Can Wearable Devices and Machine Learning Techniques Be Used for Recognizing and Segmenting Modified Physical Performance Test Items?

Yiyuan Zhang, Xiangyu Wang, Pengxuan Han, Sabine Verschueren, Wei Chen, Senior Member, IEEE, and Bart Vanrumste, Senior Member, IEEE

Abstract—Assessment of physical performance is essential to predict the frailty level of older adults. The modified Physical Performance Test (mPPT) clinically assesses the performance of nine activities: standing balance, chair rising up & down, lifting a book, putting on and taking off a jacket, picking up a coin, turning 360°, walking, going upstairs, and going downstairs. The activity performing duration is the primary evaluation standard. In this study, wearable devices are leveraged to recognize and predict mPPT items’ duration automatically. This potentially allows frequent follow up of physical performance, and facilitates more appropriate interventions. Five devices, including accelerometers and gyroscopes, were attached to the waist, wrists and ankles of eight younger adults. The system was experimented within three aspects: machine learning models, sensor placement, and sampling frequencies, to which the non-causal six-stages temporal convolutional network using 6.25 Hz signals from the left wrist and right ankle obtained the optimal performance. The duration prediction error ranged from 0.63 ± 0.29 s (turning 360°) to 8.21 ± 16.41 s (walking). The results suggest the potential for the proposed system in the automatic recognition and segmentation of mPPT items. Future work includes improving the recognition performance of lifting a book and implementing the frailty score prediction.

Index Terms—Activity recognition & segmentation, deep learning, frailty, older adults, wearable devices.

I. INTRODUCTION

Frailty level has been a standard scale for evaluating the ageing process of older adults [1]. One of its essential components is physical frailty, which is relevant to the ability to live independently [2], [3]. The assessment of physical frailty is normally dependent on the performance of physical activities (PAs) [4]. Furthermore, monitoring the performance of PA is beneficial for 1) older adults to maintain their physical health and prevent or postpone frailty; 2) doctors and physiotherapists to decide on physical treatments; 3) governments to arrange health service resources [1], [2], [5].

Physical tests are typically applied to assess the performance of PAs, such as modified physical performance test (mPPT) [6], multi-dimensional risk appraisal for older people (MRA-O) system [7], and short physical performance battery (SPPB) test [8]. Clinically, these tests were conducted sporadically, every one or three months, or only after accidents like falling. Hence, the subtle change in the PA performance cannot be monitored continuously and promptly. Furthermore, the conduction of the tests requires the supervision and recording of doctors or physiotherapists, which increases their workload and is inconvenient to patients who need to transport for the test.

This study aims to investigate whether using multiple wearable devices combined with machine learning techniques can be used to recognize and segment physical test items automatically. The mPPT is selected as the ground truth test, including nine items, which will be introduced in Section III. Three machine learning algorithms, i.e., a support vector machine (SVM) and two deep neural networks (DNNs), will be compared at the recognition and segmentation level. The segmentation result is converted to be the duration of each item, which is the primary evaluation standard of mPPT. Results will be presented in Section IV, with the discussion in Section V and the conclusion in Section VI.

II. RELATED WORK

A. Activity Recognition & Segmentation Application in Frailty

Although wearable devices have already been utilized in activity recognition, there have not been many studies employing the recognition result to measure the physical frailty of...
older adults. In the study conducted by Razjouyan et al. [9], one tri-axial accelerometer sensor (PAMSys) was adopted to collect movement signals of the chest during daily life, with 163 older adults as participants. The collected signals were used to extract features, which were grouped into four patterns: sleep quantity (sleep time, sleep onset latency, and time in bed), PA types (walking, sitting, and standing), PA intensity (sedentary, light, and moderate-to-vigorous), and step (number of steps and prolonged stepping bout). The Fried frailty index [10] was used as the ground-truth scale, classifying frailty into three levels: pre-frail, non-frail, and frail. The embedded decision tree model was trained and tested in five-fold cross-validation with all participant’s data mixed together. The results presented that the combination of PA types, PA intensity, and step features achieved the highest sensitivity (0.918) in classifying non-frail and frail.

In the study of Abril-Jiménez et al. [11], a wristband and a smartphone were combined with Bluetooth beacons to predict frailty levels. The applied sensors in the smartphone were GPS, inertial measurement unit (IMU), Bluetooth, and Wi-Fi, which were used to collect information on physical movements and daily travel trajectory. The frailty evaluation patterns were categorized into four aspects: activity movement, weekly visiting, transport daily usage, and system engagement. Their correlation values with the frailty levels were analyzed separately. The results showed that the walking speed, distance and number of steps were sensitive to the change in the frailty level. Similarly, in study [12], physical activities, such as standing, sitting, lying, gait, and standing balance were proved to be relative to frailty. Five IMUs, involving accelerometers, gyroscopes, and magnetometers, were attached to the shanks, thighs, and lower back.

