Otago Exercises Monitoring for Older Adults by a Single IMU and Hierarchical Machine Learning Models

Meng Shang, Lenore Dedeyne, Jolan Dupont, Laura Vercauteren, Nadjia Amini, Laurence Lapauw, Evelien Gielen, Sabine Verschueren, Carolina Varon, Walter De Raedt, and Bart Vanrumste, *Senior Member, IEEE*

***Abstract*— Otago Exercise Program (OEP) is a rehabilitation program for older adults to improve frailty, sarcopenia, and balance. Accurate monitoring of patient involvement in OEP is challenging, as self-reports (diaries) are often unreliable. The development of wearable sensors and their use in Human Activity Recognition (HAR) systems has lead to a revolution in healthcare. However, the use of such HAR systems for OEP still shows limited performance. The objective of this study is to build an unobtrusive and accurate system to monitor OEP for older adults. Data was collected from 18 older adults wearing a single waist-mounted Inertial Measurement Unit (IMU). Two datasets were recorded, one in a laboratory setting, and one at the homes of the patients. A hierarchical system is proposed with two stages: 1) using a deep learning model to recognize whether the patients are performing OEP or activities of daily life (ADLs) using a 10-minute sliding window; 2) based on stage 1, using a 6-second sliding window to recognize the OEP sub-classes. Results showed that in stage 1, OEP could be recognized with window-wise f1-scores over 0.95 and Intersection-over-Union (IoU) f1-scores over 0.85 for both datasets. In stage 2, for the home scenario, four activities could be recognized with f1-scores over 0.8: *ankle plantarflexors*, *abdominal muscles*, *knee bends*, and *sit-to-stand*. These results showed the potential of monitoring the compliance of OEP using a single IMU in daily life. Also, some OEP sub-classes are possible to be recognized for further analysis.**

***Index Terms*— Hierarchical activity recognition, Otago exercise program, inertial sensors, machine learning, deep learning.**

Manuscript received 9 October 2023; revised 26 November 2023 and 3 January 2024; accepted 11 January 2024. Date of publication 17 January 2024; date of current version 24 January 2024. This work was supported in part by the China Scholarship Council (CSC) and in part by the ENHANce Project (S60763) received a Junior Research Project Grant from the Research Foundation Flanders (FWO) under Grant G099721N. *(Corresponding author: Meng Shang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee Research UZ/KU Leuven under Application No. S59660 and S60763.

Meng Shang is with the Department of Electrical Engineering, STADIUS, and the e-Media Research Laboratory, KU Leuven, 3000 Leuven, Belgium, and also with IMEC, 3001 Leuven, Belgium (e-mail: meng.shang@kuleuven.be).

Lenore Dedeyne, Laura Vercauteren, Nadjia Amini, and Laurence Lapauw are with the Division of Geriatrics and Gerontology, Department of Public Health and Primary Care, KU Leuven, 3000 Leuven, Belgium.

Jolan Dupont and Evelien Gielen are with the Division of Geriatrics and Gerontology, Department of Public Health and Primary Care, KU Leuven, 3000 Leuven, Belgium, and also with the Department of Geriatric Medicine, UZ Leuven, 3000 Leuven, Belgium.

Sabine Verschueren is with the Musculoskeletal Rehabilitation Research Group, Department of Rehabilitation Sciences, KU Leuven, 3000 Leuven, Belgium.

Carolina Varon is with the Department of Electrical Engineering, STADIUS, KU Leuven, 3000 Leuven, Belgium.

Walter De Raedt is with IMEC, 3001 Leuven, Belgium.

Bart Vanrumste is with the e-Media Research Laboratory, KU Leuven, 3000 Leuven, Belgium.

Digital Object Identifier 10.1109/TNSRE.2024.3355299

I. Introduction

OLDER adults suffer from higher fall risk and consequent fall-related injuries. Annually, 24% to 40% of community-dwelling older persons fall, of which 21% to 45% fall regularly [1]. Many of these persons need to spend a long time recovering from injuries. Older adults with certain diseases such as sarcopenia and obesity could suffer from even higher fall risk [2]. With the increase of the older population, this problem increases the costs of healthcare systems [3].

To reduce fall risk, the Otago Exercise Program (OEP) was developed for older adults. OEP contains a series of balance, strength, and walking exercises, and it has been proven to reduce fall risk and mortality for community-dwelling older adults [4]. There are more than twenty types of exercises (sub-classes) in the OEP and the participants need to perform these sub-classes sequentially. Older adults participating in the OEP are requested to perform the exercises twice or three times a week within some consecutive weeks and their compliance with performing OEP is monitored by means of self-reports (diaries) [5], [6]. However, this is not an accurate method to assess their involvement in the OEP, since they sometimes inaccurately remember or report information about their past experiences. Besides, the diaries only record the duration of performing the OEP rather than which OEP sub-classes are performed.

An alternative to the self-reports is to apply wearable sensors combined with machine learning techniques. Although Internet-of-Things (IoT) sensors have been widely applied for

Human Activity Recognition (HAR), their use for recognizing the OEP has rarely been investigated. In fact, to date, only two studies have used wearable sensors for this purpose. The first study [7] applied five sensors to recognize the OEP, which were not user-friendly for older adults to wear in daily life. Although the f1-scores for most activities were high, the recruited subjects only included healthy young adults. Besides, the study did not try to distinguish between the OEP and Activities of Daily Life (ADLs). The second study [8] applied one Inertial Measurement Unit (IMU) on the waist to collect data from older adults, and two problems were investigated: 1) to distinguish between OEP and ADLs, and 2) to recognize OEP sub-classes. However, for both tasks, the f1-scores did not exceed 0.8.

Considering the gaps in the previous studies, this study investigates OEP recognition for community-dwelling older adults using a single IMU on the waist. The aim is to build an offline HAR system for the therapists to analyze their long-term compliance with OEP. The OEP was recognized in two stages: 1) using a large sliding window to recognize the occurrence of OEP and ADLs and 2) using a small sliding window to recognize OEP sub-classes. The data were collected in both a lab scenario and home scenarios from older adults to validate the proposed system. Machine learning and deep learning models were applied and compared in the two stages to optimize the recognition performance.

The contributions of this study are:

1) For both lab and home scenarios, a single wearable IMU was applied to recognize:
   a) the OEP from ADLs
   b) some sub-classes of OEP

   To date, this is the least obtrusive wearable system that could be applied in daily life.
2) A hierarchical architecture was designed for activity classification, where a large sliding window was used to recognize OEP while a following small sliding window was used to recognize OEP sub-classes.
3) A state-of-the-art deep learning model combining two different neural network architectures was built to classify OEP and ADLs with the results outperforming other models.
4) A new approach for defining OEP sub-classes. Initially, level 1 OEP sub-classes were labeled and classified. These sub-classes were further classified into level 2 sub-classes. This approach led to a reduction in the number of classes used in machine learning models, resulting in improved classification performance.

