# Deep Ensemble Learning for Human Activity Recognition Using Smartphone

Ran Zhu, Zhuoling Xiao*, Mo Cheng, Liang Zhou, Bo Yan, Shuisheng Lin
School of Information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
zhu_ran@outlook.com
{zhuolingxiao,chengmo,zlzl,yanboyu,sslin}@uestc.edu.cn

HongKai Wen
Department of Computer Science University of Warwick
Warwick, UK
hongkai.wen@gmail.com

*Abstract*—**The ubiquity of smartphones and their rich set of on-board sensors have created many exciting new opportunities. One important application is activity recognition based on smartphone inertial sensors, which is a fundamental building block for a variety of scenarios, such as indoor pedestrian tracking, mobile health care and smart cities. Though many approaches have been proposed to address the human activity recognition problem, a number of challenges still present: (i) people's motion modes are very different; (ii) there is very limited amount of training data; (iii) human activities can be arbitrary and complex, and thus hand-crafted feature engineering often fail to work; and finally (iv) the recognition accuracy tends to be limited due to confusing activities. To tackle those challenges, in this paper we propose a human activity recognition framework based on Convolutional Neural Network (CNN) using smartphone-based accelerometer, gyroscope, and magnetometer, which achieves 95.62% accuracy, and also presents a novel ensembles of CNN solving the confusion between certain activities like going upstairs and walking. Extensive experiments have been conducted using 153088 sensory samples from 100 subjects. The results show that the classification accuracy of the generalized model can reach 96.29%.**

***Keywords—Convolutional Neural Network; ensemble learning; human activity recognition; sensor data***

## I. INTRODUCTION

Human activity recognition has drawn much attention in recent years because it can be of significant use in applications including indoor pedestrian tracking [1][2], mobile health care [3], and smart cities [4]. In current research, human activity recognition can be achieved through vision-based or sensor-based approach. Compared to vision-based approach, sensor-based can identify confusing human activities with simple mathematical model by directly measuring the motion from human activities without invasion of personal privacy [5][6].

Previous studies of sensor-based activity recognition often rely on supervised machine learning approaches such as Hidden Markov Model [7], K-Nearest-Neighbor [8], Support Vector Machines (SVM) [9], and so on, using motion data collected from various types and quantities of motion sensors placed in different parts of the body. Since human activities are made of complex features, these traditional methods perform well in recognizing one activity, but badly for others [10]. In comparison, CNN, one deep learning method, has established itself as a powerful technique in speech recognition [11][12] and image recognition domains [13][14] because representations learned by CNNs can capture local dependency and scale invariance of a signal efficiently [15][16][17].

This paper presents a framework based on CNN to recognize human activities, especially those easy to be confused, using motion data collected when participants perform some typical and daily human activities [18][19]. The experiments have demonstrated the improvement of recognition accuracy with the approach proposed in this paper. In summary, the key contributions of this paper are as follows:

- A huge amount of motion data including 153088 data samples from various types of motion sensors and sports scenes with different participants and postures are collected and analyzed.
- A CNN-based recognition method has been proposed to acquire features autonomously, which decreases feature engineering and achieves 95.62% accuracy.
- A novel approach based on ensembles of CNN has been proposed to solve the confusion between going upstairs and walking, which outperforms the single CNN model and achieves 96.29%.

The reminder of this paper is organized as follows: Sec.II describes our dataset. Sec.III introduces the CNN-based human activity recognition model and the framework of ensemble learning. and Sec.IV presents our experimental results and improvements. Sec.V concludes the paper and discusses ideas for future work.

## II. DATASET

Generally speaking, for a multi-class classification problem, a large amount of training data is needed especially with the presence of a high dimension of the feature vector. In addition, rich features from training data can effectively prevent overfitting and result in a robust-behavior model.

### A. Data Collection

Many open source databases focusing on sensor-based activity recognition mainly provide a single accelerometer data. These data were collected from smartphones in participants' trousers pocket or waist at a low sampling rate. To make things worse, these data have poor quantity and unbalanced distribution in various activities, which will not be conductive to the construction of a high classification accuracy model. To improve this situation, the data of this paper come from various sports scenes with different participants and postures and have a balanced distribution.

