# Boosting is Better: Challenges Facing Advanced Neural Networks in Energy Expenditure Models

Tyler Anderton

***Abstract*—Obesity and weight gain remain unyielding public health concerns, driving a great demand for wearable devices that can accurately and conveniently monitor physical activity and energy expenditure. Recent advancements in neural networks, specifically Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models, have shown promising abilities to extract complex features from raw data signals and yield high accuracies in energy expenditure estimation. However, these studies have either relied on additional data that modern wearable devices cannot provide, or have been limited in the types of physical activities performed during data collection. This research builds upon the work of these authors by evaluating and comparing their methods of CNN feature extraction, with and without the assistance of an LSTM, against XGBoost on a dataset that reflects the capabilities of modern wearable devices and a broader range of activities. After testing a total of 192 hyperparameter combinations to rule out model architecture issues, the best RMSE that could be achieved was 5.08 kcal/min, yielding a 146%CV, which is far beyond any acceptable limit for real-world accuracy. While falling short of the best performing XGBoost model with 1.02 kcal/min RMSE (39% CV), these findings highlight the continuing challenges posed by implementing these devices and algorithms on the consumer market.**

## Introduction

### *Measuring Energy Expenditure*

For decades obesity has been a massive and growing problem in nearly all developed and developing countries [1], with obesity rates tightly linked to an increased risk for many deadly diseases including cardiovascular disease and Alzheimer's disease [2, 3]. We have known for a long time that weight gain and obesity can be largely held at bay by moderate adjustments to caloric intake and physical activity, with some authors proposing that as little as an additional 100 kcal/day in energy expenditure could be enough to prevent much of the weight gain that causes obesity [4]. We have also known for many decades that physical activity is essential to maintaining general health and well being throughout adulthood independent of weight gain and obesity [5].

Given the crucial importance of physical activity for the purposes of health and well being, the pursuit of an accurate, non-intrusive method of measuring free-living physical activity has captured much attention in the literature for many decades. The gold standard method for estimating energy expenditure (EE) from physical activity is and has been indirect calorimetry, which is usually performed in one of two ways, each with their own advantages and disadvantages.

The ultimate goal when estimating energy expenditure is often to acquire some understanding of how normal every-day behaviors affect the total energy expended. The first method of indirect calorimetry, the use of doubly-labeled water (DLW), is the most accurate choice for measuring energy expenditure in free-living scenarios [6]. Unfortunately this method requires a great deal of technical expertise and advanced equipment and is therefore prohibitively expensive for large-scale repeatable studies [7, 8].

An alternative method of indirect calorimetry that has demonstrated high accuracy involves the measurement of oxygen and carbon dioxide exchange through a ventilated mask or within a small chamber. The ratio of these gases can then be used to estimate the energy expended by the subject using a validated formula [9, 10]. Earlier methods involved whole-body calorimetry chambers, which were quite restrictive physically and also suffered from the restrictive costs associated with utilizing these chambers on a large scale [11–14]. Contemporary indirect calorimetry devices have gotten rather compact and convenient for use in the lab, usually consisting of little more than a respiratory mask that connects to a small measurement device that either straps onto the subject's body, in the case of the COSMED K4b2 [15–19] or the Oxycon Mobile [20–22], or is even self-contained within the mask itself, as in the $VO_2$ Master Analyzer [23, 24]. While these calorimeters provide an excellent solution for measuring energy expenditure in a laboratory, anything that requires subjects to wear a respiratory mask is still too intrusive to implement in free-living conditions due to the potential for the Hawthorne effect and compliance issues [25].

### *Wearable Devices*

Algorithms and methods to estimate energy expenditure need to be "accurate, fast, and comfortable" [22]. These are requirements that current indirect calorimetry procedures cannot fulfill. The ideal physical activity monitor should be in such a small and inconspicuous form-factor that users would willingly wear their device 24/7. Companies such as Apple, Whoop, and Fitbit have seen great success in selling wearable devices to consumers for the marketed purposes of activity tracking and health monitoring [26]. Such widespread adoption of these monitoring devices presents incredible opportunities, both for researchers to collect massive amounts of data on subjects engaging in free living activities, and for consumers to improve their own health and well being with empowering analytics of their own physical activity. These devices most commonly include accelerometer units and heart rate monitors at least, with some of the newest models also including temperature and blood oxygen saturation measurement capabilities [27].

The literature on estimating energy expenditure from the data available from these compact wearable devices is decades deep at this point. As early as 1979, authors have been experimenting with methods to predict energy from heart rate alone as compared to the two gold standards, whole-body indirect calorimetry [11–14], and doubly labeled water [28, 29]. Many authors also found better predictive power when combining measurements of motion through accelerometry with the heart rate information [30–35]. However, most of these studies involved explicit linear or non-linear regression, with many authors finding the need to implement complex branched equation modeling to handle the inherent non-linear relationships between heart rate, physical movement, and energy expenditure [36–42].

### *Supervised Learning and Neural Models*

These explicit techniques involve a great deal of skill and domain-knowledge to develop, and often seem to fall short in one domain of activity or another. Researchers more recently have demonstrated supervised learning models to yield greater accuracies when estimating energy expenditure [15–19, 21, 24, 43–46]. Supervised approaches are preferable in general to the methods previously discussed, as they do not require the explicit definition of the parameters for the regression equation. These methods instead learn the regression parameters automatically through special learning algorithms [47]. Furthermore, many supervised learning models are inherently non-linear and have been shown to out-perform the branched linear equations and explicit non-linear regressions previously studied [15, 19, 24, 43, 44].

While many authors validated tree-based models like random forests and XGBoost [15, 24, 43], Rothney et al. 2007 was perhaps the first group to propose an artificial neural network (ANN) for energy expenditure estimation. They showed reduced estimation errors compared to the popular regression methods of their contemporaries. Artificial neural networks (often now simply called “neural networks” or “neural models”) are models that take inspiration for their design from the neural structure of biological brains. Each “neuron” represents a linear combination of its inputs, with the “weights” and “biases” being learnable parameters that define each linear combination. The output of each neuron is fed through a non-linear “activation function” (ReLU, tanh, and sigmoid being popular choices) before being sent to the next neuron(s) in sequence or to the model output. This rather simple combination of linear and non-linear functions has proven extremely powerful in both classification and regression tasks when many layers of these neurons are connected in series and in parallel [48].

