Comparing Deep and Classical Machine Learning Methods for Human Activity Recognition using Wrist Accelerometer

Hristijan Gjoreski, Jani Bizjak, Martin Gjoreski, Matjaž Gams
Jožef Stefan Institute, Department of Intelligent Systems
Jožef Stefan International postgraduate School
firstname.lastname@ijs.si

Abstract

Motivated by the recent trends of the wristband devices and smartwatches accompanied by recent trends in deep learning, we analyzed deep learning and classical machine learning methods on human activity recognition using wrist accelerometer. In particular, we compared the recognition performance of deep learning convolutional neural networks (DL-CNN) and Random Forest with hand-crafted features (ML-RF) on two activity recognition datasets, AmI and Opportunity. The results on the first (larger) dataset showed that both methods perform similarly, achieving 74.6% (ML-RF) and 75.5% (DL-CNN) accuracy. On the second dataset we compared the results for both wrists. The results showed that the left wrist achieves higher accuracy for both methods. Additionally, the DL-CNN achieves higher accuracy for the left and lower accuracy for the right. The comparison showed that ML-RF is in general more robust and better recognizes the activities which are represented with small number of examples, e.g., transition and kneeling. However, with sufficient data (i.e., on the first dataset), DL-CNN slightly outperformed ML-RF and achieved significantly better accuracy than other ML methods: Naïve Bayes, K-Nearest Neighbors, Decision tree, and Support Vector Machines.

1 Introduction

Automatic recognition of daily activities could potentially contribute to proper management of pathologies such as obesity, diabetes and cardiovascular diseases [1]. For example, moderate to vigorous physical activity is associated with decreased risk factors for obesity, cardiovascular and pulmonary diseases, cancer, depression, and increased bone health [2]. Accurate measurement of physical activity is therefore essential in developing intervention strategies and provides rich contextual information which can be used to infer additional useful information [3, 4, 5].

The recent literature on activity recognition (AR) shows that by applying artificial intelligence (AI) methods, in particular machine learning (ML) methods, to sensors data it is possible to recognize human activities. Namely, applying ML methods on wearable accelerometers has proven to be most successful and these devices are probably the most mature sensors for recognizing single-user basic activities such as: running walking, standing, sitting, lying and similar [6,7,8]. The reason for this is that accelerometers are capable of measuring human motion (mainly by measuring the linear 3D accelerations) and estimating body postures (mainly by measuring the orientation with respect to the Earth's gravity). Multi-accelerometer systems have already shown the ability to recognize activities with high accuracies [8]. However, having multiple sensors attached is a burden to the user, which is probably the biggest reason why most such multi-sensor systems are not well-accepted and are not commercially successful regardless of the technical improvements, i.e., battery life, size and weight.

On the other hand, wristband devices (FitBit, Empaica, Microsoft band) and smartwatches (Apple watch, Android wear wristwatches) are becoming popular mainly because people are accustomed to wearing watches, which makes the wrist placement one of the least intrusive placements for wearing a device. However, developing an algorithm for a wrist device that will successfully recognize most of our daily activities is quite a challenge. The reason for this is that hand is usually the most active body part and produces more irregular movements compared to the other parts of the body (e.g. the torso). While there are some recent studies on this topic, researchers usually find this placement less informative achieving poor AR performance [7].

In recent years deep learning (DL) has emerged as a novel approach to classical machine learning. DL is capable of high-level abstraction of data, which allows for robust models capable of contending with high noise accompanying AR problems. In a typical DL architecture, each layer combines features (output) from previous layer and transforms them via non-linearity function to form new feature set. This gives the network an ability to automatically learn best features for specific problem domain, forming hierarchy where basic features are detected in first layers of the network, and in the deeper layers the abstract features from previous layers are combined to form complex feature maps.

DL is already state of the art in computer vision, voice recognition and natural language processing where it performs better than all standard methods of ML and on pair with human ability [9]. While some attempts at detecting AR problems were made with use of DL [10], this area still lacks some proof whether DL is better at solving AR problems than regular ML.

In this paper we compare DL, convolutional neural networks (DL-CNN) to standard ML methods (J48, RF, SVM, KNN, and Naïve Bayes), in the task of human AR using wrist accelerometer. The comparison was performed on two datasets: dataset recorded at our laboratory by 10 subjects, and the Opportunity dataset recorded by 4 subjects [27,28].

2 Related work

The related literature in AR field shows that wearable accelerometers are among the most suitable sensors for unobtrusive AR [12]. Accelerometers are becoming increasingly common because of their lowering cost, weight and power consumption. Currently the most exploited and probably the most mature approach to AR are wearable accelerometer ML methods [7,13,14]. This approach usually implements widely used classification methods, such as decision tree, Random Forest, SVM, KNN, Naive Bayes, and recently DL [15,16].