The paper is organized as follows: Section II reviews the state of the art in the field of HAR. Section III introduces the datasets and the implementation of the proposed system. Section IV presents the experimental results and section V discusses the results. Finally, section VI concludes the paper and proposes future work.

II. Related Works

In this section, two topics are discussed: the sliding window techniques and machine learning models that have been used in HAR systems.

A. Sliding Window

A common pre-processing method for HAR is to segment the signals into smaller pieces with the same length for feature extraction [9]. This method is called the sliding window technique. The size of the applied sliding window is important for classification results and dependent on the characteristics of the activities to be recognized. For example, small sizes such as 2s could be applied to recognize static and periodic activities (e.g. walking, sitting, etc.) [10]. For these activities, it is suggested that the window size should cover at least one cycle of the activities [9]. On the other hand, a sliding window size of as large as 1 minute could be applied for activities involving more gestures (e.g. eating) [11]. Since these activities include many sub-activities, it is harder to define the appropriate window size. Sometimes the sliding window is applied with an overlap rate to offer more segments.

B. Machine Learning Models

Traditionally, feature-based machine learning has been used in HAR systems. For this technique, hand-crafted features were important to be extracted from the raw signals. Typical hand-crafted features include time-domain and frequency-domain features [9]. These features might not capture all relevant information for classification.

Thanks to the evolution of computational power, deep learning models have been applied for the representation of raw signals. With the features extracted from the deep learning models, the recognition performance increased in many cases compared with the models based on hand-crafted features [12], [13]. However, deep learning models normally need more data for training.

Convolutional neural networks (CNN) could extract neighboring information from adjacent samples by convolutional layers to be classified by the fully-connected layers. Recently, such architecture is popular for sensor signals classification/regression by performing 1-D [14] or 2-D [15] convolution. Recurrent neural networks (RNN), on the other hand, extract temporal features from the time series. For HAR systems, Long Short-Term Memory (LSTM) as a type of RNN units has also been proven efficient for long time series [16]. Besides, Transformers have also been explored for HAR using attention mechanism [17], [18]. The mechanism searches for correlation between features by mapping a query and a set of keys, which makes it efficient for long-time series recognition.

With the development of CNN and LSTM, a new architecture called CNN-LSTM was developed for HAR systems [19], [20]. The features extracted by the convolutional layers were applied as the input for the LSTM units. Then the LSTM units further extract the temporal features for classification. This architecture has been validated for many datasets with better recognition performance than both CNN and LSTM [19], [20]. It also outperformed Transformers in recent research for HAR [18]. Based on this architecture and Bi-directional-LSTM (BiLSTM) layers, the CNN-BiLSTM was further developed [21], [22]. This architecture was applied to learn from both forward and backward time series.

TABLE I
THE LABELS FROM LEVEL 1 OEP (SIX CLASSES), LEVEL 2 OEP (15 CLASSES), AND ORIGINAL OEP

Level 1 OEP	Level 2 OEP	Original OEP
Genral walking	Marching	Marching
	Backwards walking	Backwards walking, Heel toe backwards walking
	Forwards walking	Tandem walking, Toe walking
	Walking and turn	Walking and turn
	Sideways walking	Sideways walking
	Stairs walking	Stairs walking
General standing	Back moblizer	Back moblizer
	Ankle plantarflexors	Ankle plantarflexors
	Ankle dorsiflexors	Ankle dorsiflexors
	Knee bends	Knee bends
	Static standing	Head mobilizer, Neck mobilizer, Hip abductor, Knee flexors, Tandem stance, One leg stance
Trunk moblizer	Trunk moblizer	Trunk moblizer
Abdominal muscles	Abdominal muscles	Abdominal muscles
Sit to stand	Sit to stand	Sit to stand
Sitting	Sitting	Head moblizer, Neck moblizer, Ankle moblizer, Knee extensor, Plantar and knee flexor, Hip extensor

In this work, a deep learning architecture was applied to distinguish OEP and daily activities.

III. METHODOLOGY

A. Data Collection

This study received approval from the Ethics Committee Research UZ/KU Leuven (S59660 and S60763). Written informed consent was obtained from all participants prior to study participation.

In the experiments, community-dwelling older adults (aged 65 and older) performed modified OEP while wearing a McRoberts MoveMonitor+ (McRoberts B.V., Netherlands) with a 9-axis IMU inside. To make a user-friendly system, the subjects were asked to wear the device loosely and comfortably on the waist, with possible misplacement, and without any calibration. The OEP is a rehabilitation program designed to reduce fall risk in older adults. The original program includes multiple sub-classes that need to be performed sequentially. The sub-classes are shown in Table I. Besides, subjects also performed other ADLs. Two datasets were collected in the study in different scenarios:

1) Dataset 1: Lab: The dataset was collected in the lab. The subjects performed the OEP and ADLs with instructions from the researchers certified as OEP leaders. The data of OEP and ADLs were collected either on two separate days or consecutively on one day. The ADLs included walking, walking stairs, sitting, standing, and indoor cycling.

2) Dataset 2: Home: This dataset was collected at home with videos recorded. With the camera turned on, the subjects wore the device by themselves and followed a booklet with instructions to perform OEP. Before and/or after the OEP, the subjects also randomly performed ADLs at home, out of the camera view. Therefore, the ADLs were not observed but still labeled while the subjects were wearing the IMU.

The recruited older adults were (pre-)sarcopenic or non-sarcopenic (defined by EWGSOP1 [23]). The detailed information is shown in Table II. The recruited subjects of the two datasets were different. For both datasets, each OEP sub-class was recorded and annotated based on videos. Between the sub-classes, the subjects were not instructed nor monitored. They could be practicing the exercises, sitting for a rest, or walking out of the camera view.

TABLE II
SUBJECTS INFORMATION FROM THE TWO DATASETS

	number	age	gender & sarcopenia status
dataset 1	11	84.09±5.28	5 females (2 (pre-)sarcopenia) 6 males (3 (pre-)sarcopenia)
dataset 2	7	69.43±2.92	4 females (2 (pre-)sarcopenia) 3 males (2 (pre-)sarcopenia)

B. OEP Exercises Annotation

1) Level 1 OEP: Directly classifying the OEP sub-classes was difficult for the machine learning models, due to the relatively small number of training examples and large number of classes. Therefore, the OEP sub-classes were merged according to the characteristics of the exercises, in order to build a hierarchical system, as proposed in [24]. This technique reduced the number of classes while maintaining the same number of training examples. As shown in Table I, some sub-classes were merged as *general walking*, and some were merged as *general standing*. To distinguish from *static standing*, *general standing* also included the exercises that required irregular trunk movement while standing (e.g. *ankle plantarflexors*). As a result, there were six level 1 OEP exercises.