Since 2007 a large number of authors have built upon the work of Rothney et al. 2007 in the implementation of neural models [16–19, 21, 45, 46]. Many of these works still involve non-trivial effort and technical overhead in explicit feature engineering of the raw data signals to maximize the performance of these neural models. At the very least, the acceleration signals are often “counted” by proprietary algorithms determined by the accelerometer’s manufacturer. Those counts are then aggregated over some time period, and their summary statistics, such as the mean, standard deviation, minimum, maximum, various percentile values, and interquartile range, are input as features to the neural network [18, 19, 21, 45]. For example, Staudenmayer et al. 2009 collected the second-by-second activity counts, then calculated the summary statistics for each minute of these counts. They also implemented lag one autocorrelation as a measure of temporal dynamics. An alternative method by Paraschiakos et al. 2022 simply calculated various summary statistics on the raw acceleration signals, rather than the proprietary count values. In addition to these temporal features, many authors have implemented much more complex feature engineering strategies to extract frequency-based features [15], such as calculating discrete Fast-Fourier Transform (FFT) component magnitudes [49], deriving frequency-domain entropies [49], implementing frequency filtering [50], and performing wavelet transforms [51] on the raw signals.

All of this manual effort of feature engineering clashes with the ethos of supervised learning with neural models, which should, in principle, be able to learn whatever optimal features can be extracted from the raw inputs. Indeed, recent authors have utilized Convolutional Neural Networks (CNNs) for automatic feature extraction on the raw acceleration signals with minimal to no filtering or processing [16, 17, 46]. CNNs are a specific type of neural network that have become very popular for a range of tasks including speech recognition, natural language processing (NLP), and especially computer vision. By using convolutional operations to extract features from their inputs, they have an unparalleled ability to capture spatial features and patterns, especially when implemented in a hierarchical structure [47].

### *Related Work*

Zhu et al. 2015 were the first to validate the utilization of CNNs for temporal feature extraction for the purpose of energy expenditure estimation. In this study, raw signals from a single tri-axial accelerometer were fed into a CNN with no filtering or pre-processing applied. Data were collected while subjects performed “daily ambulatory activities”, such as walking, running, and “low whole body motion” tasks like static standing, sitting, and riding an elevator. The authors found that a small Fully Connected Network (FCN) (also called a Multi-Layer Perceptron (MLP) or a Feed-Forward Neural Network (FFNN)) [48] performed better in the task of energy expenditure estimation when fed features extracted by a CNN rather than features handcrafted by explicit methods similar to those above.

Other recent authors have expanded on the work of Zhu et al. 2015 by testing Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models against or in conjunction with the CNN feature extractor [17, 18, 46]. RNNs are another type of neural model that have been demonstrated to effectively extract features from temporal information, as they sequentially process the incoming datastream, effectively “remembering” previous inputs [18]. LSTMs are an improvement on RNNs, which effectively increase their “memory” capacity to enable effective processing of much longer sequences of data [48].

Lopes et al. 2022 compared the feature extraction ability of a CNN directly against an LSTM, each feeding their features into a small FCN regressor, like Zhu et al. 2015. The authors of this study found the CNN features to yield greater accuracy than the LSTM features. However, unlike Zhu et al. 2015, Lopes et al. 2022 utilized much more data, including heart rate data, electromyography (EMG) signals, and information from five separate 6-axis inertial measurement units (IMUs) placed in various locations on the subjects' bodies. Furthermore, the activities performed by the subjects of this study were limited to just standing and walking tasks with and without exoskeleton assistance. These additional datastreams and the studied tasks are feasible and relevant for the study's intended application of exoskeleton design, but they are not applicable to current or near future commercial wearable technology.

Lee and Lee 2024 was the first group to my knowledge to test a CNN-LSTM hybrid model for energy expenditure estimation, with the LSTM placed in series after the CNN. Their design used only IMU data with minimal filtering applied, and while in total they collected data from five separate 6-axis modules placed in various locations on the subjects' bodies, they trained and evaluated their models only on individual IMU signals. Their tasks were limited to walking and running on various inclines of a treadmill. The authors found mixed results depending on which IMU was used for training and evaluation, with the hybrid CNN-LSTM model outperforming the CNN-only and LSTM-only models on the chest, foot, and shank IMU locations, and the CNN-only model returning the lowest errors from the wrist and thigh. Overall they found the worst performance on the wrist in almost all cases, which is unfortunate, given that this is the location at which most commercial devices are designed to be worn.

Given the necessity for widely-available, comfortable, and accurate activity monitors and the limitations in data collection that current commercial devices face, we must find and validate novel techniques for solving the energy expenditure estimation problem with the data available from these devices. This study uses tri-axial accelerometer and heart rate data from a wrist-worn device to estimate energy expenditure, with ground truth EE measured by a compact indirect calorimetry mask. These devices were worn by 11 subjects while engaging in three types of physical activities including resting, stationary cycling, and running on a treadmill. This study aims to test the strategies proposed by recent authors on a dataset that more accurately represents the free living conditions under which such models might be deployed on the commercial market. Specifically I will evaluate the performance of a CNN feature extraction module combined with a FCN regression module. Then I will insert an LSTM module in series with the CNN to investigate any potential for increased accuracy with the hybrid model. I will also approximate the methods described by Amarasinghe et al. 2023 in order to compare the performance of these advanced neural architectures against the well-established XGBoost algorithm with handcrafted features [24].

## Methods

### *Dataset*

This study uses the publicly available Wearable Energy Expenditure Evaluation (WEEE) Dataset provided by Gashi et al. 2022. This dataset includes measurements of 17 participants performing 3 types of physical activity: resting (sitting and standing), cycling (2 speeds), and running (2 speeds). Unacceptable performance was found while evaluating 6 of these subjects' data with preliminary models. After excluding these subjects from the dataset, the rest of the study was performed on the remaining 11. The participants wore a variety of multimodal sensors on various parts of their body, but this study focuses on the data from just two of those devices. The Empatica E4 is a small, watch-like device that was worn by the subjects on their non-dominant wrist [52]. The E4 provides tri-axial accelerometer data measured at 32 Hz and photoplethysmography (PPG) data measured at 64 Hz to estimate heart rate. Photoplethysmography is an optical measurement technique that can detect blood volume changes through the skin and has become a popular method for estimating heart rate [53]. The data streams from the E4 serve as a suitable approximation to the accelerometer and PPG data that is often provided by commercially-available, wrist-worn devices [22, 24, 54]. Subjects also wore a $VO_2$ Master Analyzer, which measured $VO_2$ consumption rates at 1 Hz. This $VO_2$ measurement serves as a basis for the calculation of ground truth energy expenditures [55]. Demographic information about each subject – their height, weight, age, and gender – were also recorded and used as additional features in the regression models, which Paraschiakos et al. 2022 showed to reduce estimation error [16–19, 46].

