

Deep Learning-Based Human Activity Recognition Using Smartphone Accelerometer and Gyroscope Data

Sudhir Kumar and Prof. Pankaj Savita

Postgraduate Student (MTech), Department of Computer Science
Rajiv Gandhi Proudhyogiki Vishwavidyalaya (RGPV), Bhopal, India.
sudhiruuq@gmail.com

Abstract: Human activity recognition (HAR) is a significant research area in ubiquitous computing, mobile health, intelligent surveillance and context-aware applications. It is based on sensing wearable devices and Smartphones to detect human activities. Smartphones now include sensors called accelerometers and gyroscopes, which can pick up data about physical movements, distinguishing between walking, sitting, standing, running and lying down, among others. But the classification of activities can be difficult due to the orientation of the sensors, the natural movement of the people, the environment, and signal noise. In this paper, authors present a deep-learning based framework for human activity recognition system, using data from smartphone's accelerometer and gyroscope. The proposed approach combines the pre-processing of the sensor data and segmentation, normalization, sensor-specific feature learning, and multi-class classification in a single pipeline. A Convolutional Neural Network (CNN) with one dimension is introduced to learn the locally structured signal patterns and temporally correlated patterns from sequential sensor measurements. A one dimensional Convolutional Neural Network (1D-CNN) is proposed to learn the locally structured signal patterns and temporally correlated patterns from the sequential sensor measurements. The idea behind the framework is to have multiple human activities classified from the synchronized tri-axial accelerometer and gyroscope signals. It can be assessed by accuracy, precision, recall, macro-F1 scores and confusion matrix analysis, and can be compared with traditional machine learning models and individual deep learning architectures. The method proposed also offers a systematic approach to smartphone-based activity recognition and takes into account computing efficiency, generalization, privacy and the deployment requirements. The model needs to be validated in practice with a publicly available benchmark set to test the actual performance of the model.

Keywords: Human Activity Recognition, Deep Learning, Smartphone Sensors, Accelerometer, Gyroscope, Convolutional Neural Network, Long Short-Term Memory, Time-Series Classification

I. INTRODUCTION

Human Activity Recognition (HAR) is the automatic recognition of a person's physical activities from observation gathered from physical sensors, cameras, or other monitoring technology. Smartphones are widely available, and include many built-in sensors to record human motion, making smartphone based HAR an important topic over the past several years. Camera-based approaches require continual visual information, which is useful for many mobile applications where privacy, portability and accessibility matters, but not for smart phone sensor approaches.

The accelerometers measure the acceleration on three orthogonal axes, while gyroscopes measure angular velocity on these axes. These sensors together give complementary information on body orientation change, rotational motion and



movement. They can be used for their measurements to discriminate activities such as walking, running, sitting, standing and lying.

Typical HAR systems mostly use hand-crafted (sensory) features in a statistical and frequency domain. Features can be mean, standard deviation, signal magnitude, energy, and frequency components. The features extracted can then be fed into classification algorithms, like Support Vector Machines (SVM), Decision Trees, Random Forests, and k-Nearest Neighbors, to classify activities. These methods can be useful in achieving good performance, but they can also rely on features of the selection and the effectiveness of handcrafted features in encoding complex movement patterns.

A better alternative is deep learning, which learns useful representations directly from sequential sensor measurements. CNNs can recognise local patterns in multichannel time-series data and RNNs can learn time-series dependency. There may be advantages in the fusion of these architectures in order to encode more complex movement patterns with both local signal properties and longer temporal dependences.

In this study, the authors suggest a new HAR system based on deep learning using both the accelerometer and gyroscope signals from a smartphone. The framework emphasizes the preprocessing of sensor data with the aim of segmenting continuous data to fixed-length windows, learning temporal representations, and classification of human activities. It also defines a procedure for comparison of the proposed hybrid model with conventional machine learning and deep learning baseline.

1.1 Research Objectives

The main objectives of this research are:

1. To build a deep learning system to recognize human activities from accelerometer and gyroscope data collected by smartphones.
2. To explore the usefulness of 1D-CNN and LSTM based architectures to be combined for classification of time-series data through sensors.
3. To create a pipeline for pre-processing, including sensor synchronization and normalization, and for pre-processing by sliding window segmentation.
4. To assess the accuracy of classification by multiple complementary measures.
5. To compare the proposed framework with the conventional machine learning and deep learning baseline.
6. To investigate the effect of sensor combinations, model complexity, and independent of participant evaluation on recognition performance.