2) Level 2 OEP: Although there were 26 original OEP sub-classes as shown in table I [8], a single IMU on the waist could not distinguish all of them. Since many OEP sub-classes do not involve trunk movements (e.g. head mobilizer, neck mobilizer), they were merged as *sitting* and *static standing*. After merging, there were 15 OEP sub-classes to be recognized as shown in Table I. These sub-classes were called level 2 OEP (top level). The recognition of level 1 OEP and level 2 OEP were cascaded to improve performance. The details will be introduced in section III-C.

3) Transition Activities: During the OEP, the subjects followed the instructions either from the researchers (dataset 1) or from the booklets (dataset 2). Before each OEP sub-class, they could be practicing the exercises, sitting for a rest, or walking out of the camera view. These activities could not be properly monitored or labeled. Therefore, the activities between each two OEP sub-classes were annotated as transition activities. Different from ADLs that happened outside the whole OEP, transition activities were short activities happening within the OEP program. These activities were not included in the training set.

In these annotation steps, if a segment contained multiple activities, it was labeled as the majority class and removed when training the models. The duration and number of segments (by a 6-second window with 50% overlap) are shown in Table III. The duration of transition activities was observed longer than OEP in dataset 1, which can be attributed to the

TABLE III
THE TOTAL DURATION (MINUTES) AND NUMBER OF SEGMENTS (BY A 6-SECOND WINDOW WITH 50% OVERLAP) IN THE DATASETS

	duration		number of segments	
	dataset 1	dataset 2	dataset 1	dataset 2
OEP	196.91	219.29	3436	1582
transition	543.94	89.46	10472	4102
ADLs	947.92	828.7	18958	23612

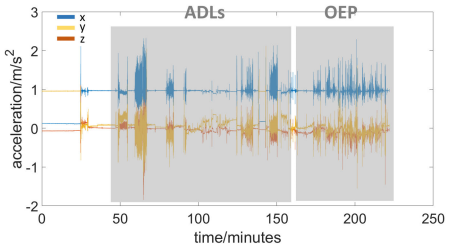


Fig. 1. An example of recorded acceleration from one subject.

older age and increased need for rest among the subjects. For dataset 2, *walking stairs* was not performed because of the limitation of camera installation.

C. System Overview

An example of collected signals from a single subject is shown in Fig. 1. The general overview of the two proposed experiments is shown in Fig. 2. Two stages of classification were proposed for different activities to be recognized.

1) Stage 1 (OEP Vs. ADLs): Considering the characteristics of OEP and ADLs, a large sliding window size was applied to segment the signals. The optimal window size and overlap rate were identified by comparing performance across window sizes of 5, 10, and 15 minutes, and overlap rates of 25%, 50%, 75%, and 80%, using a CNN-BiLSTM model. The results are shown in section IV-A where a 10-minute window with 75% overlap was selected. The segmented signals were then classified as OEP or ADLs (binary classification).

2) Stage 2 - Level 1 (six Classes of OEP Classification): The signals were segmented with a smaller sliding window. The optimal window size and overlap rate were identified by comparing performance across window sizes of 2, 4, 6, and 8 seconds, and overlap rates of 25%, 50%, 75%, and 80%, using a Random Forest model. The results are shown in section IV-B where a 6-second window with 50% overlap was selected. If the signals were classified as OEP in stage 1, they were further classified as the six classes of level 1 OEP as shown in Table I. Otherwise, they kept the ADLs labels as in stage 1.

3) Stage 2- Level 2 (15 Classes of OEP Classification): Following the results from level 1 classification, *general walking* or *general standing* were further classified by other models, whereas the other segments remained the same as level 1. There were consequently 15 classes in level 2 as shown in Table I. In level 2, the labels segmented in level 1 were further classified. Therefore, there were not any segmentation procedures.

TABLE IV
HAND-CRAFTED FEATURES EXTRACTED FROM THE IMU CHANNELS

Time-domain	Frequency-domain
Interquartile range	FFT mean coefficient
Kurtosis	Fundamental frequency
Max	Human range energy (0.6-2.5Hz)*
Mean	Max power spectrum
Median	Maximum frequency
Min	Median frequency
Root mean square	Spectral entropy
Skewness	Spectral kurtosis
Standard deviation	Spectral skewness
Variance	
Absolute energy	
Autocorrelation	
Centroid	
Entropy	
Zero crossing rate	

*The frequency band was proposed as the dominant frequency of human activity [29]

D. Pre-Processing

In this study, only the accelerometer (a_x, a_y, a_z) and gyroscope (g_x, g_y, g_z) from the IMU were used for classification, since they are more widely used than magnetometers for HAR in health care [9]. Also, adding a magnetometer did not improve the results by comparing classification performance.

For noise reduction, the raw signals were low-pass filtered by a 6-order Butterworth filter with a cut-off frequency of 10Hz. Then the signals were segmented using sliding windows as explained in III-C (10-minute window with 75% overlap for stage 1 and 6-second window with 50% overlap for stage 2).

An additional magnitude acceleration channel a_M [25], [26] was extracted according to the formula:

$$a_M = \sqrt{a_x^2 + a_y^2 + a_z^2}, \tag{1}$$

E. Hand-Crafted Features and Feature Selection

Time-domain and frequency-domain features were extracted from the seven channels (six original channels and one magnitude acceleration channel). The types of hand-crafted features are shown in Table IV. In total, 15 time-domain features and nine frequency-domain features were extracted from each channel, all implemented using TSFEL package [27]. Additionally, the subject information was included as meta-features, including the age, gender, weight, height, and health condition (sarcopenia, pre-sarcopenia, or no sarcopenia defined by EWGSOP1). In [28], it was found that the subject information had an added value for the recognition performance.

Considering that some of the OEP exercises had to be performed with a certain order in the program (e.g. marching was always at the beginning of the OEP as a warm-up exercise), a feature named relative start time was extracted for each segment in stage 2:

$$f_{rst} = \frac{t_{sseg} - t_{sOEP}}{t_{eOEP} - t_{sOEP}}, \tag{2}$$

where t_{sseg} denotes the start time of the segment (in stage 2), t_{sOEP} denotes the predicted start time of the whole OEP, and t_{eOEP} denotes the predicted end time of the whole OEP. From stage 1, t_{sOEP} and t_{eOEP} could be predicted. This feature

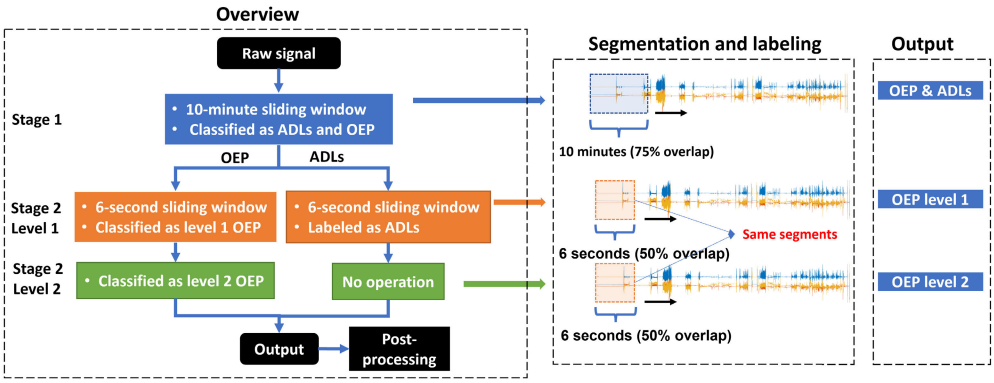


Fig. 2. An overview of the proposed system. NOTE: To better illustrate the process, the ratio of the sliding windows and signals in the figure does not correspond to the actual size. The post-processing stage is explained in the following sections.

indicates the approximate position of a certain exercise that occurred in the program.

