Recognizing Hand-based Micro Activities Using Wrist-Worn Inertial Sensors: A Zero-Shot Learning Approach

Fadi Al Machot[1 [0000−0000−0000−0000]], Habib Ullah[1 [0000−0000−0000−0000]], and Florenc Demrozi[2 [0000−0000−0000−0000]]

[1] Dep. of Data Science, Norwegian University of Life Sciences, Norway, fadi.al.machot@nmbu.no habib.ullah@nmbu.no
[2] Dep. of Electrical Engineering and Computer Science, University of Stavanger, Norway, florenc.demrozi@uis.no
[3] Corresponding Authors

Abstract. Zero-shot learning (ZSL) is a machine learning paradigm that enables models to recognize and classify data from classes they have not encountered during training. This approach is particularly advantageous in recognizing activities where labeled data is limited, allowing models to identify new, unseen activities by leveraging semantic knowledge from seen activities. In this paper, we explore the efficacy of ZSL for activity recognition using Sentence-BERT (S-BERT) for semantic embeddings and Variational Autoencoders (VAE) to bridge the gap between seen and unseen classes. Our approach leverages wrist-worn inertial sensor events to capture activity data and employs S-BERT to generate semantic embeddings that facilitate the transfer of knowledge between seen and unseen activities. The evaluation is conducted on datasets containing three seen and three unseen activity classes with an average duration of 2 seconds, as well as three seen and three unseen activity classes with an average duration of 7 seconds. The results demonstrate promising performance in recognizing unseen activities, with an accuracy of 0.84 for activities with an average duration of 7 seconds and 0.66 for activities with an average duration of 2 seconds. This highlights the potential of ZSL for enhancing activity recognition systems which is crucial for applications in fields such as healthcare, human-computer interaction, and smart environments, where recognizing a wide range of activities is essential.

Keywords: Zero-Shot Learning, Human Activity Recognition, Micro activities, Hand movements

1 Introduction

Zero-shot learning (ZSL) for Human Activity Recognition (HAR) has garnered significant attention due to its potential to classify activities without prior labeled examples [1,32,39,11]. This capability is crucial for applications in health-

care, assistive technologies, smart environments, and human-computer interaction [15,8]. HAR systems aim to quantify, classify, and interpret human activities using sensor data from wearable devices, smartphones, or environmental sensors, employing both classic Machine Learning (ML) and Deep Learning (DL) algorithms [8,5,16].

ZSL is particularly important in smart home environments, where the range of possible activities is vast and constantly evolving. Traditional HAR systems focus on broad activities such as walking, sitting, and standing but often overlook subtle hand-based micro activities that are crucial for understanding daily routines and independence [15,17]. Micro activities include tasks like brushing hair, washing hands, and chopping vegetables, which are essential for personal hygiene, food preparation, and household chores [9,8] as shown in Fig. 1. Recognizing these activities is vital for assessing individuals' functional capabilities and their ability to live independently [22,21,2]. In addition, HAR is applied in the research field related to healthcare [25,3] and Industry 4.0 regarding the safety/security of workers and production monitoring [5].

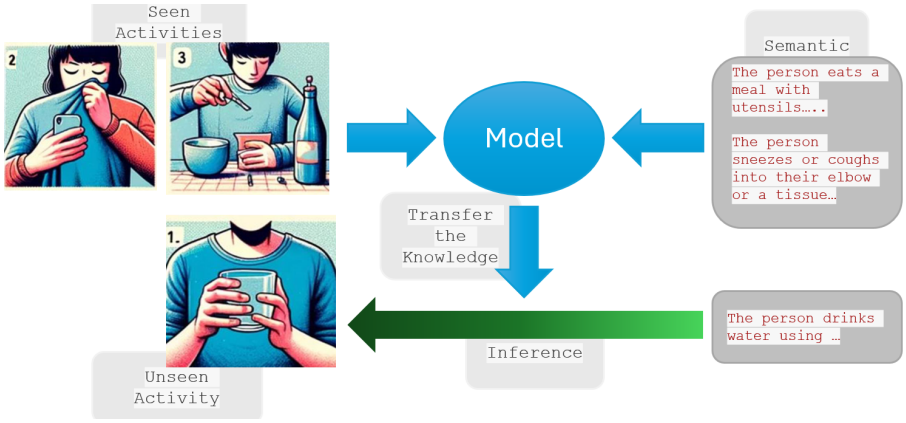

Fig. 1: An example of ZSL-based human micro activity recognition.

The primary challenge with traditional Activities of Daily Living (ADL) recognition systems is the extensive effort and cost required to collect and annotate training data for these micro activities [18]. Given the vast number of potential activities in a smart home environment, it is impractical to gather labeled examples for every possible action. These systems typically rely on predefined algorithms tailored to specific activities, which may not generalize well across different individuals or variations in movement patterns [10]. Additionally, the subjective nature of hand-based micro activities complicates the data collection and annotation process, making it difficult to develop robust models that accurately capture the nuances of human behavior.

Moreover, the varying durations of activities, such as washing hands versus opening a bottle, add to the complexity of accurate recognition [31,21]. This diversity in activity durations necessitates the development of sophisticated algorithms capable of handling these variations effectively.

This paper introduces a ZSL approach to recognize six hand-based micro ADLs. By leveraging semantic information and transfer learning, the proposed approach can generalize to new, unseen activities without requiring labeled examples. This approach addresses the limitations of existing HAR systems and enhances the recognition of micro ADLs in real-world settings. We demonstrate effective performance on six different classes (three seen and three unseen) for training and testing, using activity segments of 2 seconds and 7 seconds duration.

Sensor data is first aggregated, reshaped, and differentiated to compute the derivatives. We then calculate statistical features from the original, and the first order derivative data, forming an aggregated feature set. In addition, descriptions of actions are encoded into embeddings using the Sentence-BERT (S-BERT) model [36]. Our model architecture comprises a Variational Autoencoder (VAE) [23] that encodes the input features into a latent space and decodes them back to the original feature space. The latent space is then mapped to the S-BERT embedding space using a regressor. For zero-shot learning, we use the latent space of the VAE and the regressor to predict embeddings for unseen activities and perform nearest neighbor classification to assign action labels based on the closest S-BERT embeddings.
The main features of the proposed ZSL approach are:

- Incorporation of zero-shot learning to enable recognition of unseen micro ADLs based on semantic information and transfer learning.
- Demonstration of effective performance on six different classes (three seen and three unseen) for training and testing, using activity segments of 2 seconds and 7 seconds duration.