For the sake of the user's convenience, AR applications are often limited to a single accelerometer, even though nearly all reports find that better performance is obtained with more accelerometers. Numerous studies have shown that the performance of an AR system strongly depends on the accelerometer placement (e.g., chest, abdomen, waist, thigh, ankle) and that some placements are more suitable (in terms of AR performance) for particular activities [7,6,8]. On the other hand, the obtained accuracies strongly depend on the type of activities – micro activities demand wrist information while basic activities like standing, walking, lying, and sitting are recognized worse with wrist sensors.

The wrist was the least exploited placement for AR in the past, mainly because of our inclination towards frequent hand movements which negatively influences the AR system. The researchers were usually testing chest, waist, thighs (left and right) [13,17], ankles (left and right) and neck. For example, a recent overview of AR systems showed that only 5 out of 13 analyzed systems included wrist data in their systems [7]. The results vary a lot and cannot be compared through different studies (different datasets, different algorithm parameters, different methods, etc.). In our previous work we also tested most of these locations on two datasets. On the first one, the results showed that all of the locations perform similarly achieving around 82% accuracy [18]. On the second dataset, where the experiments were more thorough (bigger dataset, improved algorithms) the results showed that thigh and ankle perform similarly (82% and 83% respectively) and achieve higher accuracy compared to the chest (67%) [19].

However, with the penetration of the wrist-worn fitness trackers and smartwatches, it is to be expected that wrist sensor placement becomes a matter of research and application interest. Recently, Trust et al. [20] presented a study for hip versus wrist data for AR. The models using hip data slightly outperformed the wrist-data models. Similarly, in the study by Rosenberg et al. [21] for detecting a sedentary behavior, the models using hip data outperformed the wrist models. In the study by Manini et al. [11] ankle data models achieved high accuracy of 95.0% that decreased to 84.7% for wrist data models. Ellis et al. [22] presented an approach for the recognition of locomotion and household activities in a lab setting. For one subset of activities the hip-data models outperformed the wrist data, but over all activities the wrist-data models produced better results. In our study we are confirming that ankle, knee and belt sensor placement can produce better results, but the wrist produces better results compared to elbow and chest.

Garcia-Ceja et al. [23] presented person-specific wristband AR for activities such as: shopping, showering, dinner, computer-work and exercise. Similarly, Attal et al. [25] used 10-fold-cross validation to evaluate their models and additionally used 1 second window of data with 80% overlap, thus resulting in having similar instances in the training and evaluation dataset, which explains the high accuracy (99%). In our study, we are not just analyzing general models (using leave-one-subject-out evaluation technique).

Various methods of DL are more and more present in all areas of artificial intelligence. Just recently computer was finally able to beat human opponent (world champion) in game of GO [26]. Last year computer was able to learn to play old arcade games, where it again achieved master level and was able to beat human players in most of the game variants [24]. While the mentioned events were the most talked about in recent time, DL has also made large strides in areas of natural language processing, speech recognition and computer vision where it achieves close to human level accuracy.

Other areas of signal processing like AR are still mostly unexplored. In [10] authors showed that in some cases AR is even better with usage of DL compared to standard ML. They showed that by using convolutional neural networks and extensive data preprocessing that reduces influence of null class and attributes it is possible to beat standard ML methods on Opportunity dataset [27,28], using 113 attributes from various on and off body sensors.

3 Datasets

We performed our analysis on two datasets. The first dataset was recorded at our Ambient Intelligence (AmI) laboratory by ten participants. The second dataset is the Opportunity dataset, which is one of the most commonly used benchmark dataset for AR.

3.1. AmI Dataset

For the first dataset, a 120-minute scenario was designed which captures the real-life conditions of a person's behavior. The scenario was performed by ten volunteers and included eight elementary activities (the percentage of instances per class): lying (23%), standing (17%), walking (14%), sitting (12%), cycling (10), on all fours (8%), kneeling (6%), running (5%), bending (2%), and transition (3%). In particular, the walking activity was performed on a treadmill with a one-percent inclination at 4 km/h and 6 km/h, the running activity was also performed on a treadmill with a one-percent inclination at 8 km/h, and the cycling activity was performed on a stationary bicycle with 65 RPM with the difficulty set to 80 watts for the first six minutes and 160 watts for the other six minutes.

The sensor equipment included Shimmer 3-axis accelerometer placed on a wrist with adjustable strap. The data was acquired on a laptop in real-time via Bluetooth using accelerometer sampling frequency of 50 Hz. The data was manually labeled with the appropriate activities.