In summary, 173 and 174 features were extracted for each segment in stage 1 and stage 2, respectively, since f_{rst} was only extracted in stage 2.

Feature selection was performed by combining neighborhood component analysis and forward selection [30]. After splitting training and validation sets, a feature set was selected for each training set, as the details explained in Section III-H.

F. Machine Learning Models

1) Classical Machine Learning: Three classical models were applied: K-Nearest Neighbors (KNN), Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel, and Random Forest (RF). Based on (standardized) hand-crafted features, these models were widely applied for HAR [9], [31]. They were applied in both stage 1 and stage 2. In stage 1, however, they were only applied as baseline models, since deep learning models were also applied.

2) Deep Learning: Three deep learning models were applied in stage 1 classification: CNN, Transformer, and CNN-BiLSTM. They were not involved in stage 2 due to the small number of training examples. To reduce the computational cost, the input time series was down-sampled from 100Hz to 20Hz. Therefore, the input size to the models was 12000*6, where 12000 was the number of samples in 10 minutes and 6 was the number of original IMU channels.

a) CNN: The CNN architecture was introduced in [14]. It included three convolutional layers, each followed by a max-pooling layer to reduce the number of time samples and a dropout layer to avoid over-fitting. Then, the extracted features were flattened and classified by two fully-connected layers.

b) Transformer: The Transformer architecture was proposed by [17]. After normalization and position embedding, three encoder blocks were applied. Each encoder block consisted of a multi-head attention layer, a fully-connected layer, and dropout layers. Then, the classification results were generated by another fully-connected layer.

c) CNN-BiLSTM: The CNN-BiLSTM model applied two convolutional layers for local features learning and a following BiLSTM layer for temporal features learning, as shown in Fig. 3.

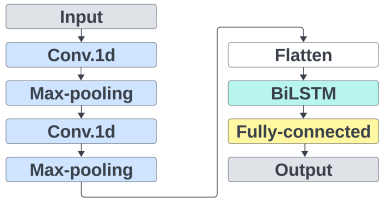


Fig. 3. The architectures of the CNN-BiLSTM model. Each convolutional layer and LSTM layer was followed by a dropout layer, which is not shown in the figure. The hyperparameters of were tuned according to Table V.

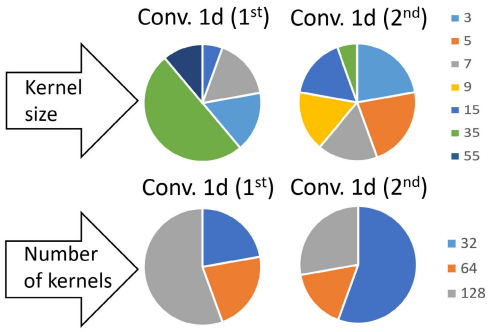


Fig. 4. Some selected hyperparameters in each outer loop from CNN-BiLSTM in stage 1 (dataset 1+ dataset 2).

The deep learning models were built on tensorflow 2.8 and were trained on a NVIDIA P100 SXM2 GPU. A batch size of 64 was applied for training. Adam algorithm was used to optimize the cross-entropy loss function. To reduce the influence of the imbalanced dataset, all models applied the "class weight" option to weight the loss function according to the number of samples for each class.

The hyperparameters were tuned using Hyperband Search [32]. The search spaces of the hyperparameters for each model are shown in Table V. Since the input time series was large in stage 1 (input length= 12000 samples), the kernel size in the convolutional layers was tuned in a large search space. Fig. 4 shows some selected hyperparameters in each outer loop from CNN-BiLSTM in stage 1. The test and validation of the models are described in Section III-H.

TABLE V
THE SEARCH SPACE OF EACH TUNED HYPERPARAMETER OF THE MODELS

Model	Hyper-parameter	Search space
SVM	Regularization parameter	[10, 1, 0.1, 0.01, 0.001]
	Kernel value (gamma)	[10, 1, 0.1, 0.01, 0.001]
KNN	Number of neighbors	[3, 5, 7, 9]
RF	Number of trees	[50, 100, 200]
	Maximum depth of the tree	[1:1:35]
Deep learning models	Conv. 1d: Kernel size	[3, 5, 7, 9, 11, 15, 35, 55]
	Conv. 1d: Number of filters	[32, 64, 128]
	BiLSTM: Number of units	[32, 64, 128]
	Attention: Number of heads	[3, 4, 5, 6]
	Fully-connected: Number of neurons	[32, 64, 128]
	Droupout rate	[0.1, 0.2, 0.5]
	Learning rate	[0.1, 0.01, 0.001, 0.0001]
	Pooling size	[2, 3, 5, 10]

G. Post-Processing

The transition activities and ADLs could be similar to the OEP sub-classes. For example, *sit-to-stand* also happened in transition activities and ADLs. However, only consecutive *sit-to-stand* labels should be classified as OEP sub-classes. To reduce the negative influence of the ADLs and transition activities in stage 1 and stage 2, post-processing was applied to improve recognition performance.

To smooth the predicted labels $O = \{o_1, o_2, \ldots, o_n\}$, as a time series, the post-processing algorithm was applied, as described in Algorithm 1. It took the predicted segments as input and returned smoothed segments $P = \{p_1, p_2, \ldots, p_n\}$. A smoothing window was moving along the time series. The values of post-processing window lengths were selected from a range of values (from 10 to 20 minutes for stage 1, and from 21 to 39 seconds for stage 2) by comparing classification performance. Finally, the window length was seven segments (i.e. 17.5 minutes) for stage 1, and 11 segments (i.e. 33 seconds) for stage 2. The label was classified as transition activities if the most adjacent labels were from different classes. The algorithm could correct some misclassified segments, which is shown in section IV-A.