The paper is organized as follows: Section 2 provides background information, Section 3 outlines the HAR pipeline, Section 4 presents experimental findings, Section 5 analyzes state-of-the-art research, and Section 6 offers concluding remarks and outlines future work.

2 Preliminaries

In this section, we outline the data collection methodology used in our study, detailing the sensors utilized, their placement on the subjects' bodies, the data collection protocol, and specifics about the dataset.

Sensors: For data collection, we used an STM Nucleo LR103FB board along with an STM X-Nucleo-IKS01A3 sensor board, which includes both inertial and environmental sensors. Figure 2.a shows the STM boards used in our study, and Figure 2.b illustrates their positioning on the subjects' wrists. These inertial

Fig. 2: Data collection device: a) board, b) on body position.

sensors provide a dependable way to capture motion data crucial for recognizing hand-based micro ADLs.

Cohort of Studied Subjects: Data was gathered from a cohort of 30 subjects, including 12 females and 18 males. The females had an average age of 24.3 years (± 4.9), an average height of 167.2 cm (± 7), and an average weight of 61.9 kg (± 13). The males had an average age of 27.5 years (± 7.2), an average height of 179.4 cm (± 6.9), and an average weight of 84.4 kg (± 17.3).

Data Collection Process: Subjects were asked to perform 24 different hand-based micro ADLs, each repeated 3 to 5 times. They conducted the data collection independently in their home environments without external supervision or training. Subjects were instructed to carry out the specified micro ADLs as they normally would in their daily routines. Figure 3 offers a detailed overview of the data collection setup used in our study, which incorporated three different sensors (accelerometer, gyroscope, and magnetometer) integrated into the STM X-Nucleo-IKS01A3 sensor board.

The sensor data was perceived at a frequency of 100 Hz, ensuring high temporal resolution and capturing subtle variations in hand movements during micro ADL performance. This sampling frequency balances data granularity and computational efficiency, enabling effective analysis of hand motion patterns while minimizing computational overhead.

Besides the raw sensor data, various derived features were generated to enhance the analysis of hand movements. These features included:

- **Pitch, Roll, and Yaw**: These angles represent the orientation of the sensor relative to a fixed reference frame and provide insights into the orientation of the subject's wrist during micro ADL performance.
- **Quaternions (Q1, Q2, Q3, and Q4)**: Quaternions offer an alternative representation of orientation, providing a compact and computationally efficient way to represent rotations in three-dimensional space.
- **Linear Acceleration**: This acceleration component excludes gravity's contribution and provides a measure of the subject's acceleration in the absence

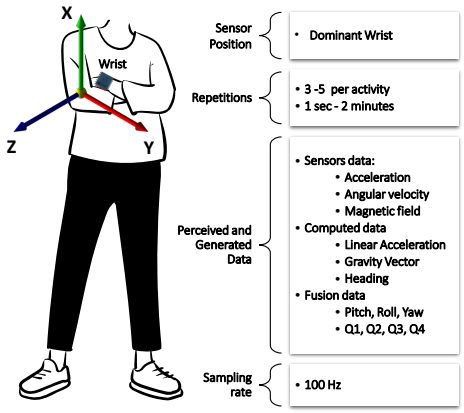

Fig. 3: Data collection setup.

of gravitational effects, offering insights into the subject's movement dynamics.

– **Gravity Vector**: The gravity vector represents the direction and magnitude of gravitational acceleration acting on the sensor, aiding in estimating the sensor's orientation relative to the Earth's gravity field.
– **Heading**: Heading refers to the direction in which the sensor points relative to the magnetic north and provides information about the subject's orientation in space.

Overall, the dataset is composed of a total of 24 features/columns, namely 1) timestamp, 2-4) accelerometer_xyz, 5-7) gyroscope_xyz, 8-10) magnetometer_xyz, 11-13) linear_accelerometer_xyz, 14-16) pitch, roll, yaw, 17-20) Q1, Q2, Q3, Q4, 21-23) gravity_xyz, and 24) heading, for a total of 10.78 hours (aka., 3882540 samples) collected activity data.

Data Characteristics: Table 1 offers a thorough summary of the micro ADLs, detailing their descriptions, the number of repetitions requested from each participant, and statistics on the duration of each activity in seconds (minimum, maximum, mean, standard deviation, 25%, 50%, 75% quantiles). It also includes the total collected data for each activity in terms of 1-second segments. Additionally, Table 2 illustrates the distribution of each activity in 1-second segments for each subject. This detailed segmentation facilitates a granular analysis of activity patterns and durations among different individuals.

The data in Tables 1 and 2 reveal significant variations between different activities and subjects, highlighting the challenges in recognizing hand-based micro ADLs. These variations emphasize the subjective nature of performed micro ADLs and the complexity involved in capturing and analyzing such data.

Uniqueness of this data: Unlike many existing datasets, such as those in [9] and [20], our data collection is unsupervised, without video/audio recording,

Table 1: Studied ADLs with Duration Statistics in seconds: in RED/BLUE the statistics lowest/highest values