Unlike the studies discussed above, study [13] focused on predicting the frailty of the arm using gyroscopes attached to the upper arm. The trauma-specific frailty index (TSFI) [14] was the golden standard and included two levels: frail and non-frail. In this study, participants were required to conduct elbow flexion continuously for 20 s. The multivariate regression was applied to predict the score, further used to predict whether a subject was frail or non-frail. The classification on the test set applied to predict the frailty using gyroscopes attached to the upper arm. The trauma-specific frailty index (TSFI) [14] was used as the ground-truth scale, classifying frailty into three levels: pre-frail, non-frail, and frail. The embedded decision tree model was trained and tested in five-fold cross-validation with all participant’s data mixed together. The results presented that the combination of PA types, PA intensity, and step features achieved the highest sensitivity (0.918) in classifying non-frail and frail.

The limitations of these studies are that:

- The frailty scale they adopted categorized frailty in two or three levels: non-frail, (pre-frail,) and frail. However, the severity level of frailty, which is also valuable information to apply the intervention to postpone the progression of frailty [15], is seldom evaluated due to the frailty assessment tool previous studies selected.
- In study [9], the model was only tested in five-fold cross-validation; hence, the robustness of the model in unseen users has not been verified yet.

B. Activity Recognition & Segmentation

In this study, mPPT items monitoring is separated into two steps, first recognizing and then segmenting activities. The segmentation results are used to predict the performance duration of every activity, which is required to calculate the mPPT test final frailty score. The segmentation performance is based on the recognition result; hence, the recognition algorithms, which are the machine learning models, are discussed here.

Machine learning models are generally categorized into two groups: traditional machine learning models and DNN models. Except SVM applied in the study of Razjouyan et al. [9], other traditional machine learning models, e.g., random forest (RF) [16], naïve Bayes (NB) [16], and artificial neural network (ANN) [16], [17], have also been used for PA recognition. One challenge of traditional machine learning models is that they typically use informative hand-crafted features as input. This requires representative features to be designed based on domain knowledge the context of the machine learning task.

In comparison, DNN models are end-to-end algorithms, which directly apply raw signals as input. Chen et al. [18] developed a convolutional neural network (CNN) to classify eight PAs, i.e., falling, running, jumping, walking, walking quickly, step walking, going upstairs, and going downstairs. Three tri-axial accelerometer sensors (sampling rate at 100 Hz) were placed in the cloth pocket, trouser pocket and to the waist. Around 86 participants’ data were used for training, and the remaining 14 participants’ for testing. The CNN model outperformed the SVM model, with an average recognition accuracy of 93.8%. However, the detailed information on the dataset was not mentioned, e.g., whether it was balanced. Accuracy may not be suitable in imbalanced data, as it mainly implies the true positive result of majority class. Instead, the F1 score is proposed, such as the micro F1 and segmentation F1@k score deployed in this study.

Besides CNN models, recurrent neural network (RNN) models have also been used for PA recognition to extract temporal features and recognize sequential activities. The RNN model is categorized into two types: the gated recurrent unit (GRU) and long short-term memory (LSTM). The GRU unit outperforms the LSTM unit because it is faster and easier to generalize while training with fewer gates [19], [20]. The RNN models were normally combined with the CNN units to extract both temporal features and features of single sensor channels or cross channels [21]–[23]. For example, in the study of Xu et al. [23], Incep V1 modules [24] were combined with the GRU units. In an inception V1 module, three CNN units were used to extract features in parallel. The combination of Inception V1 units and the GRU units is more efficient in extracting features from IMU signals for activity recognition than the normal CNN model or CNN-LSTM model.

The aforementioned models were generally applied in the fixed-length samples, with sliding window splitting signals first. One drawback of the sliding window is that the length of the window depends on the duration and the movement complexity of the classified activities. Without splitting signals, the temporal convolution network (TCN) [25] was proposed and applied to varied-length samples. The TCN model is composed of a hierarchy of temporal CNNs, which updates the prediction of every sample simultaneously. The TCN model can extract spatial and temporal information simultaneously, without RNN units. There are two types of TCN models, dilated TCN and encoder-decoder TCN (ED-TCN). Except for the temporal CNN layer, ED-TCN includes pooling and upsampling layers. Several studies have compared these two
models with the LSTM and CNN model [26], [27]. In the study of Nair et al. [26], the models were tested on the UCI HAR dataset (six PA) [28], with 70% of the data for training. The ED-TCN achieved an F1-score of 0.946, followed by the dilated TCN with an F1-score of 0.938. The CNN model achieved the highest F1-score of 0.976, but with hand-crafted features as input. Therefore, although TCN models achieved a bit lower F1-score, they were verified to be efficient in dealing with varied length sequence data without using the sliding window and selecting features.

Two limitations of these studies are that:

• their inputs were still samples segmented by the sliding window; thus, the robustness of TCN models in data with varied lengths has not been verified with IMU signals.

• the models were also mostly tested with all participants’ data mixed, which is not appropriate for the actual application, where the test data is unseen by the model.