3.2. Opportunity Dataset

The Opportunity AR dataset is a benchmark dataset, which is commonly used dataset for AR [27,28]. It contains human activities related to the breakfast scenario, which are captured by sensors configured on three subjects who perform everyday life activities. There are 4 classes in this AR task: standing (50%), walking (28%), sitting (19%) and lying (3%). The sensors include a wide variety of body-worn, object-based, and ambient sensors – 72 in total. However, for the need of our study we used only the acceleration data from the left and the right wrist. The sampling rate of the sensor signals is 30 Hz. With these sensors, each subject performed one "drill" session, which has 20 repetitions of a pre-defined sequence of activities and 4 recordings of usual daily activity consisting of 9 specific activities.

One of the biggest issues concerning the Opportunity dataset is the missing data caused by recording data over Wi-Fi or Bluetooth. Also when the activity performed by the volunteer did not belong to a targeted class or was transitioning between two activities, the class for such event was marked as *null*. This way the *null* class represents roughly 80% of all samples. Because the subjects were wearing accelerometers on both wrists, we performed analysis on both locations for the Opportunity dataset.

4 Methods

4.1. Deep Learning

Convolutional neural networks (DL-CNN) were chosen for this study. DL-CNN use multiple layers to combine features learned in previous layers into complex ones. Because the accelerometer provides time-series acceleration data, we first transformed the continuous data into windows. We used 2 and 4 s window with 1 and 2 s overlap as it is common in AR and also empirically confirmed in our previous work [29]. With the sensor sampling frequency of 50Hz this brings exactly 100 samples for AmI dataset and 120 samples for Opportunity dataset, where sampling frequency is 30Hz. Multiplying this over *n* (50-100) feature maps in *m* (at least 5) fully connected layers it can be seen that the number of calculations required at each iteration is way beyond what current computers are capable of doing. That is why we use pooling method for down-sampling. The accelerometer provides 3 data streams (x, y and z axis) which is why we use filter sizes that only pool over the window and not over the dimension (axis). The empirical analysis of the data showed that using max-pooling brings best results for this problem.

Activation function

After using convolution and max-pooling, the values need to go through activation function in order to break linearity. Rectifier Linear function (ReL) is used as activation function in each layer. Most current DL methods run on high performance GPU which are good at doing simple math operations (addition, multiplication) but bad at division and approximation functions (e.g. tan, cos). Solving ReL equation (below) can be up to 10 times faster than solving approximation for *tanh* on the GPU:

$$f(x) = Max(0, x)$$

Gradient descent Optimization

The most time-expensive part of DL methods is the gradient descent, which is used to update values of weights in the network. The most commonly used method for updating the weights is the Stochastic Gradient Descent (SGD). For each input – output pair an error is calculated and then used to update values on weights in each layer according to the gradient.

SGD is sensitive to data variance and cannot be parallelized which makes it very slow on large amounts of data. To overcome this, batched gradient descent is usually used. Instead of calculating the gradient for each input-output pair, it is calculated for a batch of input-output pairs. Then, the result of each batch is averaged and applied. This method gives us less variance sensitivity while speeding up the learning process by parallelization.

For faster learning a process called annealing can be used. This means that learning rate (α) is adjusted in real time depending on learning state. In the beginning the α is large to quickly converge toward the solution, later on α is lowered to allow for precise adjustments to reach the optimum.

Instead of adjusting α depending on current iteration, it is better to adjust it based on the strength of the gradient. Usually the gradient is larger at the beginning of learning and converges toward 0 toward the end. ADAGrad and ADADelta normalize the gradient update based on the cumulative gradients from previous iterations. This may cause the normalization parameter to become so big that it stops learning prematurely.

In our experiments (see Figure 1) it proved best to introduce decay parameter (β) to the function. Beta parameter tells the ratio between current gradient (Δf) and accumulated past gradient (v) to be used for updates.

$$v = \beta v + (1 - \beta)(\Delta f)^2$$

In our tests (see Figure 1) it can be seen that RMSProp performs the best of the mentioned methods, improving convergence speed by more than 5 times compared to SGD.

Architecture

Our DL-CNN is constructed out of 6 layers. The first three are convolution-pooling layers followed by two fully connected hidden layers, the last layer consists of softmax regression. The output of softmax layer is the probability of the input window belonging to each of the possible classes. The first three layers consist of 30 neurons (feature maps) each, the fourth layer has 40 and fifth has 50 neurons. We empirically chose this values as the best for our domain. Convolution filter sizes in convolution layers are (1, 15), (1, 10) and (1, 5), where first dimension represents accelerometer axis (x, y, z) and second represents samples in windows. Pooling filter sizes on first two layers are (1, 2). On third convolution layer a pooling over accelerometer axis is done in order to unify data stream (3, 1). Activation function used in each layer is rectifier function (ReL). Cross entropy is used as a cost function, updates in the backpropagation step are done with the RMSProp procedure.

