Comparing Feature Learning Methods for Human Activity Recognition: Performance study in new user scenario

Paula Lago
Department of Basic Science, Faculty of Engineering
Kyushu Institute of Technology
Kitakyushu, Japan
paula@mns.kyutech.ac.jp

Sozo Inoue
Graduate School of Biological Engineering, Faculty of Engineering
Kyushu Institute of Technology
Kitakyushu, Japan
sozo@brain.kyutech.ac.jp

***Abstract*—Designing robust features for human activity recognition (HAR) that perform well across a wide range of users is a hard task. Therefore, more attention is being given to feature learning techniques, to automatically learn features from raw data. In this paper, we present a comparative study among feature learning methods for HAR. Using accelerometer data, we compare four methods for feature learning from raw-sensor data (PCA-based, clustering, matrix factorization, and LSTM networks) to the traditional hand-crafted feature extraction method. We focus on performance degradation when each model is evaluated using a new user. According to our results, features learned with Principal Component Analysis are the most robust to the new user scenario. Our results evidence the importance of evaluation in unseen user, since the performance difference compared to a random split testing is big.**

***Index Terms*—feature learning, activity recognition, accelerometer, smartphones**

I. Introduction

Human Activity Recognition (HAR) is the process that enables our smartphones to know how much we have walked this week or if we went to work by bike yesterday [4]. By analyzing the data produced by their sensors (usually accelerometers and gyroscopes), smartphones can characterize the activity of the user in a given time. HAR lies at the heart of many applications in healthcare, fitness and ambient intelligence [7].

HAR is typically framed as a supervised learning problem [2]. As such, one critical step is feature engineering and selection. Feature enginnering is the process of designing and selecting the characteristics that will be fed into a classifier model to be trained. Systems designed to be used by different users need features that can handle person-to-person differences in the way of doing an activity. For example, a person might walk as fast as another person runs. There exists a trade-off between choosing discriminating features and features that are robust to such variability [2].

Descriptive statistical measures of sensor values in a time window (or their frequency spectrum) have good performance in experimental scenarios for physical activities [3], [16]. However, different studies on feature selection [5], [9] have demonstrated that no single feature performs best across all activities. Moreover, hand-crafted features usually perform lower for a new user. Even after these studies, there is still no agreement on what features should be used and researchers and practitioners need to spend time designing and selecting features.

In recent years, there has been a growing interest on feature learning [6], [12] and deep learning techniques [10] to reduce the time spent in data pre-processing and feature engineering. Feature learning is the process of *automatically learning features from raw data*. It focuses on discovering good representations for sensor data using unsupervised learning. The steps of feature engineering and feature selection in the typical Human Activity Recognition chain can be replaced by the feature learning step.

On the other hand, deep learning is used to learn activity patterns from raw sensor data, also skipping the feature engineering and selection step. Recent surveys have shown that physical activity recognition with automatic features or deep learning approaches achieve better or comparable results to that of traditional hand-crafted features [1], [12].

However, the robustness of feature learning methods and deep learning approach to a new user has not been studied before. In this paper, we focus on the empirical and comparative evaluation of their robustness. We compare the performance of four feature learning techniques and one deep learning method to the traditional hand-crafted features in a dataset with high variability.

We first review and compare feature learning techniques for accelerometer data (Section II). Then, we perform a comprehensive evaluation of the feature learning techniques, comparing their performance with different classifiers. Next, we evaluate on a public dataset (Section III) that specifically addresses variability with different subjects [14].

Our results show that features learned with Principal Component Analysis have the best performance in the new user scenario and a low performance drop compared to a scenario with no new users (Section IV). A performance decrease is seen in all methods when comparing the two scenarios. These results reveal a need to focus on model adaptability to a new user after training for better performance in real life applications (Section V).

The main contributions of this paper are the following:

- It is the first comparative study of *feature learning* techniques focusing on the *robustness to a new user* of four feature learning methods and one deep learning model.
- We evaluate each feature learning method with four different classifiers to control for the influence of the classifier method, whereas other comparative studies use a single classifier.

