# Deep Multi-Branch Two-Stage Regression Network for Accurate Energy Expenditure Estimation With ECG and IMU Data

Zhiqiang Ni, Tongde Wu, Tao Wang, Fangmin Sun, *Member, IEEE*, and Ye Li, *Senior Member, IEEE*

***Abstract*—*Objective*: Energy Expenditure (EE) estimation plays an important role in objectively evaluating physical activity and its impact on human health. EE during activity can be affected by many factors, including activity intensity, individual physical and physiological characteristics, environment, etc. However, current studies only use very limited information, such as heart rate and step count, to estimate EE, which leads to a low estimation accuracy. *Methods*: In this study, we proposed a deep multi-branch two-stage regression network (DMTRN) to effectively fuse a variety of related information including motion information, physiological characteristics, and human physical information, which significantly improved the EE estimation accuracy. The proposed DMTRN consists of two main modules: a multi-branch convolutional neural network module which is used to extract multi-scale context features from electrocardiogram (ECG) and inertial measurement unit (IMU) data, and a two-stage regression module which aggregated the extracted multi-scale context features containing the physiological and motion information and the anthropometric features to accurately estimate EE. *Results*: Experiments performed on 33 participants show that our proposed method is more accurate and the average root mean square error (RMSE) is reduced by 22.8% compared with previous works. *Conclusion:* The EE estimation accuracy was improved by the proposed DMTRN model with a well-designed network structure and new input signal ECG. *Significance:* This study verified that ECG was much more effective than HR for EE estimation and cast light on EE estimation using the deep learning method.**

***Index Terms*—Convolutional neural network, electrocardiogram, energy expenditure estimation, inertial measurement unit, two-stage regression.**

## I. Introduction

NOWADAYS, with the improvement of human living standards, more and more chronic diseases, including obesity, diabetes, hyperlipidemia, cardiovascular disease, caused by energy metabolism imbalance have become the focus of worldwide concern [1]. Active health management including scientific control of dietary energy intake and physical activity energy expenditure, provides an effective way for chronic diseases prevention and rehabilitation [2]. As the human body is a complex time-varying and nonlinear system, EE during physical activity can be affected by many factors, including activity intensity, individual physiological and psychological state, environment (e.g., temperature, humidity, barometric pressure, etc.), and anthropometric features (e.g., height, weight, age, etc.), which make real-time and accurate EE estimation a challenging study.

Although traditional clinical EE measuring methods, including direct calorimetry [3] and indirect calorimetry [4], have high accuracy, the large size, complex operation, and high cost make them unsuitable for EE measurement under free-living conditions. With the development of microelectronic technology, Micro Electro Mechanical Systems (MEMS) technology and computer technology, wearable devices with powerful sensing and computing functions have been widely used for health and activity monitoring.

However, the EE provided by most commercial wearable devices was computed from heart rate, step count, and anthropometric features. As the information contained in the discrete features is limited, the accuracy of the EE provided by commercial wearable devices is not accurate enough for some applications such as rehabilitation exercise after a heart attack [5] or heart surgery or professional sports training [6]. Besides, a systematic review [7] on the validity and reliability of commercial wearables in measuring energy expenditure published in 2020 concluded that the EE estimation function of the studied commercial wearable devices including Fitbit, Garmin, Polar, Apple Watch, Samsung, etc. were not reliable. Therefore, the

Manuscript received 24 December 2021; revised 17 February 2022; accepted 26 March 2022. Date of publication 30 March 2022; date of current version 20 September 2022. This work was supported in part by Strategic Priority CAS Project under Grant XDB38040200, in part by the National Natural Science Foundation of China under Grants 62073310 and 61902388, in part by the Basic Research Project of Guangdong Province under Grant 2021A1515011838, in part by the Joint Fund of NSFC and Shenzhen under Grant U1913210, in part by the Foundation of National Key Laboratory of Human Factors Engineering under Grant 6142222180502, and in part by the National Key R&D Program of China under Grant 2018YFB1307005. (*Corresponding authors: Ye Li; Fangmin Sun.*)

Zhiqiang Ni is with the Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China, and also with the University of Chinese Academy of Sciences, China.

Tongde Wu is with the AMSU (Shenzhen) New Technology Company Ltd, China.

Tao Wang is with the China Astronaut Research and Training Center, China.

Fangmin Sun and Ye Li are with the Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: fm.sun@siat.ac.cn; ye.li@siat.ac.cn).

Digital Object Identifier 10.1109/TBME.2022.3163429

goal of our research is to propose a new EE estimation method to improve the accuracy of EE estimation through designing new algorithms.

Numerous efforts have been done to improve the accuracy of EE estimation [8]–[24]. However, as most of the existing EE estimation methods were based on machine learning algorithms which need to manually design and select features, their EE estimation accuracy was still unsatisfactory. As the hand-crafted features used for machine learning algorithms are highly dependent on the professional knowledge of the researchers and cannot fully reflect the effective information contained in the raw signal, deep learning algorithms which can automatically extract deep features without any professional knowledge were then proposed for EE estimation [19], [20], [27].

Convolutional neural network (CNN) as one of the most widely used deep learning architecture has been proved to be an effective method to process time series signals in various applications including activity recognition [22], computer aided diagnosis [25], gait analysis [26], etc. The recent studies [19], [20] also proved the promising performance of CNN for EE estimation.

Motion signals collected by IMU sensors and HR calculated from ECG signals were the most used parameters for current EE estimation methods [13]. However, HR contains limited information compared with the raw ECG signal, which leads to a lower EE estimation accuracy of the current HR-based EE estimation methods. With the development of deep learning algorithms, comprehensive and deep-level features of ECG signals can be learned and extracted automatically.

Based on the research state and application requirement of EE estimation methods, we explored the feasibility of improving the EE estimation accuracy for application scenarios like clinical rehabilitation exercises and professional sports training monitoring by fusing multiple information. The main contributions of this study are summarized as follows:

1) Proposed a deep multi-branch CNN for automatic multi-scale feature extraction. The proposed feature extractor integrated with multiple CNN branches with different kernel sizes, by which multi-scale information was extracted from the input ECG and inertial signals.
2) Proposed a novel two-stage regression method to accurately predict EE. A soft label based ordinal regression method was first designed to realize a coarse-grained estimation of EE, then a linear regression method was implemented to further optimize the EE estimation output from the first stage.
3) To the best of our knowledge, this is the first work to make use of raw ECG signals instead of HR for high-accuracy EE estimation.
4) The experiments were performed to study the contribution of different input signals to the EE estimation model and verified that the raw ECG signal could contribute more to the performance improvement of the EE estimation in comparison with HR.

