A Hybrid Deep Neural Network for Human Activity Recognition based on IoT Sensors

Zakaria BENHAILI*, Youssef BALOUKI, Lahcen MOUMOUN
Hassan First University of Settat, Faculty of Sciences and Techniques, Mathematics
Computer Science and Engineering Sciences Laboratory (MISI), 26000 Settat, Morocco

Abstract—**Internet of things (IOT) sensors, has received a lot of interest in recent years due to the rise of application demands in domains like ubiquitous and context-aware computing, activity surveillance, ambient assistive living and more specifically in Human activity recognition. The recent development in deep learning allows to extract high-level features automatically, and eliminates the reliance on traditional machine learning techniques, which depended heavily on hand crafted features. In this paper, we introduce a network that can identify a variety of everyday human actions that can be carried out in a smart home environment, by using raw signals generated from Internet of Thing's motion sensors. We design our architecture basing on a combination of convolutional neural network (CNN) and Gated recurrent unit (GRU) layers. The CNN is first deployed to extract local and scale-invariance features, then the GRU layers are used to extract sequential temporal dependencies. We tested our model called (CNGRU) on three public datasets. It achieves an accuracy better or comparable to existing state of the art models.**

Keywords—IoT; deep learning; CNN; GRU; CNGRU; human activity recognition

I. Introduction

The Internet of Things (IoT) is a technology that has a lot of potential, it presents a platform where sensors and devices can communicate seamlessly within a smart environment. Each year, the number of IoT supporting devices increases; sectors such as transport, healthcare, security, smart cities, education, agriculture, and many others have already benefited from its development. This will result in a generation of applications capable of completing complex sensing and recognition tasks to support a new world of human-things interactions. The recognition of human activities is a field that presents an interaction between computers and humans which has been promoted recently by the expansion of artificial intelligence. This progress has reached a stage that has allowed it to integrate several fields, to the point that we find its applications in everyday life. In the field of security by making surveillance more intelligent [1]. In smart homes by improving the security and monitoring the health condition of the residents [2], and increasing the degree of independence and quality of life, especially for the elderly [3]. HAR is present as well in the field of healthcare, by the deploy of a combination of one or more techniques of recognition that notifies the medical staff once an intervention is necessary [4].

This widespread availability is owing to significant efforts to reduce the size of the electronic components and create sensors that can be included in smartphones, smart watches, and other wearable internet of things devices.

Depending on the type of sensors used, we categorize activity recognition into vision-based or sensor-based recognition. The first category deploys cameras to obtain images and videos and use it to detect and classify activities, however it faces challenges as image variation, object deformation, mobility constraints imposed by visual sensors, besides other problems related to power consumption and privacy. On the other hand, sensor based recognition which is based on acceleration sensors, gyroscope sensors, geomagnetic sensors and others, are simple to use and generate relatively accurate and reliable data. The classic approaches require a lot of data pre-processing and domain knowledge for feature engineering, which will be necessary at every change of dataset, and limit the generalization of the model.

Recently, Deep learning has achieved good performances and it has accumulated successes in image, speech, and natural language processing, and today it is introduced in human activity recognition, to profit from its capacity to learn complex movements, by abstracting features automatically from raw data without being handcrafted. Deep learning's layer-by-layer structure enables it to progressively learn features from simple to complex, which is effective in the analyse of multimodal sensory data. The various architectures of deep learning are capable of encoding these features from diverse perspectives. For example, CNNs can capture local multimodal sensory connections, where RNNs can extract each temporal dependency and learn information incrementally across multiple time intervals.

We achieve sensor-based HAR through four major steps, the first is data collection, followed by data segmentation, then feature selection or extracting features, and last the classification of the activity. Most of the previous works in HAR are based in their approaches on a manual feature engineering, which already requires an expert knowledge, the method proposed in this article does not require any design or creation of features, it exploits directly the data generated by the accelerometer and gyroscope. This is the key contributions of our work:

We propose CNGRU, an end to end Network for HAR capable of automatically extracting and learning features from raw data without pre-processing.

We deploy a combination of two types of neural networks: convolutional and gated recurrent units.

*Corresponding Author.

The network permits to recognize various activities and gestures, recorded using different types and combinations of sensors. The experience on three most widely used open datasets, proves that we reach comparable, or better results than previous methods, which demonstrates the generalization capability of the model.

We organize our paper as follows: Section II reviews related works of human activity recognition. In Section III, we propose our model for HAR. Section IV presents and examines the experimental results. And last in Section V, we draw out our conclusion.

II. Review of Literature

Prior studies on human activity recognition have been conducted utilizing open-access datasets available on the internet. Mainly the UCI HAR dataset was exploited alone or with other datasets like Opportunity[5], WISDM V1.1 [6], PAMAP2[7]. Consequently, this availability of data facilitated the design and evaluation of the activity recognition approaches based on motion sensors. Whereas some works are based on the investigation of feature selection in order to achieve higher accuracies, others attempted to avoid this design and engineering task by utilizing the capacity of deep learning models. Convolution neural network is the most common model in the approaches proposed in the literature, researchers exploit its ability to capture local connections, as well as the recurrent neural network and its variants capable of capturing temporal dependencies between signal readings. And in other works those two networks are fused or cascaded to learn the most important features.

