harAGE: A Novel Multimodal Smartwatch-based Dataset for Human Activity Recognition

Adria Mallol-Ragolta [1], Anastasia Semertzidou [1], Maria Pateraki [2,3], and Björn Schuller [1,4]

[1] EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[2] Institute of Computer Science, Foundation of Research and Technology – Hellas, Greece
[3] School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Greece
[4] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9666947

***Abstract*— This work introduces the harAGE dataset: a novel multimodal smartwatch-based dataset for Human Activity Recognition (HAR) with more than 17 hours of data collected from 19 participants using a Garmin Vivoactive 3 device. The dataset contains samples from resting, lying, sitting, standing, washing hands, walking, running, stairs climbing, strength workout, flexibility workout, and cycling activities. The resting activity, excluded from the set of activities to recognise, was explicitly conducted while avoiding stressors and external stimuli, so the data collected can be used to compute the personal, baseline heart rate at rest. We also present the HAR-based models trained using the accelerometer data to recognise different sets of activities. Specifically, we focus on different strategies to combine, fuse, and enrich the accelerometer measurements, so they can be used end-to-end. Model performances are assessed following a Leave-One-Subject-Out Cross-Validation (LOSO-CV) approach, and we use the Unweighted Average Recall (UAR) as the evaluation metric to compare the ground truth and the inferred information. The best UAR score of 98.1 % is obtained when recognising the static and the dynamic activities, excluding the samples corresponding to the washing hands, strength workout, and flexibility workout activities. When recognising the specific activities included in these two sets, the model with the best performance scores a UAR of 70.1 %. Finally, when recognising all the activities considered in the harAGE dataset, the highest UAR achieved is 64.3 %.**

I. INTRODUCTION

According to the *World Health Organization* (WHO), physical inactivity is a serious public health concern with serious implications in people's health, as it can be a risk factor for diabetes, depression, high blood pressure, or obesity. Physical activity is beneficial not only for physical health, but also for wellbeing [10], [20]. Hence, it is crucial to engage society to exercise towards a more active and healthier life. In this regard, virtual trainers [15], [22] could be part of the solution, as they could offer personalised, adaptable workout plans. One of the challenges posed by such systems, however, is the need to monitor the activities users do throughout the day. This approach could not only be used to verify whether the users have performed the planned activities, but also to detect the individual musculoskeletal weaknesses, allowing the design of specific exercises to improve them.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

Commercial smartphones are equipped with embedded sensors, including accelerometers and gyroscopes, which make them suitable to recognise human activities [6], [7], [14]. Current consumer smartwatches feature accelerometer, photoplethysmographic, and pedometer sensors, which makes them suitable for this task too [1], [17]. Furthermore, smartwatches are a high-potential wearable as their market penetration in society is increasing every year, they are non-invasive and wireless, and their placement on the wrist seems advantageous to recognise human activities. In this work, we present the harAGE dataset: a multimodal smartwatch-based dataset for *Human Activity Recognition* (HAR), collected using a Garmin Vivoactive 3 device. This dataset provides the material to train HAR-based models that could be deployed in virtual trainers for an effective monitoring of their users.

We focus our investigation on the use of the accelerometer modality to recognise different sets of human activities end-to-end. As this is the initial exploration of the presented dataset, we aim to analyse the granularity in automatically recognising the different activities of the dataset. For this, we first start with a binary classification problem, aiming at recognising the static and the dynamic activities. Next, we increase the granularity and use a subset of all the activities considered, which includes samples from the lying, sitting, standing, walking, running, stairs climbing, and cycling activities. Finally, we add an extra level of difficulty to this subset by adding samples from the washing hands, the strength workout, and the flexibility workout activities.

The models trained are end-to-end. We compare six different approaches to combine and fuse the accelerometer measurements before being fed into the models. These include computing the norm or the outer product between the accelerometer measurements in the x-, y-, and z-axes, and enriching the traces with their first and second order derivatives. To model the accelerometer-based information and to extract deep learnt representations from the input traces, we implement three common end-to-end network architectures for our initial exploration of this dataset: a) a *Recurrent Neural Network* (RNN), b) a *Convolutional Neural Network* (CNN), and c) a RNN coupled with a CNN (RNN+CNN).

The rest of the paper is laid out as follows. Section II summarises some related works in the field, while Section III presents the dataset collected. Section IV describes the

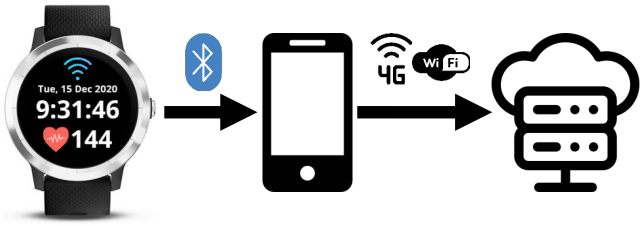

Fig. 1. Front-end design of the customised smartwatch app (left). Illustration of the system architecture implemented to collect the harAGE dataset, and the technologies used to transfer the data between the different nodes.

methodology followed in this work, and Section V details the experiments performed and analyses the results obtained. Finally, Section VI concludes the paper and suggests some future work directions.