Basic info on dataset description				Statistics on ADLs duration in seconds							1 Second
Nr	**ADLs**	**Description**	**Repeat**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**segments**
1	drink water	Drink from a glass, cup, or bottle.	5	8.8	11.3	2.0	4.2	5.8	9.4	70.8	1223
2	eat meal	Perform the gesture of eating, using a fork, a spoon, or the hands.	3	32.6	10.6	1.5	27.6	29.5	34.0	73.6	2941
3	open bottle	Open a bottle (uncap it).	5	7.1	4.7	1.6	4.2	5.3	9.0	35.5	1023
4	open box	Open a food container (e.g., Tupperware).	5	5.5	3.7	1.4	3.0	4.2	6.5	17.4	790
5	brush teeth	Brush teeth for approximately 20 seconds.	5	22.3	5.4	10.7	19.3	20.9	24.2	45.1	3216
6	brush hair	Brush hair for 10 seconds (using a comb, or the hands).	5	13.2	4.5	4.4	10.2	11.6	15.0	26.9	1903
7	take off jacket	Take off a jacket by undoing the buttons or zip.	3	8.1	5.2	2.0	4.6	5.8	11.4	27.1	721
8	put on jacket	Put on a jacket and optionally do the buttons or zip.	3	9.6	6.3	2.8	5.2	7.3	13.4	36.2	836
9	put on shoe	Put on a shoe, doing the laces, zip, etc. (if available).	3	9.5	4.8	1.8	6.2	8.4	11.0	31.9	864
10	take off shoe	Take off a shoe, by optionally undoing the laces/zip.	3	6.7	4.0	0.3	4.2	5.6	8.4	19.9	589
11	put on glasses	Put on (sun)glasses.	5	4.9	3.7	1.5	2.5	3.6	5.4	23.5	697
12	take off glasses	Take off (sun)glasses.	5	4.5	3.5	1.3	2.5	3.2	5.0	15.6	648
13	sit down	Sit down on an chair/sofa/high stool.	5	4.1	2.9	1.6	2.3	3.1	4.6	13.8	584
14	stand up	Stand up.	5	3.9	2.9	1.1	1.9	2.9	4.6	14.4	597
15	writing	Write (by hand) for 15 to 20 seconds.	5	19.5	7.9	9.4	14.7	17.5	22.1	63.3	2687
16	phone call	Pick up the (mobile) phone once (bring to ear).	5	5.9	7.2	1.3	2.7	3.3	5.9	48.7	852
17	type on keyboard	Type on a computer/laptop keyboard for 15-20 seconds	5	19.5	7.8	8.7	14.5	16.6	21.6	56.6	2715
18	salute (wave hand)	Wave the hand for 10 seconds.	5	12.3	4.7	5.4	10.0	10.8	12.7	39.4	1713
19	sneeze cough	Sneeze or cough once.	5	4.1	3.6	1.3	2.1	2.8	4.0	17.6	384
20	blow nose	Blow nose.	5	5.2	4.1	1.0	2.5	3.6	6.1	21.2	743
21	washing hands	Wash hands: apply soap, rub hands together, and rinse.	5	11.4	6.0	3.7	6.5	9.5	14.9	28.0	1649
22	dusting	Dust a surface with a rag/cloth for some time (15-20 s).	5	18.8	5.8	6.6	15.3	16.5	22.3	34.5	2577
23	ironing	Iron (a garment) for 15-20 s.	5	18.6	5.6	10.5	14.5	16.6	21.5	36.2	2565
24	washing dishes	Scrub/scour a plate, cup/glass, or fork/knife/spoon; and rinse.	5	12.9	8.6	3.1	7.3	10.1	15.6	51.2	1850

Table 2: Summary of ADLs Total Counts and Percentages per Subject

Subject	**0000**	**1125**	**1279**	**1313**	**1324**	**1358**	**1390**	**1396**	**1405**	**1435**	**1453**	**1505**	**1570**	**1697**	**1735**
1 Second segments (%)	2255 (6%)	1247 (3%)	1013 (3%)	1068 (3%)	1343 (3%)	1069 (3%)	1691 (4%)	649 (2%)	2300 (6%)	1144 (3%)	1017 (3%)	649 (2%)	1418 (4%)	1248 (3%)	918 (2%)
Subject	**1751**	**1777**	**1803**	**1825**	**1975**	**1978**	**2045**	**2056**	**2097**	**2115**	**2116**	**2136**	**2155**	**2159**	**2160**
1 Second segments (%)	2598 (7%)	1121 (3%)	872(2%)	1718 (4%)	1126 (3%)	1011 (3%)	784 (2%)	1219 (3%)	1091 (3%)	2220 (6%)	1056 (3%)	1281 (3%)	863 (2%)	1496 (4%)	1325 (3%)

and weakly annotated. This distinctive method ensures our dataset genuinely reflects everyday reality, capturing all the complexities and subtleties of human movement. The lack of supervision during data collection allows participants to interact with their environment and perform activities naturally, free from external influence or guidance. Consequently, this dataset encompasses human movements' authentic variability and complexity in real-life contexts.

Additionally, the weak annotation in our dataset mirrors the inherent subjectivity and ambiguity in human activity labeling. This provides a more realistic representation of the challenges in activity recognition, such as variations in movement patterns, participant compliance, and annotation inconsistencies.

3 Methodology

3.1 Problem Definition

ZSL aims to develop a predictive model capable of understanding sensors readings and semantic indicators to classify unseen activities. Generative ZSL becomes crucial when there are no labeled examples for all classes (activities) in question. Consequently, the micro-activity dataset is divided into a training set with seen classes denoted as $\mathrm{Y}_{\text{seen}} = y^1_{\text{seen}}, y^2_{\text{seen}}, \ldots, y^n_{\text{seen}}$ and a testing set with unseen classes represented as $\mathrm{Y}_{\text{unseen}} = y^1_{\text{unseen}}, y^2_{\text{unseen}}, \ldots, y^n_{\text{unseen}}$. It is critical to ensure that $\mathrm{Y}_{\text{seen}} \cap \mathrm{Y}_{\text{unseen}} = \emptyset$. The challenge lies in constructing a function $\mathbb{R}^d \rightarrow \mathrm{Y}_{\text{unseen}}$, which learns to predict the unseen classes using the training set and subsequently tests its efficacy on the unseen class data, maintaining the condition $\mathrm{Y}_{\text{seen}} \cap \mathrm{Y}_{\text{unseen}} = \emptyset$. ZSL aspires to mimic human adaptability to novel scenarios by leveraging models that can anticipate without prior examples.

3.2 Data Preparation

The principle of data split follows the principle of sharing contextual similarities between seen and unseen activities, enabling the model to leverage learned features from the seen classes to recognize the unseen ones. Detailed semantic descriptions are provided in the appendix. The seen classes, which the model was trained on, include activities such as "Washing hands," "Writing," and "Brushing teeth" for the 7-second duration, and "Open a bottle," "Take off a jacket," and "Put on glasses" for the 2-second duration as shown in Table 3. The unseen

classes, which the model had to generalize to, are "Washing dishes," "Typing on a keyboard," and "Brushing hair" for the 7-second duration, and "Open a box," "Take off a shoe," and "Take off glasses" for the 2-second duration. In addition, the sensor readings were downsampled from 100 Hz to 50 Hz.

Table 3: Seen and unseen Classes for Different Durations.