The authors in [8] have proposed a hybrid architecture, which combines LSTM and CNN. After preprocessing data, they fed it to two LSTM layers for temporal feature extraction, while the spatial features were extracted by two other convolution layers.

Deep et al [8] used the UCI HAR dataset to test their model composed of CNN followed by an LSTM network. They have achieved better recognition scores compared to simple LSTM architecture. On the same dataset, Hernández et al [9] presented the idea of using bidirectional LSTM networks, to recognize the six activities of this dataset. They attain a high recognition performance, except for static activities: laying and standing. Ahmad et al [10] introduced a new approach based on an architecture called multi-head CNN to recognize human activities, The fundamental idea is to employ three CNNs, each supplied by three streams: overall acceleration, body acceleration, and body gyroscope. The results of these parallel CNNs are then integrated and transmitted to another LSTM layer, resulting in a high recognition accuracy. Sikder et al [11] used frequency's and power's features of raw activity signals, and they feed each stream of them to a CNN channel, the result is concatenated for classification, finally an accuracy of 95.25% is obtained on UCI HAR.

Other works have explored the effect of deepness on recognition, the authors in [12] proposed an HDL: Hierarchical Deep Learning Model capable of recognizing activities with an accuracy of 97.95 % on the UCI HAR dataset, their model is composed of several BLSTM layers, which are used to capture information from the original data, CNN layers came afterwards to learn features from the output of the last BLSTM layer, and classification is obtained in the end using a Softmax layer. Xu et al [13] have proposed InnoHAR, a network which, takes advantage of Inception-like modules to make feature extraction, combined with GRU for sequential temporal dependencies extraction, Gao et al [14] proposed a method called DanHAR designed for challenging scenarios where there are multi-modal sensors. Their model uses a hybrid approach that fuses information using a dual-attention mechanism with CNN, which improved the ability to capture temporal and spatial patterns, resulting in a better performance while keeping the number of parameters small.

Teng et al [15] proposed a network based on convolutional neurons with a local loss after each CNN module, they compared a baseline model containing three CNN layers and one Fully Connected layer, with the same model having the first time similarity matching loss, a second time cross-entropy loss and the third time a combination between the two previous losses. Sena et al [16] divided the data into several inputs according to the type of sensor, then for each of them they built a deep CNN to extract temporal scales and features. Their method employs a DCNN, which is made up of two convolutional layers followed by a Maxpooling layer. In the end all the DCNN ensemble are merged using late fusion method. A different approach used by Bokhari et al [17], who exploited Channel State Information (CSI) to estimate and classify activities performed in an indoor environment using a deep Gated Recurrent network (DGRU).

III. Materials and Methods

Even if the conventional HAR methods have reached good scores, their reliance on handcrafted and their need to heavy data preprocessing methods limits their scalability to other datasets. Convolutional Neural Networks, Recurrent Neural Networks, and their combinations enabled for the creation of shallow and deep models in an end-to-end technique, resulting in high recognition scores in complicated task solving.

A. Convolutional Neural Network

This architecture is based on the convolutional layer, which performs the convolution operation on the input by multiplying it by the weights of a filter and then summing it to find the value corresponding to that position. The output of this linear operation is injected into a nonlinear activation function g and can be expressed as:

$$\mathrm{a_{i,j}} = \mathrm{g}(\textstyle\sum_{\mathrm{m=1}}^{\mathrm{L}} \sum_{\mathrm{n=1}}^{\mathrm{k}} \mathrm{W_{m,n}}.\mathrm{x_{i+m,j+n}} + \mathrm{b} \quad (1)$$

Where, $x_{i+m,j+n}$ is the activation of the higher neurons linked to the neuron (i, j), $W_{m,n}$ is a matrix with a size of L.K and containing the weights of the convolution filter, and b is the bias [18].

the convolutional network is a type of neural network which is mainly constituted of convolutional layer, but other layers like Maxpooling and Fully connected layers can also be present and stacked one after another to add depth and build an hierarchical network [19]. For feature extraction the convolutional layer and the Maxpooling layer can be deployed together as a single part, whereas the second part which has the role of classifying the resulting feature vectors is dedicated to the Fully connected layer, and it typically contains a number of nodes equal to the number of classes [20].

B. Gated Recurrent Unit

Conventional Recurrent Neural Network suffers from the issue of vanishing gradient when the network cannot transmit convenient gradient information back to the input layers, making the optimization difficult and prohibiting them from learning long term dependencies [21]. Short-term memory units [22] (LSTMs) and recently gated recurrent units (GRUs) [23] are two modifications of RNN designed to solve this problem. Where LSTM have the state of the art performance, it needs more inference time and processing. In our work we studied using GRUs, which are simpler than LSTM, have fewer parameters, and give a good trade-off between speed and performance [24]. The recurrent transition of GRU are obtained by:

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \tag{2}$$

$$r_t = \sigma(W_r[h_{t-1}, x_t]) \tag{3}$$

$$\widehat{h_t} = tanh(W[r_t \odot h_{t-1}, x_t]) \tag{4}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widehat{h_t} \tag{5}$$

Where $\{W_z, W_r, W\}$ designate the recurrent weights. $\boldsymbol{h_t}, \widehat{\boldsymbol{h_t}}$ are hidden states. σ denotes sigmoid function. And $\odot$ component-wise or Hadamard multiplication. z_t is the update gate and r_t is the reset gate.