II. RELATED WORKS

Researchers commonly investigate the problem of HAR using wearable sensor data [9], [18], [21], which is the focus of this work, or visual data [3], [5], [11], [13], [25], [27]. Datasets found in the literature, such as the *Human Activity Recognition Using Smartphones Data Set* [2] and the *Heterogeneity Activity Recognition Data Set* [24], are collected using smartphone and smartwatch sensors. HAR models, however, are device-dependent, as the inner properties of the sensors embedded in the devices characterise the measurements. This dependency motivated the need to collect a novel HAR dataset using the Garmin Vivoactive 3 device, so the trained models can be used to analyse the measurements obtained with this device in real-life systems and applications.

The optimal features to extract from the sensor data is one of the open questions in the field of HAR. Previous works investigated the use of hand-crafted features [4], [16], while others explored the use of deep-learnt features [12], [19]. Focusing on the problem of HAR from wristwatch data, Chernbumroong et al. [8] claim that the decision tree C4.5 obtained a better performance than neural networks using four different feature sets. Balli et al. [4] report that their best performances were obtained using a random forest method. In the approach presented by Jiang and Yin [12], the accelerometer and gyroscope signal sequences were assembled into an activity image, which was then analysed using *Deep Convolutional Neural Networks* (DCNN) to automatically learn the optimal features. Murad and Pyun [19] propose the use of *Deep Recurrent Neural Networks* (DRNN), while Shahmohammadi et al. [23] investigate the use of active learning for the task at hand.

III. DATASET

This work explores the initial version of the harAGE dataset: a new dataset for HAR from smartwatch data; specifically, from a Garmin Vivoactive 3 device. To collect the data, we implemented a customised smartwatch app, which reads information from the embedded sensors regarding the accelerometer, the heart rate, and the pedometer information. The accelerometer measurements are sensed at 25 Hz, while the heart rate and the steps information, at 1 Hz. The measurements are encapsulated into a JSON message and sent in close to real-time into a customised, encrypted and secure server via the Internet using the HTTPS protocol (cf. Figure 1). Analysing the server logs, we observed that the same device performs consecutive POST requests in a timespan of, approximately, 3 seconds.

TABLE I

SUMMARY OF THE ACTIVITIES INCLUDED IN THE HARAGE DATASET, THE NUMBER OF PARTICIPANTS COLLECTED FOR EACH ACTIVITY, AND THE AMOUNT OF DATA AVAILABLE TIME-WISE.

Activity	Participants	Duration (HH):MM:SS
Resting	19	1:25:24
Lying	19	1:39:01
Sitting	18	1:31:25
Standing	18	1:35:51
Washing Hands	18	53:40
Walking	18	2:23:59
Running	16	1:58:28
Stairs Climbing	18	2:17:23
Strength Workout	18	53:05
Flexibility Workout	18	56:50
Cycling	13	1:36:40
$\sum$	19	17:11:46

To collect the data, we designed a protocol indicating the sequence of activities for the participants to do, and the time they should spend performing each activity. We asked participants to stop the smartwatch app for at least 20 seconds between consecutive activities to simplify the segmentation and annotation of the collected data. The participants started with a resting phase during 5 minutes to collect their heart rate at rest. During this time, participants were explicitly asked to avoid stressors and external stimuli, so the collected heart rate measurements can be used to compute a personal, baseline heart rate for each participant. Nevertheless, this information is not used in this work, as we exclusively focus on the analysis of the accelerometer measurements. Then, they performed a sequence of static activities including lying, sitting and standing. These three activities were performed twice: first without moving, and then allowing reasonable free movements. Each one of these activities was performed during 3 minutes. Next, we asked participants to simulate washing their hands, without running water, also for 3 minutes. This activity was rarely included in previous HAR datasets found in the literature. However, because of the current pandemic context and the placement of the selected device in the human body, we considered washing hands as a potentially measurable interesting activity to collect.

Next in the protocol, we included the following dynamic activities: walking, running, climbing stairs (both upstairs and downstairs), and cycling. Furthermore, each one of these activities was performed three times at low, moderate and high intensities during 3 minutes each. Intensity levels are subjective, as these depend on several factors, such as the previous physical condition of the participants. Thus, to capture this variability in our dataset, we relied on the participants themselves to set their own thresholds for each intensity

level. Before the cycling set of activities, we incorporated a set of workout activities in the protocol. These activities included two sets of strength workout activities (squats and arm raising exercises), and two sets of flexibility workout activities (shoulder roll and wrist stretching exercises). These four activities were performed for 1.5 minutes each.