The rest of this paper is organized as follows: EE estimation related studies were first reviewed in Section II; The proposed DMTRN model was introduced in Section III; The designed performance evaluation experiments and corresponding test results were introduced and analyzed in Section IV; Performance discussion of the proposed model was given in Section V; Finally, Section VI concluded the whole paper.

## II. Related Works

In recent years, with the increasing application requirements of accurate EE estimation, many works have been done to improve EE estimation performance.

Motion signals collected by IMU sensors were first used for EE estimation. The earliest attempting works were proposed by Montoye et al [8] and Chen et al [9], in their studies, acceleration signals of a single fixed sensor were used to estimate EE through a linear regression model. Considering the EE level varies with physical activities, multiple regression models were more suitable for EE estimation. Choi *et al.* [10] proposed a multiple linear regression method to estimate EE during walking and running respectively. Crouter *et al.* [11] proposed a two-regression model to improve the EE estimation accuracy by recognizing physical activities firstly.

With the further understanding of the factors affecting EE, physiological parameters including HR, heart rate variability (HRV) were fused with motion signals to estimate EE. Charlot *et al.* [12] improved the accuracy of EE estimation during running by using the anthropometric parameters, HR, and running speed as the model input. Brage *et al.* [13] used HR and acceleration signals to predict EE. Their findings suggested EE estimation performance using both acceleration and HR outperforms that using either of the parameters. To reduce the effect of the inter-individual physiological differences on EE estimation accuracy, Altini *et al.* [14] proposed an HR normalization method and used the normalized HR, activity intensity, anthropometric characteristics to estimate EE.

Moreover, study [15] further suggested that ECG can provide additional information for better prediction of EE. They not only calculated heart rate from ECG but also calculated various indicators of heart rate variability (HRV) as predictors. The results showed that adding the HRV to the input parameters can improve the EE estimation accuracy. Inspired by their results, we speculate that in addition to HR and HRV, there is still other more valuable information in raw ECG signals. As a result, the raw ECG signals were taken as the input of the proposed EE estimation model.

In terms of algorithms, more and more machine learning nonlinear models were explored for EE estimation recently with the development of artificial intelligence technology. Staudenmayer *et al.* [16] proposed two artificial neural networks (ANN) for physical activity recognition and EE estimation respectively. Catal *et al.* [17] combined the boosted decision tree regression (BDTR) algorithm and the median aggregation algorithm to improve the EE estimation accuracy. Cvetković *et al.* [18] proposed a real-time activity monitoring and EE estimation algorithm with a smartphone and a wristband using the random forest (RF) algorithm which took the variations of sensors' location and orientation into considerations. However, as the manually

TABLE I
PARTICIPANTS STATISTICAL CHARACTERISTICS

| | |
|---|---|
| Number of Subjects | 33 (21 males, 12 females) |
| Age(years) | 26.7±4.0 |
| Height(cm) | 171.8±7.7 |
| Weight(kg) | 65.6±11.2 |
| Waistline(cm) | 78.5±9.1 |

designed and selected features contain very limited information, the machine learning model has low EE estimation accuracy.

Zhu *et al.* [19] were the first who proposed a deep learning method for EE estimation, raw acceleration signals were input to a CNN to estimate EE without any feature extraction and selection steps. Their experimental results showed that the CNN achieved a significant improvement in EE estimation performance compared to the activity-specific linear regression model and the ANN model. Also, the long short-term memory network [27] has also been applied in EE estimation. Nevertheless, the performances of these models still have room for improvement due to the simple network architecture and using simple HR as physiological state input.

According to the review of the previous related studies, it can be inferred that there may be more deep features in the raw ECG signals related to EE other than HR and HRV. Based on this hypothesis, we developed a deep learning architecture named DMTRN which used the raw ECG and 6-axis inertial signals for accurate EE estimation. Through ablation experiments, we verified the effectiveness of the raw ECG signal for EE estimation. Besides, the superior performance of the proposed DMTRN method was also verified through comparative studies with previous works.

## III. MATERIALS AND METHODS

### A. Data Collection

A total of 33 healthy participants were recruited to participate in the experiments, the statistical anthropometric characteristics of all participants were summarized in Table I. During the experiment, the room temperature was maintained between 25 degrees Celsius and 26 degrees Celsius. The participants were asked to do a modified Bruce treadmill test [28] to collect their EE data at different activity intensity levels ranging from rest to the individual's maximum activity intensity. The experiment process was shown in Table II, it started with a 5-minute pre-rest, during which the participants were asked to stand still on the treadmill. Then follows the exercise stage, during this stage, the participants began to run at a speed of 3 km/h, and the speed increases to the next preset value every 5 minutes until reaches the maximum preset speed (11.6 km/h), and the participants would run at this maximum preset speed until they were physically exhausted. It was not necessary to reach the maximum speed during the exercise stage and the exercise can be terminated at any time when the HR of the participant reached the maximum HR or the participant signaled that he was exhausted. After the exercise stage, a 3-minute recovery stage and a 3-minute post-rest stage followed.

TABLE II
INFORMATION OF THE MODIFIED BRUCE TREADMILL TEST

| Stage | Duration (min) | Treadmill speed (km/h) | Treadmill incline (%) |
|---|---|---|---|
| Pre-rest | 5 | 0 | 0 |
| Ex-1 | 5 | 3 | 5 |
| Ex-2 | 5 | 5 | 5 |
| Ex-3 | 5 | 6.4 | 5 |
| Ex-4 | 5 | 7.8 | 5 |
| Ex-5 | 5 | 10.2 | 5 |
| Ex-6 | Until exhausted | 11.6 | 5 |
| Recover | 3 | 3 | 0 |
| Post-rest | 3 | 0 | 0 |

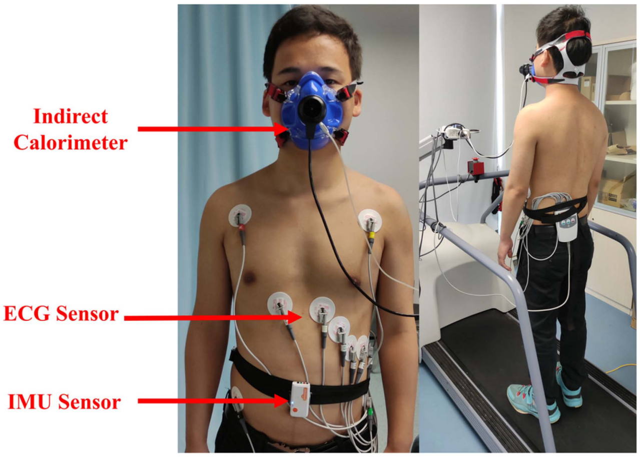

Fig. 1. The scenario of the data collection experiment.