The update gate z_t determines the degree of similarity between the hidden state $\boldsymbol{h_t}$ and the new hidden state $\widehat{h_t}$ and if the update is performed. The reset gate r_t is used to regulate how much of the prior state we wish to retain. if r_t, is equal to 1 it means that we keep information from the previous state, otherwise, this latter state is neglected.

C. Overview

Activity recognition is considered a classification problem, the signals extracted from motion sensors are time series data, in our approach Convolutional neural networks are used on these raw signals to avoid the requirement for feature engineering and to take advantage of local dependency and correlation between signal measurements [25]. The extraction of temporal features is the next stage. Because Simple RNN has a vanishing gradient problem, we opted to run signals through three consecutive GRU layers. We chose GRU because of its ability to deal with extended sequences and its time efficiency [26].

D. Proposed Architecture

Our architecture is inspired by LeNet 5 [27], it benefits from its simplicity and straightforwardness, the original network uses a pair of convolutional and average pooling layers, followed by a flattening layer, two fully-connected layers and last a Softmax classifier. It was initially designed for handwriting and printed characters' recognition. We made the following change: we divided the layers into two groups: convolution layers and dense layers. We reduced the number of units in the last layer, replaced two-dimensional convolution and two-dimensional average pooling with one-dimensional convolution and one-dimensional average pooling, and finally injected what we called a GRU block in between.

Different GRU block configurations were tested and evaluated in order to select the one with the highest accuracy. TABLE I contains the configuration of each injected block.

The first GRU block contains only one layer with 100 units, then a dropout layer of 20**%,** this architecture has the advantage of being simple, and light, its training was fast, but unfortunately it cannot recognize well all the activities. To solve this problem, we added another layer to the first one, and we kept the number of nodes for each of them at 100 nodes, then we preserved the 20% dropout after each layer, the results showed an increase in accuracy of more than 2%. In the third architecture, we wanted to test the effect of deepness on the initial network, in fact in GRU block 3 we increased the number of nodes in the first two layers to 128 nodes, then we added a third one with 64 nodes, while using Batch Normalization instead of the dropout after each layer, the experimental results for each network (CNN + GRU bloc) is presented in TABLE II. We find that the third network has the best accuracy, it means that adding three GRU layers, gives the model the capability to better extract the sequential temporal dependencies, while batch normalization layers served better in reducing Overfitting than dropout. This improvement in accuracy is also accompanied by a reduction in the number of parameters from 455,566 to 427,950. Fig. 1 illustrates the final architecture, Fig. 2 presents the diagram of the proposed solution in this paper.

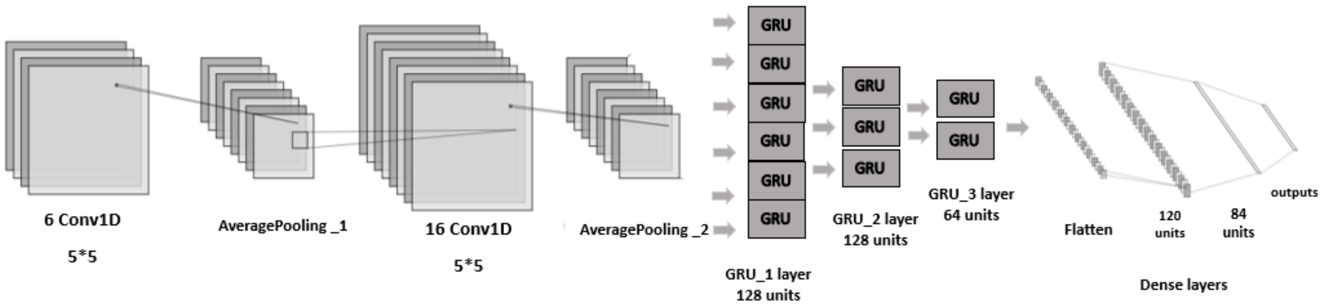


Fig. 1. The Proposed Network.

TABLE I. DEFINITION OF GRU BLOCKS

GRU block	Layers
Architecture 1	1 GRU layer (100 units) + 20% dropout.
Architecture 2	2 GRU layer (100 units) +20% dropout after each layer.
Architecture 3	2 GRU layers (128 units) +1 GRU layer (64 units) +batch Normalization after each layer.

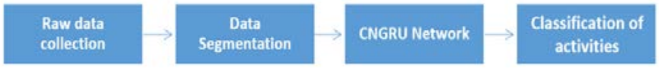


Fig. 2. Steps to Recognize Activities from Raw Data.

Several recent studies have demonstrated that a one-dimensional convolutional neural network is well suited for the analysis and extraction of discriminative features from data time series generated by sensors such as accelerometers and gyroscopes, and that it has the ability to learn an internal representation of data sequences [28]. Average pooling is often used instead of Maxpooling since it can extract features more smoothly. As mentioned earlier the 128-128-64 combination of GRU layers nodes, proved to outperform the 100-100 and 100 node combinations used in the other two architectures. We used the Adam optimizer with a learning rate fixed at 0.001, tested batch sizes of 32, 64, and 128, and finally chose 64 since it produced the best results. We trained the model for 1000 epochs and we used early stopping. TABLE III contains a definition of each layer and the parameters used in this our network.