II. Related Work

There are several studies focusing on feature selection for wearable and smartphone-based activity recognition [5], [9], [16]. However, feature selection differs from feature learning. For feature selection, first a huge set of hand-crafted features is obtained from the raw sensor data. Then, the best features are selected for training the model. Feature learning skips the step of creating features. The input data is raw sensor data.

In this section we first introduce feature learning techniques used in activity recognition and other scenarios (Section II-A). We then review previous comparison studies stating the differences with this study (Section II-B).

A. Feature Learning Methods

In the following we briefly explain the most common feature learning methods, including some that have not been tested for activity recognition using accelerometer sensors.

Codebooks: [13] consider each sensor data window as a sequence, from which subsequences are extracted and grouped into clusters. Each cluster centre is a codeword. Then, each sequence is encoded using a bag-of-words approach using codewords as features.

Principal Component Analysis (PCA) is a commonly used dimensionality reduction technique. It transforms the data by projecting it into the set of linearly independent components that best explain the total data variance. Traditionally, it is used to reduce the set of hand-crafted features of the sensor data. In this study, we use PCA on the raw sensor data as a technique to automatically learn features. The features obtained by PCA are a linear combination of the original features, which summarizes the data and enables its reconstruction.

Deep Learning: Deep learning uses Neural Networks with several layers to learn patterns from data. Recurrent Neural Networks, Long-Short Term Memory Networks and Autoencoders have been used for time series analysis. They have shown promising results in video and speech analysis and also in activity recognition [6], [10].

Matrix Factorization is technique more commonly used in recommendation systems to find "latent" features in the data. As the name suggests, the goal is to find two factor matrices W and H such that $M = W * H$, where M is the original dataset. It has been used for activity recognition from video [8], but not from sensor data.

B. Other comparison studies

Other authors have previously compared feature learning methods for Human Activity Recognition. We now review some of them.

Plötz et al. [12] investigated the suitability of feature learning for HAR using accelerometer data. They compare PCA and an autoencoder (AE) method with the traditional approach in different activity domains. Their results show that automatically learned features, both by PCA and AE, can perform as well as traditional features, even slightly better in some domains. However, they propose using cumulative distribution for representing raw data which improves the results of all methods.

Bhattacharya et al. [1] compare codebook learning to PCA and to traditional features. In contrast to the results of Plötz, in their study traditional features outperform PCA. However, their codebook approach outperforms the traditional approach.

Wang et al. [15] provide a benchmark of traditional and deep learning methods on a locomotion recognition dataset.Their results show that deep learning method can outperform a traditional approach by almost6%.

The above mentioned studies compare only a few set of feature learning methods and do not evaluate the impact of heterogeneities in sensing such as different sensing devices and how well models perform on new users. In this paper we evaluate five feature learning techniques and compare them to the traditional approach focusing on the new subject scenario.

III. Experimental Setup

As mentioned, we evaluate six different methods for feature learning. While other authors use only one classifier for all feature sets, we used 4 classifier methods and compare the performance of each feature learning method on each classifier. In this section, we describe the dataset used to perform the experiment (Section III-A), the setup of each feature learning method (Section III-B) and the evaluation metrics used for the comparison (Section III-C).

A. Data

We used the 'recording scenario' of the Heterogeneity Activity Recognition Data Set (HHAR) [14] which considers sensing heterogeneities of different devices and users. We run the experiments with only the phone accelerometer data for this study. The dataset contains six different user activities: `Biking`, `Sitting`, `Standing`, `Walking`, `Stair Up` and `Stair down`. HHAR was gathered by nine users (age range 25 30 years). The users followed a scripted set of activities conducting five minutes of each activity during dataset collection. Although the dataset contains some samples labeled as `NULL` when transitioning from one activity to the other, this class have not been considered in this paper. For each user and activity, data from 8 different smartphone models, all carried at the waist, are available.

