





# Model Evaluation Approaches for Human Activity Recognition from Time-Series Data

Lee B. Hinkle<sup>(✉)</sup>  and Vangelis Metsis 

Texas State University, San Marcos, TX 78666, USA  
{leebhinkle,vmetsis}@txstate.edu

**Abstract.** There are many evaluation metrics and methods that can be used to quantify and predict a model's future performance on previously unknown data. In the area of Human Activity Recognition (HAR), the methodology used to determine the training, validation, and test data can have a significant impact on the reported accuracy. HAR data sets typically contain few test subjects with the data from each subject separated into fixed-length segments. Due to the potential leakage of subject-specific information into the training set, cross-validation techniques can yield erroneously high classification accuracy. In this work (Source code available at: [https://github.com/imics-lab/model\\_evaluation\\_for\\_HAR](https://github.com/imics-lab/model_evaluation_for_HAR).), we examine how variations in evaluation methods impact the reported classification accuracy of a 1D-CNN using two popular HAR data sets.

**Keywords:** Model evaluation · Time-series data · Deep learning · Human activity recognition · Data resampling · Cross-validation

## 1 Introduction

With the advent of inexpensive wearable sensors in recent years, Human Activity Recognition (HAR) has been a hot topic of research both for medical applications and in human-computer interaction in general. In HAR, the methodology used for model evaluation differs from other areas such as image recognition due to the sequential nature of the data sets. HAR data sets typically consist of accelerometer and gyroscopic data recorded using a smartphone or wrist-worn device. Movement patterns specific to given activities such as running, walking, and sitting are identified using classic machine learning or newer deep learning approaches. HAR data sets typically differ from image and natural language data sets because the number of subjects is usually quite small, typical ranges are from 5 to 50 [13], with each subject contributing multiple samples while performing a range of activities. Traditional cross-fold and train/test split techniques can result in subject data from the test group being included in the training set.

The goal of trained models is generalized performance which means the performance on independent test data [7]. In the case of HAR the ability of a

model to correctly classify activities for an unknown subject. The primary issue seen in many accompanying analyses is that samples from a given subject may be present in both the train and test groups. This work examines the impact of subject assignment on two data sets. The remainder of this section briefly describes three data sets, their evaluation method, and the reported accuracy to illustrate the multiple approaches found in the literature. Section 2 describes the two data sets and processing used in this evaluation.

An example of a popular data set and evaluation with subjects preallocated into train and test groups is the UCI-HAR data set [1] which contains acceleration data captured on a waist-worn smartphone. Subjects were randomly assigned: 21 in the training set and 9 in the test set. The accompanying analysis reports an accuracy of 96% for six activities. Another example of a model evaluation with preallocated subjects is [5] which contains Android-based Smartphone data from 100 subjects. The reported accuracy *without resampling* is 93.8% for eight different activities. The authors state “the signals of the training set and test set are collected by different volunteers.” An example of hold-one-subject-out with individual results is [6] which uses multimodal motion data from the mHealth data set [2] and reports an average accuracy for 12 activities of 91.94%.

## 2 Materials and Methods

This section provides a brief overview of the two data sets used, the configuration of the 1D CNN, and the overall methodology.

The first data set used in this work is the MobiAct data set [12] which contains smartphone acquired raw accelerometer, gyroscope, and magnetometer data. 50 subjects were recorded performing nine types of activities of daily living (ADLs) and four types of falls. The accompanying analysis reports a best overall accuracy of 99.88% using 10-fold cross-validation. The authors state “we expect [the accuracy] to decrease when using leave-one-out cross-validation, which is a more realistic scenario.” For this work the timestamp ‘nanoseconds’ and accelerometer data (accel\_x/y/z) for the six Activities of Daily Living (ADL) were imported. The four types of falls, ‘sit chair’, ‘car step in’, and ‘car step out’ activities are not used as these are more events than activities. Gyro and magnetometer data are also not used for simplification. One second was discarded from the start/end of each record and the remaining data were segmented into 3-s windows. Prior works, including UCI-HAR [1] have used a 2.56-s window based on the mechanics and timing of human gait. This window length will yield multiple steps in each segment [3]. The six activity labels in y were one-hot-encoded.

The impact of the variable sampling rate and benefits of resampling were investigated using the MobiAct data. Sample timing is very consistent when using specialized equipment such as the BioRadio<sup>1</sup> or the Empatica E4 wristband<sup>2</sup>. However, when using a general-purpose device such as a smartphone preemption by other tasks results in a variation of timing between samples.

<sup>1</sup> <https://www.glneurotech.com/products/bioradio/>.

<sup>2</sup> <https://www.empatica.com/research/e4/>.

Figure 1a shows the delta time between data samples for a 30-s MobiAct walking segment. The Python’s Pandas mean resampling method was used to resample and downsample the data.