TABLE II. TEST ACCURACY, TIME PER EPOCH, AND THE NUMBER OF PARAMETERS FOR UCI-HAR

Network	Accuracy	Time	Parameters
cnn + architecture 1	94.87 %	1s	68,866
cnn + architecture 2	96.20 %	7s	455,566
cnn+ architecture 3	96.77 %	17s	427,950

TABLE III. DEFINITION OF EACH LAYER AND THE PARAMETERS USED IN THIS OUR NETWORK

Layer	Parameters
convolution_1	Kernel=5, stride=1, filters=6, activation= tanh
average pooling_1	-
convolution_2	Kernel=5, stride=1, filters=16, activation= tanh
average pooling 2	-
gru_1	128 units + batch normalization_1
gru_2	128 units + batch normalization_2
gru_3	64 units + batch normalization_3
Flatten layer	-
dense layer_1	120 units , activation= tanh
dense layer_2	84 units, activation = tanh
dense layer_3	6 units, activation = softmax

IV. RESULTS AND DISCUSSION

A. Evaluation Methodology

We ran tests on three publicly available datasets. Here is a short description of each one:

UCI HAR [29]: This dataset was gathered by 30 users aged 19-48 who wore smartphones around their waists while performing a series of activities. The information gathered is classified into five activity classes, three of which are static activities (standing, sitting, and lying) and the others are dynamic (walking, going upstairs, and going downstairs). The accelerometer and gyroscope embedded in the phone (Samsung Galaxy SII) enabled the measurement of three-axial linear acceleration as well as three-axial angular velocity.

WISDM V1.1 [6]: is a dataset collected by using only one IMU (accelerometer), the chosen activities were selected carefully, depending on their performance regularity in daily life. Those activities are Walking, Jogging, Upstairs, Downstairs, Sitting, Standing. This dataset has approximately the same activities as UCI, Fig. 3 contains a description of its activities.

SKODA [30]: this dataset has been recorded using only one type of IMU, in a manufacturing scenario and covers the problem of recognizing the activities of assembly-line workers in a car production environment. A worker carried a number of sensors while performing manual quality checks for the correct assembly of parts in newly built cars. 10 resulting hand movements are considered. TABLE IV contains various recording information about all the datasets used in this work.

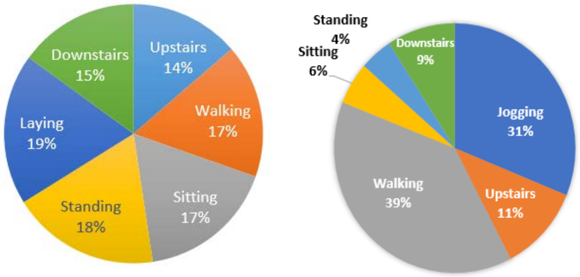


Fig. 3. Activity Description of UCI in the Left and WISDM v1.1 in the Right.

TABLE IV. DEFINITION OF THE CHARACTERISTICS OF THE DATASETS

dataset	activities	subject	place	sampling rate	samples
WISDM	6	36	thigh	20 hz	1.098.207
UCI HAR	6	30	waist	50 hz	10.298
SKODA	10	1	arms	98 hz	~701.440

B. Performance Measure

When we were evaluating our model, we noticed the lack of an evaluation standard. Various evaluation metrics are used to measure and compare the human activity recognition performance. The main ones are accuracy, recall, F-measure, Area under the Curve (AUC). Where some works use F-measure, other authors prefer accuracy. This diversity tends to make finding the state of the art model difficult. The diversity of validation protocol should also be taken into consideration when dividing data into training/test/validation since it impacts the recognition results and comparison. The parameters we used to compare the model's performance are defined as follows:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (6)$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (7)$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (8)$$

$$F - Measure = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision+Recall}} \quad (9)$$

(Where, T: True, P: Positives, F: False, N: Negatives). We use also Confusion Matrix, to have a summarized view about the performance of the classification, and to see the errors being made its type, and where the confusion occurs.

C. *Results*

We ran several tests on two other datasets to evaluate the performance and validate the efficiency of the proposed method. We used WISDM V1.1 and SKODA, the first one contains activities similar to UCI, while the second one contains a different type of gesture. We present the detailed results for UCI which was exploited in the design and tuning of our model, then we compare the results obtained with WISDM V1.1 at the level of each activity, and last we evaluate our approach on SKODA.

UCI HAR's signals were pre-processed by filtering noise then sampling in a fixed-width sliding windows of 2.56 sec and 50% overlap, again we chose to take 21 subjects for training and 9 for testing. We fed our network with data in a specific shape. Accuracy and loss over each epoch are used for evaluation. We trained the model through 1000 epochs, then we used early stopping technique to end training when the validation accuracy stops increasing. All the datasets were uploaded to Google drive, and we used for the experiment Google Colaboratory. Our model achieved an accuracy of 96.77 %. As shown in TABLE V, this value is comparable to the state of the art, and other works that use handcrafted features, classical machine learning algorithms, unsupervised machine learning algorithms or models composed of a combination of previous methods.

To show the correspondence between the predicted labels and the true ones, we used the confusion matrix illustrated in Fig. 4. It shows that we achieve good recognition for all activities. We see that the static action LAYING is easily identified, with an accuracy of 100% and it's unconfused with any other activity. The dynamic activities WALKING_UP and WALKING are also well recognized, but for STANDING and SITTING their accuracies are relatively smaller and consequently the total score of the model is reduced, furthermore we remark that they are often confused with each other's, this could be explained by the similarity of the signals of those two classes.