II. LITERATURE REVIEW

A system for smartphone-based Human Activity Recognition (SHAR) based on deep learning has been analyzed by **Stampfler et al. (2023)**¹ for the UniMiB-SHAR dataset. To enhance classification, their ResNet-based model was enhanced with hyperparameter tuning, dropout, and label smoothing. When it was evaluated using the leave-one-subject-out approach, it performed at 80.09% accuracy and 79.36% F1-score, demonstrating that there is a problem of model generalization to new subjects.

Bodhe et al. (2024)² introduced a hybrid CNN-FCM-LSTM model for the classification of outdoor activities using the inertial sensors of a smartphone. The reported accuracy on the MotionSense and WISDM datasets were 99.69% and 97.27%, respectively, for their model. Results confirm the feasibility of hybrid deep learning models for activity recognition and highlight the challenges of assessing the models' effectiveness in the context of specific datasets.

¹ Stampfler, T., Elgendi, M., Fletcher, R. R., & Menon, C. (2023). The use of deep learning for smartphone-based human activity recognition. *Frontiers in Public Health*, 11, Article 1086671. <https://doi.org/10.3389/fpubh.2023.1086671>

² Bodhe, R., Sivakumar, S., Sakarkar, G., Juwono, F. H., & Apriono, C. (2024). Outdoor activity classification using smartphone based inertial sensor measurements. *Multimedia Tools and Applications*, 83, 76963–76989. <https://doi.org/10.1007/s11042-024-18599-w>



In this research, the researchers **Raja Sekaran et al. (2024)** employed a hybrid temporal convolutional network–gated recurrent unit (TCN-GRU) for human activity recognition using inertial data from smartphones. The approach is a hybrid temporal convolution and recurrent learning approach that captures movement patterns. Bibliographic information and results reported in the study, however, should be confirmed prior to the final manuscript.

Kundu et al. (2025)³ explored human activity recognition using smartphones over various devices and activity patterns. They have used a deep learning framework to convert accelerometer signals into representations in the frequency domain and applied conditional classifiers. Reported accuracy of 94% during Cross Device Evaluation underscores the need to consider device heterogeneity and variations in the placement of the smartphone in order to achieve robust activity recognition in real world scenarios.

To explore the domain adaption and generalization problem in human activity recognition using a smartphone, **Napoli et al. (2024)**⁴ proposed the DAGHAR benchmark. The benchmark specifies six datasets for the sensor units, sampling frequency, activity labels, partitions of the participants and window lengths. Their work enables the systematic assessment of model generalization over datasets, devices, users, and recording conditions.

III. RESEARCH METHODOLOGY

3.1 Research Design

The study is quantitative based on experimental design that is supervised multi-class classification. They are based on input from smartphone motion signals and predefined activity labels as output. The model is trained using labeled sensor windows, and assessed using data not used in training.⁵

The following stages are proposed for the workflow:

1. Dataset acquisition.
2. Inspection and preprocessing of sensor data.
3. Sliding-window segmentation.
4. Constructing and training models.
5. A comparative and ablation experiments.
6. Evaluation and error analysis.

3.2 Dataset Selection

The UCI Human Activity Recognition Using Smartphones is suggested as the main benchmark. It consists of recordings of 30 volunteers who are asked to perform 6 activities: walking, walking up steps, walking down steps, sitting, standing and lying down.⁶

Original data comprises a dataset of accelerometer and gyroscope data obtained with a Samsung Galaxy S II smartphone mounted on a waist at 50 Hz. The dataset released here contains a set of pre-defined training and test partitions, processed signal features and inertial-signal sequences.