Algorithm 1 Post-Processing

INPUT: a time series of the original predicted labels $O = \{o_1, o_2, \ldots, o_n\}$; k=3 for stage 1 and k=5 for stage 2
OUTPUT: a time series of the post-processed predicted labels $P = \{p_1, p_2, \ldots, p_n\}$

Initialize P with the same length as O
for i=k+1, k+2, k+3, …, n-k-2, n-k-1, n-k **do**
p_i = the majority class of the sequence $\{o_t \in O \mid (i-k) \leq t \leq (i+k)\}$
end for

H. Dataset Split

The validation for dataset 1 applied nested Leave-One-Subject-Out Cross-Validation (LOSOCV). For each outer loop, the data from one subject was in the test set whereas the rest was in the training set. For hyperparameters tuning, in each inner loop, a single subject was excluded from the validation set. In other words, the best average weighted f1-scores of all validation sets in the inner loops were obtained. The best model hyperparameters and the best feature set were then applied to the test set in the outer loop. The final evaluation performance was calculated from all test sets.

The validation for dataset 2 was similar to dataset 1. The only difference is that dataset 1 was also included in the training set, to further improve the classification performance.

I. Evaluation

1) Window-Wise Evaluation: For each class c, the f1-score was calculated as:

$$f1_c = 2 * \frac{precision_c * recall_c}{precision_c + recall_c}, \tag{3}$$

where

$$precision_c = \frac{TP_c}{TP_c + FP_c}, and \tag{4}$$

$$recall_c = \frac{TP_c}{TP_c + FN_c}. \tag{5}$$

where TP_c, FP_c, and FN_c denote the numbers of true positive, false positive, and false negative labels.

For overall evaluation in stage 2, weighted f1-score was calculated from the f1-score and the number of labels of class c:

$$weighted\ f1 = \sum_c f1_c * N_c, \tag{6}$$

where N_c denotes the number of true labels belonging to class c.

2) Segment-Wise Evaluation: For this evaluation method, the predicted labels based on the sliding windows were reconstructed as a time series. A segment was defined as an aggregation of consecutive labels that belonged to the same class [33]. Then segmental f1-scores of Intersection over Union (IoU) were calculated for each class. The IoU matrix was defined as the ratio of intersection and the union of the predicted and true segments. After the true and predicted labels were reconstructed as a time series, a threshold of IoU values was set. Then, segment-wise true positive ($TPseg$), false positive ($FPseg$), and false negative ($FNseg$) were defined:

- $TPseg$: IoU≥threshold
- $FPseg$: IoU<threshold, true segments shorter than predicted segments
- $FNseg$: IoU<threshold, true segments longer than predicted segments

Fig. 5 illustrates the definition of IoU, TP, FP, and FN. From these values, segment-wise precision, recall, and f1-scores could be calculated by formula 3, 4, 5. Compared with the window-wise evaluation, this method could also evaluate the over-segmentation error. In this study, the overlaps of 0.5 and 0.75 were applied for stage 1 evaluation.

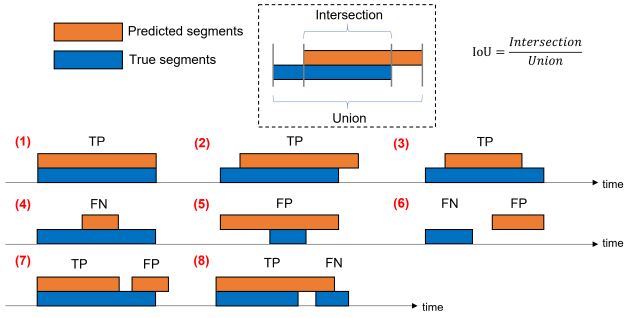


Fig. 5. The definition of IoU, TP, FP, and FN. There are eight cases shown in the figure. In case 4 and case 5, FN or FP depends on the length of the true and predicted segment. In case 7, if a true segment is predicted as some smaller segments, FP numbers increase. In case 8, if some separate true segments are predicted as one segment, FN numbers increase.

TABLE VI
OEP F1-SCORES USING DIFFERENT SLIDING WINDOW SIZES (OVERLAPS) IN STAGE 1 USING CNN-BILSTM

	5-min (50%)	10-min (75%)	15-min (85%)
dataset 1	0.924	**0.983**	0.960
dataset 2	0.916	**0.950**	0.943

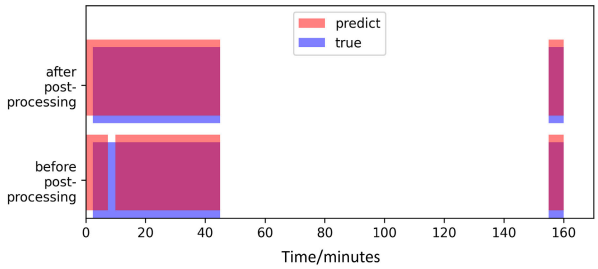


Fig. 6. An example of reconstructed time-series of labels.

TABLE VII
THE WINDOW-WISE (F1, PRECISION, RECALL) AND SEGMENT-WISE (IOU F1) EVALUATION IN STAGE 1 BY DIFFERENT MODELS AFTER POST-PROCESSING

	KNN	RF	SVM	CNN	Trans-former	CNN-BiLSTM
dataset 1						
f1	0.869	0.949	0.847	0.874	0.949	**0.983**
precision	0.952	0.984	0.959	0.862	0.984	**1.000**
recall	0.799	0.916	0.759	0.887	0.916	**0.967**
IoU f1@0.5	0.938	1.000	1.000	0.714	0.762	**1.000**
IoU f1@0.75	0.750	0.133	0.133	0.619	0.654	**0.867**
dataset 2						
f1	0.616	0.692	0.627	0.849	0.909	**0.950**
precision	0.507	0.585	0.571	0.789	0.927	**0.931**
recall	0.786	0.847	0.694	0.918	0.891	**0.969**
IoU f1@0.5	0.000	0.087	0.000	0.857	0.821	**1.000**
IoU f1@0.75	0.000	0.087	0.000	0.857	0.125	**1**

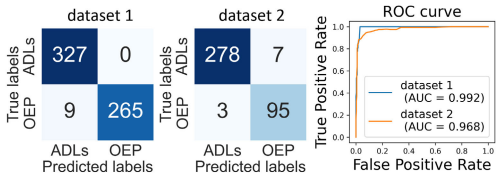


Fig. 7. Confusion matrices and ROC curves for stage 1 classification by CNN-BiLSTM after post-processing.

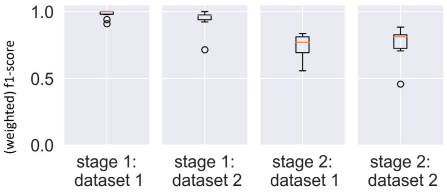


Fig. 8. Boxplots of the f1-scores in stage 1 and weighted f1-scores in stage 2 for each subject.