Duration	**Seen Classes**	**Unseen Classes**
7 seconds	Washing hands	Washing dishes
	Writing	Typing on a keyboard
	Brushing teeth	Brushing hair
2 seconds	Open a bottle	Open a box
	Take off a jacket	Take off a shoe
	Put on glasses	Take off glasses

Semantic Embedding S-BERT, or Sentence-BERT [36], is a modification of the BERT (Bidirectional Encoder Representations from Transformers) architecture [19] that is designed to generate sentence embeddings. The goal of S-BERT is to produce meaningful sentence representations that can be compared using cosine similarity for tasks like semantic textual similarity, clustering, and information retrieval. Traditional BERT models are computationally expensive for pairwise sentence comparisons because they require passing both sentences together through the model. S-BERT, however, fine-tunes BERT using a siamese and triplet network structure to derive semantically significant sentence embeddings, enabling efficient similarity comparisons.

The key improvements of S-BERT include: (a) By processing sentences independently through a shared BERT model, it generates fixed-sized sentence embeddings and (b) Uses supervised data with sentence pairs to fine-tune the model, optimizing it for tasks requiring semantic similarity. This approach significantly reduces computation time while maintaining high performance in various sentence similarity and clustering tasks.

Sensor Data Aggregation In this process, we are dealing with sensor data represented by $X \in \mathbb{R}^{N \times M}$, where N denotes the number of samples, and M indicates the number of sensor readings (measurements) per sample. To capture the dynamic changes in the sensor data, we calculate the first derivatives along the measurement axis (i.e., the time axis, assuming measurements are taken sequentially over time). The first derivative X_{diff1} is computed as:

$$X_{\text{diff1}} = \frac{\partial X}{\partial t}$$

This derivative provides insight into the rate of change of sensor readings over time, highlighting trends and variations that may not be evident from the raw data alone. For both the original sensor data X and the derived data X_{diff1}, a variety of statistical features are extracted. These features are intended to summarize key characteristics of the data distributions. The statistical features calculated include the mean (the average value of the data), standard deviation (a measure of the amount of variation or dispersion in the data), minimum (the smallest value in the data set), maximum (the largest value in the data set), median (the middle value when the data is sorted), variance (the expectation of the squared deviation of the data from its mean), range (the difference between the maximum and minimum values), interquartile range (the range within which the central 50% of the data lies, i.e., the difference between the 75th and 25th percentiles), root mean square (the square root of the mean of the squares of the data values), signal magnitude area (the sum of the absolute values of the data divided by the number of samples), and median absolute deviation (the median of the absolute deviations from the median of the data). These features can be represented as:

$$\text{features}(X) = \begin{bmatrix} \text{mean}(X) \\ \text{std}(X) \\ \text{min}(X) \\ \text{max}(X) \\ \text{median}(X) \\ \text{var}(X) \\ \text{range}(X) \\ \text{iqr}(X) \\ \text{rms}(X) \\ \text{sma}(X) \\ \text{mad}(X) \end{bmatrix}$$

To form a comprehensive feature set, we aggregate the statistical features derived from both the original sensor data X and its first derivative X_{diff1}. The aggregated feature matrix $X_{\text{aggregated}}$ is constructed as:

$$X_{\text{aggregated}} = [\text{features}(X_{\text{reshaped}}), \text{features}(X_{\text{diff1}})]$$

This combined feature set includes both the original statistical features and those computed from the first derivatives, providing a rich and informative representation of the sensor data. By preparing the data in this manner, we ensure that the most pertinent characteristics of the sensor readings are captured and ready for subsequent analysis, such as machine learning or statistical modeling. This process enhances the ability to detect patterns, trends, and anomalies within the sensor data, ultimately leading to more accurate and insightful conclusions.

3.3 Model Architecture

The Variational Autoencoder (VAE) is used as our generative model. It learns to encode data into a latent space and then decodes it back to the original space. The architecture consists of an encoder, a latent space representation, and a decoder [23].

Given the input $X_{\text{aggregated}}$, the encoder consists of dense layers with ReLU activation, batch normalization, and dropout to prevent overfitting. Specifically, the input passes through three dense layers with 512, 256, and 128 units respectively, each followed by batch normalization and dropout (0.1). The encoder outputs the mean and log variance of the latent space z, each of dimension latent_dim.

The decoder mirrors the encoder, decoding the latent variable z back to the input space through three dense layers with 128, 256, and 512 units respectively, each followed by batch normalization and dropout.

Additionally, the latent space is projected to an embedding dimension:

$$z_{\text{projected}} = \text{Dense}_{\text{embedding\_dim}}(z).$$

The total loss for the VAE combines the reconstruction loss and the KL divergence. The reconstruction loss L_{recon} measures how well the decoder reconstructs the input:

$$L_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \|x_{\text{aggregated},i} - x_{\text{recon},i}\|^2,$$

while the KL divergence L_{KL} regularizes the distribution of the latent variables:

$$L_{\text{KL}} = \frac{1}{2} \sum_{i=1}^{n} \left(\sigma_i^2 + \mu_i^2 - \log(\sigma_i^2) - 1\right),$$

where μ_i is the mean of the latent variable z_i and σ_i^2 is the variance of z_i.

The total VAE loss is:

$$L_{\text{VAE}} = \frac{1}{N} \sum_{i=1}^{N} (L_{\text{recon}} + \beta L_{\text{KL}}),$$

where β is a weighting factor.

After training the VAE, a regression model g is trained which is a neural network that maps latent representations obtained from the encoder to S-BERT embeddings. The purpose of this regressor is to translate the latent space representations, which capture the underlying structure of the input data, into the S-BERT embedding space, which is used for zero-shot learning. The regressor is trained using the Adam optimizer to minimize the mean squared error (MSE) loss between the predicted embeddings and the true S-BERT embeddings of the seen classes. Formally, the regressor model is defined as:

$$y_{\text{pred}} = g(z),$$

where z is the latent representation from the encoder, and y_{pred} is the predicted S-BERT embedding.

The regression model designed to map the latent space representations to S-BERT embeddings consists of a sequential neural network. This regressor begins with an input layer that matches the dimension of the latent space, followed by a dense layer with 256 units and ReLU activation to introduce non-linearity. To mitigate overfitting, a dropout layer with a dropout rate of 0.1 is applied, and batch normalization is used to stabilize and accelerate training. This structure is repeated in subsequent layers: a dense layer with 128 units, followed by dropout and batch normalization, and then a dense layer with 64 units, again followed by dropout and batch normalization. The final output layer is a dense layer with a number of units equal to the embedding dimension, using a linear activation function to produce the final embeddings.

