



OPEN Human activity recognition algorithms for manual material handling activities

Andreas Sochopoulos^{1,3}, Tommaso Poliero^{1,3}, Jamil Ahmad^{1,2}, Darwin G. Caldwell¹ & Christian Di Natali¹✉

Human Activity Recognition (HAR) using wearable sensors has prompted substantial interest in recent years due to the availability and low cost of Inertial Measurement Units (IMUs). HAR using IMUs can aid both the ergonomic evaluation of the performed activities and, more recently, with the development of exoskeleton technologies, can assist with the selection of precisely tailored assisting strategies. However, there needs to be more research regarding the identification of diverse lifting styles, which requires appropriate datasets and the proper selection of hyperparameters for the employed classification algorithms. This paper offers insight into the effect of sensor placement, number of sensors, time window, classifier complexity, and IMU data types used in the classification of lifting styles. The analyzed classifiers are feedforward neural networks, 1-D convolutional neural networks, and recurrent neural networks, standard architectures in time series classification but offer different classification capabilities and computational complexity. This is of the utmost importance when inference is expected to occur in an embedded platform such as an occupational exoskeleton. It is shown that accurate lifting style detection requires multiple sensors, sufficiently long time windows, and classifier architectures able to leverage the temporal nature of the data since the differences are subtle from a kinematic point of view but significantly impact the possibility of injuries.

Keywords Human activity recognition, Feedforward neural networks, Convolutional neural networks, Recurrent neural networks, Industrial wearable technologies

Work-related musculo-skeletal disorders (wMSDs) are still among the principal causes of workplace injuries, although, the ergonomic (re-)design of workplaces has contributed to the mitigation of many of the main historic risk factors (e.g., chemical^{1,2}, fall³, or electrical⁴) for the workers. Nevertheless, there are many manual material handling (MMH) activities that still represent an ergonomic risk and that cannot be easily solved through plant redesign or automation. The lower-back is the body region that is mostly affected by MMH activities and in the past 10 years, occupational back-support exoskeletons (oBSEs) have been shown - in the lab - to have potential to mitigate the risk of developing MMH injuries^{5,6}. However, these devices have been mainly tested in controlled scenarios and for short experimental cycles. Therefore, there is a growing need for extensive field studies^{7–9}. A key factor that limits the uptake of oBSEs in out of the lab scenarios is their limited versatility, i.e., they often perform only one specific task and cannot easily adapt to changing work roles. Indeed, literature studies have shown that if an oBSE is used in a task for which it was not initially designed, acceptance rate and user adoption drops^{10,11}. This prevents their generic use, but this limitation is not typically due to hardware constraints (although this is possible), rather it results from the software and controllers that are developed for only one (or at best a small number of) task(s). Key to increasing the versatility, and therefore workplace acceptance, is providing the exoskeleton with the ability to understand the user intention and the task being performed and to assist accordingly¹².

Fortunately, recent advances in wearable sensors have opened the door to accurate and portable Human Activity Recognition (HAR). In particular, inertial measurement units (IMUs) are widely accessible, affordable, and non-intrusive in various work environments, making them highly suitable for use in HAR. Moreover, their availability has resulted in an abundance of data and many open source datasets^{13–15} for everyday activities i.e., standing, sitting, walking, laying, ascending, descending etc. Subsequently, the use of machine learning and

¹Department of Advanced Robotics, Istituto Italiano di Tecnologia, 16163 Genova, Italy. ²Department of Informatics, Bioengineering, Robotics and Systems Engineering (DIBRIS), Università degli Studi di Genova (UniGe), 16145 Genova, Italy. ³Andreas Sochopoulos and Tommaso Poliero have contributed equally to this work. ✉email: christian.dinatali@iit.it