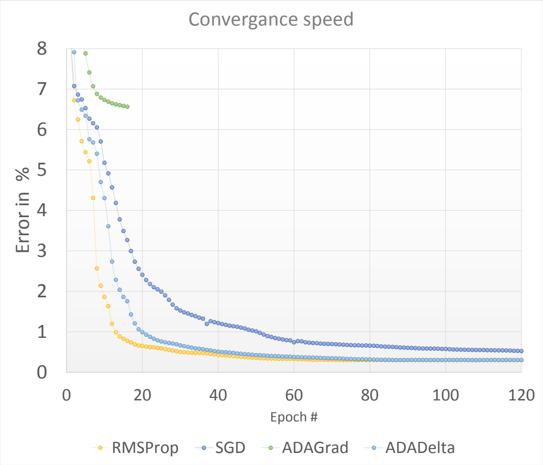


Figure 1. Convergence speed. Comparison between: RMSProp, SGD, ADAGrad, ADADelta.

4.2. Classical Machine Learning

For the classical ML approach, we used standard classification pipeline. That is: data segmentation, data filtering, feature extraction, feature selection and building a classification model. This is a result of decades of experimenting in our laboratory, resulting among others in the first place at the EvAAL competition [12,30].

The data segmentation phase uses an overlapping sliding-window technique, dividing the continuous sensor-stream data into data segments − windows. A window of a fixed size (width) is moved across the stream of data. We used 2 s windows with 1 s overlap, which was defined empirically in our previous work [29]. Once the sensor measurements are segmented, further pre-processing is performed using two simple filters: low-pass and band-pass. The feature extraction phase produces 52 features from the accelerations along the x, y and z axis. The first seven features (Mean X/Y/Z, Total mean and Area X/Y/Z) provide information about the body posture, and the rest features represent the motion shape, motion variation and motion similarity (correlation). More thorough analysis of the features can be found in Tapia's PhD thesis [31].

Once the features are extracted, a feature vector is formed. During training, features vectors extracted from training data are used by a ML algorithm to build an AR model. During classification, feature vectors extracted from test data are fed into the model, which recognizes the activity of the user. We compared five ML algorithms: J48 decision tree [32], Random Forest [33], Naive Bayes [34], SVM [35] and KNN [36].

5 Experimental Results

For the evaluation of the methods, the leave-one-person-out cross-validation technique was used. This means the model was trained on the whole dataset except for one person, and tested on the remaining person. This procedure was repeated for each person.

Four evaluation metrics, commonly used in AR, were analyzed: the recall, precision, F-measure (F1-score), and accuracy.

5.1. AmI Dataset Results

First, we evaluated the 5 ML methods: NB, J48, KNN, SVM and RF. Figure 2 shows the accuracy achieved by each of the methods. RF achieved 74.6% accuracy, which was the best achieved accuracy overall (statistical tests confirmed this). Therefore we chose RF as the ML representative (ML-RF) for the other experiments.

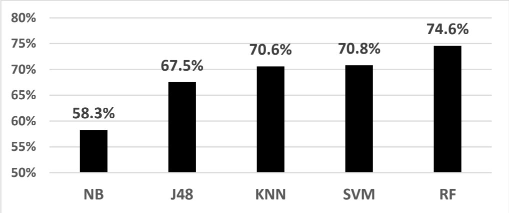


Figure 2. Comparison of the 5 ML algorithms: Naïve Bayes (NB), C4.5 Decision tree (J48), K-nearest neighbours (KNN), Support vector machines (SVM) and Randomf Forest (RF)

Figure 3 shows the comparison of the ML-RF method and the DL-CNN method for each of the subjects individually and the averaged accuracy. The results show that both methods perform similarly and that the averaged accuracy for the DL-CNN is 2 percentage points (p.p.) better than the ML-RF. The biggest improvement of the DL-CNN compared to the ML-RF is for subject 8 (for 21.5 p.p.).

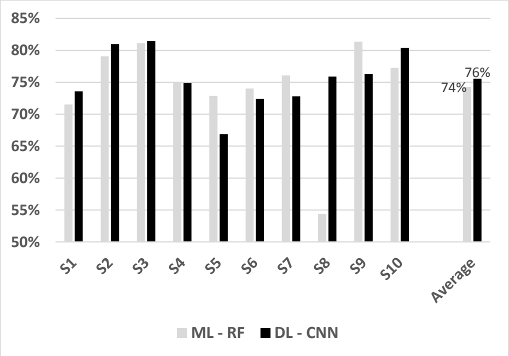


Figure 3. Comparison of the ML-RF method and the DL-CNN method for each of the subjects individually and the averaged accuracy.