The scenario of the data collection experiment is shown in Fig. 1. The participants were asked to wear 12-lead ECG sensors (GE Medical System Information Technologies, INC, Cardiac Testing System) on their body and an IMU (Inertial Measurement Unit) sensor (Shimmer, Shimmer3 IMU unit) on their waist. An indirect calorimeter (MasterScreen CPX, Jaeger, Germany) with a mask worn on the participant's face was used to collect the reference EE. Considering the high quality of signal in lead v3 of 12 leads, we decided to use v3-lead ECG as the input ECG data. The sampling rate of the indirect calorimeter, IMU sensor, and ECG sensor were 0.2 Hz, 100 Hz, and 200 Hz respectively. Each participant participated in at least 1 session and at most 3 sessions (with the interval of 1 week) data acquisition experiments. Each session lasts about 30 minutes, and a total of 60 sessions were collected.

The study was approved by the Institutional Review Board of Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. All participants signed the written informed consent before the experiments.

### B. Preprocessing

Firstly, interpolation was used to deal with some missing data. Then some filters were used to eliminate noise from the collected data. For IMU data, a Butterworth low-pass filter with a 10 Hz cutoff frequency and a Wiener filter [28] with a window size of 1 second were used. For ECG data, a low-pass filter with a cutoff

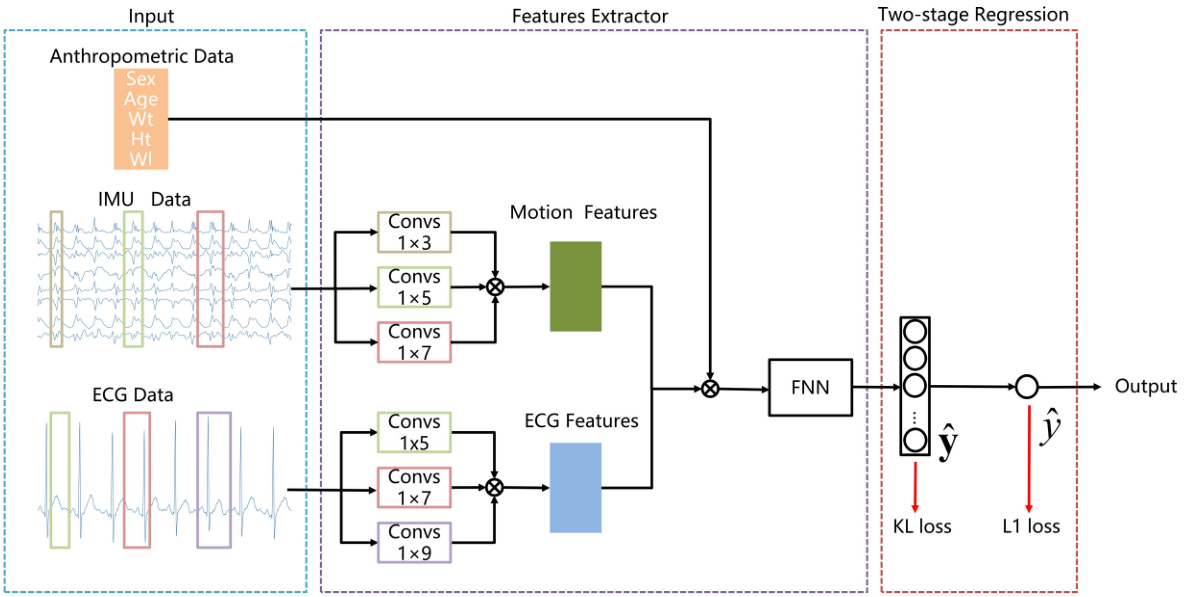

Fig. 2. Illustration of the network architecture. The network consists of a feature extractor and a two-stage regression module. The Convs components in the feature extractor module are branches with different kernel size. The two multi-branch CNNs of feature extractor module extract motion features and ECG features from IMU data and ECG data. The two-stage regression module generates the final EE based on the extracted features and the anthropometric features.

frequency of 50 Hz, and a nine-level wavelet decomposition were used.

Next, to reduce the effects of inertial sensor position changes on the EE estimation performance, the magnitude vectors of accelerometer and gyroscope signals were calculated using methods proposed by [29] and were used as the IMU data input combined with the 6-axis raw signals.

Previous studies [24] showed that the longer the sliding window used, the smaller the EE estimation error. Besides, according to the test results we found that the EE of the human body fluctuates little in one minute indicating and 1-minute window has been used in [21], [30], [31]. In order to balance the real-time and the accuracy of the model, our study adopted the 1-minute sliding window. Therefore, a 1-minute sliding window without overlap was applied on IMU data, ECG data, and reference EE data respectively for data segmentation. After the segmentation, IMU input vectors with a size of 6000×8, ECG input vectors with a size of 12000×1, and reference EE vectors with a size of 12×1 were obtained. The IMU input vectors and the ECG input vectors were directly fed into the IMU CNN branch and ECG CNN branch of the proposed model respectively, and the corresponding average values of the EE reference vectors were used for the final EE reference labels.

Moreover, five anthropometric features including sex, age, height, weight, and waistline, were also inputted into the proposed model after standardization and one-hot encoding.

**Algorithm 1:** DMTRN for EE Estimation.

```
Input: training data($X_{ECG}$, $X_{IMU}$, $X_{ANT}$, $y_i$) and testing
  data ($X'_{ECG}$ , $X'_{IMU}$, $X'_{ANT}$)
Output: the final predicted EE $\hat{y}'$ for testing data
# Training phase
Initialize hyperparameter $K$ and $\lambda$
Initialize feature extractor's weights $\mathbf{W}_{\text{extract}}$ and
  two-stage regression's weights $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$, $b_2$
Initialize maximum iterated epochs $N$
for $k = 1 \rightarrow N$ do
  Load $X_{ECG}$, $X_{IMU}$, $X_{ANT}$, $y_i$
  Calculate soft label vector $\mathbf{y}_i$ based on Eq. (1) and
    Eq. (2)
  Extracted features $\mathbf{X}$ based on Eq. (3) through the
    feature extractor
  Calculate predicted EE $\hat{\mathbf{y}}$ and $\hat{y}$ in two stages based
    on Eq. (4)
  Update the weights of DMTRN with the total loss
    function of Eq. (7)
end for
# Testing phase
Load $X'_{ECG}$ , $X'_{IMU}$, $X'_{AN}$
Load the trained weights of DMTRN
Calculate the final predicted EE $\hat{y}'$ based on Eq. (3)
  and Eq. (4)
```

## C. Proposed Method

The architecture of the proposed network model DMTRN is shown in Fig. 2, and the pseudo-code describing the process of the algorithm is shown in Algorithm 1. In this section, the proposed EE estimation method was introduced in detail. Firstly, the overall structure of the feature extractor for automatic feature extraction was introduced. Then, how to embed the two-stage regression module into a deep regression model was explained.