deep learning algorithms has started to yield ground-breaking results that are helping to solve the classification of simple everyday activities^{16–18}. Zia Uddin et al.¹⁹ proposed a body sensor-based activity modeling and recognition system using time-sequential information-based deep Neural Structured Learning (NSL). The proposed approach achieved around 99% recall rate on a public dataset for daily life activities. Tang et al.²⁰ proposed a lightweight CNN using Lego filters for HAR. The proposed algorithm was tested on five public HAR datasets that comprised of daily life activities. It was concluded that Lego CNN with local loss is smaller, faster and more accurate. Similarly, Yin et al.²¹ proposed a 1-D Convolution Neural Network (CNN)-based bi-directional Long Short-Term Memory (LSTM) parallel model with attention mechanism (ConvBLSTM-PMwA) for HAR. It was concluded that the ConvBLSTM-PMwA model performs better than the existing CNN and RNN models in both classification accuracy (96.71%) and computational time complexity (1.1 times faster). Lastly, Hassan et al.²² conducted a study to compare the performance of classical and ensemble-learning-based machine learning classifiers for HAR. It was concluded that ensemble space (97.78%) and support vector machine (95.18%) achieved high classification accuracies. However, these public datasets and algorithms are limited to daily basic activities that are not utilized in occupational scenarios which consists of different complex activities.

Classification of human activities in an occupational context has received less attention^{23–25}, however promising results for simple pulling/pushing and lowering/lifting activities have been reported^{26,27}. The application of HAR to occupational exoskeletons requires the creation of datasets on occupational activities that are commonly undertaken when using an exoskeleton. This will require efficient algorithms that can provide accurate predictions using only a few data samples, and computationally efficient algorithms able to run on embedded devices. Chen et al.²⁸ proposed a simple rule-based lift detection strategy for hip exoskeleton. The algorithm only used hip joint angles of both sides and trunk angle in the sagittal plane, which could be measured by exoskeleton embedded sensors. The developed algorithm yielded an average accuracy of $97.97 \pm 1.39\%$ during lift detection with subject-dependent model. Poliero et al.²⁹ implemented support vector machine for HAR to enhance back-support exoskeletons versatility and to introduce an automatic switching strategy. The results showed that the approach is promising with a high level of accuracy ($\sim 94\%$), precision (e.g., $\sim 94\%$ for bending) and recall (e.g., $\sim 91\%$ for walking). Similarly, Marko et al.³⁰ developed a new control scheme consisting of Gaussian mixture models (GMM) in combination with a state machine controller to identify and classify the movements of the user. The results showed an overall accuracy of $86.72 \pm 0.86\%$ for providing support to the user. However, the range of classified activities in the above studies is not diverse and is more focused on symmetric activities, thus reducing the versatility of the exoskeleton and the ability to ergonomically assess the user activities. Moreover, there is a range of hyper-parameters related to the classification problem, such as the window size, the type of IMU input data, and the sensor locations that are chosen arbitrarily in most cases. These factors do, however, heavily impact the efficiency, applicability, and accuracy of the human activity classifiers.

The literature has focused extensively on daily life activities and simple tasks, but there is a lack of analysis of complex tasks that are performed in occupational scenarios. Moreover, the increasing interest in HAR largely stems from the potential to transition traditional pen-and-paper ergonomic tools into more data-driven solutions^{31,32}. The key contributions of this work include the integration of several existing HAR algorithms, however, its primary innovation lies in the way these algorithms are being utilized to solve the problem of identifying complex activities that are carried out in occupational scenarios specifically MMH activities. Moreover, identifying the appropriate number of sensors and their placement, the type of input data to HAR algorithms and the selection of optimum window size also brings significant innovation to the research community. Nonetheless, particular focus throughout this work will be on bend detection, classifying the bending technique (stooping/squatting³³) and whether or not the movement was performed symmetrically (trunk torsion). These latter characteristics are crucial for carrying out an online ergonomic assessment of the working task. In addition, a deeper knowledge of the task can help design better controllers for versatile oBSEs that could significantly improve the daily routine of workers performing MMH activities.

The key points of this study are as follow:

1. The introduction of a dataset specifically aimed at designing and validating new HAR algorithms, with a particular emphasis on MMH activities.
2. An analysis of the design parameters (IMU sensor placement, input data, window size) that should be considered when designing and testing a HAR algorithm.
3. Training, testing and comparison of multiple HAR algorithms in the literature

Materials and methods

Experimental protocol

Dataset collection protocol

A total of 10 gender-balanced subjects (age 28.8 ± 4.2 , weight 76.6 ± 13.1 , height 178.5 ± 5.7) were asked to complete a series of randomized lifting activities and a series of activities involving walking and carrying. The experiments took place at the Istituto Italiano di Tecnologia (IIT) premises and complied with the experimental protocol approved by the Ethical Committee of Liguria, Italy, 8/10/2019, protocol number: 001/2019 and complies with the Helsinki Declaration. All the subjects belong to the IIT working population signed a consent form prior to participating, after a full explanation of the experimental procedure and all participants provided the consent to publish for this study. Each subject had to complete a total of 16 activities divided between two main classes: lifting and non-lifting activities.