### *Data Processing*

The PPG raw signals were processed and visually inspected with the Neurokit2 package for Python [56]. This package was used to apply the default cleaning techniques and extract momentary heart rate measurements at 1 Hz. These 1 Hz measurements were merged with the 1 Hz $VO_2$ data by timestamp, and any times with null values or a $VO_2$ measurement of 0 L/min were dropped. Any small gaps (2-9 seconds) in the data for each subject were interpolated, but gaps of 10 seconds or more were left alone. The ground truth energy expenditure was then calculated as EE kcal/min = $4.934 \cdot VO_2$ L/min, which has been validated and widely used throughout the literature [10].

The raw 32 Hz acceleration signals were maintained in frequency, but any times that did not have a matching EE label for that second were dropped. Each feature was standardized according to its mean and standard deviation across the entire dataset. The heart rate and EE values were then resampled to 32 Hz and merged to the acceleration data before applying half-sliding windows of 192 samples (6 seconds) [16]. Before feeding these temporal feature windows into the CNN, the average energy expenditure was calculated as the ground truth label for each 6 second window [16].

To obtain the features for XGBoost, the tsfresh Python package was used to extract the following statistics for the three accelerometer signals and the heart rate signal for each

6 second window: sum_values, median, mean, length, standard_deviation, variance, root_mean_square, maximum, absolute_maximum, and minimum [24, 57]. These feature statistics for each window were then simply concatenated with the constant demographic features before being input into the XGBoost model.

Using the subject IDs, 11 Leave-One-Out Cross Validation (LOOCV) splits were generated, one using each subject's data as the test set [15, 17–19, 21, 43]. The original dataset included 17 subjects, but P03, P04, P06, P11, P13, and P14 returned negative $R^2$ scores on preliminary neural models, and hence their data were entirely excluded moving forward.

### *Model Designs and Training Procedures*

This study primarily evaluated two types of neural model architecture, each implemented with the PyTorch Python package and compared against an XGBoost algorithm implemented with the XGBoost Python package [58, 59]. First, the CNN-FCN (shown in Figure 1) comprised a CNN feature extractor module that fed its outputs into an FCN regression module that then output the final energy expenditure prediction. The Ray Python package was implemented to perform grid search over various model hyperparameters: learning rate, batch size, number of convolution layers, number of fully connected layers, and the hidden size of the fully connected layers [60]. In total 144 parameter combinations were evaluated for the CNN-FCN model, with the optimization goal set to minimize the average validation Mean Squared Error (MSE) over all 11 LOOCV folds. Each convolution layer in the CNN had a Conv1d layer with kernel size 3, stride 1, and padding 1, a ReLU activation layer, and MaxPool1d layer, except the very last convolution layer, which had AdaptiveAvgPool1d instead of MaxPool1d. The 4 demographic features (also standardized, except for the binary gender feature) were then concatenated to the output of the CNN before input to the FCN [16, 17, 46]. Each FCN layer had one Linear layer with the specified number of hidden units, a ReLU layer, and a 50% dropout connection between each FCN layer [17, 46]. A final Linear layer reduced the dimensionality from the hidden size to a single output to represent the energy expenditure prediction. Training was performed with the Adam optimizer with MSELoss as the criterion for a maximum of 100 epochs, but with early stopping implemented with a patience of 20 epochs.

The second model architecture, the CNN-LSTM-FCN (shown in Figure 2), took the best CNN-FCN design described above and simply inserted an LSTM module between the CNN and the FCN. Ray was used to test another 48 hyperparameter combinations for this LSTM layer, defining the number of LSTM layers and the hidden size for each. 50% dropout was implemented during training in between LSTM layers if there were more than one. The demographic features were concatenated to the LSTM output this time before feeding into the FCN, and these models were trained with the same optimizer and epoch scheme.

For each parameter combination, the training and validation MSEs, Root Mean Squared Errors (RMSEs), and $R^2$ scores were averaged over all 11 LOOCV folds and reported to evaluate the models' performances.

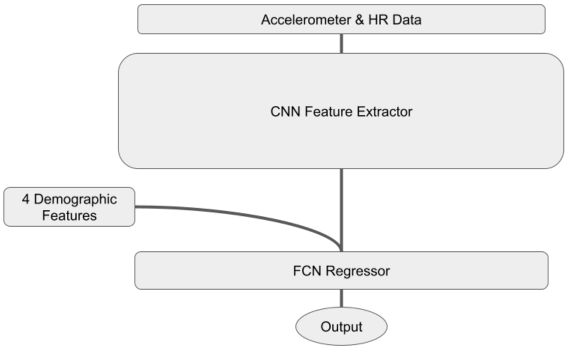

Fig. 1: CNN-FCN architecture.

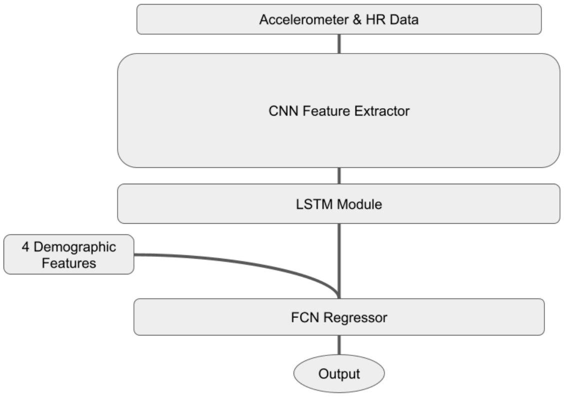

Fig. 2: CNN-LSTM-FCN architecture.

After completing the hyperparameter optimization with LOOCV, the best CNN-FCN and CNN-LSTM-FCN architectures were trained and evaluated on a traditional 80/20 train/test split over the entire dataset in order to gain some insight into how each model might perform when given the opportunity to train on a larger set of data.

The XGBoost model was also optimized using the Optuna Python package, which was set to optimize reg:squarederror over 50 trials [61]. The performance metrics of each trial were averaged over all LOOCV splits while modulating the following hyperparameters: n_estimators, learning_rate, max_depth, min_child_weight, subsample, and colsample_bytree.

## Results and Discussion

The best CNN-FCN model had 3 Conv1d layers and 3 FCN layers of 32 hidden units each. The best learning rate was 1e-2 and the best batch size was 16 windows, with training terminating after an average of 39 epochs across folds. Under LOOCV this model achieved a 5.42 kcal/min (74.0 cal/kg/min) average RMSE and a 0.52 average $R^2$ score. When trained on more data under the 80/20 split, the RMSE improved somewhat to 5.08 kcal/min (69.4 cal/kg/min) and the $R^2$ score improved a great deal to 0.72.

The best CNN-LSTM-FCN had a single LSTM layer with a hidden size of 32. The most effective learning rate was again 1e-2, but this model learned best with a batch size of 32.