The second experiment was on WISDM V1.1 using raw data again, this time we evaluated our results, using K-fold cross-validation, to allow for a reasonable comparison with preceding works. The model can predict all activities with great accuracy. The overall accuracy is (98.21%), this result is close to previous works on the same dataset done by Alsheikh et al [39] with a hybrid model using deep learning and hidden Markov models DL-HMM (98.23%). It improves accuracy over ensemble learning method [40], and slightly above the model proposed by Ravi et al [41] on the basis of shallow CNN architecture.

TABLE V. COMPARISON WITH OTHER WORKS ON UCI-HAR

Approach	**Accuracy (%)**
Ensemble method of HMM[31]	83.51
Two stage continuous HMM[32]	91.76
Hierarchical continuous HMM[33]	93.18
Our model	**96.776**
Multichannel Dilated CNN[34]	95.49
Deep Res Bidir-LSTM [35]	93.6
Handcrafted features +SVM [36]	89
FFT+1D-CNN[37]	95.75
1D CNN [37]	94.79
Stacked auto encoder +SVM [38]	92.16

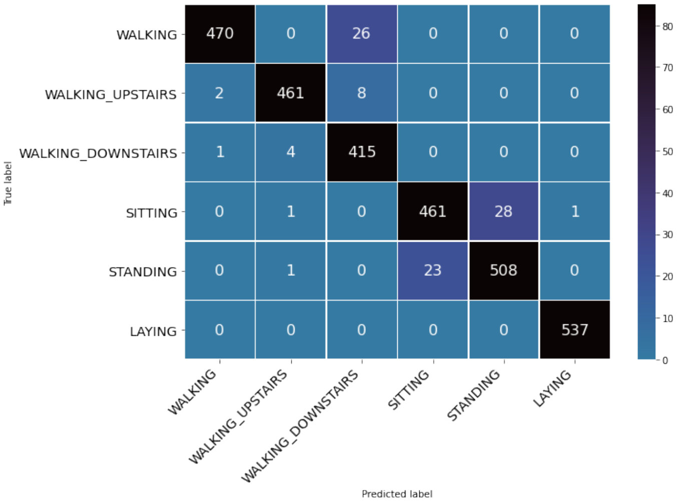


Fig. 4. Confusion Matrix for UCI HAR.

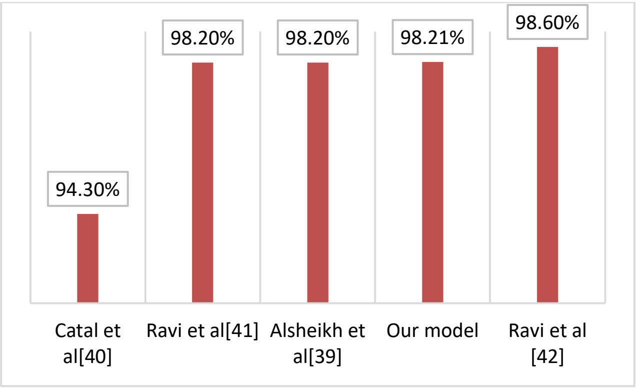


Fig. 5. Comparison between Accuracies of Previous Works on WISDM v1.1.

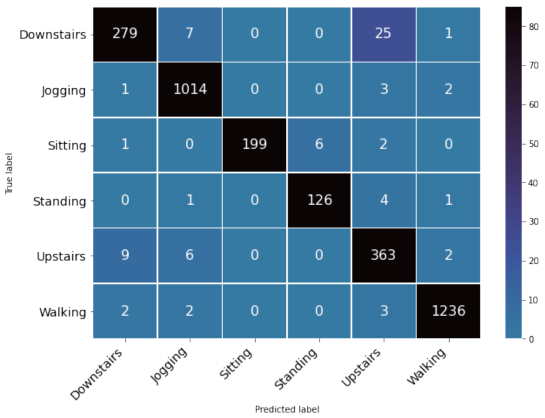


Fig. 6. Confusion Matrix for WISDM V1.1.

Fig. 5 contains a comparison with works on the same dataset. We mention that all results reported in this table are evaluated using 10-fold the cross-validation technique.

The confusion Matrix of WISDM V1.1 dataset is presented in Fig. 6 we can see that Walking and Sitting achieved a recognition close to 100%. We also note that the relative lack of sample for the two Sitting and Standing classes did not affect their recognition, which means that the change in orientation of the sensor on the thigh is easily detectable and learned, helping in result to better identify each class. Jogging is an activity that requires the movement of the whole body from point A to point B, is well identified. Where Walking Upstairs and Downstairs are often confused with each other, this indicates that the model has difficulty distinguishing between these types of movements.

In this part we will compare the ability of our model to detect each activity belonging to UCI HAR and WISDM V1.1, and compare it to other models. We chose these activities because they are the most regularly performed in daily life, and they are recorded differently in both datasets. This comparison should help us to understand the relevance of our approach.

UCI HAR and WISDM V1.1, datasets both contain 6 activities, 5 are the same, and two are different (jogging and laying). Dividing activities into two categories: static and dynamic, can lead to understand the behavior of the model. We will compare and evaluate each activity according to its F1 score, since we have an imbalance between classes.