In the following the authors present a description of the lifting activities as shown in Fig. 1. Lifting activities were further divided into symmetric and non-symmetric, according to the symmetry of the trunk - with respect to the transversal plane - during the lifting and lowering phase. The final subdivision involves the lifting style that the subjects were requested to adopt, namely a squat or a stoop. The subjects were restricted to move within



Fig. 1. The figure shows on the left how the sensors were worn by the test subjects. In the centre there is an image of the subject performing an asymmetric stoop and the rectangle defining the lifting activities area. This also shows the position of the stations according to the rules in Section "Experimental protocol". The right side of the picture shows the worker while carrying a load (flat surface), sitting, or carrying a load on an inclined surface.

a predefined rectangular area (2.40 m x 0.60 m) and lift and lower the load from 3 designated stations placed around this area for 1 minute per lifting activity. The stations' locations were modified after each lifting activity, however the new configuration had to respect the following rules:

1. A station had to be placed in the proximity of the moving area,
2. A station had to be placed far from the moving area and
3. A station had to be rotated by an angle less than 90° .

Additionally, the load to be handled at each activity was randomly selected from a predefined set of weights:

1. 3 kg (L)/ 6 kg (M)/ 9 kg (H) for males
2. 1 kg (L)/ 4 kg (M)/ 6 kg (H) for females

and a pre-defined set of boxes:

1. A large container with handles spaced 60 cm apart (LG)
2. A smaller container with handles spaced 40 cm apart (SG)
3. And a backpack (50 x 30 x 10) with no apparent grips (NG).

Finally, the load was placed at a height of:

1. 0 cm from the ground
2. 30 cm from the ground
3. 60 cm from the ground

This division gives rise to 4 different lifting styles that represent those predominantly used in MMH tasks, and together with the variations of load weights, grips, and heights this yields a diverse dataset able to capture most of the critical lifting patterns. Of course, the lifting form varies significantly from person to person and an enormous dataset from many people would be needed for a generalized classification algorithm. Therefore, the experimental protocol was designed through the randomization patterns listed above, to capture enough diversity and, hence, result in better generalization.

For the non-lifting activities, subjects were asked to walk freely in the laboratory, walk on an inclined surface on a treadmill (TRX-100, TOORX, Alessandria, IT) at self-selected inclination and speed, carry a load with randomized properties - as describe for the lifting activities -, carry a load on an inclined surface on a treadmill, and sit on a chair. All non-lifting activities were recorded for 2 minutes.

At the end of the protocol, data on the following activities were collected: (i) symmetric squat, (ii) asymmetric squat, (iii) symmetric stoop, (iv) asymmetric stoop, (v) walking, (vi) carrying, and (vii) sitting.

Data collection and labelling

For every activity, the whole-body kinematics of the subjects were acquired using a commercial motion capture system based on inertial information (MTw Awinda, Xsens Technologies B.V., Enschede, the Netherlands). The data were captured at a sampling rate of 60Hz from the 17 IMUs worn on the lower and upper legs, the lower and upper arms, the pelvis, sternum, shoulders and head. The data from each IMU include linear acceleration, angular velocity, and sensor orientation. Moreover, the joint angles extracted from the sensor orientations were also included in the dataset and will be used in the subsequent analysis.

Each performed activity was recorded independently, allowing the convenient labelling of each recording. However, the recordings of the lifting activities also included time windows in which the subject was transitioning from one station to another. To avoid mislabelling these transitions as lifting activities, the Wearable Development Toolkit³⁴ was utilised. The time windows in which a lifting activity was performed were manually segmented according to the hip angles of the subject. This occurred because transitions from station to station involved walking or carrying the load, so the hip angles were significantly smaller with respect to the ones when a lifting activity was performed, even when the load is at 60 cm from the ground. In addition, re-labelled frames were double checked using the playback animation of the avatar recreated by the Xsens software.