The contributions of the present study are:

• Activities related to the frailty of all the body parts will be monitored, which are the activities included in the validated mPPT test. In addition, this test divides frailty into more detailed levels, namely non-frail, mild frail, moderate frail, unable to be functional.

• The (MS-)TCN model is used to analyse IMU signals, compared with the SVM and Incep V1-GRU models. Models are trained in the leave-one-person-out (LOPO) mode.

• The optimal sensor placement and sampling frequencies are investigated.

• The mPPT items will be recognized and also segmented. The segmentation result will then be used to predict each activity’s duration, which is required to calculate the mPPT test score.

III. MATERIALS AND METHODS

A. Data Acquisition

Five wearable devices from Byteflies [29] were applied for collecting the movement signal. Each sensor contained one tri-axial accelerometer ($A_x$, $A_y$, $A_z$) and one tri-axial gyroscope ($G_x$, $G_y$, $G_z$), with the sampling rate at 100 Hz. Two Byteflies sensors were attached to each wrist, with one sensor in the trouser’s pocket and the other two sensors attached to the ankles.

Five males and three females, aged 24 to 33 years old, took part in the experiment. All participants were right-handed. Every participant performed the mPPT test for three rounds, and they were required to move slower than their natural movement speed to mimic the movement behaviour of older adults. Participants followed the researcher’s instruction to perform the mPPT test items in a fixed order, standing balance, chair rising up & down, lifting a book, putting on and taking off a jacket, picking up a coin, chair rising up & down, walking 50 ft, going upstairs, going downstairs, and waving hands.

B. Data Analysis

The Data analysis process is shown in Fig. 1. It includes four stages, and each stage will be explained as follows.

1) Signal Pre-processing: The collected signals are annotated and segmented based on the researcher’s annotation record. The signals of transition movements are removed. Hence, segmented signals are categorized into ten classes: standing balance, lifting a book, turning 360°, putting on and taking off a jacket, picking up a coin, chair rising up & down, walking 50 ft, going upstairs, going downstairs, and waving hands.

2) Downsampling: Proposed by Farha et al. [31], in the activity segmentation, data downsampling can overcome the over-segmentation problems with the reduced precision result of detecting the segmentation boundaries. In addition, the longest sequence of this study has 31963 samples, which is long for the TCN model, which will be introduced in Section 3.3.1. Multiple frequencies (50 Hz, 25 Hz, 12.5 Hz, and 6.25 Hz) were tested with the TCN model, and the 6.25 Hz (Decimation factor = 16) was finally selected as it achieved a similar F1-score but a higher segmentation performance compared to other frequencies. The result is shown in Table III. All models are tested with signals at 100 Hz and 6.25 Hz.

3) Sliding Window: The sliding window is applied on splitting signals, except for TCN models. If there are multiple classes within one window, the window is annotated as the majority class. The size of the window is tuned in study [32]. The optimal size is 2 s with 95% (190 samples) overlapping for the 100 Hz sampling frequency and 2 s with 92% (12 samples) overlapping for the 6.25 Hz sampling frequency. Table I lists the total performing duration based on the annotation and the number of sliding windows of 100 Hz signals of each class.

C. mPPT Items Recognition

This study compares three machine learning algorithms: dilated TCN models, Incep V1-GRU model, and SVM. The setting up of every model will be discussed as follows.

1) Machine Learning Algorithms:

a) Dilated TCN: In the Dilated TCN model, the raw signals of one sequence, refer to the data sequence of one mPPT test, in other words, data before windowing. There are in total 24 data sequences, three sequences per participant.

Two types of dilated TCN models are tested, the causal and non-causal one [25]. In the causal TCN model, while predicting the nth sample, only the previous samples are considered. In contrast, the non-causal TCN model considers all samples in the receptive field before and after the considered sample.

The multi-stage TCN (MS-TCN) model [31] is also investigated in this study. This model includes multiple stages and
every stage has the same architecture as the normal TCN and the output of the previous stage is processed by a $1 \times 1$ convolutional layer with the softmax activation to alter the dimension. The architecture is shown in Fig. 2. With this architecture, this model effectively solves the over-segmentation problem without changing the receptive field. [31]. The loss function used in the normal TCN model is the classification loss ($L_{\text{cls}}$), considering the results of each sample as shown in (1):

$$L_{\text{cls}} = \frac{1}{T} \sum_{t=1}^{T} -\log(y_{t,c})$$

(1)

where $T$ is the sequence length, $y_{t,c}$ is the predicted probability of the real annotated class for time sample $t$. In contrast, the MS-TCN model calculates the loss ($L$) of every single stage, and it combines the classification loss with a smoothing loss ($L_{T-MSE}$) to smooth over the time samples, as shown in (2),(3),(4), and the final loss is shown in (5).

$$L_{T-MSE} = \frac{1}{TC} \sum_{t,c} \Delta_{t,c}^2$$

(2)

$$\Delta_{t,c} = \begin{cases} \frac{2}{\tau}, & t_c \leq \tau \\ \frac{1}{\tau}, & t_c > \tau \end{cases}$$