TABLE III
THE ARCHITECTURE AND PARAMETERS OF THE BRANCH WITH KERNEL SIZE K

| Branch | Layer | Kernel | Stride | Output size |
|---|---|---|---|---|
| ECG | Input | - | - | 6000×8 |
| | conv | 1×k×16 | 1 | 6000×16 |
| | conv | 1×k×16 | 1 | 6000×16 |
| | maxpool | - | 5 | 1200×16 |
| | conv | 1×k×32 | 1 | 1200×32 |
| | conv | 1×k×32 | 1 | 1200×32 |
| | maxpool | - | 5 | 240×32 |
| | conv | 1×k×64 | 1 | 240×64 |
| | conv | 1×k×64 | 1 | 240×64 |
| | maxpool | - | 3 | 120×64 |
| | conv | 1×k×128 | 1 | 120×128 |
| | conv | 1×k×128 | 1 | 120×128 |
| | maxpool | - | 2 | 40×128 |
| | avgpool | - | 40 | 1×128 |
| IMU | Input | - | - | 12000×1 |
| | conv | 1×k×16 | 1 | 12000×16 |
| | conv | 1×k×16 | 1 | 12000×16 |
| | maxpool | - | 5 | 2400×16 |
| | conv | 1×k×32 | 1 | 2400×32 |
| | conv | 1×k×32 | 1 | 2400×32 |
| | maxpool | - | 5 | 480×32 |
| | conv | 1×k×64 | 1 | 480×64 |
| | conv | 1×k×64 | 1 | 480×64 |
| | maxpool | - | 3 | 160×64 |
| | conv | 1×k×128 | 1 | 160×128 |
| | conv | 1×k×128 | 1 | 160×128 |
| | maxpool | - | 2 | 80×128 |
| | conv | 1×k×128 | 1 | 80×128 |
| | conv | 1×k×128 | 1 | 80×128 |
| | maxpool | - | 2 | 40×128 |
| | avgpool | - | 40 | 1×128 |

***1) Feature Extractor:*** Two multi-branch CNNs were designed for motion and ECG features extraction respectively. Each multi-branch convolutional neural network contains three branches, and each branch employs convolutional kernels of different sizes. Since convolutional kernels of different sizes can capture information of different time scales, the multi-scale context features can be extracted through our proposed multi-branch CNNs.

The architecture and parameters of the specific branch with kernel size k are shown in Table III. The branch for ECG feature extraction consists of 8 convolution layers and 5 pooling layers, while the branch for motion feature extraction consists of 10 convolution layers and 6 pooling layers. For motion features extraction, the kernel sizes of the three branches were 3, 5, and 7, respectively; the kernel sizes of the three branches for ECG feature extraction were set 5, 7, and 9 respectively. ReLU [32] was used as the activation function, and batch normalization [33] was used to alleviate the problem of internal covariate migration and speed up the training process after each convolution layer, dropout layer [34] was added to prevent overfitting. At each layer, multiple feature maps were generated according to the specified number of filters and subsequently were fed into the next layer, deep-level features were finally learned from the feature extractor by cascading the layers.

Furthermore, we combined deep-level features extracted by the multi-branch CNNs with the anthropometric features through a feed forward neural network (FNN) containing a hidden layer with 128 neural units, which improves the generalization ability of the model for estimating EE of different subjects.

***2) Two-Stage Regression:*** The essence of ordinal regression [35] is to transform an ordinal regression task into a multi-class classification task through label discretization. With the increasing development and improvement of deep learning techniques, ordinal regression is attracting more and more attention and has been successfully applied in age estimation [36], depth estimation [37], head pose estimation [38], etc. combined with CNN.

For the first stage regression, ordinal regression was used to estimate a coarse-grained EE. The uniform discretization method was used to quantize a continuous EE value into a discrete value.

When a continuous EE interval $[a, b]$ is divided into $K$ equal parts, the discrete rank is defined as:

$$r_i = \left\lfloor \frac{(K-1)(y_i - a)}{b - a} \right\rfloor \tag{1}$$

where $y_i$ is the value of the *i-th* sample of reference EE, $r_i$ is the corresponding discretization result, and $\lfloor\rfloor$ is a down rounding function.

The discrete interval value $r_i$ of reference EE was then encoded as a soft label vector $\mathbf{y}_i$ [39] with the dimension of $1\times K$. The *j-th* element in the vector is defined as

$$\mathbf{y}_{ij} = \frac{e^{-\phi(r_i, r_j)}}{\sum_{k=1}^{K} e^{-\phi(r_i, r_k)}} \ \forall\, r_j \in [r_1, r_2, \ldots, r_K] \tag{2}$$

where $\phi(r_i,\ r_j)$ is an absolute distance between a particular reference discrete value $r_i$ and the discrete rank $r_j$.

Let $\mathbf{X}$ denote the features from the feature extractor,

$$\mathbf{X} = \Phi\left(X_{ECG}, X_{IMU}, X_{ANT}, \mathbf{W}_{extract}\right) \tag{3}$$

where $X_{ECG}$ refers to ECG data, $X_{IMU}$ refers to IMU data, $X_{ANT}$ refers to anthropometric data, $\Phi$ refers to the learned mapping from the feature extractor and $\mathbf{W}_{\text{extract}}$ refers to the weights in the feature extractor.

As can be seen from Fig. 2, the mapping from the features $\mathbf{X}$ to the final prediction of EE $\hat{y}$ can be divided into two stages: the first stage predicted EE discrete distribution $\hat{\mathbf{y}}$, and the second stage predicted EE continuous value $\hat{y}$. In detail, the whole process can be expressed as Eq. (4):

$$\begin{aligned} \hat{\mathbf{y}} &= g(\mathbf{W_1}, \mathbf{X}) + \mathbf{b_1} \\ \hat{y} &= \mathbf{W_2}\sigma(\hat{\mathbf{y}}) + b_2 \end{aligned} \tag{4}$$

where $\mathbf{W}_1$, $\mathbf{b}_1$, and $\mathbf{W}_2$, $\mathrm{b}_2$ are learned weights in the two regression stages respectively, $g$ is the mapping in the first regression stage, and $\sigma$ is the activation function ReLU.