These results show that the performance for ML and DL is similar, which on one hand confirms the years of experience and research performed in AR by standard ML (manual exhaustive feature extraction, feature selection and choosing the best ML algorithm) and on the other hand the power of the DL to automatically extract relevant features and to achieve slightly better performance.

We additionally show more detailed results (Figure 4), i.e., the confusion matrix, the recall, the precision and the F1 score for each of the activities. The results show that the best recognized activity is running with 97% F1 score. Other activities that achieve relatively high F1 score (above 75%) are: walking, sitting, lying, and cycling. Additionally, the matrix shows the mutual misclassification between sitting and lying, which is to some extent expected, since both are static activities (postures) and the orientation of the wrist is usually similar, i.e., horizontal. Similarly, walking and standing are mixed, because the orientation of the wrist is vertical. As expected, uncommon activities (such as bending, transitions and kneeling) which are represented with a lower number of examples are poorly recognized.

Accuracy = 74.6%		1	2	3	4	5	6	7	8	9	10
TRUE	Walking -1	8360	1093	41	45	346	8	4	108	117	5
	Standing-2	332	8165	184	308	406	161	74	133	1400	1004
	Sitting -3	9	309	6313	1432	0	9	0	65	12	84
	Lying-4	4	311	1617	14768	60	21	0	40	71	114
	Bending -5	97	573	1	22	808	13	0	10	36	56
	Cycling-6	99	1054	0	8	69	6064	0	6	74	31
	Running-7	42	102	4	0	1	0	3487	1	7	0
	Transition-8	181	391	48	58	23	2	0	650	148	85
	All_fours-9	144	1148	5	102	71	76	0	69	4289	92
	Kneeling-10	64	2007	253	392	139	105	0	45	448	825
Precision		90%	54%	75%	86%	42%	94%	98%	58%	65%	36%
Recall		83%	67%	77%	87%	50%	82%	96%	41%	72%	19%
F1 score		86%	60%	76%	87%	46%	87%	97%	48%	68%	25%

Figure 4. Confusion matrix for the ML-RF method.

The confusion matrix for the DL-CNN (Figure 5) shows that the best recognized activity is running, which is the same as for the ML-RF method, but with 2 p.p. higher F1 score, i.e., 93%. Similarly to ML-RF, other activities that achieve relatively high F1 score (above 75%) are: walking, sitting, lying, and cycling. Similar misclassifications as for the ML-RF, are noted also for the DL-CNN, i.e., between sitting and lying; and standing and walking. As expected, uncommon activities such as transitions and kneeling which are represented with a lower number of examples are poorly recognized.

Accuracy = 75.5%		1	2	3	4	5	6	7	8	9	10
TRUE	Walking -1	8144	489	38	37	28	33	6	0	175	2
	Standing-2	503	7836	137	325	163	720	27	4	1021	618
	Sitting -3	1	261	5635	1929	0	69	0	0	4	4
	Lying-4	7	106	957	13040	3	197	1	1	50	6
	Bending -5	133	517	0	5	701	62	0	0	25	0
	Cycling-6	139	648	4	8	1	6270	2	0	91	62
	Running-7	67	173	48	96	0	21	2999	3	19	0
	Transition-8	99	158	18	5	7	94	0	0	142	3
	All_fours-9	231	880	5	17	41	260	1	1	4275	2
	Kneeling-10	60	2123	272	311	68	949	1	0	116	190
Precision		87%	59%	79%	83%	69%	72%	99%	0%	72%	21%
Recall		91%	69%	71%	91%	49%	87%	88%	0%	75%	5%
F1 score		89%	64%	75%	87%	57%	79%	93%	0%	74%	8%

Figure 5. Confusion matrix for the DL-CNN method.

Figure 6 shows the comparison of the F1 score for each of the activities for the DL-CNN and ML-RF method. The DL-CNN achieves higher F1 score for the following activities: walking, standing, bending, and all fours. The biggest improvement is for the bending activity, which is for 11 p.p. As expected, the DL-CNN achieves significantly worse performance for the transition and the kneeling activities, which are activities that occur rarely and are not well expressed in the dataset.