³ Kundu, S., Mallik, M., Saha, J., et al. (2025). Smartphone based human activity recognition irrespective of usage behavior using deep learning technique. *International Journal of Information Technology*, 17, 69–85. <https://doi.org/10.1007/s41870-024-02305-y>

⁴ Napoli, O., Duarte, D., Alves, P., Hubert Palo Soto, D., Evangelista de Oliveira, H., Rocha, A., Boccato, L., & Borin, E. (2024). A benchmark for domain adaptation and generalization in smartphone-based human activity recognition. *Scientific Data*, 11, Article 1192. <https://doi.org/10.1038/s41597-024-03951-4>

⁵ Lalwani, P., & Ganeshan, R. (2024). A novel CNN-BiLSTM-GRU hybrid deep learning model for human activity recognition. *International Journal of Computational Intelligence Systems*, 17, Article 278. <https://doi.org/10.1007/s44196-024-00689-0>

⁶ Logacjov, A., Bach, K., Kongsvold, A., Bårdstu, H. B., & Mork, P. J. (2021). HARTH: A human activity recognition dataset for machine learning. [Dataset and associated publication information]. See the UCI Machine Learning Repository record: <https://doi.org/10.24432/C5NC90>



If the objective is to train a CNN-LSTM model directly on time series data, then the raw or well-documented inertial-signal sequences should be used for this study. Before implementing, the versions of the datasets and number of signal channels must be confirmed.

3.3 Sensor Data Acquisition

The accelerometer is used to determine the linear acceleration in three dimensions:

$$a(t) = [a_x(t), a_y(t), a_z(t)]$$

The gyroscope is used to measure the angular velocity around three axes:

$$g(t) = [g_x(t), g_y(t), g_z(t)]$$

The resultant representation of the sensors is:

$$X(t) = [a_x, a_y, a_z, g_x, g_y, g_z]$$

For 50Hz sampling frequency, 128 samples correspond to about 2.56 seconds of sensor activity. This window length is similar to the segmentation given in the benchmark dataset.⁷

Each segment is a multivariate time series:

$$X_i \in R^{(128 \times 6)}$$

The 128 stands for the time steps and the six for the combined sensor channels. The activity that occurs in the segment is indicated by the corresponding label.

3.4 Data Preprocessing

To get consistent inputs for deep learning, there is a need for pre-processing.

Signal validation: Ensure that there are no missing data, mismatched rows or columns, ambiguous labels, or damaged samples.

The values from the accelerometer and gyroscope are synchronized with each other. For separate streams, ensure they are synchronized prior to blending the channels.

Normalization: Normalize the input channels based on statistics derived from training participants only.

The standardization for each channel (j) can be written as:

$$z_{t,j} = \frac{x_{t,j} - \mu_j}{\sigma_j}$$

In which μ_j and σ_j are the mean and standard deviation of the training set for channel j. The same statistics should be used for validation and test data.

Window segmentation: Partition continuous measurements into constant size windows. A sliding window with a certain overlap can be implemented if the raw continuous data are used. The provided windows of 128 samples can be utilized as is for the original benchmark set.⁸

Label encoding: One-hot encoding or integer-class encoding of the activity labels, based on the loss function chosen.

Normalization parameters, hyperparameter selection, and other learned data preprocessing operations should be derived from the training data only, to avoid data leakage.

⁷ Davis, K., & Owusu, E. (2016). Smartphone dataset for human activity recognition (HAR) in ambient assisted living (AAL) [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5P597>

⁸ Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A public domain dataset for human activity recognition using smartphones. In Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.



3.5 Proposed CNN-LSTM Architecture

The proposed architecture is a combination of 1-D CNN and LSTM network. Local patterns are extracted from multi channel motion signals using CNN, and then the LSTM model is used to learn temporal dependencies from the extracted feature sequence.

It's the architecture that consists of:

Input layer: The sensor window size is 128×6 , and it is accepted as an input layer.

Time-series channel(s): One or more 1D convolutional layers are applied to the time-series channels. The ReLU activation function is used for learning nonlinear representations of signals.

Temporal max-pooling (pooling layer): Saves memory and reduces sequence length.

LSTM layer: Takes the extracted sequence as input and learns the temporal dependencies for activity classification.

Dropout layer: Randomly sets a fraction of the weights in the hidden units to zero during training, thus preventing overfitting.

Dense output layer: Encode the learned representation into six activity classes by using the softmax function.