Furthermore, two losses are defined for the proposed two-stage regression. The first loss is used to measure the discrepancy between the predicted EE distribution and the reference EE distribution, and finally controls the interval classification accuracy of EE. We adopt KL-Divergence

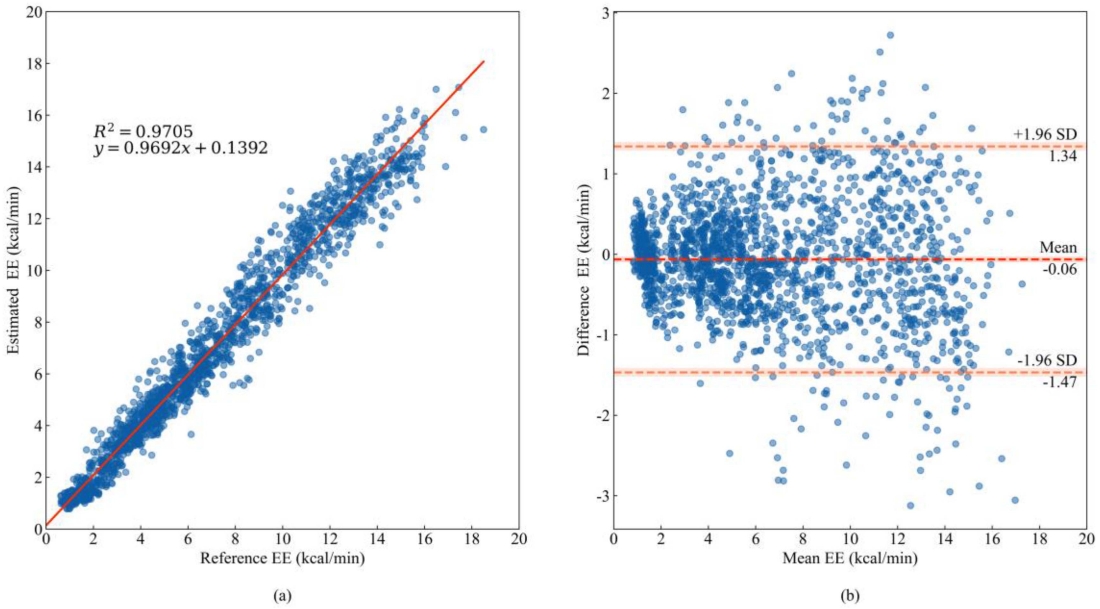

Fig. 3. Correlation (a) and Bland–Altman (b) plots comparing the estimated EE with reference EE.

(Eq. (5)) as the first loss function,

$$L_{ord} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbf{y}_{ij}log\frac{\mathbf{y}_{ij}}{\hat{\mathbf{y}}_{ij}} \qquad (5)$$

where $N$ is the total number of samples.

The second loss controls the prediction accuracy of the final EE and L1 loss (Eq. (6)) was adopted in this study.

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad (6)$$

During the training phase, both two losses of the two regression stages were merged through Eq. (7) into a total loss to train the whole model,

$$L = \lambda L_{ord} + L_{reg} \qquad (7)$$

where λ is the hyperparameter used to balance the contributions of two losses to the model in the two stages.

## IV. Experiments and Results

Extensive experiments were performed to verify and evaluate the proposed DMTRN for accurate EE estimation. Firstly, detailed ablation studies on the collected dataset were performed to verify the effectiveness of the two-stage regression module, the multi-branch module, and the extracted features. Then the EE estimation performance of our proposed model was compared with previous studies.

### A. Implementation Details

A 10-fold cross validation on the collected dataset was performed to evaluate the performance of the proposed methods. The average performance of the 10 iterations was used as the final results. Root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to evaluate the performance of the model.

Data augmentation not only can effectively increase the number of samples and enhance the generalization ability of the model but also can add random noise to the datasets and improve the robustness of the model. Two data augmentation techniques were used in this study to improve the model performance:1) Multiply the amplitude of IMU data and ECG data with a random scalar drawn from a Gaussian distribution with mean 1 and standard deviation 0.1 to change the amplitude randomly [40]; 2) Swap the 3-axis of accelerometer data or gyroscope data with random permutations and rotate them by a random angle to simulate scenarios where inertial sensors were placed on different body locations [40].

We used the deep learning framework PyTorch [41] to build the proposed model. Adam optimizer [42] was used in the training process. The maximum number of epochs was 50, and the batch size was set to 64. The initial learning rate and momentum were set to 0.001 and 0.9 respectively.

### B. Overall Performance

The overall performance of our proposed DMTRN was presented by the Correlation analysis plots and the Bland–Altman plots of the test results of the 10-fold cross validation in Fig. 3. In the Correlation plot, most of the points were lie closely to the red line, indicating a close correlation ($R^2 = 0.97$) between the estimated EE and the reference EE. In the Bland–Altman plot, more than 95% of the points lie within the limit of agreement in EE evaluation, suggesting a high EE estimation accuracy of our proposed model.

### C. Ablation Studies

***1) Multi-Branch Module and Two-Stage Regression Module:*** In order to evaluate the effect of the proposed

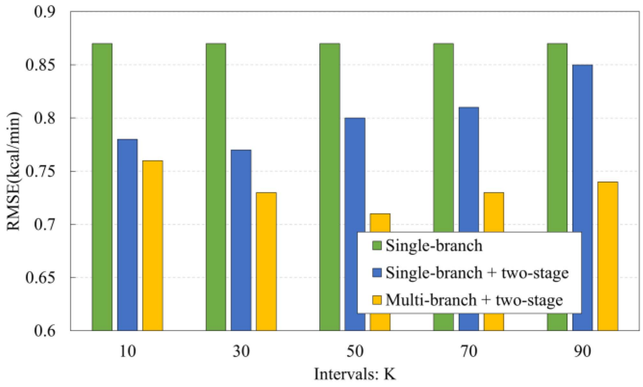

Fig. 4. Evaluation of multi-branch and two-stage regression module.

TABLE IV
PERFORMANCE EVALUATION WHEN HYPERPARAMETER WAS SET TO DIFFERENT VALUES

| $\lambda$ | RMSE (kcal/min) | MAE (kcal/min) | MAPE (%) |
|---|---|---|---|
| 0.1 | 0.75 | 0.56 | 10.67 |
| 1 | 0.71 | 0.53 | 10.35 |
| 5 | 0.81 | 0.59 | 11.65 |
| 10 | 0.90 | 0.68 | 13.37 |

multi-branch module and two-stage regression module on the EE estimation performance respectively, we set up three models with different feature extraction and regression modules: (1) Single-branch: neither of the proposed multi-branch module nor the two-stage regression module was used in this model; (2) Single-branch + two-stage: only the proposed two-stage regression module was used in the model; (3) Multi-branch + two-stage: both the proposed multi-branch module and the two-stage regression module were used.

As the results in Fig. 4 shown, the model with the two-stage regression module yields a lower EE estimation RMSE than the model without the two-stage regression module, which proved the effectiveness of the proposed two-stage regression methods. Besides, the performance of the model has been further improved when substituted single-branch module with our proposed multi-branch module, which verified that the features extracted by the multi-branch CNNs have higher quality than those extracted by the single-branch CNNs.