(3)

$$\Delta_{t,c} = |\log(y_{t,c}) - \log(y_{t-1,c})|$$

(4)

$$L = \sum_s (L_{\text{cls}} + \lambda L_{T-MSE})$$

(5)

C represents classes, $s$ for stages, $\tau$ (=4) for threshold value and $\lambda$ (=0.15) for the weight of $L_{T-MSE}$ [31]. In this study, the architecture of single stage is the same as the normal TCN, and the number of stages is tuned in the range of [2:8]. The tuned parameters of TCN models are the kernel size ($K_{size}$) ([1:1:7] for causal TCN, [1:3:7] for non-causal TCN), the number of blocks ($N_{block}$) ([1:1:4]), the number of filters ($N_{filter}$) ([16,32,64,128,256]), and the number of dilated convolutional layers of each block ($N_{layer}$) ([2:1:13]). The dilation size of convolutional layers is increased exponentially ($d_i = 2^i, i \in [0 : N_{layer} - 1]$). All these parameters determine the receptive field of the model, shown in (6):

$$\text{Receptive field} = 1 + (K_{size} - 1) \cdot N_{block} \cdot \sum d_i$$

(6)
is used to tune hyper-parameters. This method is based on the randomly-search algorithm and Bayesian theory. The combination of the hyper-parameters with the higher Bayesian probability is searched based on the recognition result of the previous combination. The hyper-parameters were tuned for five rounds. The one with the highest result was selected. The models were trained in 100 epochs. The early stopping algorithm was applied, and the training process stopped if the F1-score of the validation dataset did not increase over 0.005 in 10 epochs. The tuned results of TCN models are $K_{size}=2$ (causal TCN) \ $\setminus$ 3 (non-causal TCN), $N_{block}=1$, $N_{filter}=256$, and $N_{layer}=12$ (sampling frequency at 100 Hz) \ 7 (sampling frequency at 6.25 Hz). Therefore, for signals sampled at 100 Hz, the receptive field of causal TCN is 40.96 s and 81.91 s for non-causal TCN. For signals sampled at 6.25 Hz, the receptive field of causal TCN is 20.48 s and 40.8 s.

b) Incep V1-GRU model: In the Incep V1-GRU model, raw signals within the sliced 2s window are directly applied as the input. The model is tested on both sampling frequencies of 100 Hz and 6.25 Hz, hence the input size is $200 \times 30$ (100 Hz) / $13 \times 30$, with 200/13 samples and 30 channels ($30 = 5 \text{ sensors} \times (3 \text{ axes of accelerometer + 3 axes of gyroscope})$).

In one Incep V1 module, three CNN units at different kernel sizes are used, which are $1 \times 1$, $1 \times 3$, $1 \times 5$. Furthermore, three stacked $1 \times 1$ CNN units are used to reduce the dimension of the model. The architecture of the Incep V1-GRU model is shown in Fig. 3. This model includes three Incep V1 modules, three GRU units, one max-pooling layer, and three dense layers. The tuned hyper-parameters are the number of filters of the CNN units in Incep V1 modules, the number of neurons of dense layers, the dropout rate of the dropout layers, and the batch size. The searching range of the dropout rate is $[0.25, 0.5, 0.75]$, range of the batch size is $[16, 32, 64, 128]$, and the range of other hyper-parameters is $[2:2:256]$.

c) SVM: In the SVM model, the input is hand-crafted features extracted from the sliced 2s windows. Thirteen and six features are extracted in the time-domain and frequency domain, respectively, as listed in Table II. The features are extracted from the x-, y-, z-axis and the magnitude of the accelerometer (7) and gyroscope signals separately (8):

$$A_m = \sqrt{A_x^2 + A_y^2 + A_z^2}$$

$$G_m = \sqrt{G_x^2 + G_y^2 + G_z^2}$$

The correlation between axes (Corr) and zero-crossing times are not extracted from the magnitude of the accelerometer or gyroscope. The detailed information on the features can be found in [34]. In total, 74 ($17 \times 4 \text{ (axes + magnitude) + 2 \times 3 \text{ (axes})}$) features are extracted from one accelerometer or gyroscope sensor.

A one-versus-one multiclass SVM classifier is used with the RBF kernel. The regularisation parameter C and the $\gamma$ value of the RBF kernel are optimised. The optimized value of C is 10 (searched in $10^{[0:1:3]}$) and the optimized value of $\gamma$ is 0.01 (searched in $10^{[-5:1:-1]}$).

2) Model Training & Evaluation of Recognition Performance:
Each participant has three datasets, each representing one pass through the mPPT test. Models are trained and tested in two modes:

- Participant independent. The dataset is separated into three parts, training, validation, and testing. In each iteration, two datasets of one participant were used for testing, and the left dataset of the test participant was not used.
D. MPPT Items Segmentation

1) Segmentation Process: Based on the predicted windows/time samples, a segment is established with a predicted class at discrete time instances. This segment is compared with the ground-truth annotated segment. The items segmentation process is illustrated in Fig. 4.