The data sources of this paper are inertial sensors (i.e., accelerometer, gyroscope and magnetometer) in smartphones at a sampling rate of 50Hz. During the collection process, 100 participants aged 12-51 were asked to complete five daily human activities including: going upstairs(U), going downstairs(D), walking(W), running(R) and standing(S), as shown in Fig. 1. This study takes into account three smartphones-placements: 1) hand-swinging mode (hand holding), 2) pocket mode (trousers pocket), 3) texting mode (holding the phone in the front). Each participant collects sensor data for three times for each behavior and placement. There are a total of 4500 groups of experiments for entire dataset, of which there are 900 groups for each behavior and 1500 groups for each smartphone-placement.

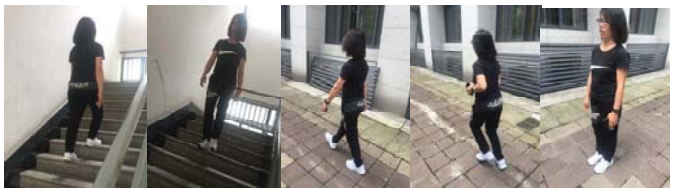

Figure. 1 Daily human activities.

There is no pre-processing of the collected data. Just to meet the format requirement of the CNN model, a sliding window segmentation approach with fixed step size is applied to each sensor data. For every second, a data sample of four seconds, including data from all three sensors, is extracted to form a (24,24,3) matrix.

### B. Training/Testing set division

The current division of training/testing set from related works [5][20][21] can lead to serious overfitting problem. The training data and testing data are divided from a completely shuffled data segmentation group, which means nearly every testing data are from the same person as the training data. This leads to an irrationally high recognition accuracy in existing testing data but poor performance in data from some different individuals. For instance, with this division method, the recognition approach proposed in this paper can reach 99.8% accuracy, which is apparently unrealistic in practical scenarios.

Instead of complete shuffle of all motion samples, this study divides the training set and testing set by individuals. From the 100 participants, the experiment randomly selects 10 participants as the testing set and the remaining 90 participants as the training set (10-fold evaluation). This method takes into account the applicability of the recognition framework to testing data from individuals totally different from the training data and thus examine whether the generalized model can be applied to real world scenarios.

## III. CNN-BASED HUMAN ACTIVITY RECOGNITION

In this section, we describe our CNN-based activity recognition model, as shown in Fig. 2, where modality-specific characteristics are integrated to handle multivariate time series measured at multiple sensors for activities recognition, and then introduces the ensemble system based on this CNN architecture.

### A. CNN-based Model

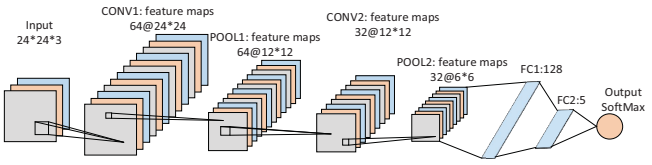


Figure. 2 Structure of CNN-based human activity recognition model. The numbers of the first and second convolutional kernels are 64 and 32 respectively.

The proposed CNN-based model has five kinds of layers: 1) an input layer, as described in Sec. II.A; 2) convolutional layers extracting features from input data; 3) max-pooling layers which reduce the size of extracted features and enhance the robustness of some detected features; 4) fully connected layers integrating all features extracted; 5) an output layer of the softmax function representing a categorical distribution over five different activities. Based on this architecture, the five motion modes are identified with an accuracy of 95.62%.

### B. Ensemble Model

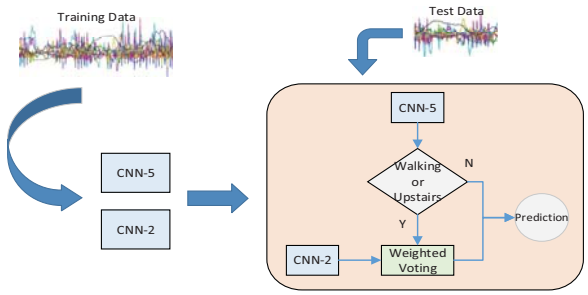


Figure. 3 Ensemble of CNN for human activity recognition – system overview.