3.4 Zero-Shot Learning

Embedding Prediction Once the regressor is trained, it is used to predict the S-BERT embeddings for the unseen classes data. The latent representations z_{unseen} of the unseen classes data are first obtained using the encoder. These latent representations are then passed through the regressor to generate the predicted embeddings:

$$y_{\text{pred}} = g(z_{\text{unseen}})$$

This step translates the latent space of unseen classes data to the S-BERT embedding space, enabling the comparison with the embeddings of unseen classes.

Nearest Neighbor Classification After obtaining the predicted embeddings y_{pred} for the unseen classes data, the next step is to classify these embeddings. This is done using a nearest neighbor search in the S-BERT embedding space. The embeddings of the unseen classes $y_{\text{unseen}}(k)$ are precomputed using S-BERT. For each predicted embedding, the nearest neighbor among the unseen class embeddings is found by minimizing the Euclidean distance:

$$\hat{y} = \arg\min_{k} \|y_{\text{pred}} - y_{\text{unseen}}(k)\|$$

where $\hat{y}$ is the predicted label for the unseen classes data, and k indexes the unseen class embeddings. The nearest neighbor search effectively assigns the activity labels to the test data based on the closest S-BERT embeddings, leveraging the semantic similarity captured by S-BERT.

4 Results

The performance metrics for the selected activities demonstrate the effectiveness and limitations of the ZSL approach in recognizing both seen and unseen activities. Table 4 evaluates activities "washing hands," "typing on a keyboard,"

and "brushing hair". For "washing hands," the activity shows a precision of 0.72, recall of 0.91, and F1-score of 0.80, indicating high sensitivity but slightly lower precision. "Typing on a keyboard" exhibits excellent performance with a precision of 0.94, recall of 0.98, and F1-score of 0.96, showcasing the model's accuracy in recognizing this activity. "Brushing hair" demonstrates lower performance with a precision of 0.87, recall of 0.57, and F1-score of 0.69, highlighting the challenges in recognizing this unseen activity. Overall, the model achieves an accuracy of 0.85, a macro average F1-score of 0.82, and a weighted average F1-score of 0.84.

Table 4: Performance Metrics for Selected Activities. Seen classes: washing hands, writing, brushing teeth. Unseen classes: washing dishes, typing on a keyboard, brushing hair (7 seconds).

Activity	**Precision**	**Recall**	**F1-Score**	**Support**
Washing hands	0.72	0.91	0.80	195
Typing on a keyboard	0.94	0.98	0.96	311
Brushing hair	0.87	0.57	0.69	190
Accuracy		0.85		
Macro avg	0.84	0.82	0.82	696
Weighted avg	0.86	0.85	0.84	696

Table 5 focuses on the activities "open a box," "take off a shoe," and "take off glasses". "Open a box" shows moderate performance with a precision of 0.70, recall of 0.60, and F1-score of 0.64, indicating balanced but not outstanding recognition capability. "Take off a shoe" achieves a precision of 0.85, recall of 0.48, and F1-score of 0.61, suggesting high precision but lower recall, likely due to variability in how this activity is performed. "Take off glasses" presents a precision of 0.58, recall of 0.92, and F1-score of 0.71, revealing high recall but lower precision. The overall accuracy for this set of activities is 0.66, with a macro average F1-score of 0.66 and a weighted average F1-score of 0.66. These results reflect the challenges in recognizing activities with high variability and those that were unseen during training.

The analysis reveals that the zero-shot learning model demonstrates the ability to generalize to new, unseen activities, although with varying degrees of success.

These results emphasize the importance of continuous improvement in zero-shot learning techniques to enhance the recognition of diverse and subtle activities in smart home environments. The current data is limited, focusing on a small set of activities, and should be extended to include more classes to improve the model's generalization capabilities. The ability to accurately recognize a wide range of human activities is crucial for applications in healthcare, assistive technologies, and beyond, where understanding and adapting to a wide range of human activities is essential. Expanding the dataset to include more activities

Table 5: Performance Metrics for Selected Activities. Seen classes: open a bottle, take off a jacket, put on glasses. Unseen classes: open a box, take off a shoe, take off glasses (2 seconds).

Activity	**Precision**	**Recall**	**F1-Score**	**Support**
Open a box	0.70	0.60	0.64	324
Take off a shoe	0.85	0.48	0.61	253
Take off glasses	0.58	0.92	0.71	262
Accuracy		0.66		
Macro avg	0.71	0.67	0.66	839
Weighted avg	0.71	0.66	0.66	839

with varying contexts and semantics will further enhance the robustness and applicability of the ZSL approach in real-world settings.

5 Related Work

5.1 Non-ZSL Approaches

In this section, to conduct a thorough review of the literature, we employed a systematic approach utilizing the research query shown in Table 6 on Scopus. The query is specifically designed to capture relevant studies that a) focus on accelerometers, gyroscopes, magnetometers, or inertial measurement unit (IMU) sensors, b) mounted on the wrist, and c) aim to recognize a set of micro ADLs.

Table 6: Defined Research Query (2015-2024)

Operator	**Keywords**
	(accelerometer OR gyroscope OR magnetometer OR IMU)
AND	(wrist-mounted OR wrist)
AND	(drink water OR eat meal OR open a bottle OR open a box OR brush teeth OR brush hair OR take off a jacket OR put on a jacket OR put on a shoe OR take off a shoe OR put on glasses OR take off glasses OR sit down OR stand up OR writing OR phone call OR type on a keyboard OR salute OR wave hand OR sneeze cough OR blow nose OR washing hands OR dusting OR ironing OR washing dishes)

Through this approach, 30 relevant studies were identified discussing the challenges of recognizing micro ADLs using wearable sensors and proposing potential solutions. Out of these studies, only 12 delve into the recognition of hand-related micro-ADLs. In particular, they are linked to the PAAL-ADL (Performance in an Active and Assisted Living-ADL) dataset [9] and the HTAD (Home-Tasks Activities Dataset) dataset [20]. The methodologies related to PAAL-ADL, such as those outlined in [8] and [4], propose HAR methodologies achieving accura-

cies of nearly 86% and 91%, respectively, in recognizing the 24 ADLs[4] within the PAAL-ADL dataset. Both methods employ data filtering, feature extraction in both time and frequency domains, and the utilization of different algorithms, precisely the Nondominated Sorting Genetic Algorithm III (NSGA-III) [8] and Locally Weighted Random Forest (LWRF) [4].