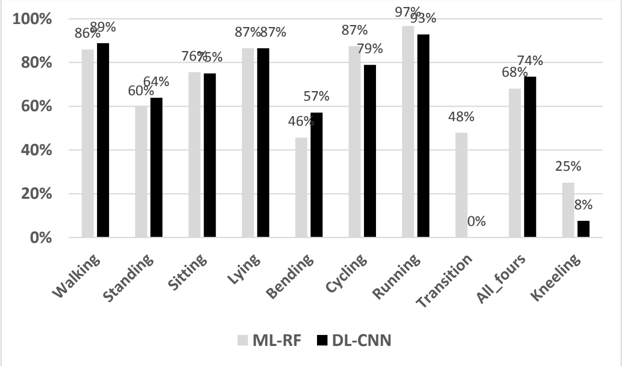


Figure 6. Comparison of the F1 scores for each of the activities for both methods.

5.2. Opportunity

For the Opportunity dataset we performed analysis for both wrists: left and right. Figure 7 shows the accuracies achieved by the ML-RF and DL-CNN methods for both sensor locations respectively.

The results show that the left sensor placement achieves higher accuracy compared to the right-one for both methods. Additionally, the DL-CNN achieves higher accuracy (compared to the ML-RF) for all of the subjects for the left wrist. However, the outcome is opposite when we analyze the results for the right wrist, i.e., the ML-RF achieves better accuracy for each of the subjects. Therefore, one can only conclude that the left sensor placement (non-dominant hand) is better for AR for the right-handed (all test subjects).

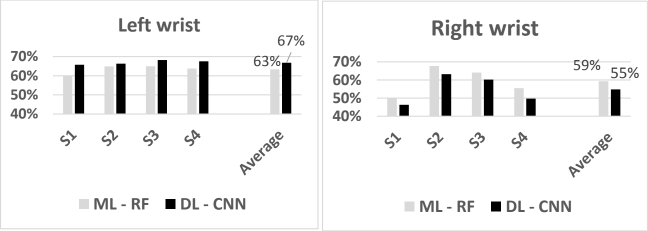
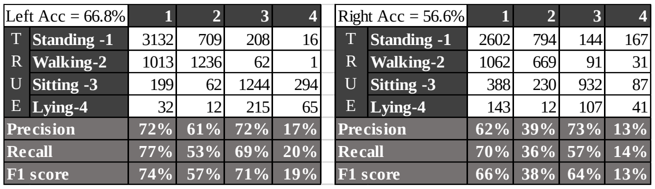


Figure 7. The accuracies achieved by both methods for both wrists.

Similar as for the AmI dataset, we also analyze the confusion matrices – Figure 8. Similar conclusions can be made for both wrists, i.e., standing and sitting are better recognized compared to the walking and lying. Moreover, lying is poorly recognized achieving F score of 19% and 13% for the left and the right wrist respectively. This is probably due to the small number of lying examples in the dataset (approximately 4%).

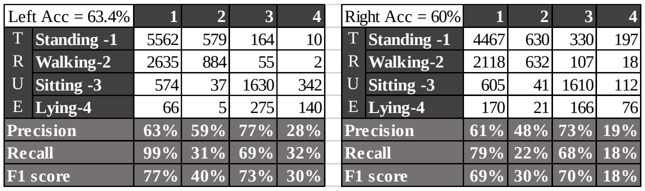


Left Acc = 66.8%		1	2	3	4
T	Standing -1	3132	709	208	16
R	Walking-2	1013	1236	62	1
U	Sitting -3	199	62	1244	294
E	Lying-4	32	12	215	65
Precision		72%	61%	72%	17%
Recall		77%	53%	69%	20%
F1 score		74%	57%	71%	19%

Right Acc = 56.6%		1	2	3	4
T	Standing -1	2602	794	144	167
R	Walking-2	1062	669	91	31
U	Sitting -3	388	230	932	87
E	Lying-4	143	12	107	41
Precision		62%	39%	73%	13%
Recall		70%	36%	57%	14%
F1 score		66%	38%	64%	13%

Figure 8. Confusion matrices for the DL-CNN method for both wrists.

Similar results are achieved by the ML-RF method (Figure 9). That is, for both wrists standing and sitting are better recognized compared to walking and lying. Also, lying is poorly recognized achieving F score of 30% and 18% for the left and the right wrist, respectively. However, lying is in general better recognized by the ML-RF compared to the DL-CNN for both wrists.



Left Acc = 63.4%		1	2	3	4
T	Standing -1	5562	579	164	10
R	Walking-2	2635	884	55	2
U	Sitting -3	574	37	1630	342
E	Lying-4	66	5	275	140
Precision		63%	59%	77%	28%
Recall		99%	31%	69%	32%
F1 score		77%	40%	73%	30%

Right Acc = 60%		1	2	3	4
T	Standing -1	4467	630	330	197
R	Walking-2	2118	632	107	18
U	Sitting -3	605	41	1610	112
E	Lying-4	170	21	166	76
Precision		61%	48%	73%	19%
Recall		79%	22%	68%	18%
F1 score		69%	30%	70%	18%

Figure 9. Confusion matrices for the DL-CNN method for both wrists.