We observe that the static activities sitting standing and laying, are differentiable by the model among the others even if we change the dataset, this indicates its aptitude to detect those movements despite using only an accelerometer instead of its combination with a gyroscope. We deduct also that the location of sensors does not affect the detection of those activities. the other remaining activities "walking downstairs", “Jogging”, “walking Upstairs”, and” Walking” are dynamic and they present the vast majority of the data in WISDM V1.1, and almost half of UCI HAR dataset. Jogging and Walking are well identified 99% of the time in WISDM V1.1, and 96% in UCI HAR (Walking). The lowest score achieved is 93% in WISDM V1.1, it indicates that the model does not manage to detect with ease the Downstairs class.

TABLE VI. PER ACTIVITY COMPARISON

Activity	**F1 score**					
	WISDM V1.1			***UCI HAR***		
	Our method	Ravi et al [42]	Ronao et al [37]	Lin et al [34]	Zhao et al [35]	Our method
Downstairs	92.99	95.14	99.49	97.16	93.7	95.37
Jogging	99	99.50	-	-	-	-
Sitting	98.50	98.14	87.68	91.14	89.15	94.50
Standing	98.50	97.64	91.37	93.47	90.87	95
Upstairs	94.50	95.30	99.50	96.65	93.96	98.50
Walking	99	99.30	99.44	95.09	94.53	96.96
Laying	-	-	90.55	99.26	99.75	1

Considering the number of sensors, we remark that the use of a single accelerometer alone did not provide the necessary information to identify the dynamic actions which are related to climbing or descending, specifically moving downstairs or upstairs because they obtain the lowest score among classes and even for the other works presented in TABLE VI. On the other hand, we note that the recording in UCI HAR realized with both a gyroscope and an accelerometer allowed a good detection despite the small number of samples, as indicated in TABLE IV.

In WISDM V1.1 dataset the most recognized classes are jogging and walking, followed by walking upstairs in UCI HAR dataset, and the lowest score is for walking downstairs which reaches 93%.

Comparing our results with other approaches, we see that our network can classify activities in a similar way or better than other works using feature engineering, like the spectrogram domain of the time series signal, or hierarchical continuous hidden Markov model or using complex end to end deep learning networks.

In this part we want to test our model on a dataset that does not contain the same characteristics of the two previous ones. As previously mentioned Skoda contains gestures made with the hand in an assembly environment. Performed by a single subject and one type of sensors, it contains 10 gesture classes, to evaluate our work and compare it with others we used the 10-fold cross-validation process. The accuracy of our network is 96%. Fig. 7 shows that it outperforms other works previously done on the same dataset. The classification results are shown in Fig. 8 as a form of a confusion matrix. In this matrix we visualize that the model recognizes all the activities with a high score, except for the activity “close both left Front door” which is confused with “opening left front door” and “closing left front door”. We see also that the NULL class causes the largest confusion.

Class names: 0:'Null Class',1:'Write on Notepad',2:'Open Hood', 3:'Close Hood', 4:'Check Gaps on the Front Door', 5:'Open Left Front Door',6:'Close Left Front Door',7:'Close Both Left Front Door',8:'Check Trunk Gaps',9:'Open and Close Trunk', 10:'Check Steering Wheel'.

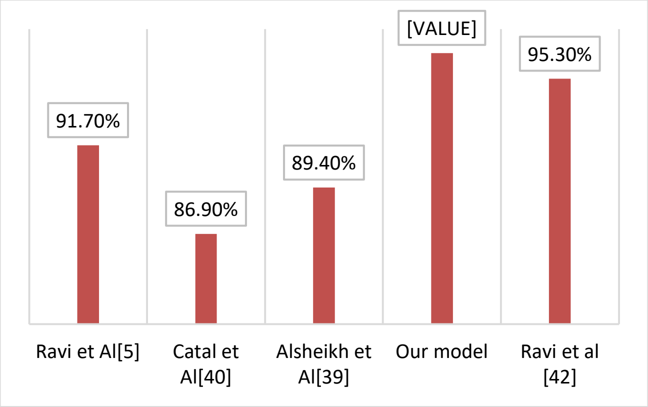
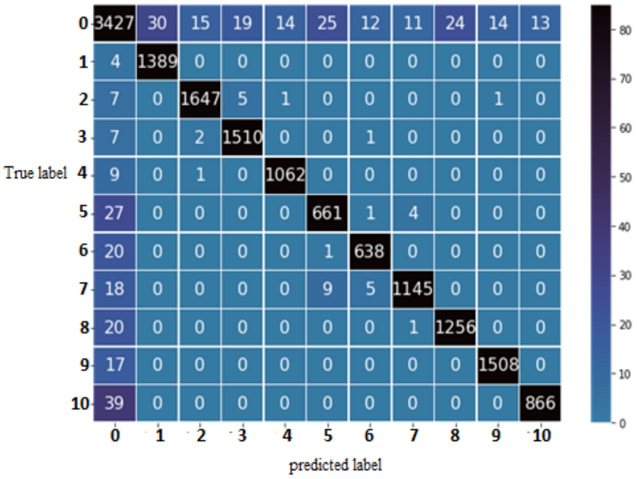


Fig. 7. Comparison between Accuracies of other Works on Skoda.