HAR classification description

Sensor configurations

One of the most crucial and impactful choices to be made in Human Activity Recognition, especially in the occupational domain, is the numbers of sensors that will be used and their locations. This study aims to uncover the impact that legs, arms, one sided body kinematics and full body kinematics have on the classification quality and generalisation capability of the considered models. The single sensor configuration that is very often considered²⁶ is also tested to show that a single sensor cannot provide adequate information regarding the lifting styles of the subject, although kinematically dissimilar tasks can be distinguished. Consequently, the sensor configurations that will be assessed in this study are 6, namely:

- Full Body (FB): Sensors to reconstruct lower and upper body kinematics (no wrist kinematics) are used. This uses a total of 12 IMUs (2 per arm, 3 per leg, 1 for the pelvis, and 1 for the sternum)
- Right Side (RS): Sensors on the Right Shank, Right Thigh, Pelvis, Sternum and Right Forearm are used. This uses a total of 5 sensors
- Different Sides (DS): Sensors on the Left Shank, Left Thigh, Pelvis, Sternum and Right Forearm are used. This uses a total of 5 sensors
- No Arms (NA): Sensors on the Right and Left Shank, Right and Left Thigh, Pelvis and Sternum are used. This uses a total of 6 sensors
- No Arms Right Side (NARS): Sensors on the Right Shank, Right Thigh, Pelvis and Sternum are used. This uses a total of 4 sensors
- Single Sensor (SS): A single sensor placed on the Pelvis.

Given the great number of sensors involved, the Full Body configuration is infeasible in “out of the lab” environments, and only serves as an indicator of the best possible performance that could be achieved, since kinematics from the entirety of the human body are made available. Following similar arguments to those in²⁶, the configuration involving one side serves in understanding the possibilities for sensor reductions. Configurations with no arms are motivated by the fact that lifting styles considered in this study are determined using legs in the lifting motion and usually the hand kinematics should not significantly change across different lifting motions.

Input data used from the IMUs

IMUs provide measurements of the tri-axial linear acceleration and angular velocity. Moreover, when they incorporate magnetometers, the orientation can also be estimated. When referring to human body kinematics, it is often more convenient to work with the joint angles deduced from the estimated orientations. Indeed, joint angles encode information about the relative position of the IMUs and give much more intuitive understanding of the human pose. Thus, it is logical to ask if they can provide a more informative input to the classification algorithms used. The types of representations of the human kinematics described above require an increasing computational complexity, but also encode an increasing amount of information. Thus, it is crucial to assess their effect in the classification performance.

In this study, therefore, the performance of models relying on 3 inputs are analysed: (a) (AV) sensing of linear Acceleration and angular Velocity, (b) (AVO) sensing of linear Acceleration, angular Velocity and Orientation and (c) (AVJ) sensing of linear Acceleration, angular Velocity and estimated Joint angles.

Time series segmentation using fixed sliding windows

Intuitively, as the window size increases, so does the recognition delay, but the computational needs decrease³⁵. It follows that defining the optimal window size is non-trivial^{35,36}. In addition, large data windows are normally considered for the recognition of complex activities. As with any time series classification problem, the choice of the window sizes that will be processed is of paramount importance.

It has been shown³⁵ that the best trade-off between performance and classification capabilities lies between 1–2 seconds for a specific dataset and that further enlargement could even deteriorate the classification results. However, these results have been obtained for a specific dataset that does not contain lifting activities, rather activities with a periodic and static nature. Moreover, the models in³⁵ accepted features extracted from the segmented time series as inputs, while in this study Deep Learning models able to process raw time series are mainly considered.

For the reasons described, 6 different time windows will be tested, namely windows with (a) 5 samples, (b) 15 samples, (c) 30 samples, (d) 60 samples, (e) 120 samples and (f) 240 samples. The time interval depends on the sampling frequency, which is 60 Hz in this specific case.

