is more beneficial.

As we mentioned before, CNN with heart rate was the best single signal after minute ventilation, but using it with the (right and left) ankle acceleration (RMSE: 1.49 and 1.51 W/kg) improved the performance by almost 17%. Another effective pair was left ankle acceleration

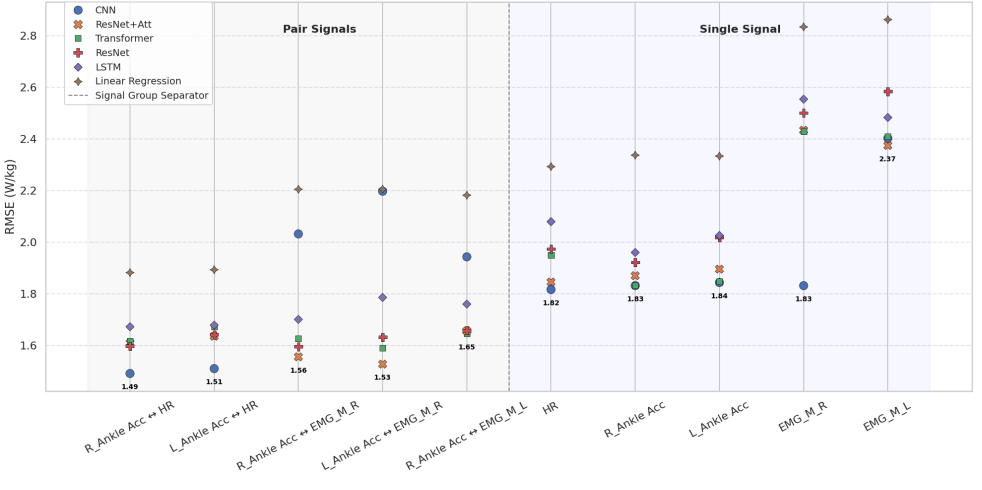

Figure 4: Model performance using alternative input signals for Minute Ventilation. The left panel shows results from paired signal combinations, while the right panel shows single-signal inputs. Different models are represented by distinct colors and markers.

and EMG magnitude (right) with the ResNet+Attention, yielding an RMSE of 1.53 W/kg.

When analyzing the most effective pair for each signal (in the absence of minute ventilation), heart rate and ankle acceleration (left or right) frequently emerged as the strongest partners. Across the majority of best pairs cases, CNN delivered the best predictive performance, achieving the lowest RMSE relative to other models.

While CNN was effective for certain signal combinations (grouped or pairs), poor signal selection led to significantly worse results. For example, pairing EMG with electrodermal activity produced the highest RMSE (8.05 W/kg), underscoring the poor suitability of these signals for this task. Similarly, electrodermal and temperature signals, whether considered individually or in pairs, consistently yielded high errors (e.g., 3.15–3.23 W/kg), across diverse models, highlighting their limited predictive value.

The complete tables for the best partner(pair) of each signal and for the least effective pairs, along with the corresponding models, are provided in the supplementary materials.

3.4 Per-Activity Evaluation

Next, we evaluated model performance per activity, including different speeds and resistance conditions. The models were trained on all activity types, as in the previous experiments, but testing was carried out separately for each activity.

We observed two central findings. First, Figure 3 showed that Transformer and CNN performed best with single inputs. Linear regression with minute ventilation achieved the lowest single-signal RMSE of 0.29 W/kg in the backward walking at 1 m/s. Grouped signals with CNN-, LSTM-, and Transformer-based methods consistently improved over single inputs. The best overall result, also 0.29 W/kg, was obtained by the Transformer with Local+Global data during walking at 0.6 m/s.

Activity	Condition	NRMSE_single	Signal	NRMSE_group	group
Walking	0.6 m/s	0.14	Min_Vent	0.08	Loc+Glob
	0.9 m/s	0.13	Min_Vent	0.11	Loc+Glob
	1.2 m/s	0.14	Min_Vent	0.14	Loc+Glob
Incline	0.6 m/s (4°)	0.13	Min_Vent	0.09	Loc+Glob
	1.2 m/s (4°)	0.09	L_Wrist_Elec	0.13	Hexoskin
	0.6 m/s (9°)	0.10	Min_Vent	0.13	Hexoskin
	1.2 m/s (9°)	0.12	Min_Vent	0.11	Hexoskin
Backwards	0.4 m/s	0.15	Min_Vent	0.14	Loc+Glob
	0.7 m/s	0.13	Min_Vent	0.10	Loc+Glob
	1.0 m/s	**0.04**	Min_Vent	0.08	Hexoskin
Running	1.2 m/s	0.09	Min_Vent	0.10	Global
	1.8 m/s	0.08	Min_Vent	0.11	Hexoskin
	2.2 m/s	0.12	R_Ankle_ACCL	0.12	Hexoskin
	2.7 m/s	0.09	Min_Vent	0.10	Local
Cycling	70 rpm (R1)	0.13	Chest_ACC	0.11	Global
	70 rpm (R3)	0.06	SpO_2	0.08	Loc+Glob
	70 rpm (R5)	0.09	Min_Vent	0.07	Global
	100 rpm (R1)	0.11	Min_Vent	0.10	Local
Stairs Climbing	60 Watts	0.12	R_Ankle_ACCL	0.11	Local+Global
	75 Watts	0.11	Min_Vent	0.10	Global
	90 Watts	0.11	Min_Vent	0.11	Hexoskin