Each line in the dataset corresponds to a sensor reading. For our experiments, we *segmented* the dataset in windows of 200

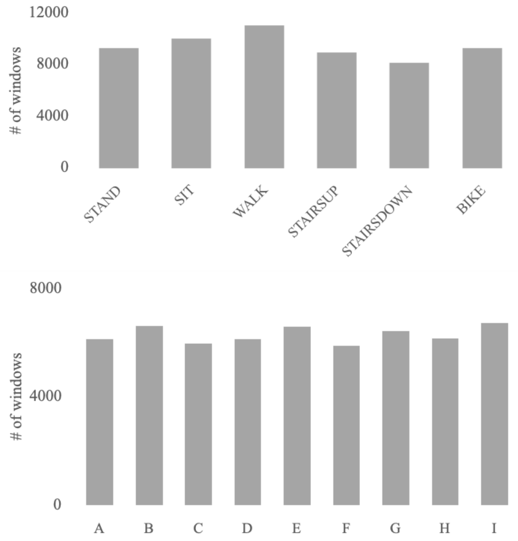

Fig. 1. The window distribution by class and by user is balanced. The mean number of windows by class is about 9442 and by user it is about 6200.

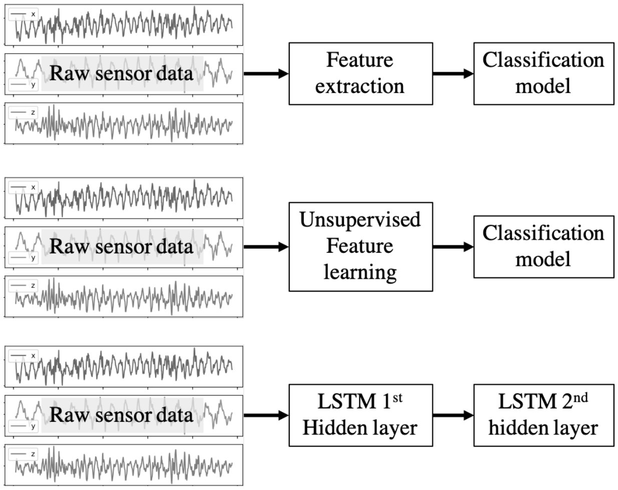

Fig. 2. General pipelines of the evaluation for traditional approach (top), feature learning techniques (middle) and LSTM model (bottom)

sensor observations from the same smartphone with no overlap. One observation corresponds to a triaxial accelerometer reading. A summary of the distribution of number of windows by activity and by user in the dataset is shown in Figure 1

B. Feature learning

The purpose of this study is to benchmark different feature learning methods. The general pipeline followed for the study is shown in Figure 2.

As a baseline, we use a traditional activity recognition approach using time-domain and frequency domain features. For each window, we calculate the following set of statistical features in each axis and domain: mean, minimum value, maximum value, variance, standard deviation, skew, kurtosis, median value and interquartile range (54 features in total). Then, we evaluated the unsupervised feature learning methods presented in Section II with the following considerations:

- *Principal Component Analysis (PCA) features:* Using each raw sensor observation as features (i.e. 600 features per sample), we use PCA to obtain 100, 50 and 25 principal components.
- *Matrix Factorization:* Using each raw sensor observation as features (i.e. 600 features per sample), we use Non-negative matrix factorization to obtain 100, 50, 25 and 10 components. Since only positive values can be used for this method, we add the minimum value observed to all sensor readings.
- *Time series clustering:* Using the Dynamic Time Wrapping distance, we use K-Means to obtain 15 clusters, used as codes. This process was done on sub-windows of 25 sensor observations, so each window is encoded in 8 codes.
- *Deep learning features:* We used a model with 2 fully connected and 2 LSTM layers with 64 hidden units each. One layer is equivalent to the unsupervised feature learning and the second layer is equivalent to the classifier model of the other 5 techniques evaluated. We used each raw observation as features separated by axis (i.e. each input is a 3*200 matrix)

For each evaluation fold, we learned the features and the classifier model with all-but-one subject and then applied both the feature extraction model and classifier model to the left-out subject. For each feature set, except the LSTM, we evaluated a linear SVM, RBF Kernel SVM, Naive Bayes and Random Forest Classifier [1].

C. Classification and Performance Metric

We used the $F_{1-Score}$ as an evaluation metric for all methods[2] Because we used leave-one-subject-out cross-validation, we average the F1-Score of each fold to obtain a final evaluation score for each model.

$$Avg.F_{1-Score} = \sum_{i=1}^{s} \frac{F^{i}_{1-Score}}{s} \qquad (1)$$

where $F^{i}_{1-Score}$ is the average $F^{i}_{1-Score}$ obtained when using subject i as test data and s is the total number of subjects, in this case 9.