On the other hand, the methodology [20] investigates the recognition of 7 activities (i.e., eating chips, mopping the floor, sweeping, brushing teeth, washing hands, typing on the keyboard, and watching TV) using data from a wrist accelerometer and audio stream provided in the HTAD dataset. The methodology proposes a Multilayer Perception (MLP) approach that takes as input a set of 16 statistical features related to acceleration and 36 Mel Frequency Cepstral Coefficients (MFCCs) related to audio, achieving an F1-Score of 0.91.

However, in [8,4,20], during the training and testing phases, the authors apply a classic k-fold approach over the dataset. This implies that the proposed method includes data from the same subject and the same data collection session of that subject in both the training and testing phases, thereby posing a potential risk of overfitting to specific individuals and sessions, which may limit the generalization capability of the model to broader populations or different contexts.

In [12], the authors presented a multi-level segmentation approach for recognizing a set of 24 hand-related micro activities, achieving higher accuracy results for activities longer than 7 seconds of average duration.

In [37], authors present a transfer learning methodology that recognizes seven toilet-related activities (i.e., dressing, undressing, brushing teeth, using the toilet, washing face, and washing hands), achieving an F1-Score of 0.84. Other hand-related micro-activities recognition includes digit recognition [24,35], handwritten signature [34,40], finger movements [6], and hand washing [41] through wrist movements. Moreover, from the grey literature, various approaches were proposed during the *IEEE COINS 2023 Contest for In Sensor Machine Learning Computing* [33]. However, regardless of the highly accurate results achieved, the details of the proposed methods are missing.

5.2 ZSL Approaches

Finally, by updating the Scopus research query from Table 6 to include an element related to ZSL (see Table 7), it was revealed that no prior work on hand-related micro ADL has been proposed in the literature. This makes the present work the first of its kind.

Table 7: Updated Research Query (2015-2024)

Operator	**Keywords**
AND	(zsl OR zero shot learning OR zero-shot learning OR zero-shot OR zero shot)

[4] The PAAL-ADL and our dataset present the same ADLs.

Based on our analysis, the existing literature mainly concentrates on: *a)* reduced sets of hand-based ADLs presenting similar temporal and functional characteristics, *b)* data collection performed in a laboratory environment, undergoing strict constraints, *c)* testing approaches that are not subject-independent, and *d)* to the best of our knowledge, recognizing hand-based micro activities using wrist-worn inertial sensors with ZSL has not been proposed in the state-of-the-art.

6 Conclusions

This study focused on recognizing 12 hand-based Activities of Daily Living (ADLs) using inertial sensor data, introducing a two-level segmentation strategy.

We examined the efficacy of Zero-Shot Learning (ZSL) for activity recognition, using Sentence-BERT (S-BERT) for semantic embeddings and Variational Autoencoders (VAE) to link seen and unseen classes. Our approach leveraged sensor data and S-BERT-generated embeddings to transfer knowledge effectively between seen and unseen activities. Future work will focus on expanding the number of recognized activities beyond the current 12, ensuring that seen and unseen classes share contextual semantic similarities to facilitate a more effective transfer learning process. We will also explore the use of large language models to represent the semantic space, potentially enhancing the accuracy and robustness of the ZSL framework.

7 Celebrating Prof. Tiziana Margaria

Tiziana has been a pivotal figure in my (Florenc Demrozi's) academic journey, as well as that of my colleague, Fadi al Machot. Our collaboration spans various research activities, project involvements, and organizational endeavors, all of which have been profoundly enriching.

We indirectly met in 2022 when my paper [14] and her paper [7] were the best paper candidates at the IFIP International Internet of Things Conference. After that, we meet within the IFIP Working Group 10.5 and the AWS Fellowship context. This collaboration has led to notable contributions in the fields of Human Activity Recognition (HAR) and educational methodologies in software engineering, as well as a three-month visiting period at the Immersive Software Engineering (ISE) program at the University of Limerick. During this visiting period, we started our collaboration, which led to the first two joint publications. In our paper *Experiences from the First Delivery of a New Immersive Software Engineering Course: Mathematical Foundations and Data Analytics* [13], we explore the integration of mathematical foundations and data analytics into the ISE course. This work underscores the transformative potential of innovative teaching methodologies in software engineering education.

Another notable contribution is our research on *CNN-based HAR on Edge Computing Devices* [38]. This study delves into the application of Convolutional Neural Networks (CNN) for HAR on edge computing devices. It highlights the

potential of edge computing in enhancing real-time data processing and activity recognition accuracy, paving the way for more efficient and effective HAR systems. My collaboration with Tiziana also extends to organizing significant conferences, such as Very Large Scale Integration - System on Chip (VLSI-SoC 2026), and prolific discussions towards European projects focused on predictive health technologies dedicated to developing a human digital twin for health status prediction and Alzheimer's disease prevention. These endeavors showcase our collective commitment to advancing research and development in crucial areas of technology and health. Looking ahead, we are excited about potential future collaborations in the context of the Research at Immersive Software Engineering ($R@ISE$) project. This initiative, along with similar projects [30], promises to further our exploration into immersive and practical aspects of software engineering education. The $R@ISE$ project exemplifies our forward-thinking approach to integrating immersive technologies into educational frameworks, enhancing the learning experience, and preparing students for the evolving demands of the software engineering industry. Additionally, the philosophy of simplicity [28,29] underpins our methodologies, as articulated in several influential publications. This philosophy advocates for streamlined, efficient approaches to complex problems, ensuring that solutions are both effective and accessible. The foundational concepts of Low-Code/No-Code (LCNC) [26] development have been a cornerstone of our research, fostering innovative approaches to software and HAR model creation. These concepts promote the use of visual development environments and pre-built components, enabling faster and more flexible software development processes [27].