The second data set used is the Smartphone Human Activity Recognition data set from the University of Milan Bicocca (UniMiB SHAR) [8] which contains both fall and ADL data from 30 subjects that have been preprocessed into 3-s samples. The subjects are not preallocated into train/test and the accompanying analysis reports results for both component and total acceleration using 5-fold cross-validation and hold-one-subject-out validation. The highest performing RNN classifier achieves an accuracy of 88.41% using component acceleration and 5-fold cross-validation. Each classifier showed a decrease in accuracy in the Leave-One-Subject-Out validation. The accuracy drops to 73.17% using Leave-One-Subject-Out and 72.67% using total acceleration. The authors state that human subjects perform tasks in unique ways. The UniMiB SHAR acceleration data were transformed into a  $153 \times 3$  array and the total acceleration was calculated. The nine ADL class labels were one-hot-encoded.

A fixed 1D-CNN Keras [11] model shown to have good performance on time-sliced accelerometer data [4] was used for all experiments for consistency. Minimal tuning was performed, the primary change was increasing the convolution kernels to span one second of activity time. For a brief description of the layer functions with respect to time-series see [9]. The topology of the 1D-CNN is shown in Table 1.

**Table 1.** Keras sequential model 1D-CNN layers

Type	Input	Conv1D	Conv1D	Dropout	Max P1	Flatten	Dense	Dense
Params	$[60 \times 1]$	#f = 50 size = 1 s	#f = 50 size = 1 s	rate = 0.5	size = 2,		act = relu	act = softmax

All subject allocation experiments use total acceleration; MobiAct was resampled 20 Hz, UniMiB SHAR remains the 50 Hz. The next section describes how subjects were allocated to the training, validation, and test groups.

**Allocation Using Stratification:** While is easy to implement using the Scikit-learn [10] `train_test_split` method with stratification enabled a single subject’s samples are likely to be present in each of the groups.

**Allocation of Subjects by Attributes:** The UCI-HAR data set preallocates subjects but the UniMIB SHAR and MobiACT data sets do not. To generate a baseline each subject was allocated to the train, validate, *or* test group in a 60%/20%/20% ratio. Assuming that height would affect the mechanics of motion more than weight for the ADLs, subjects were sorted by height and manually allocated. Swaps were made to preserve the male to female ratio and a mix of age and weight<sup>3</sup>. The subject allocation is shown in Table 2.

<sup>3</sup> Several MobiAct subjects did not complete all ADLs were dropped resulting in a non-contiguous subject list. E.g. there is no subject number 14.

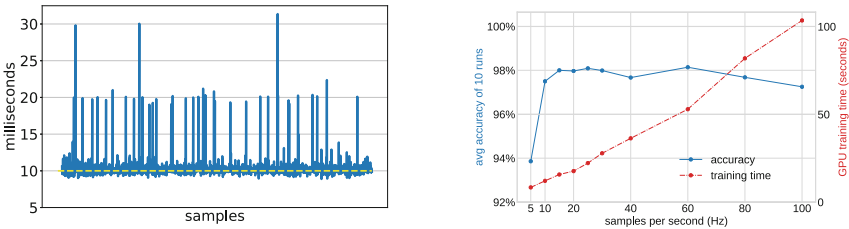
**Table 2.** Subject numbers: attribute based assignment

Data Set	Training	Validation	Test
MobiAct	[2,4,5,9,10,16,18,20,23-28,32,34-36,38,42,45-54,57]	[3,6,8,11,12,22,37,40,43,56]	[7,19,21,25,29,33,39,41,44,55]
UniMiB SHAR	[4-8,10-12,14,15,19-22,24]	[1,9,16,23,25,28]	[2,3,13,17,18,30]

**Subject Aware Cross-Validation:** Each subject was placed into the test group with the remaining subjects used for training and validation for hold-one-subject-out. The process was repeated with two, three, five, and ten subjects held out. To establish a range of possible results, the best and worst classified hold-one-out subjects were placed into min and max test groups.

### 3 Results and Discussion

Figure 1a shows the variation in sampling time for a walking sample. Figure 1b shows 5 Hz sampling results in reduced accuracy 10 Hz and above were largely the same. Reducing the sampling frequency significantly reduced the GPU-based<sup>4</sup> training time. This was even more pronounced when using CPU-based training where the 20 Hz data required just 4.5% of the training time required for 100 Hz data. For MobiAct the accuracy increased from 95.3% to 97.5% when using total acceleration and the attribute-based subject allocation with negligible impact on GPU training time. Table 3 shows that the accuracy when using stratification is extremely high at 99.3% (average of 10 runs, 200 epochs). Using the same model but with subjects allocated based by attribute, the accuracy drops to 96.9% for an error rate of 3.1% versus the stratified error of only 0.7%. The allocation of individual subject’s data into both the train and test groups results in erroneously increased accuracy when using stratified split. The UniMiB SHAR data results show the same trend.