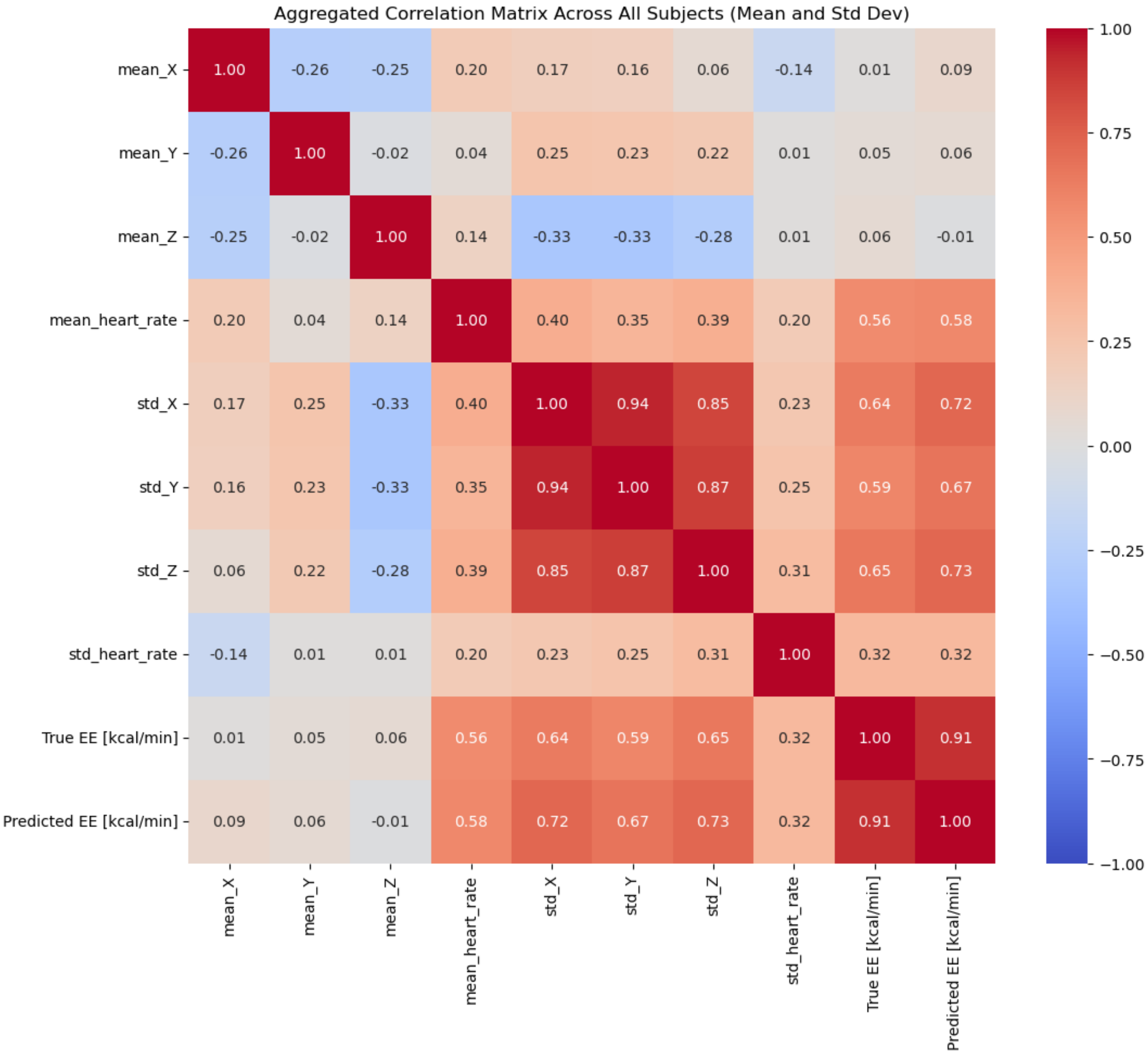

Fig. 3: Correlation matrix of the best CNN-FCN architecture, showing the correlations between the mean and standard deviations of each temporal feature and the true and predicted energy expenditures for each window.

This time early stopping ended training at an average of 38 epochs. Overall I did not find that the LSTM module added any benefit to the CNN-FCN design, with the CNN-LSTM-FCN only achieving 5.49 kcal/min (75.0 cal/kg/min) average RMSE and 0.48 average $R^2$ under LOOCV, and 5.09 kcal/min (69.5 cal/kg/min) RMSE and also 0.72 $R^2$ under the 80/20 split.

The mean and standard deviation of the energy expenditures in the dataset are 3.48 kcal/min and 2.75 kcal/min, respectively. Unfortunately with coefficients of variation (CVs) on the order of 146-158%, these neural models cannot be said to produce results accurate enough for academic nor commercial use. In comparison, the best XGBoost model had 181 estimators, a learning rate of  0.02, a max depth of 5, a minimum child weight of 4,  0.53 subsample, and column sample by tree of  0.80. While this model yielded only 0.11 average $R^2$, it massively reduced the errors in the neural models by achieving 1.02 kcal/min (13.9 cal/kg/min) average RMSE (only 39% CV). Zhu et al. 2015 claimed a similar 1.12 kcal/min overall RMSE with comparable features input to their CNN model. Likewise, Lee and Lee 2024 claimed promising results from their experiment, but their results were reported in terms of the Normalized Root Mean Squared Error (NRMSE) and the Mean Absolute Percentage Error (MAPE), which are difficult to interpret without further knowledge of their data.

After many iterations on the methods presented here to more closely match those presented by Zhu et al. 2015 and Lee and Lee 2024 and to optimize model performance, the most reasonable explanation for the poor accuracies of these neural