Table 2: NRMSE for different activities and different conditions.

Second, performance varied with activity intensity. RMSEs were lower for low-intensity activities, while higher-intensity tasks produced larger RMSEs. However, Table 2 showed that normalization (NRMSE) reduced these differences. Several high-intensity conditions (e.g., running at 1.8 m/s) also achieved comparable NRMSE. This indicates that while intensity increases error, models scale proportionally.

Table 2 also indicated that, as expected, minute ventilation emerged as the strongest single input and obtained the overall best NRMSE of 0.04, while among grouped signals, Local+Global and Hexoskin consistently delivered the best performance across activities.

3.5 Per-Subject Evaluation

Lastly, we evaluated the effectiveness of different physiological signals and models per subject to examine how results fluctuate with individual differences. Figure 5 compares CNN and Transformer models for single and grouped signals. While overall trends were consistent across architectures, we highlighted the most informative results here (further plots are available in the supplementary materials). Each boxplot shows the distribution of RMSE values across 10 subjects for a given input signal. As expected, minute ventilation consistently yielded the lowest RMSE with minimal inter-subject variability, confirming its role as the most robust predictor of energy expenditure. In contrast, signals such as SpO_2 and EMG magnitude (left and right) showed both higher RMSE and greater variance, reflecting weak predictive power and strong inter-individual differences in signal quality. Heart rate achieved a low average RMSE but displayed high variance across subjects.

Interesting insights arose from comparing the variance differences between chest and ankle accelerations. Chest acceleration, which reflects global body motion, presumably benefited from the Transformer's ability to capture smooth, long-range dependencies, resulting in lower variance. In contrast, CNNs, which rely on local temporal filters, may have failed to capture these patterns. On the other hand, ankle acceleration signals are periodic and

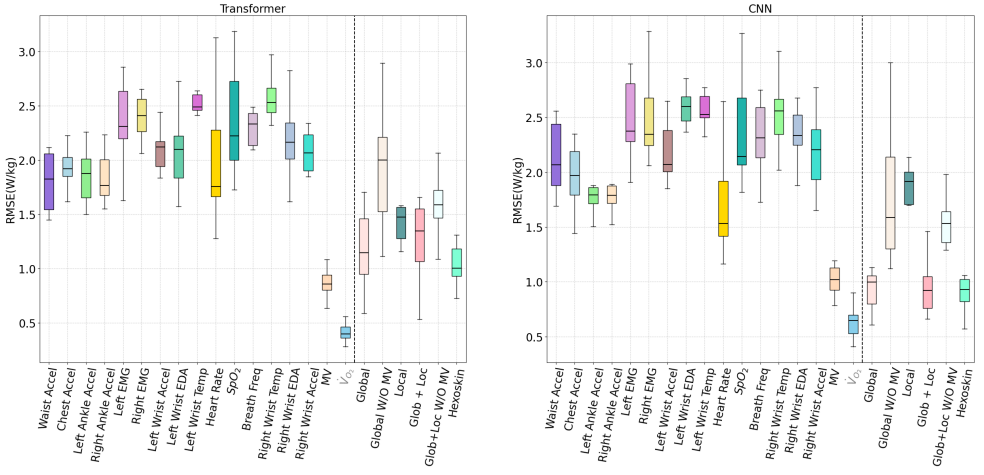

Figure 5: Performance of Transformer and CNN for single and grouped signals across 10 subjects. The dashed line separates single- from grouped-signals in each plot. Boxplots represent the distribution of RMSE values across subjects: median (line), 25th–75th percentiles (box), and whiskers to 1.5×IQR. (MV: Minute Ventilation)

structured patterns that were well-suited to the CNN's short-window convolutional architecture. Here, the Transformer may have been overly sensitive to small subject-specific gait variations, leading to higher variance. Finally, we observed that removing minute ventilation from the Global and Local+Global signal groups led to a noticeable increase in both median RMSE and variance.

4 Discussion and Conclusion

In this work, we implemented and systematically compared different neural network-based architectures for energy expenditure prediction from wearable physiological signals across diverse activities. Our main objectives were to: (1) compare ML and DL models, (2) assess generalizability across signals and activities, and (3) analyze subject-specific variability.