	0	1	2	3	4	5	6	7	8	9	10
0	3427	30	15	19	14	25	12	11	24	14	13
1	4	1389	0	0	0	0	0	0	0	0	0
2	7	0	1647	5	1	0	0	0	0	1	0
3	7	0	2	1510	0	0	1	0	0	0	0
4	9	0	1	0	1062	0	0	0	0	0	0
5	27	0	0	0	0	661	1	4	0	0	0
6	20	0	0	0	0	1	638	0	0	0	0
7	18	0	0	0	0	9	5	1145	0	0	0
8	20	0	0	0	0	0	0	1	1256	0	0
9	17	0	0	0	0	0	0	0	0	1508	0
10	39	0	0	0	0	0	0	0	0	0	866

Fig. 8. Confusion Matrix for Skoda.

V. Conclusion

In this paper we aimed to integrate Internet of Things (IoT) technology and deep learning to recognize human activities. We presented CNGRU, a new structure that combines convolution layers with GRU. This architecture is able to learn features automatically from raw data, unlike previous works based on handcrafted features. The effectiveness of this architecture is proved by experimenting on three datasets containing a variety of activity classes and recorded using different sensors. We achieved 96.77% on UCI-HAR, 98.21% on WISDM V1.1, and 96.70% on SKODA. This final result is superior than or close to existing state-of-the-art approaches that use shallow or deep designs or classical methods.

Future works will investigate a resource efficient implementation of this network for IoT devices, and explore other datasets that contains more complex activities.

References

[1] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, “A Review on Video-Based Human Activity Recognition,” Computers, vol. 2, no. 2, Art. no. 2, Jun. 2013, doi: 10.3390/computers2020088.

[2] H. D. Mehr, H. Polat, and A. Cetin, “Resident activity recognition in smart homes by using artificial neural networks,” in 2016 4th International Istanbul Smart Grid Congress and Fair (ICSG), Istanbul, Turkey, Apr. 2016, pp. 1–5. doi: 10.1109/SGCF.2016.7492428.

[3] G. Sebestyen, I. Stoica, and A. Hangan, “Human activity recognition and monitoring for elderly people,” in 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP), Sep. 2016, pp. 341–347. doi: 10.1109/ICCP.2016.7737171.

[4] S. Ranasinghe, F. Al Machot, and H. C. Mayr, “A review on applications of activity recognition systems with regard to performance and evaluation,” Int. J. Distrib. Sens. Netw., vol. 12, no. 8, p. 1550147716665520, Aug. 2016, doi: 10.1177/1550147716665520.

[5] D. Roggen et al., “Collecting complex activity datasets in highly rich networked sensor environments,” in 2010 Seventh International Conference on Networked Sensing Systems (INSS), Jun. 2010, pp. 233–240. doi: 10.1109/INSS.2010.5573462.

[6] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, “Activity recognition using cell phone accelerometers,” ACM SIGKDD Explor. Newsl., vol. 12, no. 2, pp. 74–82, Mar. 2011, doi: 10.1145/1964897.1964918.

[7] A. Reiss and D. Stricker, “Introducing a New Benchmarked Dataset for Activity Monitoring,” Jun. 2012, pp. 108–109. doi: 10.1109/ISWC.2012.13.

[8] S. Deep and X. Zheng, “Hybrid Model Featuring CNN and LSTM Architecture for Human Activity Recognition on Smartphone Sensor Data,” in 2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), Dec. 2019, pp. 259–264. doi: 10.1109/PDCAT46702.2019.00055.

[9] F. Hernández, L. F. Suárez, J. Villamizar, and M. Altuve, “Human Activity Recognition on Smartphones Using a Bidirectional LSTM Network,” in 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), Apr. 2019, pp. 1–5. doi: 10.1109/STSIVA.2019.8730249.

[10] W. Ahmad, B. M. Kazmi, and H. Ali, “Human Activity Recognition using Multi-Head CNN followed by LSTM,” in 2019 15th International Conference on Emerging Technologies (ICET), Peshawar, Pakistan, Dec. 2019, pp. 1–6. doi: 10.1109/ICET48972.2019.8994412.

[11] N. Sikder, M. S. Chowdhury, A. S. M. Arif, and A. Nahid, “Human Activity Recognition Using Multichannel Convolutional Neural Network,” in 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), Sep. 2019, pp. 560–565. doi: 10.1109/ICAEE48663.2019.8975649.

[12] T. Su, H. Sun, C. Ma, L. Jiang, and T. Xu, “HDL: Hierarchical Deep Learning Model based Human Activity Recognition using Smartphone Sensors,” in 2019 International Joint Conference on Neural Networks (IJCNN), Jul. 2019, pp. 1–8. doi: 10.1109/IJCNN.2019.8851889.

[13] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, “InnoHAR: A Deep Neural Network for Complex Human Activity Recognition,” IEEE Access, vol. 7, pp. 9893–9902, 2019, doi: 10.1109/ACCESS.2018.2890675.

[14] W. Gao, L. Zhang, Q. Teng, J. He, and H. Wu, “DanHAR: Dual Attention Network for multimodal human activity recognition using wearable sensors,” Appl. Soft Comput., vol. 111, p. 107728, Nov. 2021, doi: 10.1016/j.asoc.2021.107728.

[15] Q. Teng, K. Wang, L. Zhang, and J. He, “The Layer-Wise Training Convolutional Neural Networks Using Local Loss for Sensor-Based Human Activity Recognition,” IEEE Sens. J., vol. 20, no. 13, pp. 7265–7274, Jul. 2020, doi: 10.1109/JSEN.2020.2978772.