In response to the confusing activities and with the aim of improving the robustness of activity recognition system we have developed a novel framework based on Ensembles of CNN, as shown in Fig. 3. This system trains a five-class network (i.e., CNN-5) and a two-class (going upstairs and walking) network (i.e., CNN-2) based on the CNN architecture described above. When assessing the performance of the model on the testing set, two CNN model perform weighted voting to predict unknown activities. It has been shown that such ensemble learning approach can be very beneficial for the confusion between certain activities like going upstairs and walking.

## IV. Experimental Analysis

This section presents a systematical analysis and evaluation of human activities recognition based on phone sensor data. We analyze the impact of smartphone-placement on model performance and investigate the impact of the difference between individuals. Then we compare the model performance via various model hyper-parameters. We also propose and evaluate a novel approach based on ensembles of CNN to tackle the confusion between going upstairs and walking.

### A. Classification Accuracy

In the first experiment, we evaluate activities recognition on the dataset we collected. Our model consists of two convolutional layers and max-pooling layers with a filter size of 7 and pooling kernel size of 2. The top two full connected layers have 128 nodes and 5 nodes respectively. An additional softmax top layer is used to generate the state posterior probability. The architecture of the CNN used has 64 and 32 feature maps in two convolutional layers, as shown in Fig. 2.

TABLE I. Detailed Accuracy By Activity

| Activity | Model Evaluation | | |
|---|---|---|---|
| | *Precision* | *Recall* | *F1-score* |
| U | 0.90 | 0.92 | 0.91 |
| D | 0.95 | 0.95 | 0.95 |
| S | 1.00 | 0.98 | 0.99 |
| R | 0.99 | 0.99 | 0.99 |
| W | 0.93 | 0.93 | 0.93 |
| Avg/all | 0.96 | 0.96 | 0.96 |

Training data from 90 participants are used to train the proposed CNN network. It turns out that the generalized model can reach 95.62% on the testing data of 10 participants. Table I shows the performance of proposed model in terms of precision, recall rate and F1-score for each activity.

To analyze the results in more details, we show the confusion matrix for activities recognition. From Fig. 4, we can see the prediction error mainly comes from the confusion between going upstairs and walking. This is because the motion signals of "going upstairs" and "walking" are highly similar, which will be effectively solved in Sec. IV.C.

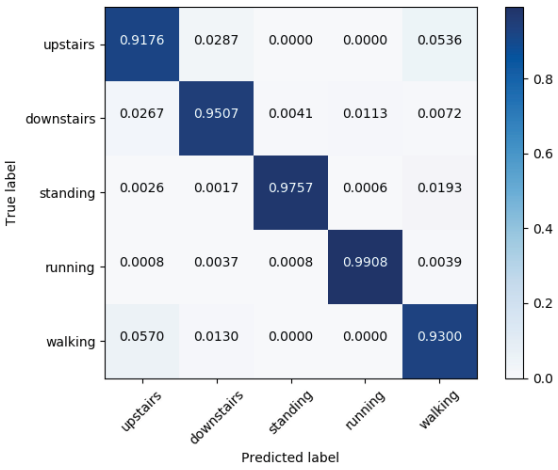


Figure. 4 The normalized confusion matrix

*1)* Different smartphone-placements will affect model performance, even with the same motion mode. In Fig. 5, we calculate the accuracy in four placement modes: (i) hand-swinging mode; (ii) pocket mode; (iii) texting mode; (iv) mix mode where data from all previous three modes are mixed.


Figure. 5 The influence of smartphone-placement on accuracy

From Fig. 5, we can see that the accuracy in pocket mode keeps consistent above 96.5% in all motions while the accuracies in hand-swinging and texting mode behave less consistently. The reason lies in the fact the motion of phone in the pocket is restricted, thus regular and easy to predict.