The dataset was mainly collected outdoors in order to guarantee the safety measures against the COVID-19 pandemic. The downside is that data transfer between the smartwatch and the server was performed via the 4G connection of the smartphone with which the smartwatch was paired. The 4G connection slowed down the data transmission, occasionally causing the loss of sensor measurements. The lost measurements were discarded by the back-end of the smartwatch app as a preventive measure to avoid running out of memory because of an overflow of the internal buffers implemented to temporarily store the sensed measurements before the transmission. Thus, the measurements received from each activity occasionally contained discontinuities. As a preprocessing stage, we trimmed the data into segments of at least 20 seconds of consecutive sensor measurements, which are then used to populate the dataset.

The current version of the harAGE dataset contains 17 h 11 min 46 sec of data from 19 participants (9 f, 10 m), with a mean age of 41.7 and a standard deviation of 8.0. Before the data collection, participants read and signed an *Informed Consent Form* (ICF), which was previously approved by the competent ethics committee. A summary of the different activities considered in the dataset, and the amount of data available for each activity is provided in Table I. Some participants partially completed the activities included in the protocol because of data transmission issues, or the impossibility to get access to a bike for the cycling-related activities.

IV. METHODOLOGY

This section describes the methodology followed in this work (cf. Figure 2). Section IV-A describes the processing applied to the accelerometer measurements, Section IV-B details the network architectures implemented, and Section IV-C summarises the parameters and procedures used to train them.

A. Data Processing

First, we read the accelerometer measurements in the x-, y-, and z-axes. To normalise the measurements, we follow a personal debiasing approach. We compute for each participant and axis the mean of the accelerometer measurements among all the activities the current participant performed. Then, we subtract the three mean values calculated for each participant from the corresponding axis of the current measurements. With this approach, we aim to remove potential personal biases in the movements, obtaining their intrinsic representations for a more robust activity recognition. As sensor measurements are prone to errors, we filter the debiased measurements using a 1-dimensional Gaussian filter with a standard deviation of 1 to smooth them and reduce noise [28].

This work implements an end-to-end system, and, therefore, we do not extract hand-crafted features from the processed measurements. Instead, we investigate the following six approaches to feed the models with accelerometer-based information.

i) f_{xyz} – This approach feeds into the networks the filtered, debiased accelerometer measurements. The input sequences have a dimensionality $\in \mathbb{R}^3$.

ii) $f_{xyz}^{'}$ – In addition to the filtered, debiased accelerometer measurements, this approach computes the first and second order derivatives of the x-, y-, and z-axes, separately, and feeds them into the networks. The first and second order derivatives are estimated using the first and second discrete differences of the original traces, respectively. The input sequences have a dimensionality $\in \mathbb{R}^9$.

iii) f_{norm} – This approach fuses the accelerometer information in the x-, y-, and z-axes by computing their norm; i. e. ,

$$norm = \sqrt{x^2 + y^2 + z^2}. \quad (1)$$

The input sequences have a dimensionality $\in \mathbb{R}^1$.

iv) $f_{norm}^{'}$ – In addition to the norm of the accelerometer measurements computed as defined in Equation (1), this approach calculates the first and second order derivatives of the norm and feeds them into the networks. The input sequences have a dimensionality $\in \mathbb{R}^3$.

v) $f_{x \otimes y \otimes z}$ – Inspired by the tensor fusion layer proposed by Zadeh et al. [26], this approach fuses the filtered, debiased accelerometer measurements in the three different axes at each time step by computing the outer product between them. Before this, we add an extra constant dimension with value 1 into the accelerometer measurements, so that the outer product produces a 3-dimensional tensor. The relevant property of this tensor is that it contains the original filtered, debiased measurements unaltered in addition to their 2-dimensional and 3-dimensional fusion. The result is flattened, so the input sequences to be fed into the networks have a dimensionality $\in \mathbb{R}^8$.

vi) $f_{x \otimes y \otimes z}^{'}$ – This approach also follows the outer product-based fusion described above. The difference is that, in this case, not only the filtered, debiased accelerometer measurements in the x-, y-, and z-axes are fused, but also their first and second order derivatives. The resulting tensor is flattened, so the input sequences to be fed into the networks have a dimensionality $\in \mathbb{R}^{64}$.

Each activity segment contained in the dataset has a different duration. Nevertheless, neural networks need to be trained using fixed-length data sequences. Thus, we window the data with sequences of 20 seconds length. As the sampling frequency of the accelerometer sensors is 25 Hz, each window of accelerometer data contains 500 data points. We label each windowed sequence with the activity from which it is extracted. As a form of data augmentation for training the models, the activity segments are windowed using an overlap of 50 %. For testing, the activity segments are windowed