IV. Results and Discussion

In this section, we present the results of the comparison evaluation comparing the results on two scenarios: no new user and the new user scenario. We discuss the implications of this results and the limitations of this study.

[1] The source code used for this evaluation will be publicly available

[2] We use the micro average $F_{1-Score}$ of all classes given by the sklearn library. See scikit-learn documentation for details.

A. On the robustness of feature learning techniques to new user scenario

For the first scenario, we evaluated all models with a randomized train-test split in which samples from all subjects can be found in the training and test set. Because the windows do not overlap, the random split does not provide test data during training.

From the data in Figure 3, it is evident that when the models are trained with data from all users, the performance is higher than when the evaluation is done in a new user. This result is in line with knowledge about model evaluation, in which the new user scenario is the hardest scenario for evaluating a model. What is surprising is that their relative performances change when the evaluation considers a new user. Such results indicate that the methods with a better rank in the second scenario are more robust to inter-person variability.

The method with the highest drop in performance in the new user scenario was the LSTM. Possibly the LSTM model is overfitting to the data in the first scenario, with no new users in the test data. In this scenario, the training data contains examples of how each user performs each activity and the LSTM can learn how to recognize on each user. However, in the second scenario, if the network overfits to seen users, it explains the low performance on a new user. This is one of the main drawbacks of deep approaches as they need high volume of data to overcome overfitting. This result also shows possible dangers of the evaluation of machine learning models on time series data, which may be overconfident if the test data is not adequately selected. Also surprising, is that only the PCA learned features, and the codebook features outperform or have a comparable performance to the traditional hand-crafted features on the leave-one-subject-out cross-validation. Both the PCA and codebook have the lowest drop in performance on the new user scenario. This shows that the features learned by these methods are robust and generalize well.

One advantage of the clustering method is that it enables a deeper insight on the sensor data by providing the cluster centers and the description of each activity by cluster density. However, due to the complexity of the distance function (O(n^2) it can take a long time to train and to test as the distance of the values to all cluster centers needs to be obtain to transform the data. On the other hand, there are incremental implementations of the PCA methods [3] that can be used to create an update-able model.

To better understand these results, we further explore the average precision and recall for each class on the PCA (best performance) and LSTM (lowest performance) methods in Figure 4. We can see that the LSTM has the lowest performance on the `walk` class, with both precision and recall being below 0.4. Possibly, this class has the greatest variability among users, due to different speeds and rhythms. It is possible to see also, that among the static activities (`stand` and `sit`) the difference in performance between PCA and LSTM is not big, but it grows on the dynamic activities (`walk`, `stairsup` and `bike`).

B. On the general performance of each method and the classifier influence

Our results reveal that some classifiers are better suited for each feature set. For instance, for PCA learned features, non-linear classifiers work best. In contrast, linear classifier had the best performance with the hand-crafted features. Studies that use only one classifier method may penalize one feature set over the other when the classifier is not appropriate for the learned features. A closer analysis to the results on each feature learning method is presented next.

As in the study by Plötz et al. the PCA-based features outperformed the traditional features. There seems to be a sweet spot in terms of how many components to choose as the performance decreases with both a high number of features and a low number of features (Figure 5).

The matrix factorization technique had a low performance in general. However, it is interesting that its performance increases as the number of components decreases (Figure 6). As for dimensionality reduction, this is a desired behavior as it better compacts the data.

Finally, although the traditional features have a performance decrease on the new user scenario, they are fast to compute and they can be easily implemented.

C. Limitations of this work

The results on this study show that feature learning can be used to obtain activity classifier robust to new user scenario. However, the dataset used for the evaluation is rather small with respect to number of users (9), their age range, and the number of activities (6). New studies with larger datasets are needed to confirm our findings.

A second limitation is that we have only considered one deep learning method while recently new methods have been proposed. These methods could obtain different results as the ones shown in this paper. One example is the Deep Convolutional LSTM [11] as an upgrade to the LSTM used in this study.

V. Conclusions

This paper contributes to a growing literature in feature learning from raw accelerometer data for robust activity recognition using wearable sensors. In this work, we have focused on the evaluation of feature learning methods for new user scenario. We compared five different feature learning methods to traditionally used hand-crafted features. Our evaluation focused on two factors that impact the performance of an activity recognition model. First,the impact of a new subject for the final classification model. Second, the interaction between the feature learning model and the classifier model.