6 Conclusion

We compared the recognition performance of DL-CNNs and ML-RF method on two datasets, AmI and Opportunity, for determining basic activities such as standing or sitting.

On the first dataset, AmI, the results showed that RF performs best when compared to other ML methods: NB, J48, KNN, and SVM. When compared to the DL-CNN, it achieved for 2 p.p. worse accuracy. This performance similarity confirms the years of research experience in AR performed by standard ML procedure: manual exhaustive feature extraction, feature selection and choosing the best ML algorithm. The slightly better performance achieved by the DL-CNN shows the power of DL to automatically extract relevant features and to learn a classification model.

On the second dataset we compared both methods for the left and the right wrist. The results show that the left sensor placement achieves higher accuracy compared to the right-one for both methods for the right-handed. Additionally, the DL-CNN achieves higher accuracy (compared to the ML-RF) for all of the subjects for the left wrist. However the situation is opposite when we analyze the results for the right wrist, i.e., the ML-RF achieves better accuracy for each of the subjects. Therefore, one can only conclude that the left sensor placement (non-dominant hand) is better for AR.

One of the main problems of using DL-CNN and DL in general is the large amount of data they require in order to learn. This was also shown empirically on both datasets. That is, transition and kneeling activities in the AmI dataset are poorly recognized mainly because of this reason, i.e., they are represented by 6% and 3% of the whole dataset. Similarly, lying is poorly recognized in the Opportunity dataset because it is only represented by 4% in the whole dataset. In these cases, it seems that ML-RF is more robust and achieves better results for these activities.

For future work we plan to apply the RF algorithm on the features learned by the DL-CNN method. This way we would be able to directly compare the accuracy achieved by the automatically learned features and the hand-crafted ones. Combining both methods is also considered for future work, e.g., combining the features, combining the outputs of the two methods using voting techniques, meta-learning (Stacking) and similar.

Acknowledgments

The authors would like to thank Mitja Luštrek, Simon Kozina and Božidara Cvetković for their help in programming the data collection software and initial AR algorithm, and all the people that took part in the recording of the datasets. Also, the authors would like to thank Tadej Magajna for proofreading the paper. This work was partly supported by the Slovene Human Resources Development and Scholarship funds, partly by the IN LIFE project – H2020 (grant agreement no: 643442), and partly by the CHIRON project - ARTEMIS JU (grant agreement No. 2009-1-100228).

References

1. Plasqui, G.; Westerterp, K.R. Physical Activity Assessment With Accelerometers: An Evaluation Against Doubly Labeled Water. Obesity. 2007; 15(10): 2371–9.
2. Pedersen, B.K.; Saltin, B. Evidence for prescribing exercise as therapy in chronic disease. Scand J Med Sci Sports. 2006; 16: 3–63.
3. Gregory, D.A.; Anind, K. D.; Peter J. B.; Nigel, D.; Mark, S.; Pete, S. Towards a better understanding of context and context-awareness. 1st International Symposium Handheld and Ubiquitous Computing, 1999; 304-307.
4. Gjoreski, H.; Kaluža, B.; Gams, M.; Milić, R.; Luštrek, M. Context-based Ensemble Method for Human Energy Expenditure Estimation. Applied Soft Computing, 2015; 37: 960-970.
5. Vyas, N.; Farringdon, J.; Andre, D.; Stivoric, J. I. Machine learning and sensor fusion for estimating continuous energy expenditure. Innovative Applications of Artificial Intelligence Conference. 2012; 1613-1620.
6. Atallah, L.; Lo, B.; King, R; Yang, GZ. Sensor Placement for Activity Detection Using Wearable Accelerometers. In Proceedings BSN. 2010; 24–29.
7. Cleland, I.; Kikhia, B.; Nugent, C.; 1, Boytsov, A.; Hallberg, J.; Synnes, K.; McClean, S.; Finlay, D. 1 Cleland, I. Optimal Placement of Accelerometers for the Detection of Everyday Activities. Sensors. 2013; 13 (7).

8. Gjoreski, H.; Luštrek, M.; Gams, M. Accelerometer Placement for Posture Recognition and Fall Detection. 7th International Conference on Intelligent Environments (IE). 2011; 47–54.

9. LeCun Y.;Bengio Y.; Hinton G. Deep learning. Nature. 2015; 521: 436–444

10. Yang J.B.; Nguyen M.N.; San P.P.; Li X.L.; Krishnaswamy S. Deep convolutional neural networks on multichannel time series for human activity recognition. International Conference on Artificial Intelligence (IJCAI). AAAI Press 2015; 3995-4001.