The neighbouring windows with the same predictions are considered one group, which is one mPPT item. The item’s duration is calculated by counting the number of covered windows or time samples and converting it into seconds.

2) Evaluation Metrics of Segmentation Performance:

- \( f_1 \): the \( f_1 \) score is adopted as the evaluation metric for comparing the performance of the same model using signals with different sampling frequencies.
- Computation duration: the training and prediction duration of the whole dataset is calculated separately.

E. Optimal Sensor Placement

The optimal sensor placement is investigated with the best machine learning model. Firstly, signals of every Byteflies sensor are applied separately. At first, signals of one sensor is added until all the five sensors are included. The combination of sensors is evaluated based on the \( f_1 \) score (\( k = 0.5 \)). If the \( f_1 \) values are similar, the \( F_1 \)-score will be used as the second evaluation criterion.

IV. RESULTS

A. Activity Recognition & Segmentation in the Participant Independent Mode

Table III presents the activity recognition & segmentation results of all models using signals at frequencies of 100 Hz and 6.25 Hz. Among the models, TCN models performed better than other models in both activity recognition and segmentation, with the non-causal TCN model (6.25 Hz) achieving the highest \( f_1 \) (\( k = 0.5 \)) of 0.277. While comparing TCN models, the non-causal TCN outperforms the causal TCN model in both recognition and segmentation results.

B. Activity Recognition & Segmentation With the Transfer Learning Method

The MS-TCN (6 stages) is trained with the transfer learning method, shown in Table IV. The performance of the MS-TCN (1 stage) is better than the normal TCN, especially its \( f_1 \) value (\( k = 0.5 \)) of 0.06 higher than the normal TCN. Among the MS-TCN models, MS-TCN (6 stages) obtains the highest segmentation performance, with the \( f_1 \) value (\( k = 0.5 \)) of 0.705 in the participant independent mode.

After comparison, the non-causal TCN models using signals sampled at 6.25 Hz are applied to test the MS-TCN model with the transfer learning method, shown in Table IV. The performance of the MS-TCN (1 stage) is better than the normal TCN, especially its \( f_1 \) value (\( k = 0.5 \)), around 0.05 higher than the normal TCN. Among the MS-TCN models, MS-TCN (6 stages) obtains the highest segmentation performance, with the \( f_1 \) value (\( k = 0.5 \)) of 0.705 in the participant independent mode.

B. Activity Recognition & Segmentation With the Transfer Learning Method

The MS-TCN (6 stages) is trained with the transfer learning method and investigates the best sensor placement because it obtains the highest performance compared to other models. As listed in Table IV, including part of the test participant’s data for training the model, the robustness of the model is improved in unseen participants, with \( F_1 \) scores increasing from 0.709 to 0.758 and \( f_1 \) scores increasing over 0.1.
Fig. 4. Example of items segmentation process and the definition of TP, FP, and FN while calculating f1@k (k = 0.5) of the first time mPPT test performed by participant 2 (P2). The figures are: (a) the predicted data sequence using the non-causal MS-TCN model (6 stages) using 6.25 Hz; (b) The example of the TN, FP, FN segments of waving hands; (c) The real annotated sequence. The red dotted lines represent the actual segments range, and the green dotted lines represent the related predicted segments range. The grey blocks (+ yellow blocks) represent the intersection of segments. The interaction over union (IOU) is compared with the k value. The predicted segment is a true positive (TP) if its corresponding IOU is larger than 0.5; otherwise, it is a false positive (FP). The unpaired real segments are false negative (FN). Abbreviations are: standing balance feet side-by-side (B-1), standing balance semi-tandem (B-2), standing balance full-tandem (B-3), waving hands (WA), chair rising up & down (CH), lifting a book (LB), putting on and taking off a jacket (JA), picking up a coin (CO), turning 360° (R), walking 50 ft (WK), going upstairs (US), and going downstairs (DS).

Table III

| Model (frequency) | F1-score | F1@k | Training duration(s) | Prediction duration (s) |
|-------------------|----------|------|----------------------|-------------------------|
|                   |          | 0.1  | 0.25 | 0.28 | 0.272 | 0.218 | 2461.938 | 147.237 |
| SVM (100 Hz)      | 0.610    | 0.281 | 0.172 | 0.218 | 2461.938 | 147.237 |
| SVM (6.25 Hz)     | 0.559    | 0.117 | 0.103 | 0.007 | 2236.421 | 142.326 |
| Incep V1-GRU (100 Hz) | 0.338 | 0.094 | 0.088 | 0.044 | 4983.747 | 93.675 |
| Incep V1-GRU (6.25 Hz) | 0.398 | 0.094 | 0.084 | 0.044 | 1174.086 | 14.101 |
| Causal TCN (100 Hz) | 0.585 | 0.145 | 0.133 | 0.089 | 2076.542 | 13.615 |
| Causal TCN (6.25 Hz) | 0.545 | 0.340 | 0.311 | 0.214 | 251.702 | 7.578 |
| Non-causal TCN (100 Hz) | 0.647 | 0.125 | 0.116 | 0.097 | 1648.311 | 14.791 |
| Non-causal TCN (6.25 Hz) | 0.646 | 0.397 | 0.375 | 0.277 | 261.856 | 7.796 |