Model and signal configurations: The Transformer and ResNet+Attention consistently outperformed other models, while CNNs offered a strong balance between accuracy and computational efficiency. Across all models, minute ventilation was the most reliable predictor, achieving an RMSE of $0.87\,\mathrm{W/kg}$ with the Transformer. Since it is difficult to measure in practice, we examined alternatives: heart rate was the strongest single signal, and pairing or grouping signals further improved accuracy. For example, pairing heart rate with ankle acceleration or EMG signals across limbs reduced RMSE substantially. Signal fusion using Hexoskin and Local+Global inputs also outperformed single-signal baselines.

Activity effects: Performance varied with activity intensity. Low-intensity tasks (e.g., backward walking at $1\,\mathrm{m/s}$) yielded very low errors, while higher-intensity activities showed greater RMSE. Normalization reduced these differences, with some intense activities achieving competitive NRMSE. This scaling of the error with intensity highlights the value of activity-specific refinements over universal models.

Individual differences: A central finding of this study is the extent of inter-subject variability. While minute ventilation provided stable performance for all participants, other signals such as heart rate and EMG were highly variable, likely reflecting physiological differences and variations in sensor quality. Model choice also interacted with signal type: Transformers captured smoother, whole-body dynamics (e.g., chest acceleration) more consistently across subjects, whereas CNNs better handled periodic patterns (e.g., ankle movement). These results highlight that robust EE estimation requires not only choosing the right signals but also matching model architecture to signal characteristics as well as individual variability.

Further practical recommendations: Taken together, our findings suggest several guidelines for real-world applications. When minute ventilation is available and processing time is less critical, the Transformer is the optimal choice. If faster inference is required and signals captured by the Hexoskin shirt are accessible, CNNs offer a good balance of efficiency and accuracy. In cases where both minute ventilation and EMG magnitude are available, ResNet+Attention provides the best overall accuracy. Finally, when minute ventilation cannot be measured, pairing heart rate with ankle acceleration and applying a CNN yields a strong and practical alternative.

Neural network–based approaches for EE prediction, particularly considering diverse physiological signals, have been understudied. Our results demonstrate both the potential of these methods and the substantial inter-subject variability that remains. This variability highlights the need for future work on activity-specific and personalized models. To encourage further research, we will release our code and models at this GitHub repository.

5 Acknowledgment

The project was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2075 - 390740016). We acknowledge the support of the Stuttgart Center for Simulation Science (SimTech) and the International Max Planck Research School for Intelligent Systems (IMPRS-IS). The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (https://www.nhr-verein.de/en/our-partners). HoreKa is partly funded by the German Research Foundation (DFG).

References

[1] Hexoskin (carré technologies inc.): Hexoskin smart shirts. URL https://hexoskin.com/.

[2] J.M. Brockway. Derivation of formulae used to calculate energy expenditure in man. *Human Nutrition: Clinical Nutrition*, 41(6):463–471, 1987.

[3] Bozidara Cvetković, Radoje Milić, and Mitja Luštrek. Estimating energy expenditure with multiple models using different wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, 19(5):1574–1581, 2015. doi: 10.1109/JBHI.2015.2432911.

[4] Tiziana Falcone, Simona Del Ferraro, Vincenzo Molinaro, Loredana Zollo, and Paolo Lenzuni. Estimation of the metabolic rate in the occupational field: a regression model using accelerometers. *International Journal of Industrial Ergonomics*, 96:103454, 2023. doi: 10.1016/j.ergon.2023.103454.

[5] Wyatt Felt, Jessica C. Selinger, J. Maxwell Donelan, and C. David Remy. "body-in-the-loop": Optimizing device parameters using measures of instantaneous energetic cost. *PLoS ONE*, 10(8):e0135342, 2015. doi: 10.1371/journal.pone.0135342.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016(1):770–778, 2016.

[7] Aya Houssein, Jacques Prioux, Steven Gastinger, Brice Martin, Fenfen Zhou, and Di Ge. Energy expenditure estimation from respiratory magnetometer plethysmography: A comparison study. *IEEE Journal of Biomedical and Health Informatics*, 27 (5):2345–2352, 2023. doi: 10.1109/JBHI.2023.3252173.

[8] Kimberly A. Ingraham, Daniel P. Ferris, and C. David Remy. Evaluating physiological signal salience for estimating metabolic energy cost from wearable sensors. *Journal of Applied Physiology*, 126(3):717–729, 2019. doi: 10.1152/japplphysiol.00714.2018.

[9] Gayatri Kasturi, Pragya Shrestha, Scott J Strath, and Rohit J Kate. Estimating physical activity energy expenditure from video. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8. IEEE, 2024.