Models

As discussed in the introduction, it is possible to choose among many different approaches. In most daily HAR tasks, those methods may rely heavily on heuristic handcrafted feature extraction, which is usually limited by

human domain knowledge. Furthermore, only shallow features can be learnt by those approaches, leading to undermined performance for unsupervised and incremental tasks. Due to those limitations, the performances of conventional methods are restricted regarding classification accuracy and model generalization³⁷. Moreover, the feature extraction and model building procedures are often performed simultaneously in the deep learning models. The features can be learnt automatically through the network instead of being manually designed. Additionally, the deep neural network can also extract high-level representation in deep layers, which makes it more suitable for complex activity recognition tasks.

Some of the most popular deep learning approaches for time series that are considered are: 1-Dimensional CNNs and Recurrent Architectures (ie. RNN, GRU, LSTM). Recurrent Architectures are recommended for the recognition of short activities that have a natural order, while CNN is better at inferring long term repetitive activities. The reason is that Recurrent Architectures can make use of the time-order relationship between sensor readings, and CNN is more capable of learning deep features contained in recursive patterns. Both architectures can identify patterns in temporal data successfully however it is not evident on how window sizes and different types of signals impact their learning capabilities.

In the following analysis there will be a presentation of results from deep learning approaches in addition to evaluating methods according to growing complexity, a comparison with simple machine learning methods (hand-crafted features) is discussed. The analysed models include (a) a feedforward Neural Network (NN), (b) a Convolutional Neural Network (CNN), (c) a Recurrent Neural Network (RNN), (d) a network with Convolutional and LSTM layers (CRNN), (e) a Bidirectional LSTM network (BiLSTM) and (f) a combination of Convolutional and Bidirectional LSTM layers (BiCRNN). Table 1 shows the architecture of the HAR models i.e., layer types and key features.

Dataset labels

The classification problem is formulated as a single-class multilabel classification one, meaning a single activity is performed at each time window. This is a reasonable assumption considering the type of activities that are to be classified. The problem could also be formulated as a multi-class classification problem (activity and symmetry), however since the total number of distinct activities is not large, a single-label setting is deemed preferable. Eventually the classification problem consists of 7 classes as reported in Section "Experimental protocol".

To convert the output probabilities from the classifier to activity predictions, the class with the maximum posterior probability is chosen. This is the most straightforward way to convert probabilities to class predictions, however depending on the specific examined problem, more methodologies could be explored. For this specific problem, the simple strategy yields satisfactory results.

Training parameters and evaluation metrics

The dataset collected consists of time-series of accelerations, velocities, and orientations from all the IMUs and all the subjects. The first step in the training pipeline, involves the windowing process for all the time-series. Most of the architectures explored in this work, operate directly on time-series, because they have either a convolutional or a recurrent layer. However, a simple feedforward neural network cannot process raw time-series, hence features must be extracted from them. There are a variety of features to choose from in the literature, and previous works show that simple statistical features are enough for HAR^{38–40}. More precisely, the statistical features extracted from each time window are: (a) mean, (b) standard deviation, (c) min, (d) max, (e) kurtosis, (f) skewness. These features are then fed as an input to the feedforward neural network. The mathematical formulas and description of features are shown in Table 2.

All the models have been trained and tested with a 50% overlap of samples and based on a Categorical Cross-entropy loss. The framework used is Tensorflow⁴¹ and the optimizer utilized for the training process is the Adam optimizer with default parameters. The maximum number of training epochs is set to 150.

To avoid overfitting and allow the models to properly generalize, several measures have been taken. Firstly, regularization is enabled by Dropout layers with a dropout probability of 0.3 in all the neural networks. Moreover, data from 1 of the 10 subjects is kept as a validation set. This set allows us to use an Early Training Stopping mechanism. This latter stops the training process, before the 150 training epochs are over, if no improvement is witnessed *n* training epochs after the best training epoch at the moment (with respect to the minimum validation loss). In our analysis, *n* was set to 10.