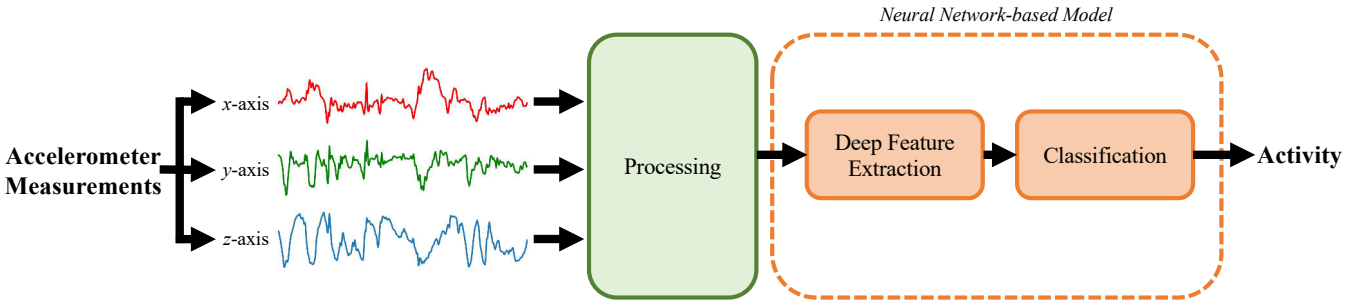

Fig. 2. Block diagram illustrating the end-to-end system implemented. The system receives the accelerometer measurements in the x-, y-, and z-axes as input and processes them to enrich and/or fuse their representations. These are then fed into a neural network-based model composed of two blocks: the first block extracts deep learnt representations from the input traces, while the second block is responsible for the actual classification.

TABLE II

SUMMARY OF THE DESCRIPTIVE STATISTICS (μ: MEAN, σ: STANDARD DEVIATION) COMPUTED FROM THE UAR SCORES OBTAINED WHEN ASSESSING THE BINARY CLASSIFICATION-BASED END-TO-END MODELS USING LOSO-CV FOR THE DIFFERENT APPROACHES AND NETWORK ARCHITECTURES CONSIDERED TO COMBINE, FUSE, AND ENRICH THE ACCELEROMETER INFORMATION, AND TO EXTRACT DEEP LEARNT REPRESENTATIONS FROM THE INPUT TRACES.

UAR [%]	**RNN**		**CNN**		**RNN+CNN**	
	μ	σ	μ	σ	μ	σ
f_{xyz}	80.8	16.2	97.2	3.1	79.1	16.6
f'_{xyz}	**91.7**	13.7	**98.0**	2.2	92.9	12.1
f_{norm}	73.8	14.3	97.4	3.2	78.6	10.3
f'_{norm}	88.7	16.5	94.7	11.2	**98.1**	2.0
$f_{x\otimes y\otimes z}$	63.2	13.7	93.1	11.3	71.0	14.0
$f'_{x\otimes y\otimes z}$	82.3	15.4	91.5	11.9	92.5	12.2

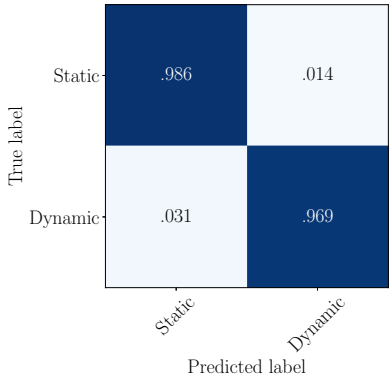

Fig. 3. Confusion matrix computed by comparing the ground truth and the inferred activities corresponding to the windowed sequences of accelerometer information using the best binary classification-based end-to-end model.

without overlap, and each sequence is analysed and assessed independently. The purpose to include the resting activity as part of the dataset was to capture the participants' heart rate at rest (cf. Section III). We exclude the samples corresponding to this activity for training the models, and we do not consider the personal, baseline heart rate at rest in this work, as it focuses on the accelerometer modality.

B. Models Description

To model the windowed sequences with accelerometer-related information, we tackle the task as a sequence modelling problem. The networks implemented are composed of two main blocks: the first block is responsible for extracting embedded representations from the input data, while the second block performs the actual classification. The classification block is composed of two-stacked *Fully Connected* (FC) layers, preceded by two dropout layers with probability 0.3. The first FC layer contains 32 neurons and uses the *Rectified Linear Unit* (ReLU) as the activation function. The second layer has as many neurons as classes we need to classify our samples into and uses Softmax as the activation function, so that the outputs of the network can be interpreted as probability scores.

The feature extraction block of the networks plays a vital role, as it is in charge of extracting the salient information embedded in the input windowed sequences for the task at hand. For this, as laid out, we investigate three different configurations: a) a RNN, b) a CNN, and c) a RNN+CNN. For the RNN-based configuration, we use a single-layer, bidirectional *Gated Recurrent Unit - Recurrent Neural Network* (GRU-RNN) with 128 hidden units, which translates into a total of 256 hidden units. Only the hidden representation extracted at the last time step of the input sequence is used for the actual classification. For the CNN-based configuration, we use a 1-dimensional CNN with 128 output channels, a kernel size of 2, and a stride of 1. The number of input channels depends on the dimensionality of the input traces. Batch normalisation is applied to the output of the convolution, the resulting representation is transformed using a ReLU function, and a 1-dimensional adaptive average pooling is applied using a kernel size of 2 to obtain 256 values as a result of the feature extraction. Finally, for the RNN+CNN configuration, we merge the aforementioned configurations and adjust them for a smooth integration. Specifically, we also use a single-layer, bidirectional GRU-RNN with 128 hidden units, but, this time, all the hidden representations extracted throughout the input sequence are fed into a 1-dimensional CNN with 256 and 128 input and output channels, respectively.