11. Mannini, A; Intille, S.S; Rosenberger, M.; Sabatini, A.M.; Haskell, W. Activity recognition using a single accelerometer placed at the wrist or ankle. Med Sci Sports Exerc. 2013; 45(11): 2193–2203.

12. Gjoreski, H.; Kozina, S.; Gams, M.; Lustrek, M. Competitive Live Evaluation of Activity-recognition Systems. IEEE Pervasive Computing. 2015; 14(1): 70 – 77.

13. Kwapisz J.R; Weiss G.M; Moore, S.A. Activity Recognition using Cell Phone Accelerometers. Human Factors. 2010; 12: 74–82.

14. Wu, H.; Lemaire, E.D.; Baddour, N. Activity Change-of-state Identification Using a Blackberry Smartphone. Journal of Medical and Biological Engineering. 2012; 32: 265–272.

15. Ordóñez, F.J.; Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors. 2016; 16: 115.

16. Wang, L. Recognition of Human Activities Using Continuous Autoencoders with Wearable Sensors. Sensors. 2016; 16: 189.

17. Ravi, N; Dandekar N.; Mysore, P.; Littman, M.L. Activity Recognition from Accelerometer Data. In Proceedings of the 17th conference on Innovative applications of artificial intelligence. 2005; 1541–1546.

18. Gjoreski, H. Adaptive Human Activity Recognition and Fall Detection Using Wearable Sensors. Jozef Stefan International Postgraduate School. Master Thesis. 2011

19. Kozina, S.; Gjoreski, H.; Gams, M.; Luštrek, M. Three-layer activity recognition combining domain knowledge and meta-classification. Journal of Medical and Biological Engineering. 2013; 33(4): 406-414.

20. Trost, S.G.; Zheng, Y.; Weng-Keen Wong. Machine learning for activity recognition: hip versus wrist data. Physiol Meas. 2014; 35(11): 2183-9.

21. Rosenberger, M.; Haskell, W.L.; Albinali, F.; Mota, S.; Nawyn, J.; Intille, S. Estimating Activity and Sedentary Behavior From an Accelerometer on the Hip or Wrist. Med Sci Sports Exerc. 2013; 45(5): 964–975.

22. Ellis, K.; Kerr, J.; Goodboole, S.; Lanckriet, S.; Winq, D.; Marshall, S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol. Meas. 2014; 35: 2191–2203.

23. Garcia-Ceja, E.; Brena R. F.; Carrasco-Jimenez J. C.; Garrido, L. Long-Term Activity Recognition from Wristwatch Accelerometer Data. Sensors 2014; 14: 22500-22524;

24. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; Hassabis, D. Human-level control through deep reinforcement learning, Nature, 2015; 518: 529–533

25. Attal, F.; Dedabrishvili, M.; Mohammed, S.; Chamroukhi, F.; Oukhellou, L.; Amirat, Y. Physical Human Activity Recognition Using Wearable Sensors. Sensors 2015; 15(12): 31314-31338

26. Google AI beats GO champion: http://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/ . 2016.

27. Roggen, D.; Calatroni, A.; Rossi, M.; Holleczek, T.; Tröster, G.; Lukowicz, P.; Pirkl, G.; Bannach, D.; Ferscha, A.; Doppler, J.; Holzmann, C.; Kurz, M,; Holl, G,; Chavarriaga, R,; Sagha, H,; Bayati, H,; Millán, J. Collecting complex activity data sets in highly rich networked sensor environments. In Seventh International Conference on Networked Sensing Systems (INSS). 2010.

28. Chavarriaga, R.; Sagha, H.; Calatroni, A.; Digumarti, S.; Tröster, G.; Millán, J.; Roggen, D. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. Pattern Recognition Letters. 2013.

29. Gjoreski, M.; Gjoreski, H.; Luštrek, Gams, M. Recognizing atomic activities with wrist-worn accelerometer using machine learning. Information Society. 2015.

30. Gjoreski, H.; Kozina, S.; Gams M.; Luštrek, M. RAReFall — Real-time activity recognition and fall detection system. IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). 2014; 145-147.

31. Tapia, E. M. Using Machine Learning for Real-time Activity Recognition and Estimation of Energy Expenditure. Ph.D. Thesis. Massachusetts Institute of Technology. 2008.

32. Quinlan, J.R. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research. 1996; 4: 77-90.

33. Tin Kam, Ho. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition Montreal. 1995; 278–282.

34. Stuart, R.; Peter, N;. Artificial Intelligence: A Modern Approach. Second Edition, Prentice Hall.

35. Aha, D.; Kibler, D. Instance-based learning algorithms. Machine Learning. 1991; 6: 37-66.

36. Cristianini, N; Shawe-Taylor, J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press. 2000.