As expected, the performance of all methods decreases when evaluated with a new user. This has two main implications. First, that activity recognition evaluation needs to be closely inspected for possible overfits when using random

[3] See for example https://scikit-learn.org/stable/modules/decomposition.html#incremental-pca

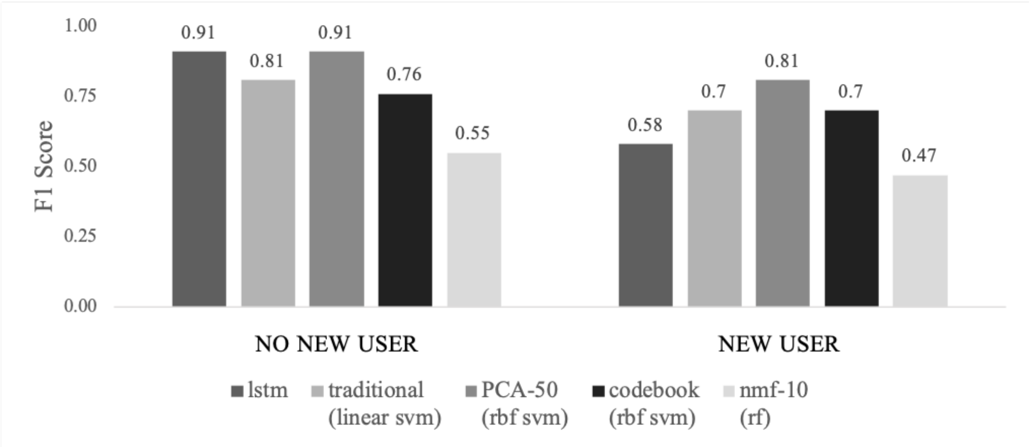

Fig. 3. F1-Score for each model for the two scenarios. The classifier with the best performance is shown in parenthesis (rf: random forest, svm: support vector machine, nb: naive bayes, knn: nearest neighbors). The F1-Score for all models drops when the model is evaluated on a new user. Their relative performances also change when the comparison is done considering the new user scenario.

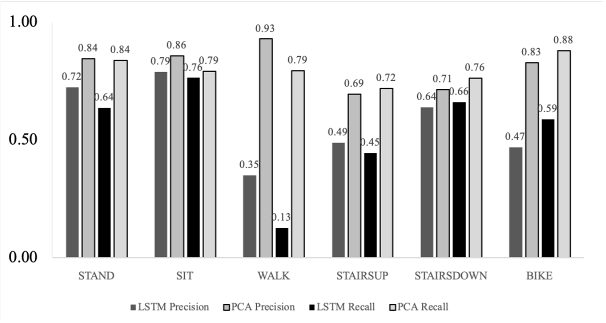

Fig. 4. Average precision and recall for each class using LSTM and PCA learned features in the new user scenario. LSTM performance is lower notably for dynamic activities, which have higher variability among users.

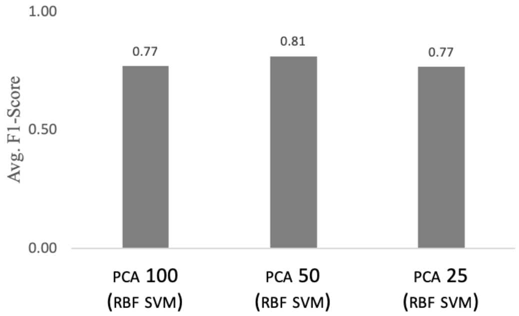

Fig. 5. Average F1-Score for the PCA features according to the number of components selected.