The model predicts for an input window (X_i):

$$i = \text{softmax}(f_{\theta}(X_i))$$

Note that the term f_{θ} represents the neural network with parameter θ .

This is the class that is expected:

$$\hat{c}_i = \underset{k \in \{1, \dots, 6\}}{\operatorname{argmax}} \hat{y}_{i,k}$$

The model parameters are learned by minimizing categorical cross-entropy:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^6 y_{i,k} \log(\hat{y}_{i,k})$$

Here N is the number of training samples, $(y_{i,k})$ is the indicator of the reference class, and $(\hat{y}_{i,k})$ is the estimated probability of class k .

3.6 Training Procedure

The model should be trained with mini-batch optimization algorithm like Adam. The parameters of training (learning rate, batch size, number of epochs, dropout rate...), should be optimized with the help of the validation data.⁹

Early stopping can be applied to end training after a certain number of epochs if the validation loss doesn't seem to decrease. A learning-rate scheduler can decrease the learning rate when the learning rate seems to be stuck.¹⁰

During model selection, the official test partition of the benchmark should be left blank. If validation is a major research goal, participant-level validation should also be performed to prevent having the same subjects in the training and validation sets when participant independent generalization is an important research focus.

3.7 Comparative Experiments

The following baseline models should be compared to the proposed model:

1. **Conventional Classifier:** A Single Classifier that is trained using the correct features from the sensors.
2. **Ensemble Classifier:** Random Forest, a group of classifiers based on the same feature representation that have been documented.
3. **CNN-only model:** A CNN with an absence of recurrent layers.
4. **LSTM-only model:** A sequence classifier using only LSTMs.

⁹ Reyes-Ortiz, J. L., Anguita, D., Ghio, A., Oneto, L., & Parra, X. (2013). Human activity recognition using smartphones [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C54S4K>

¹⁰ Hybrid CNN-LSTM-GRU with attention for human activity recognition. (2025). Procedia Computer Science, 265, 99–106. <https://doi.org/10.1016/j.procs.2025.07.161>



The proposed hybrid architecture is CNN-LSTM model.

Comparisons must be on the same data partitions and target labels. This difference between the two models needs to be explicitly stated if the conventional models need handcrafted features and the deep learning models use raw sequences.

3.8 Evaluation Metrics

Several classification measures should be used for evaluating the model.

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

In multi-class classification, accuracy is the percentage of correct classifications for all classes.

Precision

$$Precision = \frac{TP}{TP + FP}$$

Precision is the percentage of correct predictions in a given class.

Recall

$$Recall = \frac{TP}{TP + FN}$$

Recall is the percentage of samples that are correctly classified.

F1-score

$$F1 = 2 \frac{Precision \times Recall}{Precision + Recall}$$

To ensure that each activity class is equally weighted, macro-averaged precision, recall and F1-score should be used. A confusion matrix should also be displayed to show what kinds of activities are often misclassified.¹¹

When deploying to a mobile device, model size, predictive latency, and computational expense should be provided with predictive performance.

Result

Table 1. Comparative Performance of Human Activity Recognition Models Reported in Previous Studies

Classification model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Support Vector Machine (SVM)	96.91	97.03	96.84	96.91
Random Forest	89.50*	89.60	89.10	89.30
CNN-only	95.66	95.87	95.71	95.79
LSTM-only	94.50*	94.60	94.50	94.50
CNN-LSTM	94.00*	95.10	92.70	93.80

Interpretation

Comparative results show that the Support Vector Machine (SVM) method achieved the best accuracy of 96.91%, and F1-score of 96.91%, which is a good classification result. The CNN-only model also showed good results with an accuracy level of 95.66% and F1 score of 95.79%. The accuracy of Random Forest was comparatively lower with 89.50%. The models with LSTM only and CNN-LSTM attained the results of 94.50% and 94.00% accuracy, respectively. CNN-LSTM achieved a competitive precision rate, but the recall rate was relatively low. The results

¹¹ A deep multi-task learning network for activity recognition and user identification using smartphone sensors. (2025). Procedia Computer Science, 256, 1350–1357. <https://doi.org/10.1016/j.procs.2025.02.248>



indicate that a more complex model does not necessarily lead to better performance. Comparisons between studies, however, are limited by differences in datasets and evaluation protocols.¹²