*2)* Characteristics of motions, such as range and speed corresponding to age, gender, habits, etc., also have impact on the performance of model. For example, comparing the young, the elder's step frequency is lower and step length is shorter; men often step over two steps of staircase; adults have longer steps than children. We investigate the impact of individual information in Table II. Each individual has different classification accuracy due to its own biometric characteristics. Even for the same activity, different individuals have distinct motion mode. Therefore, more data from different individuals would be of great use to gain better activities recognition accuracy.

TABLE II. Classification Accuracy Comparison Among Various Smartphone-placements

| Partici-pants | Classification Accuracy (%) | | | |
|---|---|---|---|---|
| | *Hand-swinging mode* | *Pocket mode* | *Texting mode* | *Mix mode* |
| P_1 | 84.25 | 98.79 | 95.42 | 92.85 |
| P_2 | 91.15 | 97.25 | 95.93 | 94.97 |
| P_3 | 94.86 | 99.77 | 94.32 | 96.34 |
| P_4 | 99.38 | 98.91 | 98.21 | 98.85 |
| P_5 | 98.15 | 99.33 | 90.58 | 96.16 |
| P_6 | 94.17 | 97.29 | 95.84 | 95.81 |
| P_7 | 92.51 | 97.38 | 97.84 | 95.86 |
| P_8 | 98.83 | 99.10 | 94.77 | 97.64 |
| P_9 | 95.89 | 98.13 | 93.72 | 95.95 |
| P_10 | 93.38 | 97.33 | 96.72 | 95.83 |

### B. Parameters of mode

In the following, we observe the test accuracy changes as filter size, pooling size and dropout respectively while keeping the other parameters as the best settings.

*1) Filter size:* CNN uses filters to extract features from input data in convolutional layers. In order to evaluate the impact of the filter size on model performance, we have tested the proposed CNN model with filter size of 3,5,7,9. Fig. 6 shows each activity recognition accuracy for different

filter size. It can be seen that the is achieved when the filter size is set to be 7.

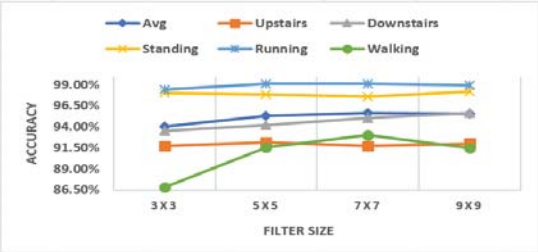


Figure. 6 The influence of filter size on accuracy

*2) Pooling size:* In the following, we evaluate the sensitivity of pooling size of the CNN configuration with size from 1 to 5, where 1 corresponds to the case of no max-pooling. Fig. 7 shows that max-pooling achieves its best performance with size of 2.

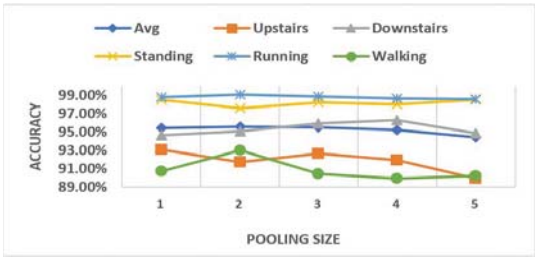


Figure. 7 The influence of pooling size on accuracy.

*3) Dropout:* In the Deep Neural Network(DNN), over-fitting of models (weak generalization) has always been a tricky problem. As a simple and effective solution widely used in DNN, Dropout randomly discards units to improve the generalization of the model. Moreover, dropping out is done independently for each hidden unit for each training case. Thus, dropout can be regarded as a method of making bagging practical for neural networks. Dropout has a tune-able hyper-parameter *p* representing the probability of retaining a hidden unit in the network. Fig. 8 shows the impact of varying *p* on the generalized model(network cannot converge when *p*=0.1). The dropout is applied in the top fully-connected layer of our model.


Figure. 8 The influence of dropout on accuracy.