Table V lists the results of a single sensor and optimal combinations of sensors when the number of sensors increased from two to five. The optimal sensor combination is sensors on the left hand and right ankle, with the f1@k (k = 0.5) of 0.850, 0.11 higher than using all five sensors and only 0.18 lower than the highest result, including the sensor on the left ankle. In addition, the number of worn sensors is reduced. The recognition confusion matrix is presented in Fig. 6. The absolute duration predicted errors of the non-causal MS-TCN model (6 stages) using signals of the left hand and right ankle are calculated, as shown in Table VI. The duration prediction error averaged over all mPPT items participants 4-7 are about 1 s, which is acceptable as it is the duration precision in the mPPT test. The most significant errors of these participants are going upstairs (16.16 s) and going downstairs (10.48 s) for P1, walking 50 ft (48.16 s) and standing balance (28.24 s) for P2. An example of the poor result is shown in Fig. 5 for P2.

V. DISCUSSION

A. Machine Learning Models

Different sampling rates have various impacts on different models. For TCN models, downsampling results in a higher segmentation performance, as proposed in study [31]. However, for the SVM model, signals at 100 Hz achieved higher segmentation results. This is because, within the same length (2 s), the larger the sample size, the less spread one gets on the extracted statistical features. Therefore, the window at 100 Hz has more samples, which is more informative, resulting in a better segmentation performance. However, the impact of
frequency on the Incep V1-GRU model is insignificant, which could be because it is more influenced by the dataset size than the sampling rate.

The MS-TCN (six stages) outperformed other models in recognizing and segmenting mPPT items. In addition, it does not require the sliding window method and extracting hand-crafted features. However, there are still limitations of the MS-TCN models: 1) it applies the non-causal TCN model to consider the complete sequence during prediction. Hence, the non-causal TCN model can be only applied for offline data analysis; 2) all the participants performed mPPT test items following the same order. Thus, whether the model can be used for a random order sequence has not been verified yet; 3) Although participants were asked to slow their speed, their performance is still different from older adults, even frail older adults, such as the increased step instability and reduced ability of keep balance [36]; 4) The physical meaning of the features extracted by the MS-TCN has not been explored yet.

### B. The Impact of the Transfer Learning Method

To overcome the impact of subjective difference in the participant independent mode, the transfer learning method is proposed to be effective to improve the generalizability of the model. In a real-life application, it can be applied with a data pre-collection session for new users. During this session, movement signals can be collected while users learn how to perform mPPT test items and wear the devices.

### C. Sensor Placement

The optimal sensor combination is a sensor on the left hand and right ankle. The sensor on the ankle is more sensitive to the movement of the lower extremity, such as walking, as proposed in study [37]. In addition, the non-dominant hand is better than the dominant hand when combined with the right ankle in recognizing mPPT items, which is the same as in study [38]. As a result, with only two sensors, the segmentation performance is even higher than with five sensors. Moreover, the experimental setup is simpler, which is more convenient for users to perform the mPPT test by themselves. However, from the confusion matrix we note that this combination is not helpful to differentiate going upstairs, going downstairs, and walking, which were similar activities but performed in different directions. The features extracted from the transitional acceleration and IMU signals between knee and ankle [39] can be considered in future work.

### D. MPPT Items

The confusion matrix of the non-causal MS-TCN (6 stages) model is shown in Fig. 6. Waving hands is confused with other activities, except for going downstairs. Among the mPPT items, lifting a book achieved the lowest recognition performance. The reason could be that this activity has the smallest numbers of samples, not enough for the model to learn properly. In addition, participants were required to mimic the behaviour of older adults, like slowing down their movement speed. Hence, the collected movement signals of this activity could be similar to waving hands without moving the arms. In addition, going upstairs and downstairs are confused with each other.

Regarding segmentation results, P2 obtained the largest duration predicted errors in walking 50 ft and going upstairs. While checking the performance of P2, (s)he performed the mPPT test at a much slower speed than others, resulting in a much longer duration of each activity. Moreover, (s)he stopped multiple times during the walking and going upstairs test, which can happen with older adults and is misclassified with standing balance, shown in Fig. 5. This misclassification implies the limitation of the dataset not including all the variability of performing these items; hence the model did not learn that stopping could occur during the walking/going upstairs test. In addition, a larger size of receptive field of the TCN model may be needed to grasp the context.
This study predicted the duration of most mPPT items with the error of less than 4 s, except for walking 50 ft, going upstairs, and going downstairs; however, the mPPT test score can not be automatically predicted yet. What is still required to be able to calculate the mPPT score will be: 1) detecting the steadiness and continuity of turning 360°; 2) counting the number of finished steps of going up-/downstairs; 3) detecting different feet positions of standing balance. The solution can be using multiple types of sensors, like a pressure sensor to record the feet positions and movement. Furthermore, the zero velocity updates (ZUPT) method, which measures the velocity and location of the feet, could be applied to detect the feet positions and movement. The future works can mainly focus on solving the problems of increasing the recognition performance of some mPPT items, like lifting a book, and testing the feasibility of the model on older adults.