Fig. 4 also illustrated the sensitivity of the proposed model to the EE discretization intervals *K* in the first regression stage. When *K* increases from 10 to 90, the EE estimation RMSE of our model ranges from 0.71 kcal/min to 0.76 kcal/min, indicating DMTRN's good robustness to a long range of discrete EE interval numbers. As too few discretizations intervals would cause large quantization error of the first-stage regression, while too large intervals would reduce the effects of the first-stage regression, one can also see that the RMSE increased when *K* was set smaller or larger than 50.

Further, we studied the effect of hyperparameter $\lambda$ on the EE estimation performance. The RMSE, MAE, and MAPE of EE estimation were evaluated when $\lambda$ were set to 0.1, 1, 5, 10. The test results listed in Table IV showed that the best performance was achieved when $\lambda = 1$. Since $\lambda$ adjusted the contributions of the two regression tasks in the two stages, too small or too large $\lambda$ could break the balance of their contributions.

TABLE V
PERFORMANCE EVALUATION OF DIFFERENT INPUT DATA

| NO. | Input data | RMSE (kcal/min) | MAE (kcal/min) | MAPE (%) |
|---|---|---|---|---|
| 1 | IMU | 1.09 | 0.81 | 14.99 |
| 2 | ECG | 1.16 | 0.87 | 18.25 |
| 3 | IMU + HR | 0.88 | 0.66 | 13.03 |
| 4 | IMU + ECG | 0.79 | 0.60 | 11.88 |
| 5 | ANT + IMU + HR | 0.84 | 0.61 | 11.74 |
| 6 | ANT + IMU + ECG | 0.71 | 0.53 | 10.35 |

In the following experiments, the discretization interval *K* was set to 50 and the hyperparameter $\lambda$ was set to 1 if there was no special declaration.

***2) Input Data:*** Many parameters including anthropometric data (ANT), inertial data (IMU), ECG data (ECG), heart rate (HR) were considered to be related to EE. In this section, we studied the effects of different input data on the proposed EE estimation model. The test results were listed in Table V. First, we can see from NO.1, 2, and 4 that using IMU or ECG alone as the model input leads to inferior EE estimation performance to the combination of ECG and IMU as the model input. Then, by further adding ANT to the model input we can see form NO.6 that the anthropometric features are useful for EE estimation performance improvement, although the improvement is not very significant.

To compare the contribution of ECG and HR to the EE estimation model, we deleted the ECG feature extractor module from our DMTRN and used the manually calculated HR to replace the extracted ECG features. The comparison results of NO.3 & 4 and NO. 5 & 6 proved the superiority of ECG to HR on EE estimation.

## *D. Comparison Studies*

In order to verify the advantage of our proposed EE estimation method, we compared our method with other machine learning or deep learning algorithms including linear regression (LR) [15], [43], boosted decision tree regression (BDTR) [17], extreme gradient boosting (XGBoost) [21], random forest (RF) [18], convolutional neural network (CNN) [19] and densely connected convolutional network (DenseNet) [20] on our dataset. For machine learning algorithms, anthropometric features, motion features designed by [18], and HRV features designed by [15] were used to train the model with default parameters.

The test results were shown in Table VI, we can see from the results that RF had the best performance among all the compared algorithms. However, compared with the best algorithm, our proposed DMTRN model reduced the EE estimation error by 22.8% in terms of RMSE respectively.

Aimed at further evaluating the superiority of the proposed model, we also compared our method with other related studies. For easy comparison of various methods, various EE units and

TABLE VI
COMPARISON WITH OTHER MACHINE LEARNING OR DEEP LEARNING ALGORITHMS ON OUR DATASET

| Reference | Method | Feature type | RMSE (kcal/min) | MAE (kcal/min) | MAPE (%) |
|---|---|---|---|---|---|
| [15] [43] | LR | hand-crafted | 1.30 | 0.97 | 21.84 |
| [17] | BDTR | hand-crafted | 1.22 | 0.98 | 27.76 |
| [21] | XGBoost | hand-crafted | 0.93 | 0.64 | 11.32 |
| [18] | RF | hand-crafted | 0.92 | 0.65 | 11.82 |
| [19] | CNN | learned | 1.54 | 1.18 | 22.66 |
| [20] | DenseNet | learned | 1.23 | 0.92 | 17.88 |
| This work | DMTRN | learned | 0.71 | 0.53 | 10.35 |

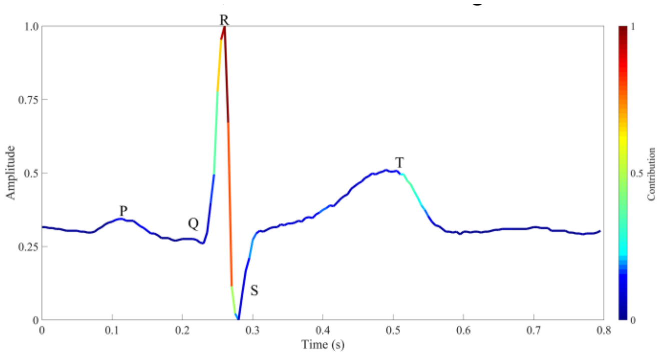

Fig. 5. Visualization of the features learned mapped to ECG from proposed model.

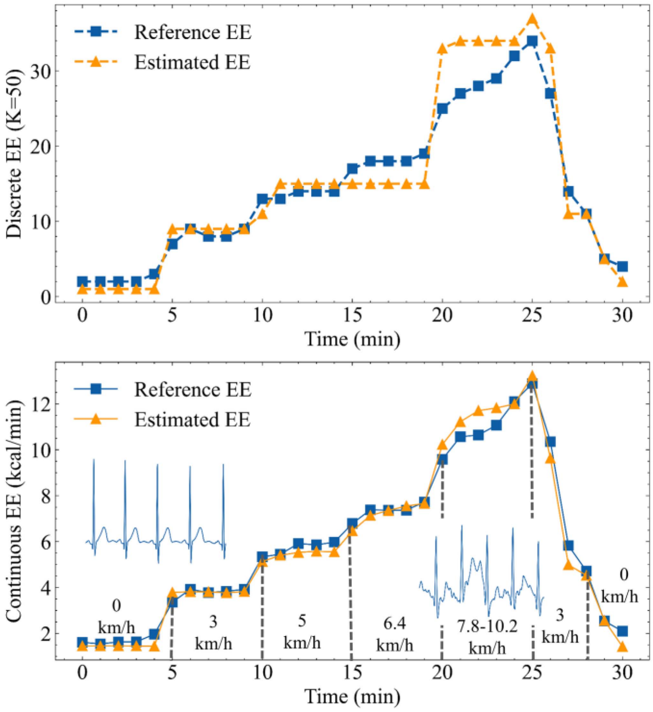

Fig. 6. EE tracking performance. Top: EE estimation results after the first stage regression. Bottom: EE estimation results after the second stage regression.

evaluation metrics were used. Comparison results of the input data, method, and EE estimation performance were listed in Table VII. Obviously, one great advantage of our model is that it is based on ECG and IMU, while other studies are usually based on HR and IMU. As the raw ECG signals contain more EE related information and the proposed DMTRN model can estimate EE more accurately. We can observe that the proposed DMTRN model achieves state-of-the-art performance in terms of RMSE, MAE, and MAPE compared with previous studies.