It is essential to highlight the academic journey of Fadi Al Machot, a former student of Tiziana at Potsdam University. Under her mentorship, Fadi developed a deep understanding of software engineering and cyber-physical systems. He created the Machine Learning and Neurocomputing group as an Associate Professor at the Norwegian University of Life Sciences, where he applies his expertise to advance the field and mentor the next generation of researchers. Tiziana's mentorship has been instrumental in shaping Fadi's career, and her influence is evident in the quality and impact of his work. Our collaboration with Tiziana has shaped our research directions and achievements. Her mentorship and contributions have left an indelible mark on our professional paths, and we look forward to continuing this fruitful partnership in future endeavors. Prof. Margaria's unwavering commitment to advancing the field of software engineering, combined with her innovative approach to research and education, makes her a truly remarkable and inspirational figure in our academic community.

Appendix: Detailed Descriptions (Semantics) of Seen and Unseen Activities used for ZSL

Seen Classes

1. The person washes their hands with soap and water for hygiene, typically in a bathroom or kitchen. This involves rubbing their hands together with

soap under running water, often for at least 20 seconds to ensure cleanliness and reduce the risk of infection.
2. The person writes notes or a letter using a pen or pencil, typically at a desk or table. This involves holding the writing instrument and making marks on paper, often focusing on conveying thoughts clearly and legibly.
3. The person brushes their teeth with a toothbrush and toothpaste, typically in a bathroom. This involves applying toothpaste to the brush and moving it back and forth against the teeth, aiming to remove plaque and maintain oral hygiene, and often results in a fresh minty taste in their mouth.
4. The person opens a plastic bottle by unscrewing the cap, typically to drink or pour the contents. This involves gripping the bottle with one hand and twisting the cap with the other, sometimes hearing a popping sound as the seal breaks.
5. The person removes a jacket they are wearing by pulling it off, usually when entering a warm indoor space. This involves unzipping or unbuttoning the jacket and sliding it off their arms, often feeling relief from the heat as they do so.
6. The person puts on glasses to improve their vision, typically done in a well-lit area. This involves lifting the glasses and positioning them on their nose and ears, allowing them to see more clearly and reduce eye strain.

Unseen Classes

1. The person washes dishes in the sink or a dishwasher after a meal, typically in a kitchen. This involves scrubbing dishes with a sponge or loading them into a dishwasher, often ensuring that all food residue is removed and the dishes are clean and ready for future use.
2. The person types on a keyboard of a computer or laptop, typically sitting at a desk. This involves pressing keys to input text or commands, often focusing on accuracy and speed to complete a task or communicate online.
3. The person brushes their hair using a hairbrush or a comb, usually in front of a mirror. This involves running the brush or comb through their hair to detangle and smooth it, often making their hair look neat and presentable.
4. The person opens a cardboard box to retrieve an item inside, usually by cutting or tearing the tape. This involves pulling open the flaps and reaching inside the box, often feeling a sense of anticipation and curiosity about the contents.
5. The person takes off a shoe they are wearing by pulling it off, often when returning home. This involves loosening any laces or straps and sliding the shoe off their foot, often feeling a sense of relief and comfort as their feet are freed.
6. The person removes glasses they were wearing to see better, usually to clean them or switch to contact lenses. This involves taking hold of the frames and lifting them off their face, often feeling a temporary blur in their vision.

References

1. Al Machot, F., R. Elkobaisi, M., Kyamakya, K.: Zero-shot human activity recognition using non-visual sensors. Sensors **20**(3), 825 (2020)
2. Ali, M.T., Turetta, C., Demrozi, F., Pravadelli, G.: ICT-based solutions for alzheimer's disease care: A systematic review. IEEE Access **12**, 13944–13961 (2024). https://doi.org/10.1109/ACCESS.2024.3356348
3. Ali, M.T., Turetta, C., Pravadelli, G., Demrozi, F.: Ict-based solutions for alzheimer's disease care: A systematic review. IEEE Access (2024)
4. Aşuroğlu, T.: Complex human activity recognition using a local weighted approach. IEEE Access **10**, 101207–101219 (2022)
5. Boldo, M., Bombieri, N., Centomo, S., De Marchi, M., Demrozi, F., Pravadelli, G., Quaglia, D., Turetta, C.: Integrating wearable and camera based monitoring in the digital twin for safety assessment in the industry 4.0 era. In: International Symposium on Leveraging Applications of Formal Methods. pp. 184–194. Springer (2022)
6. Chandel, V., Ghose, A.: Demo abstract - nntrak: Real-time wrist tracking using smartwatch with cnn. In: SenSys 2022 - Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems. p. 754 – 755 (2022). https://doi.org/10.1145/3560905.3568047
7. Chaudhary, H.A.A., Guevara, I., John, J., Singh, A., Margaria, T., Pesch, D.: Low-code internet of things application development for edge analytics. In: IFIP International Internet of Things Conference. pp. 293–312. Springer (2022)
8. Climent-Pérez, P., Florez-Revuelta, F.: Privacy-preserving human action recognition with a many-objective evolutionary algorithm. Sensors **22**(3), 764 (2022)
9. Climent-Pérez, P., Muñoz-Antón, Á.M., Poli, A., Spinsante, S., Florez-Revuelta, F.: Dataset of acceleration signals recorded while performing activities of daily living. Data in Brief **41**, 107896 (2022)
10. Compagnon, P., Lefebvre, G., Duffner, S., Garcia, C.: Learning personalized adl recognition models from few raw data. Artificial Intelligence in Medicine **107**, 101916 (2020)
11. Deelaka, P.N., De Silva, D.Y., Wickramanayake, S., Meedeniya, D., Rasnayaka, S.: Tezarnet: Temporal zero-shot activity recognition network. In: International Conference on Neural Information Processing. pp. 444–455. Springer (2023)
12. Demrozi, F., AL Machot, F.: An enhanced subject-independent approach for hand-based micro activities recognition. In: 2024 IEEE International Conference on Omni-layer Intelligent Systems (COINS). pp. 1–4 (2024)
13. Demrozi, F., Marchisio, M., Margaria, T., Sacchet, M.: Experiences from the first delivery of a new immersive software engineering course: mathematical foundations and data analytics. In: 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). pp. 1576–1581. IEEE (2023)
14. Demrozi, F., Pravadelli, G.: Shpia: a low-cost multi-purpose smart home platform for intelligent applications. In: IFIP International Internet of Things Conference. pp. 217–234. Springer (2022)
15. Demrozi, F., Pravadelli, G., Bihorac, A., Rashidi, P.: Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. IEEE access **8**, 210816–210836 (2020)
16. Demrozi, F., Serlonghi, N., Turetta, C., Pravadelli, C., Pravadelli, G.: Exploiting bluetooth low energy smart tags for virtual coaching. In: 2021 IEEE 7th World Forum on Internet of Things (WF-IoT). pp. 470–475. IEEE (2021)