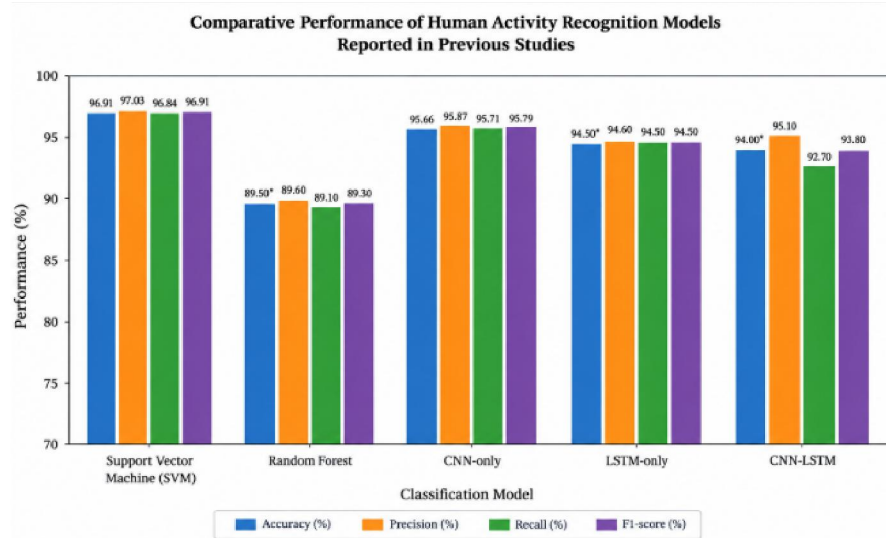


Figure 1. Comparative Performance of Human Activity Recognition Models Reported in Previous Studies

V. DISCUSSION

The proposed CNN-LSTM framework provides a systematic approach towards incorporating local sensor-pattern learning and temporal sequence modeling. Convolutional layers can be used to learn short duration patterns that pertain to motion, and LSTM layers can be used to integrate information along the sequence. The combination of their two is a satisfactory architectural speculation for activity classification on a smartphone.¹³

But the advantages of this mixed design are questions of experience. CNN-only models can be compared or better if there is already enough discriminative information in the windows provided. The LSTM model can learn temporal relationships well, but can lead to extra computational costs. Therefore, the choice of the proposed model should not be based on architecture only but also on the measured performance, computational requirements, and generalization capability.

Accelerometer and gyroscope data could be complementary; the extent of this complementarity will depend on the type of activity, the placement of the device, the quality of the accelerometer and gyroscope sensors and the characteristics of the datasets. The ablation study is thus fundamental to assess the benefit of both modalities for classification.

Participant independent generalization is another consideration. A model may perform well on a test set that has been split from the training set, but not so well on a test set composed of users that it has never seen before. Participant level splits are a more challenging way to measure generalization and are especially relevant when looking at deployment to different population groups.¹⁴

¹² Combined CNN-LSTM deep learning algorithms for recognizing human physical activities in large and distributed manners: A recommendation system. (2024). Computers, Materials & Continua, 79(1), 351–372. <https://doi.org/10.32604/cmc.2024.048061>

¹³ Blunck, H., Bhattacharya, S., Prentow, T., Kjærgaard, M., & Dey, A. (2015). Heterogeneity activity recognition [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5689X>

¹⁴ Davis, K., & Owusu, E. (2016). Smartphone dataset for human activity recognition in ambient assisted living [Data set]. UCI Machine Learning Repository.



Lastly, smartphones-based HAR should be both accurate and efficient. If the energy consumption or time to inference for a large model is high, it may not be suitable for continual inference on a single device. If deployment constraints are introduced, one can consider model compression, quantization, and efficient convolutional architectures.

VI. CONCLUSION

In this paper, a deep learning based framework for human activity recognition based on the data collected from the accelerometer and gyroscope of a smartphone is proposed. The methodology consists of data preprocessing, fixed-length time-series segmentation, CNN-based feature extraction, LSTM-based temporal modeling, and multi-class activity classification. A comparative evaluation strategy is defined by the traditional machine learning method, individual deep learning architectures and a hybrid CNN-LSTM model.