## V. Discussion

### A. Feature Visualization

For a better understanding of the properties of the proposed model, we visualized the learned features by mapping them to the raw ECG signals using the guided backpropagation approach [44]. As can be seen from Fig. 5, the contribution of every part of the ECG signals to the final estimation of EE is presented in different colors. The closer the color is to dark red, the greater the signal contributes to the EE estimation model, while the closer the color is to dark blue, the less.

First, it is obvious that the R wave of the ECG signal attracts most of the attention, and its contribution is greater than that of other parts. This can be explained by the fact that the proposed deep learning model learned and extracted the HR related features, which is closely related to the task of EE estimation. Besides the R wave, the T wave also has a part of the contribution to the model. This indicates that the model not only learned the HR related information but also learned morphological information near the T wave. Previous research [45] has found that the amplitude of T wave decreased significantly during exercise and increased significantly after exercise, which probably represented the anoxic and anaerobic myocardial metabolism. Therefore, the T wave plays an auxiliary role in the EE estimation, which also explains why the raw ECG signal is superior to HR in EE estimation.

### B. The Tracking of Individual EE Changes

To evaluate the performance of the proposed model in tracking the EE changes, we provided our model estimated EE and the reference EE of one participant's test session in Fig. 6. Fig. 6 showed the EE estimation results of the first stage regression and the second stage regression respectively. By comparing the test results shown in Fig. 6, we can observe that the EE estimated value in the second stage was closer to the reference EE than that in the first stage. It is exactly the purpose of our designed two-stage regression module: in the first stage, a coarse-grained prediction of EE to determine the range is made; in the second stage, a further fine-grained prediction to determine the final value is generated.

Moreover, the final EE estimation results shown in Fig. 6 after the two-stage regression module demonstrated that the proposed model can accurately track the large changes of an individual's EE, even in the case of poor ECG signal quality caused by motion artifacts at a high speed.

TABLE VII
COMPARISON WITH PREVIOUS STUDIES

| Year | Reference | Input data | Method | Feature type | RMSE (kcal/min) | MAE (kcal/min) | RMSE (MET) | MAE (MET) | MAPE (%) |
|---|---|---|---|---|---|---|---|---|---|
| 2015 | [19] | IMU, HR, ANT | CNN | learned | 1.12 | - | - | - | - |
| 2016 | [43] | IMU | LR | hand-crafted | 1.15 | - | - | - | - |
| 2017 | [15] | IMU, ANT, HRV | LR | hand-crafted | 0.89 | - | - | - | - |
| 2018 | [17] | IMU, HR, ST, GSR | BDTR | hand-crafted | - | - | 0.76 | 0.53 | - |
| 2018 | [18] | IMU, HR, ST, GSR | RF | hand-crafted | - | - | 0.757 | 0.526 | 23 |
| 2021 | [20] | IMU, HR, FP | DenseNet | learned | 1.05 | 0.83 | - | - | - |
| 2021 | [21] | IMU | XGBoost | hand-crafted | - | - | 0.891 | - | - |
| **-** | This work | IMU, ECG, ANT | DMTRN | learned | 0.71 | 0.53 | 0.64 | 0.47 | 10.35 |

ST = Skin temperature; GSR = Galvanic skin response; BVP = Blood volume pulse; FP = Foot pressure

TABLE VIII
THE COMPLEXITY OF THE PROPOSED MODEL

| | |
|---|---|
| Parameters | 1,977,461 |
| Model size | 35.00MB |
| FLOPs | 472.71M |
| Inference time | <154ms |

## C. Complexity Analysis

As our DMTRN used a 1-D convolutional neural network for EE estimation, it is lightweight and can be implemented on mobile devices for real-time EE estimation. As shown in Table VIII, the parameter amount and model size was not large, which means low memory-consuming and the model can run on mobile systems. Apart from parameter amount and model size, the number of floating point operations (FLOPs) is also important. [46] showed that current mobile systems on market need about 154 ms to finish 569 MFlops when using MobileNet, as our model need 472.71 MFlops to finish EE estimation, it can be deduced that it would take less than 154 ms to finish EE estimation.

## D. Limitations

Although the proposed approach demonstrated the feasibility of improving the accuracy of EE monitoring using inertial sensors and ECG signals, some uncertainties remain. First, our dataset was collected under a controlled laboratory environment. Data preprocessing procedures were implemented before the model's development to reduce the noise and remove signals with very poor quality. However, the signals collected in a real environment feature more noise, which may influence the stability of the proposed model. Second, the distribution of the participants was narrow and the number of participants was limited. Multiple factors such as diseases, age, and individual differences may affect the EE variations, resulting in uncertainty in the performance of the proposed model.

# VI. CONCLUSION

In this paper, we proposed a DMTRN model for accurate EE estimation using multiple sensor information. The multi-branch CNN module and two-stage regression module were developed to improve the EE estimation performance. The low memory-consuming and the short inference time showed the feasibility of the proposed model for real-time processing on mobile systems. The experiments show that DMTRN obtains the state-of-the-art performance, with the RMSE of 0.71 kcal/min, reduced by 22.8% compared with traditional RF model respectively. Besides, our study demonstrated that the raw ECG signals contained more other EE related information in addition to HR for the first time.

In future work, we will first transfer our model to wearable EE estimation scenario where ECG and IMU data were collected by wearable devices. and then we can improve the model's robustness and generality by enhancing the dataset through collecting data from subjects with a large range of age when doing different types of physical activities.

# REFERENCES

[1] K. Strong *et al.*, "Preventing chronic diseases: How many lives can we save?," *Lancet*, vol. 366, no. 9496, pp. 1578–1582, 2005.

[2] F. W. Booth *et al.*, "Lack of exercise is a major cause of chronic diseases," *Comprehensive Physiol.*, vol. 2, no. 2, pp. 1143–1211, 2011.