## C. Ensembles of CNN

From Sec. III.A, we can see that the prediction error mainly comes from the confusion between going upstairs and walking. In order to solve this problem, we use the Ensemble model introduced in Sec.II.B to predict unknown activity during testing. The detailed process is described in Algorithm 1 below.

Two CNN models are trained in the ensemble model. One is used to identify five activities as described in Sec.II.A (i.e., CNN-5),the other is a two-class classification that only

**Algorithm 1** Framework of Ensembles of CNN

**Input:** testing data:$X_{i=1}^{MaxSample}$
**Output:** human activity (upstairs,downstairs,running,walking or standing)
**while** $i$ to $MaxSample$ **do**
  CNN-5 network gives a predicted activity
  **if** activity is upstairs or walking **then**
    normalize the probability of upstairs and walking $P_u^1, P_w^1$
    define weights:$\alpha_1 = P_u^1, \beta_1 = P_w^1$
    the CNN-2 network gives another prediction $P_u^2, P_w^2$
    define weights:$\alpha_2 = P_u^2, \beta_2 = P_w^2$
    **if** activity is upstairs **then**
      $P_u = \beta_1 \times P_u^1 + \alpha_1 \times P_u^2, P_w = \beta_1 \times P_w^1 + \alpha_1 \times P_w^2$
    **else**
      $P_u = \beta_2 \times P_u^1 + \alpha_2 \times P_u^2, P_w = \beta_2 \times P_w^1 + \alpha_2 \times P_w^2$
    **end if**
    **return** the predicted activity from $max(P_u, P_w)$
  **else**
    **return** the predicted activity from CNN-5
  **end if**
**end while**

classifies the confusing "going upstairs" and "walking" (i.e., CNN-2). Both models are trained on the network architecture proposed in this paper. When assessing the performance of the model on the testing set, two CNN models are integrated as ensembles of CNN. Specifically, test data are feed into CNN-5 to predict the motion mode. If the predicted result is going downstairs, running or standing, this output is directly used as the final prediction result. If the predicted result is going upstairs or walking, the current test data are then feed into the CNN-2 for prediction. Combining two predictions to make a weighted voting, which is used as the final prediction result of going upstairs and walking. The weights (i.e., $\alpha$, $\beta$) of two CNN network are selected from the recognition accuracy of going upstairs and walking, which can be regarded as the prior probability.

When judging confusing activities, two models are combined to make decisions, which can effectively solve the over-fitting problem caused by relying on a single model. Fig. 9 shows the confusion matrix for human recognition accuracy of ensembles of CNN. The recognition accuracy reaches as high as 96.29%. Comparing to the single CNN model, the recognition rate of going upstairs and walking has increased, which indicates that the proposed ensemble model is quite a promising classification method in the confusing activities.

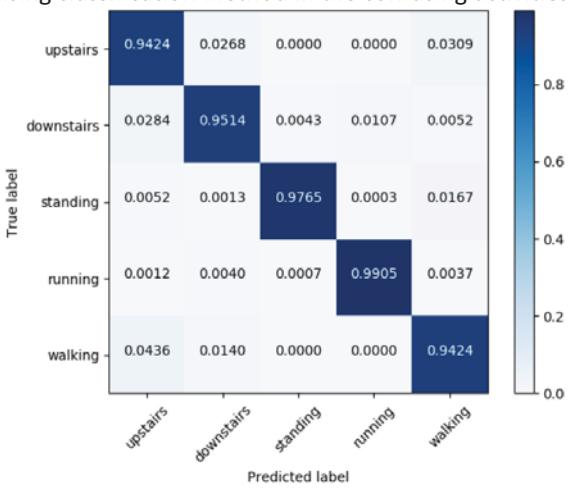


Figure. 9 The normalized confusion matrix of ensembles of CNN

## V. Conclusion

This paper has proposed a CNN-based human activity recognition with the three-axis signal of accelerometer, gyroscope, and magnetometer embedded in smartphones. We have analyzed and compared the performance of the proposed model with five daily activities and three different smartphone-placements. In order to further improve the recognition accuracy, this paper also proposes a new ensembles of CNN approach. Experiment results have shown the effectiveness and efficiency of the proposed method in real world scenarios.