17. Demrozi, F., Turetta, C., Kindt, P.H., Chiarani, F., Bacchin, R.A., Valè, N., Pascucci, F., Cesari, P., Smania, N., Tamburin, S., et al.: A low-cost wireless body area network for human activity recognition in healthy life and medical applications. IEEE Transactions on Emerging Topics in Computing **11**(4), 839–850 (2023)
18. Demrozi, F., Turetta, C., Machot, F.A., Pravadelli, G., Kindt, P.H.: A comprehensive review of automated data annotation techniques in human activity recognition. arXiv preprint arXiv:2307.05988 (2023)
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
20. Garcia-Ceja, E., Thambawita, V., Hicks, S.A., Jha, D., Jakobsen, P., Hammer, H.L., Halvorsen, P., Riegler, M.A.: Htad: A home-tasks activities dataset with wrist-accelerometer and audio features. Lecture Notes in Computer Science **12573 LNCS**, 196 – 205 (2021). https://doi.org/10.1007/978-3-030-67835-7_17
21. Ishihara, Y., Ozaki, H., Nakagata, T., Yoshihara, T., Natsume, T., Kitada, T., Ishibashi, M., Deng, P., Yamada, Y., Kobayashi, H., et al.: Association between daily physical activity and locomotive syndrome in community-dwelling japanese older adults: a cross-sectional study. International Journal of Environmental Research and Public Health **19**(13), 8164 (2022)
22. Issa, M.E., Helmi, A.M., Al-Qaness, M.A., Dahou, A., Abd Elaziz, M., Damaševičius, R.: Human activity recognition based on embedded sensor data fusion for the internet of healthcare things. In: Healthcare. vol. 10, p. 1084. MDPI (2022)
23. Kingma, D.P., Welling, M., et al.: An introduction to variational autoencoders. Foundations and Trends® in Machine Learning **12**(4), 307–392 (2019)
24. Leong, L., Wiere, S.: Digit recognition from wrist movements and security concerns with smart wrist wearable iot devices. In: Proceedings of the Annual Hawaii International Conference on System Sciences. vol. 2020-January, p. 6448 – 6455 (2020)
25. Mantovani, E., Demrozi, F., Hertz, D.L., Turetta, C., Ferro, O., Argyriou, A.A., Pravadelli, G., Tamburin, S.: Wearables, sensors, and smart devices for the detection and monitoring of chemotherapy-induced peripheral neurotoxicity: Systematic review and directions for future research. Journal of the Peripheral Nervous System **27**(4), 238–258 (2022)
26. Margaria, T.: Knowledge management for inclusive system evolution. Transactions on Foundations for Mastering Change I pp. 7–21 (2016)
27. Margaria, T., Chaudhary, H.A.A., Guevara, I., Ryan, S., Schieweck, A.: The interoperability challenge: building a model-driven digital thread platform for cps. In: International Symposium on Leveraging Applications of Formal Methods. pp. 393–413. Springer (2021)
28. Margaria, T., Floyd, B.D.: Simplicity in it: a chance for a new kind of design and process science. Journal of Integrated Design and Process Science **17**(3), 1–7 (2013)
29. Margaria, T., Hinchey, M.: Simplicity in it: The power of less. Computer **46**(11), 23–25 (2013). https://doi.org/10.1109/MC.2013.397
30. Margaria, T., Steffen, B.: Extreme model-driven development (xmdd) technologies as a hands-on approach to software development without coding. Encyclopedia of Education and Information Technologies pp. 732–750 (2020)
31. Martins, L.M., Ribeiro, N.F., Soares, F., Santos, C.P.: Inertial data-based ai approaches for adl and fall recognition. Sensors **22**(11), 4028 (2022)
32. Matsuki, M., Lago, P., Inoue, S.: Characterizing word embeddings for zero-shot sensor-based human activity recognition. Sensors **19**(22), 5043 (2019)

33. Pau, D., Korobitsyn, A., Proshin, D., Zherebtsov, D., Bianco, M.: Ieee coins 2023 contest for in sensor machine learning computing. Authorea Preprints (2023)
34. Ramachandra, R., Venkatesh, S., Raja, K., Busch, C.: Handwritten signature and text based user verification using smartwatch. In: Proceedings - International Conference on Pattern Recognition. p. 5099 – 5106 (2020). https://doi.org/10.1109/ICPR48806.2021.9412048
35. Rattray, J.M., Ujhazy, M., Stevens, R., Etienne-Cummings, R.: Assistive multimodal wearable for open air digit recognition using machine learning. In: International IEEE/EMBS Conference on Neural Engineering, NER (2023). https://doi.org/10.1109/NER52421.2023.10123870
36. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
37. Shang, M., Zhang, Y., Ali Amer, A.Y., D'Haeseleer, I., Vanrumste, B.: Bathroom activities monitoring for older adults by a wrist-mounted accelerometer using a hybrid deep learning model. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. p. 7112 – 7115 (2021). https://doi.org/10.1109/EMBC46164.2021.9630659
38. Singh, A., Margaria, T., Demrozi, F.: Cnn-based human activity recognition on edge computing devices. In: 2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS). pp. 1–4 (2023). https://doi.org/10.1109/COINS57856.2023.10189270
39. Wang, W., Li, Q.: Generalized zero-shot activity recognition with embedding-based method. ACM Transactions on Sensor Networks **19**(3), 1–25 (2023)
40. Xu, C., Pathak, P.H., Mohapatra, P.: Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch. In: HotMobile 2015 - 16th International Workshop on Mobile Computing Systems and Applications. p. 9 – 14 (2015). https://doi.org/10.1145/2699343.2699350
41. Zhang, Y., Xue, T., Liu, Z., Chen, W., Vanrumste, B.: Detecting hand washing activity among activities of daily living and classification of who hand washing techniques using wearable devices and machine learning algorithms. Healthcare Technology Letters **8**(6), 148 – 158 (2021). https://doi.org/10.1049/htl2.12018