IV. RESULTS

A. Stage 1 Classification

For both datasets, a 10-minute sliding window was selected from 5, 10, and 15 minutes, and an overlap rate of 75% was selected from 25%, 50%, 75%, and 80%, using a CNN-BiLSTM model. They were hence applied for stage 1 classification. Table VI shows the impacts of some values of window size and overlap rate.

The window-wise and segment-wise f1-scores of OEP classified by different machine learning models are shown in Table. VII. The window-wise f1-scores of the CNN-BiLSTM architecture were higher than the other models in both datasets (0.983 and 0.950, respectively). Although the f1-scores dropped from dataset 1 to dataset 2, such a drop is smaller using the CNN-BiLSTM models. For dataset 2, the IoU f1-scores of classical machine learning models significantly decreased. The IoU f1-scores (75%) of CNN-BiLSTM were found to be 0.867 and 1.000, respectively, which outperformed the other models.

An example of the predicted time series of stage 1 is shown in Fig 6, which shows the impacts of post-processing.

The confusion matrices and receiver operating characteristic (ROC) curves by the CNN-BiLSTM models are shown in Fig. 7. For dataset 1, all the ADLs segments were classified correctly. On the other hand, nine OEP segments were classified as ADLs. For dataset 2, seven ADLs segments were classified as OEP while three OEP segments were classified as ADLs segments. Fig. 8 shows the boxplot illustrating the f1-scores for each subject in stage 1. According to the figure, there were three outliers in total. Apart from this, the f1-scores among subjects showed low variance.

B. Stage 2 Classification

For both datasets, a 6-second sliding window was selected from 2, 4, 6, and 8 seconds, and an overlap rate of 50% was selected from 25%, 50%, 75%, and 80%, using a RF model. They were hence applied for stage 2 classification. Table VI shows the impacts of some values of window size and overlap rate.

As feature selection was performed on each iteration in LOSOCV, each training set resulted in a different feature set. The ten most frequently selected features were: Relative start time, Entropy (a_y), Entropy (a_M), Entropy (a_x), Mean (g_z), Centroid (a_M), Median (g_x), Centroid (g_x), Kurtosis (a_y), Centroid (a_x).

TABLE VIII
WEIGHTED F1-SCORES USING DIFFERENT SLIDING WINDOW SIZES IN STAGE 2 (WITH 50% OVERLAP) USING RF

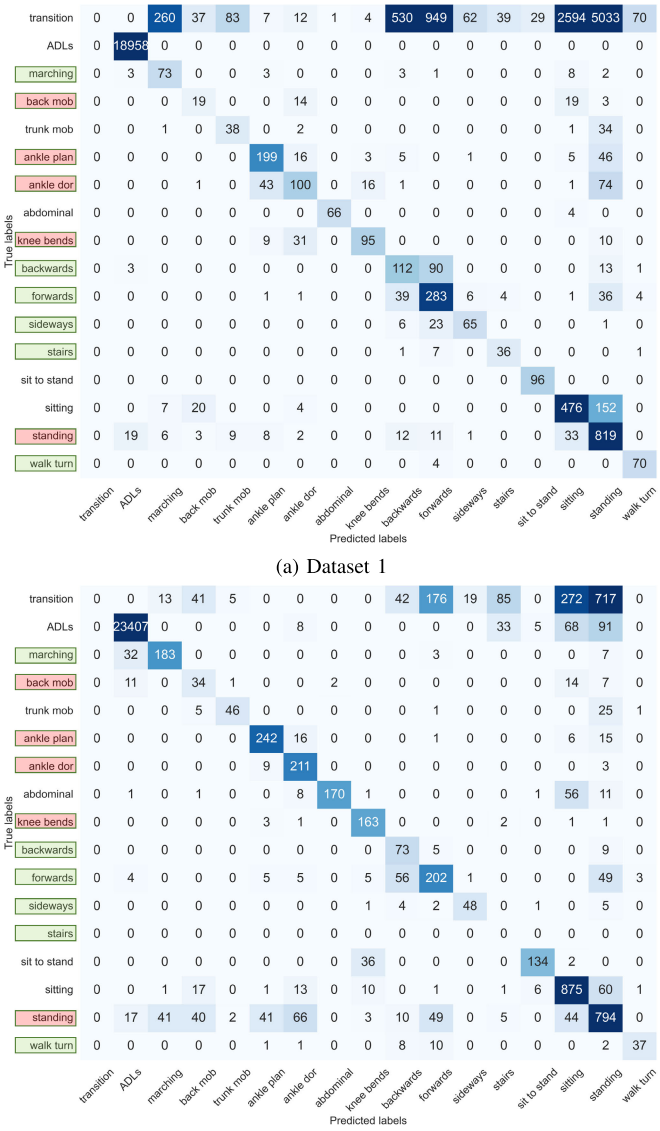
	2s	4s	6s	8s
dataset 1	0.733	0.741	**0.743**	0.739
dataset 2	0.722	0.775	**0.798**	0.766

TABLE IX
WEIGHTED F1-SCORES IN STAGE 2 BY DIFFERENT MODELS

	KNN	RF	SVM
dataset 1	0.653	**0.743**	0.689
dataset 2	0.699	**0.798**	0.724

The weighted f1-scores in stage 2 by different models are shown in Table IX. For both datasets, RF models obtained the best weighted f1-scores of 0.743 and 0.798, respectively. The confusion matrices for both datasets by RF are shown in Fig. 9. Compared with stage 1, ADLs in stage 2 showed less confusion because of the further post-processing in stage 2. On the other hand, the transition activities led to false positive predictions for most of the classes.

Fig. 10 shows the f1-scores for some activities. As illustrated by the light colors, in dataset 2 at home, there were four activities beyond the threshold of 0.8: *ankle plantarflexors*, *abdominal muscles*, *knee bends*, and *sit-to-stands*. With transition classified as FP, there were three activities with f1-scores decreased by more than 0.1: *sideways walking*, *stairs walking*, and *walking and turn*. Fig. 8 shows the boxplot illustrating the weighted f1-scores for each subject in stage 2, showing larger variance than stage 1.



(a) Dataset 1

(b) Dataset 2

Fig. 9. Confusion matrices for stage 2 classification by RF after post-processing. The green labels denote *general walking* and the red labels denote *general standing* in level 1. (abbreviations: plan= plantarflexors, dor= dorsiflexors, mob= mobilizer). There are no FP labels for transition activities class, because the labels were not in the training set.

V. DISCUSSION

A. Stage 1

This study reported high f1-scores for OEP recognition in both lab and home environments. In stage 1, a large sliding window size (10 minutes) was applied to segment the signals. Compared with the previous study applying a 4-second window [8], the f1-scores in this study were higher than 0.95 for both datasets. With a small window size, each activity was segmented individually. However, most individual OEP activity was almost the same as ADLs. For example, *forwards walking* in the OEP was the same as walking in daily life.