ACKNOWLEDGMENT
The assistance provided by Ahmed Youssef Ali Amer and Benjamin Filtjens (KU Leuven) was greatly appreciated. In addition, the authors thank all the participants who volunteered in the experiment.

REFERENCES
[1] A. Clegg, J. Young, S. Iliffe, M. O. Rikkert, and K. Rockwood, “Frailty in elderly people,” Lancet, vol. 381, no. 9868, pp. 752–762, Mar. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0140673612621679
[2] F. Landi, G. Onder, I. Carpenter, M. Cesari, M. Soldato, and R. Bernabei, “Physical activity prevented functional decline among frail community-living elderly subjects in an international observational study,” J. Clin. Epidemiol., vol. 60, no. 5, pp. 518–524, May 2007.
[3] J. S. McPhee, D. P. French, D. Jackson, J. Nazroo, N. Pendleton, and H. Degens, “Physical activity in older age: Perspectives for healthy ageing and frailty,” Biogerontology, vol. 17, no. 3, pp. 567–580, Jun. 2016, doi: 10.1007/s10522-016-9641-0.
[4] G. Kojima, A. E. M. Liljas, and S. Iliffe, “Frailty syndrome: Implications and challenges for health care policy,” Risk Manage. Healthcare Policy, vol. 12, pp. 23–30, Feb. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6385767/
[5] F. Lobelo et al., “Routine assessment and promotion of physical activity in healthcare settings: A scientific statement from the American Heart Association,” Circulation, vol. 137, no. 18, May 2018. [Online]. Available: https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000005559
[6] M. Brown, D. R. Sinacore, E. F. Binder, and W. M. Kohrt, “Physical and performance measures for the identification of mild to moderate frailty,” Journals Gerontol. A, Biol. Sci. Med. Sci., vol. 55, no. 6, pp. M350–M355, Jun. 2000.
[7] K. Walters et al., “Promoting independence, health and well-being for older people: A feasibility study of computer-aided health and social risk appraisal system in primary care,” BMC Fam. Pract., vol. 18, no. 1, p. 47, Mar. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5366113/
[8] S. Perera, S. H. M. PharmD, R. C. Woodman, and S. A. Studenski, “Meaningful change and responsiveness in common physical performance measures in older adults,” J. Amer. Geriatrics Soc., vol. 54, no. 5, pp. 743–749, May 2006, doi: 10.1111/j.1532-5415.2006.00701.x.
[9] J. Razjouyan et al., “Wearable sensors and the assessment of frailty among vulnerable older adults: An observational cohort study,” Sensors, vol. 18, no. 5, p. 1336, Apr. 2018. [Online]. Available: https://www.mdpi.com/1424-8240/18/5/1336
[10] L. P. Fried et al., “Frailty in older adults: Evidence for a phenotype,” J. Gerontol. A, Biol. Sci. Med. Sci., vol. 56, no. 3, pp. M146–M157, Mar. 2001, doi: 10.1093/gerona/56.3.M146.
[11] P. Abril-Jímenez et al.,”Aging-friendly cities for assessing older adults’ decline: IoT-based system for continuous monitoring of frailty risks using smart city infrastructure,” Aging. Clin. Exp. Res., vol. 32, no. 4, pp. 663–671, Apr. 2020, doi: 10.1007/s40520-019-01238-y.
[12] M. Schwenk et al., “Wearable sensor-based in-home assessment of gait, balance, and physical activity for discrimination of frailty status: Baseline results of the Arizona frailty cohort study,” Gerontology, vol. 61, no. s3, pp. 258–267, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC445218/
[13] N. Toosizadeh, C. Wendel, C.-H. Hsu, E. Zamrini, and J. Mohler, “Frailty assessment in older adults using upper-extremity function: Index development,” BMC Geriatrics, vol. 17, no. 1, p. 117, Jun. 2017, doi: 10.1186/s12877-017-0509-1.
[14] B. Joseph et al.,”Validating trauma-specific frailty index for geriatric trauma patients: A prospective analysis,” J. Amer. College Surgeons, vol. 219, pp. 10–17, Jul. 2014.
[15] M. Marcucci et al., “Interventions to prevent, delay or reverse frailty in older people: A journey towards clinical guidelines,” BMC Med., vol. 17, no. 1, p. 193, Dec. 2019. [Online]. Available: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1434-2
[16] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, “Smartwatch-based activity recognition: A machine learning approach,” in Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI), Feb. 2016, pp. 426–429.
[17] A. Chelli and M. Pätzold, “A machine learning approach for fall detection and daily living activity recognition,” IEEE Access, vol. 7, pp. 38670–38687, 2019.
[18] Y. Chen and Y. Xue, “A deep learning approach to human activity recognition based on single accelerometer,” in Proc. IEEE Int. Conf. Syst. Man, Cybern., Oct. 2015, pp. 1488–1492.