[10] Min-Seo Kim and Ju-Hyeon Seong. A personalized energy expenditure estimation method using modified met and heart rate-based dqn. *Sensors*, 25(11), 2025. ISSN 1424-8220. doi: 10.3390/s25113416. URL https://www.mdpi.com/1424-8220/25/11/3416.

[11] Jeffrey R. Koller, Deanna H. Gates, Daniel P. Ferris, and C. David Remy. 'body-in-the-loop' optimization of assistive robotic devices: A validation study. In *Proceedings of Robotics: Science and Systems (RSS) XII*, pages 1–10, Ann Arbor, MI, USA, 2016.

[12] Chang June Lee and Jung Lee. Imu-based energy expenditure estimation for various walking conditions using a hybrid cnn–lstm model. *Sensors*, 24:414, 01 2024. doi: 10.3390/s24020414.

[13] João M. Lopes, Joana Figueiredo, Pedro Fonseca, João J. Cerqueira, João P. Vilas-Boas, and Cristina P. Santos. Deep learning-based energy expenditure estimation in assisted and non-assisted gait using inertial, emg, and heart rate wearable sensors. *Sensors*, 22 (20):7913, 2022. doi: 10.3390/s22207913.

[14] Marco Marena, Neethan Ratnakumar, Rachel Jones, Xianlian Zhou, Sanchoy Das, and Bo Shen. Predicting metabolic rate for firefighting activities with worn loads using a heart rate sensor and machine learning. In *Proceedings of the IEEE International Conference on Body Sensor Networks (BSN)*, pages 1–4, Boston, MA, USA, 2023. IEEE. doi: 10.1109/BSN58485.2023.10331063.

[15] Alessandro Masullo, Tilo Burghardt, Dima Damen, Sion Hannuna, Víctor Ponce-López, and Majid Mirmehdi. Calorinet: From silhouettes to calorie estimation in private environments. *British Machine Vision Conference (BMVC)*, 2018.

[16] Sara Monteiro, Joana Figueiredo, and Cristina Santos. Towards a more efficient human-exoskeleton assistance. In *Proceedings of the IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 181–186, Tomar, Portugal,, 2023. IEEE. doi: 10.1109/ICARSC58346.2023.10129556.

[17] Sara Monteiro, Joana Figueiredo, Pedro Fonseca, J. Paulo Vilas-Boas, and Cristina P. Santos. Human-in-the-loop optimization of knee exoskeleton assistance for minimizing user's metabolic and muscular effort. *Sensors*, 24(11):3305, 2024. doi: 10.3390/s24113305.

[18] Zhiqiang Ni, Tongde Wu, Tao Wang, Fangmin Sun, and Ye Li. Deep multi-branch two-stage regression network for accurate energy expenditure estimation with ecg and imu data. *IEEE Transactions on Biomedical Engineering*, 69(10):3224–3233, 2022. doi: 10.1109/TBME.2022.3163429.

[19] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Should i take a walk? estimating energy expenditure from video data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2075–2085, 2022.

[20] Patrick Slade, Rachel Troutman, Mykel J. Kochenderfer, Steven H. Collins, and Scott L. Delp. Rapid energy expenditure estimation for ankle assisted and inclined loaded walking. *Journal of NeuroEngineering and Rehabilitation*, 16(1):67, 2019. doi: 10.1186/s12984-019-0535-7.

[21] Patrick Slade, Mykel J. Kochenderfer, Scott L. Delp, and Steven H. Collins. Sensing leg movement enhances wearable monitoring of energy expenditure. *Nature Communications*, 12(1):4312, 2021. doi: 10.1038/s41467-021-24173-x.

[22] Patrick Slade, Christopher Atkeson, J. Maxwell Donelan, Han Houdijk, Kimberly A. Ingraham, Myunghee Kim, Kyoungchul Kong, Katherine L. Poggensee, Robert Riener, Martin Steinert, Juanjuan Zhang, and Steven H. Collins. On human-in-the-loop optimization of human–robot interaction. *Nature*, 633(8031):779–788, 2024. doi: 10.1038/s41586-024-07697-2.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, Long Beach, CA, USA, 2017. Curran Associates, Inc.

[24] Wenjin Xu, Wei Meng, Chang Zhu, Jingjing Kong, Quan Liu, and Qingsong Ai. Spatial-temporal fusion network with hybrid attention for energy expenditure prediction based on multi-sensor. In *Proceedings of the 2024 30th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pages 1–6, 2024. doi: 10.1109/M2VIP62491.2024.10746155.

[25] Jinfeng Yuan, Yuzhong Zhang, Shiqiang Liu, and Rong Zhu. Wearable leg movement monitoring system for high-precision real-time metabolic energy estimation and motion recognition. *Research*, 6:0214, 2023. doi: 10.34133/research.0214.