A large window size, on the other hand, included multiple exercises instead of a single one. Therefore, one segment contained more information. As shown in Fig. 1, each OEP sub-class took a shorter duration than ADLs. Also, the signals of OEP showed higher variance than ADLs. Therefore, a large sliding window could better identify these characteristics. A 10-minute sliding window with 75% overlap resulted in a delay of 2.5 minutes, which was applicable for offline systems. In real life, the system could also be applied as a reminder to perform OEP for older adults.

The annotation of OEP was from the beginning of the first OEP activity to the end of the last OEP activity. During the program, the subjects could take a rest on a chair or go to the bathroom. These behaviors are also expected in real life. Therefore, the smoothing algorithm was applied to also take non-OEP activities into account and hence, the recognition performance was improved.

In stage 1, CNN-BiLSTM was proven to be the best classification model for both datasets. The scenarios using a large window (long time series) for OEP recognition could take advantage of CNN-BiLSTM. Similar results were also reported by a previous study, reporting that CNN-LSTM outperformed CNN, LSTM, and Transformer for IMU-based activity recognition [18]. However, deep learning models require more training examples than classical models. Thus, the models still encountered overfitting problems due to dataset size. Although the usage of dropout layers improved this problem and made the models more efficient than hand-crafted features, the lack of training examples was still a limitation of the study.

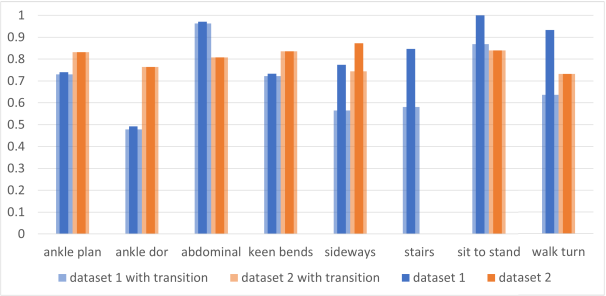


Fig. 10. F1-scores of some classes in stage 2 by RF after post-processing. The light colors illustrates the results with counting the transition activities.

TABLE X
COMPARISON: OEP F1-SCORES OF THE PREVIOUS STUDY AND THIS STUDY

previous study [8]	Dataset 1	Dataset 2
0.777	0.983	0.950

The amount of ADLs collection was limited. In dataset 1, the ADLs included *walking*, *standing*, *sitting*, and *cycling*. In dataset 2, the duration of ADLs was no more than two hours. In real life, ADLs happen much more than OEP. Therefore, in the future, the methods should be validated on the dataset with more types and longer duration of ADLs. Another limitation was that the study did not consider calibration and misplacement of the sensor, since the older adults could not comply with these tasks alone. Although it would result in decreased f1-scores, the system still showed robustness in generalization according to Fig. 8.

B. Stage 2

The size of the training set was small compared with the number of OEP classes. To reduce the number of classes for each model, a hierarchical classification method was applied. Considering the transition activities between each two OEP sub-classes, the f1-scores of some sub-classes were decreased such as *sideways walking*, *walking stairs*, *walking turn*, and *sit-to-stand*. These activities could also happen in ADLs. Therefore, some transition activities were classified as the ones in OEP sub-classes, although such influence was improved by post-processing in stage 2. Since the results of stage 2 were based on the output of stage 1, they showed higher variance according to Fig. 8.

The f1-scores of *knee bends* and *ankle plantarflexors* were increased in dataset 2. The first reason was that dataset 1 was also applied for training to test dataset 2. With more training data, the models were less over-fitting. The second reason was that the subjects in dataset 2 were younger than in dataset 1, as shown in Table II. Therefore, the subjects in dataset 2 performed the exercises with less intra-class variance.

On the other hand, compared to dataset 1, exercises such as *abdominal muscles* and *sit to stand* had decreased f1-scores in dataset 2. Such a decrease was due to the fact that the subjects followed the booklet rather than the direct instructions from the therapists. Therefore, they were unable to adhere to the instructions effectively.

TABLE XI
COMPARISON: F1-SCORES FOR OEP SUB-CLASSES OF THE PREVIOUS STUDY (WITH THE WAIST-MOUNTED IMU) AND THIS STUDY

	ankle dor	knee bends	sideways	sit to stand
previous study [7]	0.370	0.310	0.790	0.490
This study: dataset 1	0.478	0.722	0.565	**0.869**
This study: dataset 2	**0.764**	**0.836**	**0.744**	0.840

Because of the lack of training data, deep learning models were not applied in stage 2, which limited this study. In future studies, more data will be collected so that deep learning models can be applied to improve performance.

C. Comparison With Previous Studies

There were only two studies applying wearable sensors for OEP recognition. The first study applied a waist-mounted IMU to recognize OEP and ADLs for (sarcopenic) older adults [8]. The study categorized OEP exercises into strength, balance, and other exercises, without classifying the OEP sub-classes. The f1-scores for classifying OEP and ADLs of the study are shown in Table X. The results show that our proposed 10-minute sliding window and CNN-BiLSTM model obtained higher f1-scores than the previous study using a 4-second sliding window and a RF model. Besides the advantages of the deep learning models, the large sliding window also shows its ability to capture more information, leading to improved results.

The second study applied five IMUs on the four limbs and waist worn by young adults [7]. Since the study categorized OEP exercises differently, only four classes were comparable to this study. Also, although the study applied five IMUs, only the results with the waist-mounted IMU were compared to this study. Table XI compares the f1-scores of the previous study and this study. The results showed that the previous study (only validated on four healthy young subjects) did not obtain f1-scores over the threshold (0.8, which is a common threshold value for OEP [8]) for the four classes. On the other hand, our proposed method obtained higher f1-scores for these classes, with two of them (*knee bends* and *sit to stand*) exceeding the threshold. Additionally, the previous study did not take into account the transition activities between OEP sub-classes. The superior performance of our proposed method may be attributed to the two-level labeling of OEP sub-classes and relevant hand-crafted features such as relative start time.

VI. CONCLUSION

This study proposes a hierarchical system to recognize OEP from a waist-mounted IMU. The system was tested on the older adults in both lab and home environments. In stage 1, using the CNN-BiLSTM architecture, the system could distinguish OEP and ADLs with f1-scores over 0.95. The results showed the capability of a single IMU to evaluate the compliance of OEP for older adults with and without sarcopenia. Besides, the system could distinguish four OEP sub-classes with f1-scores over 0.8: *ankle plantarflexors*, *abdominal muscles*, *knee bends*, and *sit-to-stand*. The results

showed the potential of monitoring the compliance of OEP using a single IMU in daily life. Furthermore, the proposed system demonstrated its capability to recognize and analyze OEP sub-classes.

In the future, the size of the dataset should be improved, since the amount of training examples did not support deep learning models in stage 2. Also, the hyperparameters of this system could be reduced for better generalization. For example, end-to-end deep learning models such as temporal convolutional networks (TCN) [33] could be applied without post-processing.