[16] J. Sena, J. Barreto, C. Caetano, G. Cramer, and W. R. Schwartz, “Human activity recognition based on smartphone and wearable sensors using multiscale DCNN ensemble,” Neurocomputing, vol. 444, pp. 226–243, Jul. 2021, doi: 10.1016/j.neucom.2020.04.151.

[17] S. M. Bokhari, S. Sohaib, A. R. Khan, M. Shafi, and A. ur R. Khan, “DGRU based human activity recognition using channel state information,” Measurement, vol. 167, p. 108245, Jan. 2021, doi: 10.1016/j.measurement.2020.108245.

[18] J. Gu et al., “Recent advances in convolutional neural networks,” Pattern Recognit., vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[19] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” ArXiv151108458 Cs, Dec. 2015.

[20] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, “Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks,” IEEE Trans. Geosci. Remote Sens.,

vol. 54, no. 10, pp. 6232–6251, Oct. 2016, doi: 10.1109/TGRS.2016.2584107.

[21] T. Mikolov, A. Joulin, S. Chopra, M. Mathieu, and M. Ranzato, “Learning Longer Memory in Recurrent Neural Networks,” ArXiv14127753 Cs, Apr. 2015.

[22] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Comput, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[23] K. Cho et al., “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Oct. 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.

[24] S. Khandelwal, B. Lecouteux, and L. Besacier, “COMPARING GRU AND LSTM FOR AUTOMATIC SPEECH RECOGNITION,” LIG, Research Report, Jan. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01633254.

[25] N. Y. Hammerla, S. Halloran, and T. Plötz, “Deep, convolutional, and recurrent models for human activity recognition using wearables,” in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, New York, USA, Jul. 2016, pp. 1533–1540.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” NIPS 2014 Workshop Deep Learn. Dec. 2014, 2014.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[28] J. Brownlee, “1D Convolutional Neural Network Models for Human Activity Recognition,” Machine Learning Mastery, Sep. 20, 2018. https://machinelearningmastery.com/cnn-models-for-human-activity-recognition-time-series-classification/.

[29] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A Public Domain Dataset for Human Activity Recognition using Smartphones,” presented at the 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, Apr. 2013.

[30] P. Zappi et al., “Activity Recognition from On-Body Sensors: Accuracy-Power Trade-Off by Dynamic Sensor Selection,” in Wireless Sensor Networks, Berlin, Heidelberg, 2008, pp. 17–33. doi: 10.1007/978-3-540-77690-1_2.

[31] Y.-J. Kim, B. Kang, and D. Kim, “Hidden Markov Model Ensemble for Activity Recognition Using Tri-Axis Accelerometer,” 2015 IEEE Int. Conf. Syst. Man Cybern., 2015, doi: 10.1109/SMC.2015.528.

[32] C. A. Ronao and S. B. Cho, “Human activity recognition using smartphone sensors with two-stage continuous hidden markov models: 2014 10th International Conference on Natural Computation, ICNC 2014,” 2014 10th Int. Conf. Nat. Comput. ICNC 2014, pp. 681–686, 2014, doi: 10.1109/ICNC.2014.6975918.

[33] C. A. Ronao and S.-B. Cho, “Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models,” Int. J. Distrib. Sens. Netw., vol. 13, no. 1, p. 1550147716683687, Jan. 2017, doi: 10.1177/1550147716683687.

[34] Y. Lin and J. Wu, “A Novel Multichannel Dilated Convolution Neural Network for Human Activity Recognition,” Math. Probl. Eng., vol. 2020, p. e5426532, Jul. 2020, doi: 10.1155/2020/5426532.

[35] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang, “Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors,” Math. Probl. Eng., vol. 2018, p. e7316954, Dec. 2018, doi: 10.1155/2018/7316954.

[36] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine,” in Ambient Assisted Living and Home Care, Berlin, Heidelberg, 2012, pp. 216–223. doi: 10.1007/978-3-642-35395-6_30.

[37] C. A. Ronao and S.-B. Cho, “Human activity recognition with smartphone sensors using deep learning neural networks,” Expert Syst. Appl., vol. 59, pp. 235–244, Oct. 2016, doi: 10.1016/j.eswa.2016.04.032.

[38] Y. Li, D. Shi, B. Ding, and D. Liu, “Unsupervised Feature Learning for Human Activity Recognition Using Smartphone Sensors,” in Mining Intelligence and Knowledge Exploration, Cham, 2014, pp. 99–107.

[39] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, “Deep Activity Recognition Models with Triaxial Accelerometers,” ArXiv151104664 Cs, Oct. 2016, [Online]. Available: http://arxiv.org/abs/1511.04664.

[40] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, “On the use of ensemble of classifiers for accelerometer-based activity recognition,” Appl. Soft Comput., vol. 37, pp. 1018–1022, Dec. 2015, doi: 10.1016/j.asoc.2015.01.025.

[41] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, “Deep learning for human activity recognition: A resource efficient implementation on low-power devices,” in 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), Jun. 2016, pp. 71–76. doi: 10.1109/BSN.2016.7516235.

[42] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, “A Deep Learning Approach to on-Node Sensor Data Analytics for Mobile or Wearable Devices,” IEEE J. Biomed. Health Inform., vol. 21, no. 1, pp. 56–64, Jan. 2017, doi: 10.1109/JBHI.2016.2633287.