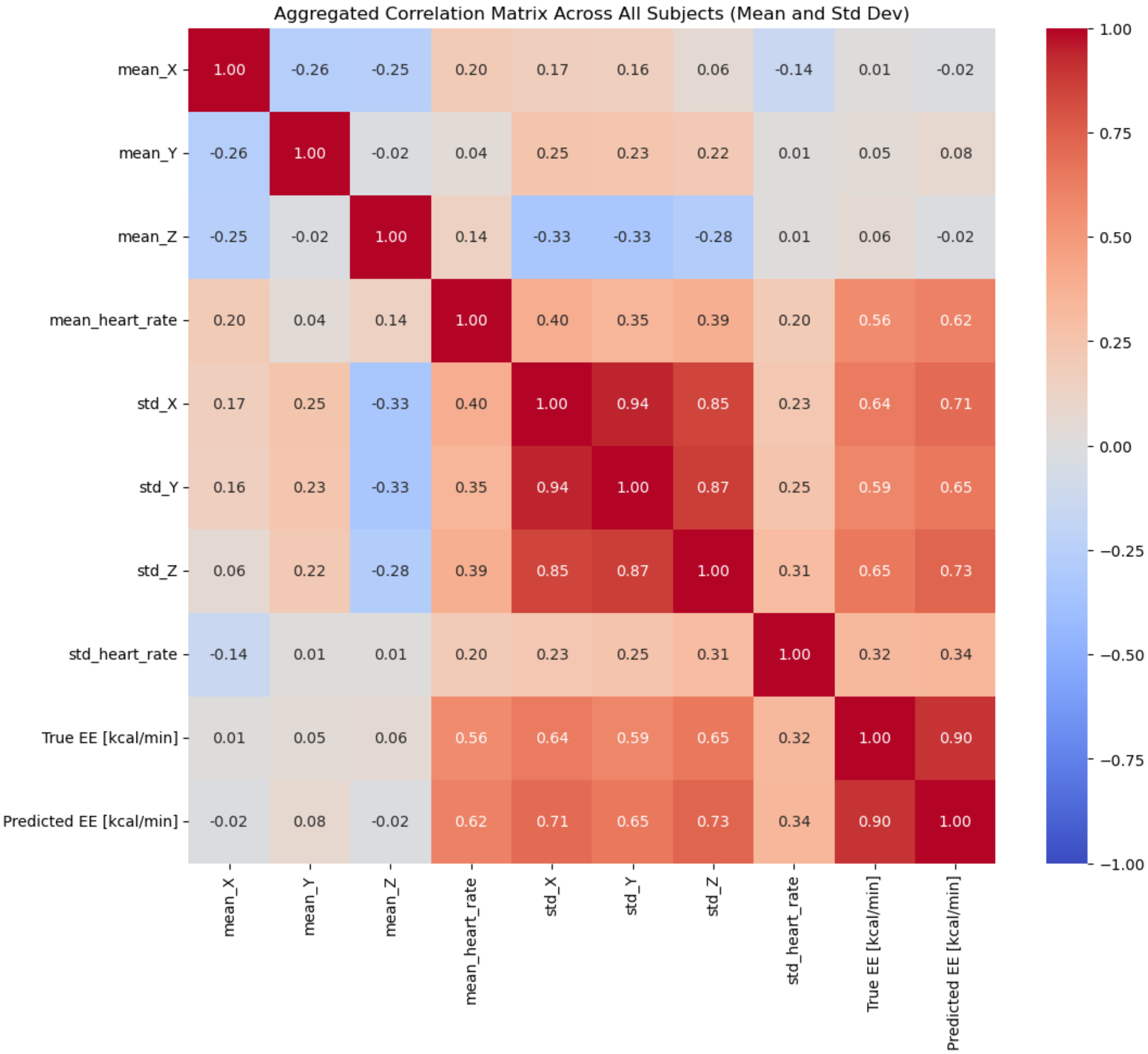

Fig. 4: Correlation matrix of the best CNN-LSTM-FCN architecture, showing the correlations between the mean and standard deviations of each temporal feature and the true and predicted energy expenditures for each window.

models is due to limitations of these models to manage a more complex dataset. The main difference between the data used in these studies and the WEEE Dataset of this study is the inclusion of the stationary cycling activity here. Zhu et al. 2015 only engaged their subjects in “daily ambulatory activities”, and Lee and Lee 2024 limited their subjects’ activity to walking and running on various inclines. Although these datasets presented their own challenges, the monitoring of the cycling activity through predominantly accelerometer data poses a unique obstacle. As indicated by the correlation matrices in Figures 3 and 4, the models learned to depend greatly on the variability of the accelerometer signals for their predictions. However, as shown by low accelerometer standard deviation values during the cycling phase in Figure 5 (roughly the middle third of the graph), the wrists were relatively stable during this exercise, even while heart rate and energy expenditure increased. This decoupling of the accelerometer signals from the EE ground truth likely caused significant confusion during the training of the model, leading to poor outcomes overall. This is also indicated in Figure 5, where the energy expenditure predictions for each subject during the resting phase (roughly the left third of the graph) are almost invariably over-estimations of the true energy expenditure. It is plausible that the higher true EE values during the cycling phase pulled up the walking phase EE predictions due to their very similar accelerometer variances.

Followup work could eliminate the cycling events from the dataset and repeat these training and evaluation methods to verify this hypothesis. In the event that this problem is confirmed, it could highlight a massive challenge in implementing neural models to the task of wearable device energy expenditure estimation. While Amarasinghe et al. 2023 included data from more devices in their XGBoost model, they used the same WEEE Dataset (with the same activities) that was used in this study. My confirmation of their acceptable error scores in comparison to the unacceptable inaccuracies of the neural models indicates that current neural architectures still do not promise any measurable benefit over the established lightweight algorithms like XGBoost, especially when considering the noisy and limited datastreams that are available from consumer devices.

Furthermore the poor accuracy values achieved by the neural models in this study indicate the potentially model-agnostic problems that arise when just a single wrist-independent activity, cycling, is thrown into the mix. Most popular companies currently market their wearable devices' abilities to accurately measure energy expenditure during an extremely wide range of activities. Whoop for example, allows the user to log dozens of activities in their app, ranging from golf to gymnastics to video gaming [62]. However, independent studies repeatedly reveal less than ideal accuracies in commercial devices' energy expenditure estimates, illustrating a clear demand for improvement [54]. This demand suggests the necessity for both further hardware engineering research toward higher resolution, multimodal data collection techniques and additional innovation in data processing and modeling strategies that are compatible with the devices that are appropriate for the consumer market.

## References

[1] C. Koliaki, M. Dalamaga, and S. Liatis, "Update on the obesity epidemic: After the sudden rise, is the upward trajectory beginning to flatten?" *Current Obesity Reports*, vol. 12, no. 6, pp. 514–527, 2023.

[2] K. C. Koskinas, E. M. Van Craenenbroeck, C. Antoniades, M. Blüher, T. M. Gorter, H. Hanssen, N. Marx, T. A. McDonagh, G. Mingrone, A. Rosengren, E. B. Prescott, and the ESC Scientific Document Group, "Obesity and cardiovascular disease: an esc clinical consensus statement," *European Heart Journal*, vol. 45, no. 38, pp. 4063–4098, 08 2024. [Online]. Available: https://doi.org/10.1093/eurheartj/ehae508

[3] Y. Ding, T. Ge, J. Shen, M. Duan, C. Yuan, Y. Zhu, and D. Zhou, "Associations of metabolic heterogeneity of obesity with the risk of dementia in middle-aged adults: Three prospective studies," *Alzheimer's Research & Therapy*, vol. 16, no. 220, 2024.