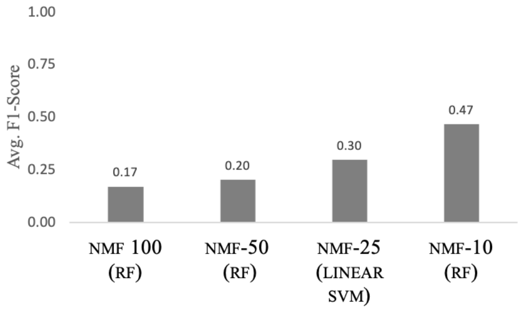

Fig. 6. Average F1-Score for the PCA learned features according to the number of components selected.

test splits or cross-validations. Second, it reveals the need of adapting classifiers to their new users. This can be done either by a short training setup or by obtaining labeled instances by experience sampling. This might be more important in applications with very different groups of users. For example, with different age groups or different abilities in performing the activity. Providing the feature to adapt the model to each new user in real applications is important to defeat the performance drop.

Feature learning can help reduce the need of large amounts of labeled data. As feature learning techniques use unsupervised learning, unlabeled data can be used to learn good features. Another advantage of feature learning techniques is that they can provide insights on activity characteristics that may help in other applications such as qualitative assessment or learning new activities.

Principal Component Analysis showed both a good performance in the recognition process and strong robustness to new user. In future work, we want to explore variants of the method such as the incremental implementation to provide model adaptation with short training sessions.

In summary, feature learning techniques can reduce time of the feature engineering process, but their robustness to inter-person variability and other variability sources in the data need to be considered. We hope this assessment will contribute to a better assessment of the techniques for feature learning from accelerometer data.

References

[1] Sourav Bhattacharya, Petteri Nurmi, Nils Hammerla, and Thomas Plötz. Using unlabeled data in a sparse-coding framework for human activity

recognition. *Pervasive and Mobile Computing*, 15:242–262, 2014.

[2] Andreas Bulling, U Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 1(June):1–33, 2014.

[3] Davide Figo, Pedro C. Diniz, Diogo R. Ferreira, and João M. Cardoso. Preprocessing techniques for context recognition from accelerometer data. *Personal Ubiquitous Comput.*, 14(7):645–662, October 2010.

[4] Arindam Ghosh and Giuseppe Riccardi. Recognizing Human Activities from Smartphone Sensor Signals. In *Proceedings of the ACM International Conference on Multimedia - MM '14*, pages 865–868, Orlando, Florida, 2014.

[5] Tâm Huynh and Bernt Schiele. Analyzing features for activity recognition. In *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies*, sOc-EUSAI '05, pages 159–163, New York, NY, USA, 2005. ACM.

[6] Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42(1):11–24, 2014.

[7] Jonathan Lester, Tanzeem Choudhury, and Gaetano Borriello. A Practical Approach to Recognizing Physical Activities. In Kenneth P. Fishkin, Bernt Schiele, Paddy Nixon, and Aaron Quigley, editors, *Pervasive Computing*, pages 1–16. Springer Berlin Heidelberg, 2006.

[8] C. Lin, B. Chen, W. Wu, W. Lin, and C. Tsai. Human action recognition based on non-negative matrix factorization. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1091–1093, Dec 2015.

[9] Syed Agha Muhammad, Bernd Niklas Klein, Kristof Van Laerhoven, and Klaus David. A feature set evaluation for activity recognition with body-worn inertial sensors. In Reiner Wichert, Kristof Van Laerhoven, and Jean Gelissen, editors, *Constructing Ambient Intelligence*, pages 101–109, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[10] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-garadi, and Uzoma Rita Alo. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, 105:233–261, 2018.

[11] Francisco Javier Ordez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 2016.

[12] Thomas Plötz, Nils Y. Hammerla, and Patrick Olivier. Feature learning for activity recognition in ubiquitous computing. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1729–1734. AAAI Press, 2011.

[13] Kimiaki Shirahama and Marcin Grzegorzek. On the Generality of Codebook Approach for Sensor-Based Human Activity Recognition. *Electronics*, 6(2):44, 2017.

[14] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, SenSys '15, pages 127–140, New York, NY, USA, 2015. ACM.

[15] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Sami Mekki, Stefan Valentin, and Daniel Roggen. Benchmarking the shl recognition challenge with classical and deep-learning pipelines. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, pages 1626–1635, New York, NY, USA, 2018. ACM.

[16] Mi Zhang and Alexander Sawchuk. A Feature Selection-Based Framework for Human Activity Recognition Using Wearable Multimodal Sensors. *Proceedings of the 6th International ICST Conference on Body Area Networks*, 1, 2011.