TABLE III

SUMMARY OF THE DESCRIPTIVE STATISTICS (μ: MEAN, σ: STANDARD DEVIATION) COMPUTED FROM THE UAR SCORES OBTAINED WHEN ASSESSING THE STANDARD HAR-BASED END-TO-END MODELS USING LOSO-CV FOR THE DIFFERENT APPROACHES AND NETWORK ARCHITECTURES CONSIDERED TO COMBINE, FUSE, AND ENRICH THE ACCELEROMETER INFORMATION, AND TO EXTRACT DEEP LEARNT REPRESENTATIONS FROM THE INPUT TRACES.

UAR [%]	**RNN**		**CNN**		**RNN+CNN**	
	μ	σ	μ	σ	μ	σ
f_{xyz}	53.4	17.5	65.4	15.9	56.7	19.7
f'_{xyz}	**65.4**	14.8	**70.1**	13.7	**68.2**	15.7
f_{norm}	32.9	6.7	63.2	9.6	37.9	10.9
f'_{norm}	51.3	14.8	64.8	8.6	63.0	10.7
$f_{x\otimes y\otimes z}$	28.7	9.5	50.1	9.7	36.4	8.7
$f'_{x\otimes y\otimes z}$	40.8	13.1	50.5	14.5	57.5	13.1

C. Networks Training

All the models investigated in this work are trained under the exact same conditions for a fair comparison. For reproducibility purposes, the pseudorandom number generator is seeded at the initialisation of the models. The networks are trained to minimise the Categorical Cross-Entropy Loss, using Adam as the optimiser with a fixed learning rate of 10^{-3}. The metric selected to compare the inferred and the ground truth information is the *Unweighted Average Recall* (UAR) to account for the potential imbalance of the windowed sequences of accelerometer measurements generated for the different activities. Hence, we define $(1 - \text{UAR})$ as the validation error to monitor the training progress. Network parameters are updated in batches of 64 samples and trained during a maximum of 100 epochs. We implement an early stopping mechanism to stop training when the validation error does not improve for 20 consecutive epochs. To assess the models, we follow a *Leave-One-Subject-Out Cross-Validation* (LOSO-CV) approach. Iteratively, data from $N-1$ participants is used for training the model –where N corresponds to the total number of participants in the dataset–, while the data from the excluded participant is used for testing its performance. With this approach, we test the performance of the trained models on all the participants available in the dataset. We obtain a UAR score from each participant separately and then compute the descriptive statistics of the overall UAR scores, as reported in Section V. Each fold is trained during a specific number of epochs. Therefore, when modelling all training material and to prevent overfitting, the training epochs are determined by computing the median of the training epochs processed in each fold.

V. EXPERIMENTAL RESULTS

This section analyses and interprets the experimental results obtained. Section V-A tackles the task as a binary classification problem. Section V-B and Section V-C investigate the task as a multi-class classification problem. The former includes the set of standard activities commonly considered in HAR datasets (lying, sitting, standing, walking, running, stairs climbing, and cycling), while the latter includes all the activities, excluding the resting one, from the harAGE dataset (cf. Section III).

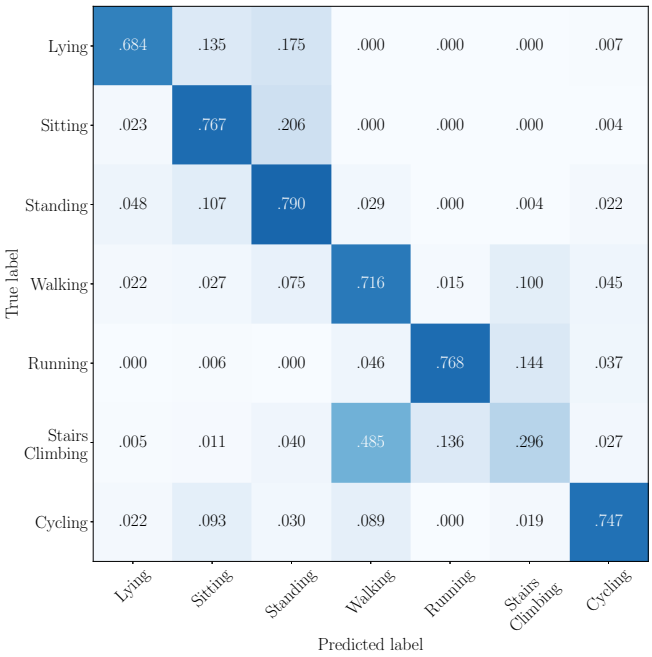

Fig. 4. Confusion matrix computed by comparing the ground truth and the inferred activities corresponding to the windowed sequences of accelerometer information using the best standard HAR-based end-to-end model.