[4] J. O. Hill, H. R. Wyatt, G. W. Reed, and J. C. Peters, "Obesity and the environment: Where do we go from here?" *Science*, vol. 299, no. 5608, pp. 853–855, 2003. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1079857

[5] W. L. Haskell, I.-M. Lee, R. R. Pate, K. E. Powell, S. N. Blair, B. A. Franklin, C. A. Macera, G. W. Heath, P. D. Thompson, and A. Bauman, "Physical activity and public health: Updated recommendation for adults from the american college of sports medicine and the american heart association," *Circulation*, vol. 116, no. 9, pp. 1081–1093, 2007.

[6] P. Klein, W. James, W. Wong, C. Irving, P. Murgatroyd, M. Cabrera, H. Dallosso, E. Klein, and B. Nichols, "Calorimetric validation of the doubly-labelled water method for determination of energy expenditure in man," *Human nutrition. Clinical nutrition*, vol. 38, no. 2, pp. 95–106, March 1984. [Online]. Available: http://europepmc.org/abstract/MED/6423577

[7] E. L. Melanson, T. Swibas, W. M. Kohrt, V. A. Catenacci, S. A. Creasy, G. Plasqui, L. Wouters, J. R. Speakman, and E. S. F. Berman, "Validation of the doubly labeled water method using off-axis integrated cavity output spectroscopy and isotope ratio mass spectrometry," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 314, no. 2, pp. E124–E130, 2018, pMID: 28978547. [Online]. Available: https://doi.org/10.1152/ajpendo.00241.2017

[8] S. Brage, K. Westgate, P. W. Franks, O. Stegle, A. Wright, U. Ekelund, and N. J. Wareham, "Estimation of free-living energy expenditure by heart rate and movement sensing: A doubly-labelled water study," *PLOS ONE*, vol. 10, pp. 1–19, 09 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0137206

[9] W. R. Leonard, "Laboratory and field methods for measuring human energy expenditure," *American Journal of Human Biology*, vol. 24, no. 3, pp. 372–384, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajhb.22260

[10] K. J. Kaiyala, B. E. Wisse, and J. R. B. Lighton, "Validation of an equation for energy expenditure that does not require the respiratory quotient," *PLOS ONE*, vol. 14, pp. 1–15, 02 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0211585

[11] M. J. Dauncey and W. P. T. James, "Assessment of the heart-rate method for determining energy expenditure in man, using a whole-body calorimeter," *British Journal of Nutrition*, vol. 42, no. 1, p. 1–13, 1979.

[12] G. Spurr, A. Prentice, P. Murgatroyd, G. Goldberg, J. Reina, and N. Christman, "Energy expenditure from minute-by-minute heart-rate recording: comparison with indirect calorimetry," *The American Journal of Clinical Nutrition*, vol. 48, no. 3, pp. 552–559, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002916523162885

[13] S. M. Ceesay, A. M. Prentice, K. C. Day, P. R. Murgatroyd, G. R. Goldberg, W. Scott, and G. B. Spurr, "The use of heart rate monitoring in the estimation of energy expenditure: a validation study using indirect whole-body calorimetry," *British Journal of Nutrition*, vol. 61, no. 2, p. 175–186, 1989.

[14] M. Garet, G. Boudet, C. Montaurier, M. Vermorel, J. Coudert, and A. Chamoux, "Estimating relative physical workload using heart rate monitoring: A validation by whole-body indirect calorimetry," *European Journal of Applied Physiology*, vol. 94, no. 1-2, pp. 46–53, 2005.

[15] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers," *Physiological Measurement*, vol. 35, no. 11, p. 2191, oct 2014. [Online]. Available: https://dx.doi.org/10.1088/0967-3334/35/11/2191

[16] J. Zhu, A. Pande, P. Mohapatra, and J. J. Han, "Using deep learning for energy expenditure estimation with wearable sensors," in *2015 17th International Conference on E-health Networking, Application & Services (HealthCom)*, 2015, pp. 501–506.

[17] J. M. Lopes, J. Figueiredo, P. Fonseca, J. J. Cerqueira, J. P. Vilas-Boas, and C. P. Santos, "Deep learning-based energy expenditure estimation in assisted and non-assisted gait using inertial, emg, and heart rate wearable sensors," *Sensors*, vol. 22, no. 20, p. 7913, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/20/7913

[18] S. Paraschiakos, C. Rebelo de Sá, J. Okai, P. E. Slagboom, M. Beekman, and A. Knobbe, "A recurrent neural network architecture to model physical activity energy expenditure in older people," *Data Mining and Knowledge Discovery*, vol. 36, pp. 477–512, 2022.

[19] J. Staudenmayer, D. Pober, S. Crouter, D. Bassett, and P. Freedson, "An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer," *Journal of Applied Physiology*, vol. 107, no. 4, pp. 1300–1307, 2009, pMID: 19644028. [Online]. Available: https://doi.org/10.1152/japplphysiol.00465.2009

[20] K. Lyden, S. L. Kozey, J. W. Staudenmeyer, and P. S. Freedson, "A comprehensive evaluation of commonly used accelerometer energy expenditure and met prediction equations," *European Journal of Applied Physiology*, vol. 111, no. 2, pp. 187–201, 2011.

[21] A. H. K. Montoye, M. Begum, Z. Henning, and K. A. Pfeiffer, "Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data," *Physiological Measurement*, vol. 38, no. 2, p. 343, jan 2017. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/38/2/343

[22] K. A. Ingraham, D. P. Ferris, and C. D. Remy, "Evaluating physiological signal salience for estimating metabolic energy cost from wearable sensors," *Journal of Applied Physiology*, vol. 126, no. 3, pp. 717–729, 2019, pMID: 30629472. [Online]. Available: https://doi.org/10.1152/japplphysiol.00714.2018

[23] S. Gashi, C. Min, A. Montanari, S. Santini, and F. Kawsar, "A multidevice and multimodal dataset for human energy expenditure estimation using wearable devices," *Scientific Data*, vol. 9, no. 537, pp. 1–14, 2022.

[24] Y. Amarasinghe, D. Sandaruwan, T. Madusanka, I. Perera, and L. Meegahapola, "Multimodal earable sensing for human energy expenditure estimation," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2023, pp. 1–4.