[19] L. Jing et al., “Gated orthogonal recurrent units: On learning to forget,” Neural Comput., vol. 31, no. 4, pp. 765–783, Apr. 2019.
[20] L. Kaiser and I. Sutskever, “Neural GPUs learn algorithms,” 2015, arXiv:1511.08228.
[21] F. Li, K. Shirahama, M. A. Nisar, L. Köpíng, and M. Grzegorzek, “Comparison of feature learning methods for human activity recognition using wearable sensors,” Sensors, vol. 18, no. 2, p. 679, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/2/679
[22] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, “Deepsense: A unified deep learning framework for time-series mobile sensing data processing,” in Proc. 26th Int. Conf. World Wide Web, 2017, pp. 351–360.
[23] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, “InnoHAR: A deep neural network for complex human activity recognition,” IEEE Access, vol. 7, pp. 9893–9902, 2019.
[24] C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.
[25] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks for action segmentation and detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 156–165.
[26] N. Nair, C. Thomas, and D. B. Jayagopi, “Human activity recognition using temporal convolutional network,” in Proc. 5th Int. Workshop Sensor-Based Activity Recognit. Interact., New York, NY, USA, Sep. 2018, pp. 1–8, doi: 10.1145/3266157.3266221.
[27] F. Aparecido Garcia, C. Mazzoni Ranieri, and R. Aparecida Franceline Romero, “Temporal approaches for human activity recognition using inertial sensors,” in Proc. Latin Amer. Robot. Symp. (LARS), Brazilian Symp. Robot. (SBR) Workshop Robot. Educ. (WRE), Oct. 2019, pp. 121–125.
[28] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A public domain dataset for human activity recognition using smartphones,” in Proc. ESAW, 2013, pp. 437–442.
[29] BVBA, Byteflies. (2019). Byteflies. We Enable Wearable Health. Antwerp, Belgium. Accessed: Mar. 1, 2019. [Online]. Available: https://www.byteflies.com/
[30] Y. Zhang et al., “Automatically segmenting physical performance test items for older adults using a Doppler radar: A proof of concept study,” IEEE Access, vol. 9, pp. 152765–152779, 2021.
[31] Y. A. Farha and J. Gall, “MS-TCN: Multi-stage temporal convolutional network for action segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3575–3584.
[32] P. Han and X. Wang, “Indoor activity recognition for modified physical performance test via continuous wave radar sensor,” in Proc. 24th Int. Conf. Neural Inf. Process. Syst. (NIPS), Dec. 2011, pp. 2546–2554.
[33] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyper-parameter optimization,” in Proc. Adv. Neural Inf. Process. Syst., J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2011, pp. 2546–2554. [Online]. Available: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf
[34] Y. Zhang, J. Wullems, I. D’Haeseleer, V. A. Abeele, and B. Vanrumste, “Bathroom activity monitoring for older adults via wearable device,” in Proc. IEEE Int. Conf. Healthcare Informat. (ICHI), Nov. 2020, pp. 1–10. [Online]. Available: https://www.computer.org/csdl/proceedings-article/ichi/2020/09374396/1rUJ0Nk7Ufm
[35] E. Bisong, “Google colaboratory,” in Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, E. Bisong, Ed. Berkeley, CA, USA: Apress, 2019, pp. 59–64, doi: 10.1007/978-1-4842-4470-8_7.
[36] M. Y. Osoba, A. K. Rao, S. K. Agrawal, and A. K. Lalwani, “Balance and gait in the elderly: A contemporary review,” Laryngoscope Investigative Otolaryngol., vol. 4, no. 1, pp. 143–153, Feb. 2019.
[37] T. Sztyler and H. Stuckenschmidt, “On-body localization of wearable devices: An investigation of position-aware activity recognition,” in Proc. IEEE Int. Conf. Perspasive Comput. Commun. (PerCom), Mar. 2016, pp. 1–9.
[38] H. Gjoreski, J. Bizjak, M. Gjoreski, and M. Gams, “Comparing deep and classical machine learning methods for human activity recognition using wrist accelerometer,” in Proc. Workshop Deep Learn. Artif. Intell. (IJCAI), New York, NY, USA, 2016, p. 7.
[39] R. Stolyarov, G. Burnett, and H. Herr, “Translational motion tracking of leg joints for enhanced prediction of walking tasks,” IEEE Trans. Biomed. Eng., vol. 65, no. 4, pp. 763–769, Apr. 2018.
[40] F. Gu, K. Khoshelham, J. Shang, F. Yu, and Z. Wei, “Robust and accurate smartphone-based step counting for indoor localization,” IEEE Sensors J., vol. 17, no. 11, pp. 3453–3460, Jun. 2017.