Model Name	Layer Types	Key Features
BiCRNN	Conv1D, LeakyReLU, Dropout, Bidirectional LSTM, Dense, BatchNormalization	Combines convolutional and recurrent layers, uses Dropout and Batch Normalization for regularization
BiLSTM	Bidirectional LSTM, Dropout, Dense, BatchNormalization	Uses Bidirectional LSTM for temporal dependencies, with Dropout and Batch Normalization for regularization
CRNN	Conv1D, LeakyReLU, Dropout, LSTM, Dense, BatchNormalization	Combines Conv1D and LSTM layers, uses Dropout and Batch Normalization for regularization
CNN	Conv1D, LeakyReLU, Dropout, Dense, BatchNormalization	Uses Conv1D layers for feature extraction, with Dropout and Batch Normalization for regularization
RNN	LSTM, Dropout, Dense, BatchNormalization	Uses LSTM for temporal dependencies, with Dropout and Batch Normalization for regularization
NN	Dense, Dropout, BatchNormalization	Simple feedforward neural network with Dropout and Batch Normalization for regularization

Table 1. HAR algorithms along with architecture and key features.

Feature	Formula	Description
Mean (μ)	$\mu = \frac{1}{n} \sum_{i=1}^n x_i$	The arithmetic average of all values in a segment.
Standard deviation (σ)	$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$	The spread of data points from the mean in a segment.
Min	$\min(x_1, x_2, \dots, x_n)$	The smallest value in the segment.
Max	$\max(x_1, x_2, \dots, x_n)$	The largest value in the segment.
Skewness (S)	$S = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^3$	The asymmetry of the data distribution in a segment.
Kurtosis (K)	$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$	The “tailedness” of the data distribution in a segment.

Table 2. Mathematical formulas for statistical features used in the analysis.

The method chosen for the evaluation of the each of the models is Leave-A-Subject-Out (LASO) instead of k-fold cross validation. More specifically, 9 out of 10 subjects are chosen as the training set and the data of the subject left is used as the validation set. The models are trained with the aforementioned training set and then evaluated on the validation set by computing the 4 metrics described below. This process is repeated until data from every subject has been used as a validation set and the metrics are finally averaged. Since the data are naturally sorted according to each subject and every subject has a distinct way of performing certain activities, the LASO method was chosen because it is believed that it will offer a better insight on the generalization capabilities of the classifiers in unseen data from subjects outside the test sample.

In conclusion, 648 combinations were analysed, i.e. the cartesian product of the sets of the 6 different models, the 6 different window sizes, the 6 different sensor configurations and the 3 different input signal configurations.

For each model, the following metrics were computed (note that T stands for True, F for False, P for Positive and N for Negative):

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

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

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

$$F1 - score : \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

These metrics are by definition computed considering a binary classification. The metrics have been extracted for each class (against all the other classes) and then weighted-averaged putting more significance on the classes with more instances.

Results

Cross validation on the dataset

The time required to train all the 648 combinations of classifiers was 37 hours and 19 minutes. This training was performed using a machine with the following specifications: 2x Intel(R) Xeon(R) Silver 4210 CPU @ 2.20 GHz Sky Lake, NVIDIA Tesla V100 16 GB.

The data collection resulted in more than 300k samples. This corresponds to almost 2 hours of recordings. Of the total samples, 58% were labelled as lifting activities and 42% as non-lifting. Specifically, the labelled activities were symmetric stoops (28%), asymmetric stoops (24%), symmetric squats (24%), and asymmetric squats (24%). Considering non-lifting activities, walking was performed in 40% of the samples, followed by carrying (39%) and sitting (21%). It is possible to conclude that the dataset is balanced and, hence, no big differences between F1-score and accuracy are expected.

The heatmaps from Fig. 2 can guide a first understanding of the macro results. In particular, Fig. 2 shows how accuracy changed according to the sensor configuration (z axis), the classifier model (x axis), and the window size (y axis), according to a given set of inputs (AVJ, AVO, AV, respectively). Each heatmap represents 216 different models (6 sensor configurations x 6 window sizes x 6 classifier models). A visual inspection shows that the addition of joint angles to the inputs results in better performance regardless of the model. Instead, if few inputs are selected, satisfactory results are mainly associated with simpler models.