[3] G. P. Kenny *et al.*, "Direct calorimetry: A brief historical review of its use in the study of human metabolism and thermoregulation," *Eur. J. Appl. Physiol.*, vol. 117, no. 9, pp. 1765–1785, 2017.

[4] E. Ferrannini, "The theoretical bases of indirect calorimetry: A review," *Metabolism*, vol. 37, no. 3, pp. 287–301, 1988.

[5] M. Luštrek, B. Cvetković, and S. Kozina, "Energy expenditure estimation with wearable accelerometers," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2012, pp. 5–8.

[6] K. R. Westerterp, "Physical activity and physical activity induced energy expenditure in humans: Measurement, determinants, and effects," *Front. Physiol.*, vol. 4, 2013, Art. no. 90.

[7] D. Fuller *et al.*, "Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: Systematic review," *JMIR mHealth uHealth*, vol. 8, no. 9, 2020, Art. no. e18694.

[8] H. J. Montoye *et al.*, "Estimation of energy expenditure by a portable accelerometer," *Med. Sci. Sports Exercise*, vol. 15, no. 5, pp. 403–407, 1983.

[9] K. Y. Chen and M. Sun, "Improving energy expenditure estimation by using a triaxial accelerometer," *J. Appl. Physiol.*, vol. 83, no. 6, pp. 2112–2122, 1997.

[10] J. H. Choi *et al.*, "Estimation of activity energy expenditure: Accelerometer approach," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, 2006, pp. 3830–3833.

[11] S. E. Crouter *et al.*, "A novel method for using accelerometer data to predict energy expenditure," *J. Appl. Physiol.*, vol. 100, no. 4, pp. 1324–1331, 2006.

[12] K. Charlot *et al.*, "Improvement of energy expenditure prediction from heart rate during running," *Physiol. Meas.*, vol. 35, no. 2, 2014, Art. no. 253.

[13] S. Brage *et al.*, "Reliability and validity of the combined heart rate and movement sensor actiheart," *Eur. J. Clin. Nutr.*, vol. 59, no. 4, pp. 561–570, 2005.

[14] M. Altini *et al.*, "Personalization of energy expenditure estimation in free living using topic models," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1577–1586, Sep. 2015.

[15] H. Park *et al.*, "The role of heart-rate variability parameters in activity recognition and energy-expenditure estimation using wearable sensors," *Sensors*, vol. 17, no. 7, 2017, Art. no. 1698.

[16] J. Staudenmayer *et al.*, "An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer," *J. Appl. Physiol.*, vol. 107, pp. 1300–1307, 2009.

[17] C. Catal and A. Akbulut, "Automatic energy expenditure measurement for health science," *Comput. Methods Programs Biomed.*, vol. 157, pp. 31–37, 2018.

[18] B. Cvetković *et al.*, "Real-time activity monitoring with a wristband and a smartphone," *Inf. Fusion*, vol. 43, pp. 77–93, 2018.

[19] J. Zhu *et al.*, "Using deep learning for energy expenditure estimation with wearable sensors," in *Proc. 17th Int. Conf. E-health Netw., Application Serv. (HealthCom)*, 2015, pp. 501–506.

[20] H. Eom *et al.*, "Deep learning-based optimal smart shoes sensor selection for energy expenditure and heart rate estimation," *Sensors*, vol. 21, no. 21, 2021, Art. no. 7058.

[21] M. T. Mardini *et al.*, "Age differences in estimating physical activity by wrist accelerometry using machine learning," *Sensors*, vol. 21, no. 10, 2021, Art. no. 3352.

[22] Y. Liu *et al.*, "From action to activity: Sensor-based activity recognition," *Neurocomputing*, vol. 181, pp. 108–115, 2016.

[23] E. Valanou *et al.*, "Methodology of physical-activity and energy-expenditure assessment: A review," *J. Public Health*, vol. 14, no. 2, pp. 58–65, 2006.

[24] J. A. Álvarez-García *et al.*, "A survey on energy expenditure estimation using wearable devices," *ACM Comput. Surv. (CSUR)*, vol. 53, no. 5, pp. 1–35, 2020.

[25] A. Y. Hannun *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, 2019.

[26] A. S. Alharthi, S. U. Yunas, and K. B. Ozanyan, "Deep learning for monitoring of human gait: A review," *IEEE Sensors J.*, vol. 19, no. 21, pp. 9575–9591, Nov. 2019.

[27] M. Sevil *et al.*, "Determining physical activity characteristics from wristband data for use in automated insulin delivery systems," *IEEE Sensors J.*, vol. 20, no. 21, pp. 12859–12870, Nov. 2020.

[28] R. A. Bruce *et al.*, "Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease," *Amer. Heart J.*, vol. 85, no. 4, pp. 546–562, 1973.

[29] P. Alinia *et al.*, "Impact of sensor misplacement on estimating metabolic equivalent of task with wearables," in *Proc. IEEE 12th Int. Conf. Wearable Implantable Body Sensor Netw.*, 2015, pp. 1–6.

[30] C.-W. Lin *et al.*, "A wearable sensor module with a neural-network-based activity classification algorithm for daily energy expenditure estimation," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 5, pp. 991–998, Sep. 2012.

[31] K. Ellis *et al.*, "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers," *Physiol. Meas.*, vol. 35, no. 11, 2014, Art. no. 2191.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[33] G. Huang *et al.*, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[34] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] F. E. Harrell, Jr., *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed., Cham, Switzerland: Springer International AG, 2015, pp. 282–294.

[36] Z. Niu *et al.*, "Ordinal regression with multiple output cnn for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4920–4928.

[37] H. Fu *et al.*, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2002–2011.

[38] H.-W. Hsu *et al.*, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, Apr. 2019.

[39] R. Diaz and A. Marathe, "Soft labels for ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4738–4747.

[40] T. T. Um *et al.*, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," in *Proc. 19th ACM Int. Conf. Multimodal Interaction*, 2017, pp. 216–220.

[41] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, pp. 8026–8037, 2019.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR (Poster)*, 2015, pp. 1–15.

[43] P. Alinia *et al.*, "A reliable and reconfigurable signal processing framework for estimation of metabolic equivalent of task in wearable sensors," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 5, pp. 842–853, Aug. 2016.

[44] J. T. Springenberg *et al.*, "Striving for simplicity: The all convolutional net," in *Proc. ICLR (Workshop)*, 2015, pp. 1–14.

[45] A. Kitchin and J. Neilson, "The T wave of the electrocardiogram during and after exercise in normal subjects," *Cardiovasc. Res.*, vol. 6, no. 2, pp. 143–149, 1972.

[46] Y. Chen *et al.*, "Deep learning on mobile and embedded devices: State-of-the-art, challenges, and future directions," *ACM Comput. Surv. (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.