A. Binary Classification

The results obtained when tackling the task as a binary classification problem are reported in Table II. The best UAR of 98.1 % is obtained using the RNN+CNN network architecture with the f'_{norm} approach, closely followed by the CNN network using the f'_{xyz} approach, which scores a UAR of 98.0 %. In terms of the RNN-based architecture, the highest UAR of 91.7 % is achieved using the f'_{xyz} approach.

In most of the cases investigated, we observe that the approaches including the first and second order derivatives of the accelerometer measurements obtain a better performance. This result suggests that the first two derivatives capture relevant information for the recognition of the considered activities. The outer product-based fusion implemented does not seem to be effective for this task, as it usually obtains the lowest UAR scores regardless of the network architecture considered. While the worst performance of the CNN-based model is achieved by the $f'_{x\otimes y\otimes z}$ approach, 91.5 %, the $f_{x\otimes y\otimes z}$ approach scores the lowest UAR scores using the RNN and the RNN+CNN network configurations, with a UAR of 63.2 % and 71.0 %, respectively.

Figure 3 depicts the confusion matrix of the best model; i. e., the RNN+CNN network configuration using the f'_{norm} approach. As it can be observed, 98.6 % and 96.9 % of the windowed sequences corresponding to the static and dynamic activities, respectively, are correctly classified. Hence, the percentage of misclassified windowed sequences is low.

TABLE IV

SUMMARY OF THE DESCRIPTIVE STATISTICS (μ: MEAN, σ: STANDARD DEVIATION) COMPUTED FROM THE UAR SCORES OBTAINED WHEN ASSESSING THE MULTI-CLASS HARAGE -BASED END-TO-END MODELS USING LOSO-CV FOR THE DIFFERENT APPROACHES AND NETWORK ARCHITECTURES CONSIDERED TO COMBINE, FUSE, AND ENRICH THE ACCELEROMETER INFORMATION, AND TO EXTRACT DEEP LEARNT REPRESENTATIONS FROM THE INPUT TRACES.

UAR [%]	**RNN**		**CNN**		**RNN+CNN**	
	μ	σ	μ	σ	μ	σ
f_{xyz}	51.1	16.0	58.4	15.2	52.8	16.8
f'_{xyz}	**58.4**	15.0	**64.1**	14.8	**64.3**	14.7
f_{norm}	24.4	4.6	45.1	8.6	25.8	7.3
f'_{norm}	33.2	8.1	48.4	8.3	55.9	9.3
$f_{x\otimes y\otimes z}$	18.4	6.7	34.9	6.7	25.5	7.2
$f'_{x\otimes y\otimes z}$	34.0	9.8	47.2	11.7	53.7	13.1

B. Standard HAR Classification

The results obtained when tackling the task as a 7-class classification problem are summarised in Table III. The best UAR of 70.1 % is obtained using the CNN network architecture with the f'_{xyz} approach, followed by a UAR of 68.2 % and 65.4 % obtained using the RNN+CNN and the RNN network configurations, respectively, both using this same approach.

In all cases investigated, the approaches including the first and second order derivatives of the accelerometer measurements obtain a better performance. This result supports their use, as they seem to capture suitable information for the recognition of the considered activities. The worst UAR scores are obtained using the $f_{x\otimes y\otimes z}$ approach, highlighting their limited performance for the task at hand. Nevertheless, it is worth emphasising the performance improvement experienced by considering the first and second order derivatives of the accelerometer measurements, $f'_{x\otimes y\otimes z}$, especially using the RNN and the RNN+CNN network configurations. In these cases, the model performances improve from a UAR of 28.7 % to 40.8 % and from a UAR of 36.4 % to 57.5 %, respectively.

Figure 4 depicts the confusion matrix of the best model; i. e. , the CNN network configuration using the f'_{xyz} approach. As it can be observed, most of the activities are correctly classified with a likelihood greater than 70 %. Nonetheless, the results obtained point out that the stairs climbing activity is problematic, as the windowed sequences corresponding to this activity are misclassified into the walking activity with a likelihood of 48.5 %.

C. Multi-class harAGE Classification

The results obtained when tackling the task as a 10-class classification problem are synthesised in Table IV. The best UAR of 64.3 % is obtained using the RNN+CNN network architecture with the f'_{xyz} approach, followed by a UAR of 64.1 % and 58.4 % obtained using the CNN and the RNN network configurations, respectively, both using this same approach.