On the other hand, the boxplots from Fig. 3 can be used to understand underlying trends in the hyperparameters under analysis. It is worth mentioning that these boxplots represent the distribution of 108 models and so it is not surprising to find big variability. As the dataset is balanced, accuracy and F1-score show similar trends. Starting from the classification models, Fig. 3 shows that, NN has, by far, the highest accuracies considering up to the 75th percentile. The comparison between models becomes more homogeneous also considering the values up to the upper whisker (for all models > 92%). NN also has the biggest variance, underlying that some of the models work very well (accuracy around 95%), while others are extremely poor (accuracy around 55%).

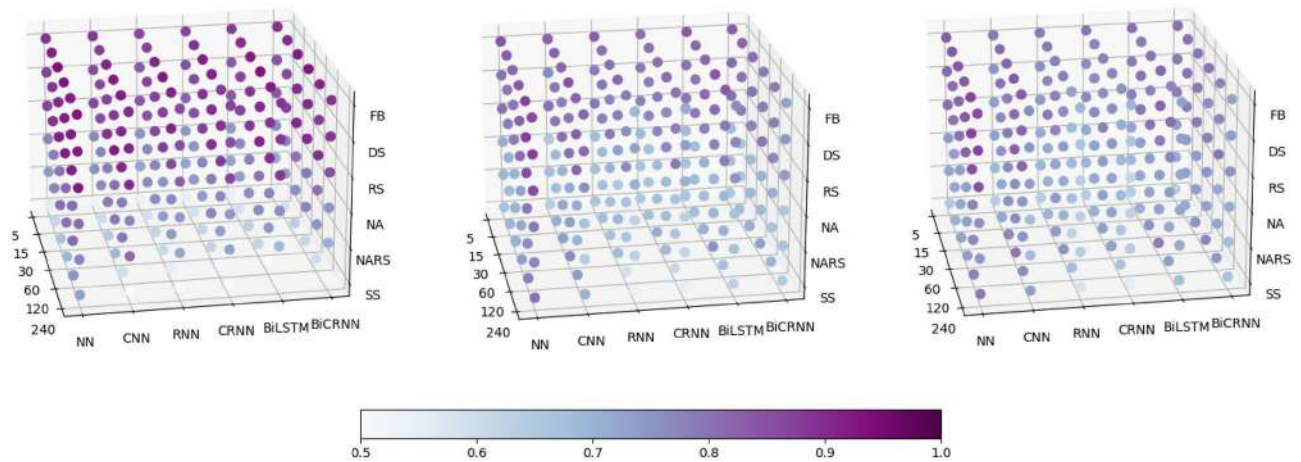


Fig. 2. The heatmaps in the figure show how accuracy changed according to the sensor configuration (z axis), the classifier model (x axis), and the window size (y axis) when considering input AVJ (left), AVO (middle) and AV(right). See Section “HAR classification description” for acronyms details.