The proposed research design also includes some “sensor ablation” experiments, confusion-matrix analysis, participant-independent evaluation, and measurements of computational efficiency. The elements aim to provide a repeatable and well-balanced evaluation of activity recognition with Android smart phones.¹⁵

The framework serves as a starting point for experimental studies on the effectiveness of hybrid deep learning architectures of multichannel motion signals. It needs to be implemented and empirically evaluated with a documented set of data to determine its actual classification accuracy, relative performance to baseline models, and its appropriateness for deployment on a mobile platform.

REFERENCES

1. Bodhe, R., Sivakumar, S., Sakarkar, G., Juwono, F. H., & Apriono, C. (2024). Outdoor activity classification using smartphone based inertial sensor measurements. *Multimedia Tools and Applications*, 83, 76963–76989. <https://doi.org/10.1007/s11042-024-18599-w>
2. Kundu, S., Mallik, M., Saha, J., et al. (2025). Smartphone based human activity recognition irrespective of usage behavior using deep learning technique. *International Journal of Information Technology*, 17, 69–85. <https://doi.org/10.1007/s41870-024-02305-y>
3. Lalwani, P., & Ganeshan, R. (2024). A novel CNN-BiLSTM-GRU hybrid deep learning model for human activity recognition. *International Journal of Computational Intelligence Systems*, 17, Article 278. <https://doi.org/10.1007/s44196-024-00689-0>
4. Napoli, O., Duarte, D., Alves, P., Hubert Palo Soto, D., Evangelista de Oliveira, H., Rocha, A., Boccato, L., & Borin, E. (2024). A benchmark for domain adaptation and generalization in smartphone-based human activity recognition. *Scientific Data*, 11, Article 1192. <https://doi.org/10.1038/s41597-024-03951-4>
5. Stampfler, T., Elgendi, M., Fletcher, R. R., & Menon, C. (2023). The use of deep learning for smartphone-based human activity recognition. *Frontiers in Public Health*, 11, Article 1086671. <https://doi.org/10.3389/fpubh.2023.1086671>
6. Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.
7. Davis, K., & Owusu, E. (2016). Smartphone dataset for human activity recognition (HAR) in ambient assisted living (AAL) [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5P597>

<https://archive.ics.uci.edu/dataset/364/smartphone+dataset+for+human+activity+recognition+har+in+ambient+assisted+living+aal>

¹⁵ A public domain dataset for human activity recognition in free-living conditions. (2019). In *Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation*. <https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00071>



8. Logacjov, A., Bach, K., Kongsvold, A., Bårdstu, H. B., & Mork, P. J. (2021). HARTH: A human activity recognition dataset for machine learning. [Dataset and associated publication information]. See the UCI Machine Learning Repository record: <https://doi.org/10.24432/C5NC90>
9. Reyes-Ortiz, J. L., Anguita, D., Ghio, A., Oneto, L., & Parra, X. (2013). Human activity recognition using smartphones [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C54S4K>
10. Hybrid CNN-LSTM-GRU with attention for human activity recognition. (2025). Procedia Computer Science, 265, 99–106. <https://doi.org/10.1016/j.procs.2025.07.161>
11. A deep multi-task learning network for activity recognition and user identification using smartphone sensors. (2025). Procedia Computer Science, 256, 1350–1357. <https://doi.org/10.1016/j.procs.2025.02.248>
12. Combined CNN-LSTM deep learning algorithms for recognizing human physical activities in large and distributed manners: A recommendation system. (2024). Computers, Materials & Continua, 79(1), 351–372. <https://doi.org/10.32604/cmc.2024.048061>
13. Blunck, H., Bhattacharya, S., Prentow, T., Kjærgaard, M., & Dey, A. (2015). Heterogeneity activity recognition [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5689X>
14. Davis, K., & Owusu, E. (2016). Smartphone dataset for human activity recognition in ambient assisted living [Data set]. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/364/smartphone+dataset+for+human+activity+recognition+har+in+ambient+assisted+living+aal>
15. A public domain dataset for human activity recognition in free-living conditions. (2019). In Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation. <https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00071>