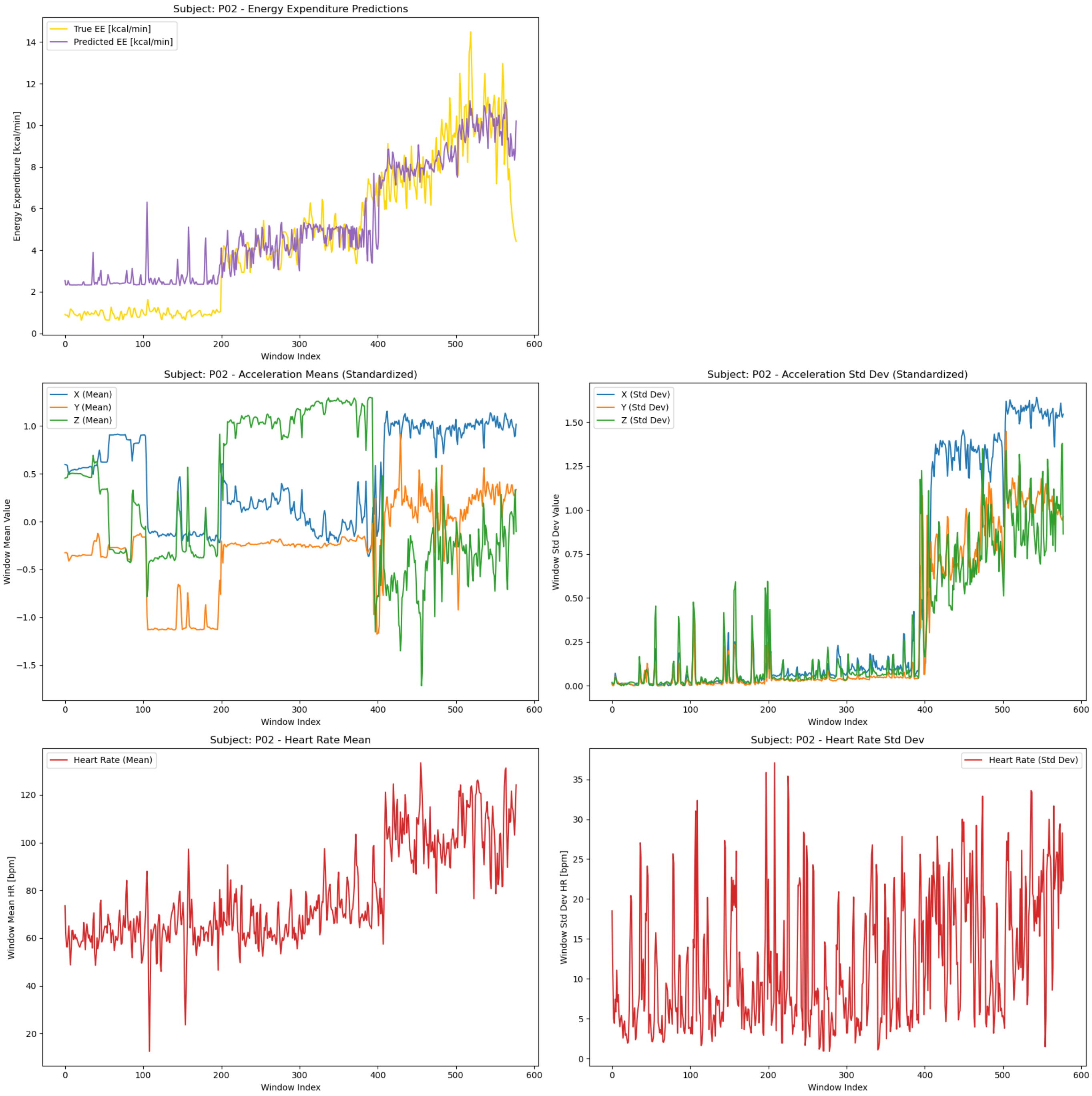

Fig. 5: Time-series graphs of the temporal features and energy expenditure labels and predictions for subject P02. Temporal features include the mean and standard deviation values for each sliding window. Energy expenditures include the ground truch mean expenditure for each window and the CNN-FCN model's prediction. The three activity phases can be seen quite clearly here, with roughly the left third of each graph representing the resting phase (which can also be seen as two distinct sitting and standing sub-phases), roughly the middle third representing the cycling phase, and roughly the right third representing the most vigorous running phase.

[25] A. G. Bonomi, G. Plasqui, A. H. Goris, and K. R. Westerterp, "Estimation of free-living energy expenditure using a novel activity monitor designed to minimize obtrusiveness," *Obesity*, vol. 18, no. 9, pp. 1845–1851, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1038/oby.2010.34

[26] Future Market Insights, "Wearable fitness technology market size, share, trends, industry analysis 2023–2033," n.d., retrieved December 1, 2024, from https://www.futuremarketinsights.com/reports/wearable-fitness-technology-market.

[27] Y. Jiang, C. Spies, J. Magin, S. J. Bhosai, L. Snyder, and J. Dunn, "Investigating the accuracy of blood oxygen saturation measurements in common consumer smartwatches," *PLOS Digital Health*, vol. 2, no. 7, pp. 1–13, 07 2023. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000296

[28] M. Livingstone, A. Prentice, W. Coward, S. Ceesay, J. Strain, P. McKenna, G. Nevin, M. Barker, and R. Hickey, "Simultaneous measurement of free-living energy expenditure by the doubly labeled water method and heart-rate monitoring," *The American Journal of Clinical Nutrition*, vol. 52, no. 1, pp. 59–65, 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000291652316847X

[29] L. Davidson, G. McNeill, P. Haggarty, J. S. Smith, and M. F. Franklin, "Free-living energy expenditure of adult men assessed by continuous heart-rate monitoring and doubly-labelled water," *British Journal of Nutrition*, vol. 78, no. 5, p. 695–708, 1997.

[30] P. Avons, P. Garthwaite, H. Davies, P. Murgatroyd, and W. James, "Approaches to estimating physical activity in the community: calorimetric validation of actometers and heart rate monitoring," *European journal of clinical nutrition*, vol. 42, no. 3, pp. 185–196, March 1988. [Online]. Available: http://europepmc.org/abstract/MED/3383823

[31] W. Haskell, M. Yee, A. Evans, and P. Irby, "Simultaneous measurement of heart rate and body motion to quantitate physical activity," *Medicine and science in sports and exercise*, vol. 25, no. 1, pp. 109–115, January 1993. [Online]. Available: https://journals.lww.com/acsm-msse/abstract/1993/01000/simultaneous_measurement_of_heart_rate_and_body.15.aspx

[32] K. L. Rennie, T. Rowsell, S. A. Jebb, D. Holburn, and N. J. Wareham, "A combined heart rate and movement sensor: Proof of concept and preliminary testing study," *European Journal of Clinical Nutrition*, vol. 54, pp. 409–414, 2000.

[33] S. J. Strath, D. R. Bassett, D. L. Thompson, and A. M. Swartz, "Validity of the simultaneous heart rate-motion sensor technique for measuring energy expenditure," *Medicine and science in sports and exercise*, vol. 34, no. 5, pp. 888–894, May 2002. [Online]. Available: https://journals.lww.com/acsm-msse/Fulltext/2002/05000/Validity_of_the_simultaneous_heart_rate_motion.25.aspx

[34] S. Brage, N. Brage, P. W. Franks, U. Ekelund, and N. J. Wareham, "Reliability and validity of the combined heart rate and movement sensor actiheart," *European Journal of Clinical Nutrition*, vol. 59, no. 5, pp. 561–570, 2005.