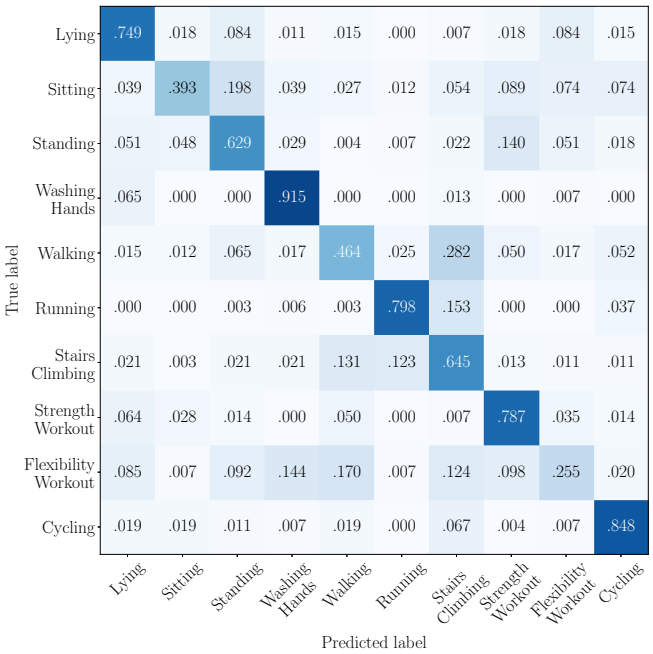

Fig. 5. Confusion matrix computed by comparing the ground truth and the inferred activities corresponding to the windowed sequences of accelerometer information using the best multi-class harAGE -based end-to-end model.

Analogously to the results observed in Section V-B, the approaches including the first and second order derivatives of the accelerometer measurements obtain a better performance in all the cases investigated. The worst UAR scores are also obtained using the $f_{x\otimes y\otimes z}$ approach, with a UAR of 18.4 %, 34.9 %, and 25.5 % for the RNN, CNN, and RNN+CNN network architectures, respectively.

Figure 5 depicts the confusion matrix of the best model; i. e. , the RNN+CNN network configuration using the f'_{xyz} approach. In this case, in addition to the misclassification of the stairs climbing activity, we observe some confusion with the sitting activity, as the windowed sequences corresponding to this activity are classified into the standing activity with a likelihood of 19.8 %. Furthermore, we also observe that the flexibility workout activities are difficult to recognise, as a noticeable percentage of windowed sequences corresponding to this activity are misclassified into the washing hands, walking, and stairs climbing activities. This result could be attributed to the shoulder roll and wrist stretching exercises defined as the flexibility workout activities, which might be difficult to capture using the accelerometer sensors embedded in the smartwatch.

VI. CONCLUSIONS

This work presented the harAGE dataset and the initial HAR-based models trained using the accelerometer modality. This investigation focused on the analysis of different information fusion approaches and neural network architectures to extract salient embedded representations for the task at hand. Model performances varied depending on the granularity with which

the different activities wanted to be recognised, the input information fed into the models, and the network architectures. The results obtained highlighted the suitability of the f'_{xyz} approach for this task, as it achieved the highest UAR scores in most of the cases investigated, regardless of the network architectures implemented. The outer product-based fusion implemented at an early stage did not seem to be effective. Nevertheless, it provided competitive results when the fusion considered the first and second order derivatives of the accelerometer measurements.

As a follow up of this study, we aim to engage new participants into the data collection in order to increase the sample size. Further research directions include the fusion of the three modalities available for this task, and the exploration of the outer product-based approach at a later stage, for instance, to fuse the embedded representations learnt by modality-specific models before performing the final classification. Moreover, future works can consider investigating the performance of the presented methodology in other human activity recognition datasets, and implementing more sophisticated network architectures to learn embedded representations from the accelerometer traces.

VII. ACKNOWLEDGMENTS

The authors would like to sincerely thank the participants who took part in the collection of the investigated dataset. We would also like to thank Georgios Athanassiou and Michalis Maniadakis for their contributions in the protocol design for collecting the data.

REFERENCES

[1] M. Adjeisah, G. Liu, D. O. Nyabuga, and R. N. Nortey. Multi-Sensor Information Fusion and Machine Learning for High Accuracy Rate of Mechanical Pedometer in Human Activity Recognition. In *Proceedings of the International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking*, pages 1064–1070, Xiamen, China, 2019. IEEE.

[2] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the* 21^{st} *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 437–442, Bruges, Belgium, 2013. i6doc.

[3] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan, and M. Zaharadeen. Automated Daily Human Activity Recognition for Video Surveillance Using Neural Network. In *Proceedings of the* 4^{th} *International Conference on Smart Instrumentation, Measurement and Application*, Putrajaya, Malaysia, 2017. IEEE. 5 pages.

[4] S. Balli, E. A. Sağbaş, and M. Peker. Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm. *Measurement and Control*, 52(1–2):37–45, 2019.

[5] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor. Glimpse Clouds: Human Activity Recognition From Unstructured Feature Points. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 469–478, Salt Lake City, UT, USA, 2018. IEEE.

[6] A. Bayat, M. Pomplun, and D. A. Tran. A Study on Human Activity Recognition Using Accelerometer Data from Smartphones. *Procedia Computer Science*, 34:450–457, 2014.

[7] Y. Chen and C. Shen. Performance Analysis of Smartphone-Sensor Behavior for Human Activity Recognition. *IEEE Access*, 5:3095–3110, 2017.