Acknowledgment

The funding provider did not contribute or influence the design of the study and data collection, analysis, and interpretation in writing this manuscript.

References

[1] A. J. Milat, W. L. Watson, C. Monger, M. Barr, M. Giffin, and M. Reid, "Prevalence, circumstances and consequences of falls among community-dwelling older people: Results of the 2009 NSW falls prevention baseline survey," *New South Wales Public Health Bull.*, vol. 22, no. 4, p. 43, 2011.

[2] F. Landi et al., "Sarcopenia as a risk factor for falls in elderly individuals: Results from the ilSIRENTE study," *Clin. Nutrition*, vol. 31, no. 5, pp. 652–658, Oct. 2012.

[3] *World Population Prospects 2022: Summary of Results*, United Nations, New York, NY, USA, 2022.

[4] S. Thomas, S. Mackintosh, and J. Halbert, "Does the 'Otago exercise programme' reduce mortality and falls in older adults: A systematic review and meta-analysis," *Age Ageing*, vol. 39, no. 6, pp. 681–687, Nov. 2010.

[5] S. Mat et al., "Effect of modified otago exercises on postural balance, fear of falling, and fall risk in older fallers with knee osteoarthritis and impaired gait and balance: A secondary analysis," *PM&R*, vol. 10, no. 3, pp. 254–262, Mar. 2018.

[6] R. Almarzouki et al., "Improved balance in middle-aged adults after 8 weeks of a modified version of otago exercise program: A randomized controlled trial," *PLoS ONE*, vol. 15, no. 7, Jul. 2020, Art. no. e0235734.

[7] A. Bevilacqua, K. MacDonald, A. Rangarej, V. Widjaya, B. Caulfield, and T. Kechadi, "Human activity recognition with convolutional neural networks," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2019, pp. 541–552.

[8] L. Dedeyne, J. A. Wullems, J. Dupont, J. Tournoy, E. Gielen, and S. Verschueren, "Exploring machine learning models based on accelerometer sensor alone or combined with gyroscope to classify home-based exercises and physical behavior in (Pre)sarcopenic older adults," *J. Meas. Phys. Behav.*, vol. 4, no. 2, pp. 174–186, Jun. 2021.

[9] Y. Wang, S. Cang, and H. Yu, "A survey on wearable sensor modality centred human activity recognition in health care," *Expert Syst. Appl.*, vol. 137, pp. 167–190, Dec. 2019.

[10] Y. Li and L. Wang, "Human activity recognition based on residual network and BiLSTM," *Sensors*, vol. 22, no. 2, p. 635, Jan. 2022.

[11] K. Ellis, J. Kerr, S. Godbole, and G. Lanckriet, "Multi-sensor physical activity recognition in free-living," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct Publication*, Sep. 2014, pp. 431–440.

[12] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "Hand-crafted features vs residual networks for human activities recognition using accelerometer," in *Proc. IEEE 23rd Int. Symp. Consum. Technol. (ISCT)*, Jun. 2019, pp. 153–156.

[13] S. R. Shakya, C. Zhang, and Z. Zhou, "Comparative study of machine learning and deep learning architecture for human activity recognition using accelerometer data," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 6, pp. 577–582, 2018.

[14] S.-M. Lee, S. Min Yoon, and H. Cho, "Human activity recognition from accelerometer data using convolutional neural network," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 131–134.

[15] D. Wagner, K. Kalischewski, J. Velten, and A. Kummert, "Activity recognition using inertial sensors and a 2-D convolutional neural network," in *Proc. 10th Int. Workshop Multidimensional (nD) Syst. (nDS)*, Sep. 2017, pp. 1–6.

[16] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang, "Deep residual bidir-LSTM for human activity recognition using wearable sensors," *Math. Problems Eng.*, vol. 2018, pp. 1–13, Dec. 2018.

[17] I. Dirgová Luptáková, M. Kubovč ík, and J. Pospíchal, "Wearable sensor-based human activity recognition with transformer model," *Sensors*, vol. 22, no. 5, p. 1911, Mar. 2022.

[18] M. F. Trujillo-Guerrero, S. Román-Niemes, M. Jaén-Vargas, A. Cadiz, R. Fonseca, and J. J. Serrano-Olmedo, "Accuracy comparison of CNN, LSTM, and transformer for activity recognition using IMU and visual markers," *IEEE Access*, vol. 11, pp. 106650–106669, 2023.

[19] R. Mutegeki and D. S. Han, "A CNN-LSTM approach to human activity recognition," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2020, pp. 362–366.

[20] S. Mekruksavanich and A. Jitpattanakul, "Smartwatch-based human activity recognition using hybrid LSTM network," in *Proc. IEEE Sensors*, Oct. 2020, pp. 1–4.

[21] S. K. Challa, A. Kumar, and V. B. Semwal, "A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data," *Vis. Comput.*, vol. 38, no. 12, pp. 4095–4109, Dec. 2022.

[22] N. T. Hoai Thu and D. S. Han, "HiHAR: A hierarchical hybrid deep learning architecture for wearable sensor-based human activity recognition," *IEEE Access*, vol. 9, pp. 145271–145281, 2021.

[23] A. J. Cruz-Jentoft et al., "Sarcopenia: European consensus on definition and diagnosisreport of the European working group on sarcopenia in older peoplea," *Age Ageing*, vol. 39, no. 4, pp. 412–423, 2010.

[24] K. Kondo and T. Hasegawa, "Sensor-based human activity recognition using adaptive class hierarchy," *Sensors*, vol. 21, no. 22, p. 7743, Nov. 2021.

[25] P. Siirtola and J. Röning, "Ready-to-use activity recognition for smartphones," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, Apr. 2013, pp. 59–64.

[26] P. Esfahani and H. T. Malazi, "PAMS: A new position-aware multi-sensor dataset for human activity recognition using smartphones," in *Proc. 19th Int. Symp. Comput. Archit. Digit. Syst. (CADS)*, Dec. 2017, pp. 1–7.

[27] M. Barandas et al., "TSFEL: Time series feature extraction library," *SoftwareX*, vol. 11, Jan. 2020, Art. no. 100456.

[28] T. Maekawa and S. Watanabe, "Unsupervised activity recognition with user's physical characteristics data," in *Proc. 15th Annu. Int. Symp. Wearable Comput.*, Jun. 2011, pp. 89–96.

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

[30] Y. Zhang, J. Wullems, I. D'Haeseleer, V. V. 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.

[31] N. R. Nurwulan and G. Selamaj, "Random forest for human daily activity recognition," *J. Phys., Conf. Ser.*, vol. 1655, no. 1, Oct. 2020, Art. no. 012087.

[32] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6765–6816, 2017.

[33] 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. 3570–3579.