[35] H. P. Johansson, L. Rossander-Hulthén, F. Slinde, and B. Ekblom, "Accelerometry combined with heart rate telemetry in the assessment of total energy expenditure," *British Journal of Nutrition*, vol. 95, no. 3, p. 631–639, 2006.

[36] S. Brage, U. Ekelund, N. Brage, M. A. Hennings, K. Froberg, P. W. Franks, and N. J. Wareham, "Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity," *Journal of Applied Physiology*, vol. 103, no. 2, pp. 682–692, 2007, pMID: 17463305. [Online]. Available: https://doi.org/10.1152/japplphysiol.00092.2006

[37] D. Thompson, A. M. Batterham, S. Bock, C. Robson, and K. Stokes, "Assessment of low-to-moderate intensity physical activity thermogenesis in young adults using synchronized heart rate and accelerometry with branched-equation modeling1,2," *The Journal of Nutrition*, vol. 136, no. 4, pp. 1037–1042, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022316622081937

[38] S. Brage, N. Brage, P. W. Franks, U. Ekelund, M.-Y. Wong, L. B. Andersen, K. Froberg, and N. J. Wareham, "Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure," *Journal of Applied Physiology*, vol. 96, no. 1, pp. 343–351, 2004, pMID: 12972441. [Online]. Available: https://doi.org/10.1152/japplphysiol.00703.2003

[39] S. J. Strath, S. Brage, and U. Ekelund, "Integration of physiological and accelerometer data to improve physical activity assessment," *Medicine and science in sports and exercise*, vol. 37, no. 11 Suppl, pp. S563–71, November 2005. [Online]. Available: https://doi.org/10.1249/01.mss.0000185650.68232.3f

[40] M. A. Calabró, J.-M. Lee, P. F. Saint-Maurice, H. Yoo, and G. J. Welk, "Validity of physical activity monitors for assessing lower intensity activity in adults," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 11, p. 119, 2014.

[41] K. Y. Chen and M. Sun, "Improving energy expenditure estimation by using a triaxial accelerometer," *Journal of Applied Physiology*, vol. 83, no. 6, pp. 2112–2122, 1997, pMID: 9390989. [Online]. Available: https://doi.org/10.1152/jappl.1997.83.6.2112

[42] S. E. Crouter, K. G. Clowers, and D. R. Bassett, "A novel method for using accelerometer data to predict energy expenditure," *Journal of Applied Physiology*, vol. 100, no. 4, pp. 1324–1331, 2006, pMID: 16322367. [Online]. Available: https://doi.org/10.1152/japplphysiol.00818.2005

[43] R. O'Driscoll, J. Turicchi, M. Hopkins, G. W. Horgan, G. Finlayson, and J. R. Stubbs, "Improving energy expenditure estimates from wearable devices: A machine learning approach," *Journal of Sports Sciences*, vol. 38, no. 13, pp. 1496–1505, 2020, pMID: 32252598. [Online]. Available: https://doi.org/10.1080/02640414.2020.1746088

[44] M. P. Rothney, M. Neumann, A. Béziat, and K. Y. Chen, "An artificial neural network model of energy expenditure using nonintegrated acceleration signals," *Journal of Applied Physiology*, vol. 103, no. 4, pp. 1419–1427, 2007, pMID: 17641221. [Online]. Available: https://doi.org/10.1152/japplphysiol.00429.2007

[45] A. H. K. Montoye, L. M. Mudd, S. Biswas, and K. A. Pfeiffer, "Energy expenditure prediction using raw accelerometer data in simulated free living," *Medicine and science in sports and exercise*, vol. 47, no. 8, pp. 1735–1746, August 2015. [Online]. Available: https://doi.org/10.1249/MSS.0000000000000597

[46] C. J. Lee and J. K. Lee, "Imu-based energy expenditure estimation for various walking conditions using a hybrid cnn-lstm model," *Sensors*, vol. 24, no. 2, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/2/414

[47] M. Krichen, "Convolutional neural networks: A survey," *Computers*, vol. 12, no. 8, 2023. [Online]. Available: https://www.mdpi.com/2073-431X/12/8/151

[48] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM - a tutorial into long short-term memory recurrent neural networks," *CoRR*, vol. abs/1909.09586, 2019. [Online]. Available: http://arxiv.org/abs/1909.09586

[49] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing*, A. Ferscha and F. Mattern, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 1–17.

[50] A. Mannini, S. S. Intille, M. Rosenberger, A. M. Sabatini, and W. Haskell, "Activity recognition using a single accelerometer placed at the wrist or ankle," *Medicine & Science in Sports & Exercise*, vol. 45, no. 11, pp. 2193–2203, 2013.

[51] A. Graps, "An introduction to wavelets," *IEEE Computational Science and Engineering*, vol. 2, no. 2, pp. 50–61, 1995.

[52] M. Garbarino, M. Lai, D. Bender, R. W. Picard, and S. Tognetti, "Empatica e3 — a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," in *2014 4th International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, 2014, pp. 39–42.

[53] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, p. R1, feb 2007. [Online]. Available: https://dx.doi.org/10.1088/0967-3334/28/3/R01

[54] G. Hajj-Boutros, M.-A. Landry-Duval, A. S. Comtois, G. Gouspillou, and A. D. Karelis, "Wrist-worn devices for the measurement of heart rate and energy expenditure: A validation study for the apple watch 6, polar vantage v and fitbit sense," *European Journal of Sport Science*, vol. 23, no. 2, pp. 165–177, 2023, pMID: 34957939. [Online]. Available: https://doi.org/10.1080/17461391.2021.2023656

[55] A. H. K. Montoye, J. D. Vondrasek, and J. B. I. Hancock, "Validity and reliability of the vo2 master pro for oxygen consumption and ventilation assessment," *International Journal of Exercise Science*, vol. 13, no. 4, pp. 1382–1401, 2020. [Online]. Available: http://www.intjexersci.com

[56] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[57] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh - a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[59] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p. 785–794.

[60] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, and I. Stoica, "Ray: A Distributed Framework for Emerging AI Applications," in *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*. USENIX Association, 2018, pp. 561–577. [Online]. Available: https://www.usenix.org/conference/osdi18/presentation/moritz

[61] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, p. 2623–2631.

[62] WHOOP, "List of whoop activities," October 31 2024, retrieved December 1, 2024, from https://support.whoop.com/s/article/List-of-WHOOP-Activities.