[8] S. Chernbumroong, A. S. Atkins, and H. Yu. Activity classification using a single wrist-worn accelerometer. In *Proceedings of the* 5^{th} *International Conference on Software, Knowledge Information, Industrial Management and Applications*, Benevento, Italy, 2011. IEEE. 6 pages.

[9] G. De Leonardis, S. Rosati, G. Balestra, V. Agostini, E. Panero, L. Gastaldi, and M. Knaflitz. Human Activity Recognition by Wearable Sensors: Comparison of different classifiers for real-time applications. In *Proceedings of the International Symposium on Medical Measurements and Applications*, Rome, Italy, 2018. IEEE. 6 pages.

[10] K. R. Fox. The influence of physical activity on mental well-being. *Public Health Nutrition*, 2(3a):411–418, 1999.

[11] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognition*, 61:295–308, 2017.

[12] W. Jiang and Z. Yin. Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks. In *Proceedings of the* 23^{rd} *International Conference on Multimedia*, pages 1307–1310, Brisbane, Australia, 2015. ACM.

[13] H.-J. Jung and K.-S. Hong. Versatile Model for Activity Recognition: Sequencelet Corpus Model. In *Proceedings of the* 13^{th} *International Conference on Automatic Face & Gesture Recognition*, pages 325–332, Xi'an, China, 2018. IEEE.

[14] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim. Human Activity Recognition via an Accelerometer-Enabled-Smartphone Using Kernel Discriminant Analysis. In *Proceedings of the* 5^{th} *International Conference on Future Information Technology*, Busan, South Korea, 2010. IEEE. 6 pages.

[15] I. Kouris, M. Sarafidis, T. Androutsou, and D. Koutsouris. HOLOBALANCE: An Augmented Reality virtual trainer solution forbalance training and fall prevention. In *Proceedings of the* 40^{th} *Annual International Conference of the Engineering in Medicine and Biology Society*, pages 4233–4236, Honolulu, HI, USA, 2018. IEEE.

[16] M.-C. Kwon and S. Choi. Recognition of Daily Human Activity Using an Artificial Neural Network and Smartwatch. *Wireless Communications and Mobile Computing*, 2018, 2018. 9 pages.

[17] O. D. Lara, A. J. Pérez, M. A. Labrador, and J. D. Posada. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and Mobile Computing*, 8(5):717–729, 2012.

[18] F. Li, K. Shirahama, M. A. Nisar, L. Köping, and M. Grzegorzek. Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors. *Sensors*, 18(2):679, 22 pages, 2018.

[19] A. Murad and J.-Y. Pyun. Deep Recurrent Neural Networks for Human Activity Recognition. *Sensors*, 17(11):2556, 17 pages, 2017.

[20] F. J. Penedo and J. R. Dahn. Exercise and well-being: a review of mental and physical health benefits associated with physical activity. *Current Opinion in Psychiatry*, 18(2):189–193, 2005.

[21] S. Rosati, G. Balestra, and M. Knaflitz. Comparison of Different Sets of Features for Human Activity Recognition by Wearable Sensors. *Sensors*, 18(12), 2018.

[22] B. Schooley, D. Akgun, P. Duhoon, and N. Hikmet. Persuasive AI Voice-Assisted Technologies to Motivate and Encourage Physical Activity. In *Advances in Computer Vision and Computational Biology*, pages 363–384. Springer, 2021.

[23] F. Shahmohammadi, A. Hosseini, C. E. King, and M. Sarrafzadeh. Smartwatch Based Activity Recognition Using Active Learning. In *Proceedings of the International Conference on Connected Health: Applications, Systems and Engineering Technologies*, pages 321–329, Philadelphia, PA, USA, 2017. IEEE.

[24] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen. Smart Devices Are Different: Assessing and MitigatingMobile Sensing Heterogeneities for Activity Recognition. In *Proceedings of the* 13^{th} *Conference on Embedded Networked Sensor Systems*, pages 127–140, Seoul, South Korea, 2015. ACM.

[25] G. Vaquette, A. Orcesi, L. Lucat, and C. Achard. The DAily Home LIfe Activity Dataset: A High Semantic Activity Dataset for Online Recognition. In *Proceedings of the* 12^{th} *International Conference on Automatic Face & Gesture Recognition*, pages 497–504, Washington, DC, USA, 2017. IEEE.

[26] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, 2017. ACL.

[27] N. Zhuang, T. Yusufu, J. Ye, and K. A. Hua. Group Activity Recognition with Differential Recurrent Convolutional Neural Networks. In *Proceedings of the* 12^{th} *International Conference on Automatic Face & Gesture Recognition*, pages 526–531, Washington, DC, USA, 2017. IEEE.

[28] Z. Zhuang and Y. Xue. Sport-Related Human Activity Detection and Recognition Using a Smartwatch. *Sensors*, 19(22):5001, 21 pages, 2019.