(a)  $\Delta$ -time between samples      (b) Accuracy & training time vs. sampling rate

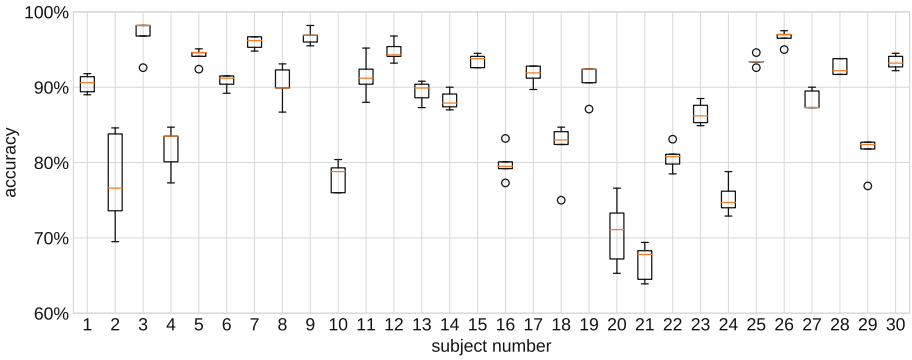
**Fig. 1.** Smartphone data sample fluctuation and impact of resample.

<sup>4</sup> GPU model Tesla P100-PCIE-16 GB at <https://www.colab.research.google.com>.

Figure 2 is a box plot of five runs for each subject and shows the large variation in accuracy among individual subjects. The overall by-subject cross-validation results are shown in Table 4.

**Table 3.** Stratified versus attribute-based subject split accuracy

Data set: Train/Validate/Test Split Method	Avg	Error	Delta
MobiAct: Stratified (incorrect)	99.3%	0.7%	–
MobiAct: Manual by Subject Attributes	96.9%	3.1%	2.4%
UniMiB SHAR: Stratified (incorrect)	93.9%	6.1%	–
UniMiB SHAR: Manual by Subject Attribute	92.3%	7.7%	1.6%



**Fig. 2.** Accuracy of each UniMiB Subject Tested Individually

**Table 4.** Accuracy based on X-fold and best/worst subjects. The delta between best/worst vs average accuracy narrows as more subjects are placed in a fold.

#subj/fold	MobiAct			#subj/fold	UniMiB SHAR		
	All	Min	Max		All	Min	Max
1	95%	78%	100%	1	87%	67%	97%
2	95%	84%	98%	2	86%	72%	94%
3	95%	89%	98%	3 (10-fold-CV)	85%	70%	93%
5 (10-fold-CV)	95%	92%	98%	6 (5-fold-CV)	86%	74%	91%
10 (5-fold-CV)	95%	92%	97%	–	–	–	–

## 4 Conclusion

In this work, we have shown that re-sampling smartphone acceleration data does not improve accuracy but downsampling can substantially reduce training time. This is important because consistent with prior work, stratified random allocation where samples from a single subject are present in both the training and testing groups generated higher accuracy than can be expected given an unknown subject. Hold-one-subject out is recommended but requires a train/test pass for each subject. We have shown that individual subject accuracies can vary greatly in a hold-one-out scenario and as the number of subjects in each fold increases the delta between possible min and max folds is reduced. Group-based 5-fold cross-validation can be used and closely matches the accuracy reported by averaging hold-one-subject-out.

## References

1. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: A public domain dataset for human activity recognition using smartphones. In: Esann, vol. 3, p. 3 (2013)
2. Banos, O., et al.: mHealthDroid: a novel framework for agile development of mobile health applications. In: Pecchia, L., Chen, L.L., Nugent, C., Bravo, J. (eds.) IWAAL 2014. LNCS, vol. 8868, pp. 91–98. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-13105-4\\_14](https://doi.org/10.1007/978-3-319-13105-4_14)
3. BenAbdelkader, C., Cutler, R., Davis, L.: Stride and cadence as a biometric in automatic person identification and verification. In: Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 372–377. IEEE (2002)
4. Brownlee, J.: 1D convolutional neural network models for human activity recognition, July 2020. <https://machinelearningmastery.com/cnn-models-for-human-activity-recognition-time-series-classification/>
5. Chen, Y., Xue, Y.: A deep learning approach to human activity recognition based on single accelerometer. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1488–1492. IEEE (2015)
6. Ha, S., Choi, S.: Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 381–388. IEEE (2016)
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. SSS. Springer, New York (2009). <https://doi.org/10.1007/978-0-387-84858-7>. <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
8. Micucci, D., Mobilio, M., Napoletano, P.: UniMiB SHAR: a dataset for human activity recognition using acceleration data from smartphones. Appl. Sci. **7**(10), 1101 (2017)
9. Nils: Introduction to 1d convolutional neural networks in keras for time sequences. <https://blog.goodaudience.com/introduction-to-1d-convolutional-neural-networks-in-keras-for-time-sequences-3a7ff801a2cf>
10. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
11. François, C.: Keras. GitHub repository (2015). <https://github.com/fchollet/keras>

12. Vavoulas, G., Chatzaki, C., Malliotakis, T., Pediaditis, M., Tsiknakis, M.: The mobiact dataset: recognition of activities of daily living using smartphones. In: ICT4AgeingWell, pp. 143–151 (2016)
13. Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L.: Deep learning for sensor-based activity recognition: a survey. *Pattern Recogn. Lett.* **119**, 3–11 (2019)