There are several potential research topics for future work. Experiments with more activity categories and smartphone-placements will be conducted to verify the robustness and practicality of proposed model in the real world. Another topic would be the parameter optimization of CNN and ensemble model.

## Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. The authors also are grateful to each participant involved in the data collection.

## References

[1] M. Elhoushi, J. Georgy, A. Noureldin, and M. J. Korenberg, "Motion Mode Recognition for Indoor Pedestrian Navigation Using Portable Devices," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 1, pp. 208–221, Jan. 2016.

[2] W. Kang and Y. Han, "SmartPDR: Smartphone-Based Pedestrian Dead Reckoning for Indoor Localization," *IEEE Sensors Journal*, vol. 15, no. 5, pp. 2906–2916, May 2015.

[3] P. Wu, H.-K. Peng, J. Zhu, and Y. Zhang, "SensCare: Semi-automatic Activity Summarization System for Elderly Care," in *Mobile Computing, Applications, and Services*, Springer, Berlin, Heidelberg, 2011, pp. 1–19.

[4] G. Sagl, B. Resch, and T. Blaschke, "Contextual Sensing: Integrating Contextual Information with Human and Technical Geo-Sensor Information for Smart Cities," *Sensors (Basel)*, vol. 15, no. 7, pp. 17013–17035, Jul. 2015.

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

[6] M. Elhoushi, J. Georgy, A. Noureldin, and M. J. Korenberg, "A Survey on Approaches of Motion Mode Recognition Using Sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1662–1686, Jul. 2017.

[7] Y.-S. Lee and S.-B. Cho, "Activity Recognition Using Hierarchical Hidden Markov Models on a Smartphone with 3D Accelerometer," in *Proceedings of the 6th International Conference on Hybrid Artificial Intelligent Systems - Volume Part I*, Berlin, Heidelberg, 2011, pp. 460–467.

[8] F. Foerster, M. Smeja, and J. Fahrenberg, "Detection of posture and motion by accelerometry: A validation study in ambulatory monitoring," *Computers in Human Behavior - COMPUT HUM BEHAV*, vol. 15, pp. 571–583, Sep. 1999.

[9] S. Wang, J. Yang, N. Chen, X. Chen, and Q. Zhang, "Human activity recognition with user-free accelerometers in the sensor networks," in *2005 International Conference on Neural Networks and Brain*, 2005, vol. 2, pp. 1212–1217.

[10] T. Huynh and B. 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, New York, NY, USA, 2005, pp. 159–163.

[11] D. Palaz, M. Magimai-Doss, R. Collobert, in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH(International Speech and Communication Association, 2015), vol. 2015–January, pp. 11–15.

[12] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[13] G. E. Hinton, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[14] X. S. Wei, C. W. Xie, and J. Wu, "Mask-CNN: Localizing Parts and Selecting Descriptors for Fine-Grained Image Recognition," 2016.

[15] M. Zeng *et al.*, "Convolutional Neural Networks for human activity recognition using mobile sensors," *6th International Conference on Mobile Computing, Applications and Services*, Austin, TX, 2014, pp. 197-205.

[16] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 381–388.

[17] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition," in *Proceedings of the 24th International Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp. 3995–4001.

[18] T. Zebin, P. J. Scully, and K. B. Ozanyan, "Human activity recognition with inertial sensors using a deep learning approach," in *2016 IEEE SENSORS*, 2016, pp. 1–3.

[19] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors (Basel)*, vol. 14, no. 6, pp. 10146–10176, Jun. 2014.

[20] F. Gu, K. Khoshelham and S. Valaee, "Locomotion activity recognition: A deep learning approach," 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications

[21] I. Bisio, A. Delfino, F. Lavagetto and A. Sciarrone, "Enabling IoT for In-Home Rehabilitation: Accelerometer Signals Classification Methods for Activity and Movement Recognition," in IEEE Internet of Things Journal, vol. 4, no. 1, pp. 135-146, Feb. 2017.