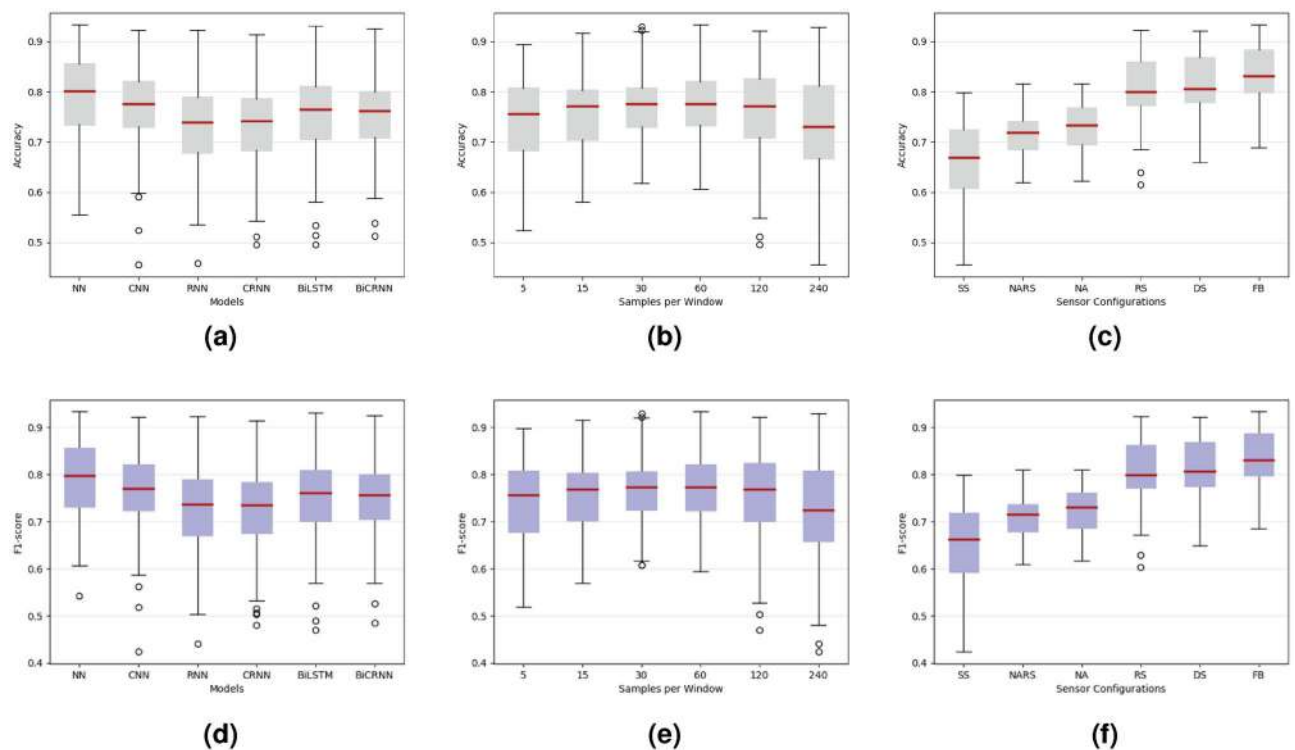


Fig. 3. This figure shows the boxplots of how accuracies (top row) and F1-scores (bottom row) vary according to the considered hyperparameter: classification models (left column), window sizes (central row) and sensors configurations (right column). Refer to Section 2.3 for acronym details.

Figure 3 shows that the biggest time window (240 samples, corresponding to 4 seconds) is associated with the lowest accuracy. Also, the shorter the time windows, the lowest the variance. As presented in³⁵ enlarging the window size does not necessarily imply improved performance. In fact, in many cases the classifiers are unable to learn the underlying mapping, resulting in the increased variance shown in Fig. 3. Interestingly, the smallest window (5 samples, about 0.08 seconds) never attains satisfactory results (accuracy < 90%). Considering the median value of the distribution, the best result is obtained with a 1 second window (60 samples).

In Fig. 3c the effect of arm kinematics in the classification accuracy can be clearly seen. Indeed, configurations that use the kinematics of at least one arm have a notably higher median than those that don't. Moreover, the SS configuration results in a maximum of 80% accuracy but fails in many instances.

Figure 4 shows the accuracy boxplots for every model, window size and sensor configuration for each individual class, namely symmetric squat, asymmetric squat, symmetric stoop, asymmetric stoop, walking, carrying, and sitting. It can help identify which classes pose a problem to the classifiers and what is the purpose behind it. More precisely, the feedforward Neural Network seems to perform poorly mainly on the sitting class, while the two Bidirectional models are able to achieve 100% accuracy on every class for some hyperparameter instances. Figure 4b shows that the variance of the accuracies is increasing as the window size increases and the performance dramatically drops for the symmetric squat class when 240 sample windows are used (approximately 50% accuracy). Figure 4c reveals how kinematics from different parts of the body affect the classification of the activities. Configurations with no arms seem to perform poorly on lifting tasks (accuracy for each lifting class is less than 90%).

Table 3 lists the 5 configurations with highest accuracy and f1-score (note that all the configurations generated during this study are available from the corresponding author on reasonable request). It can be seen that the highest accuracy is 93.36% and the highest f1-score is 93.41%. However, a window size of 120 samples, or 2 seconds, a full body sensor configuration and data on accelerations, velocities and joint angles are required.

Discussion

First and foremost it should be underlined that the objective of this work was: (1) presenting a dataset that could be used to validate HAR algorithms when dealing with MMH activities, and (2) analyse the design parameters that influence such algorithms (3) Training, testing and comparison of multiple HAR algorithms.

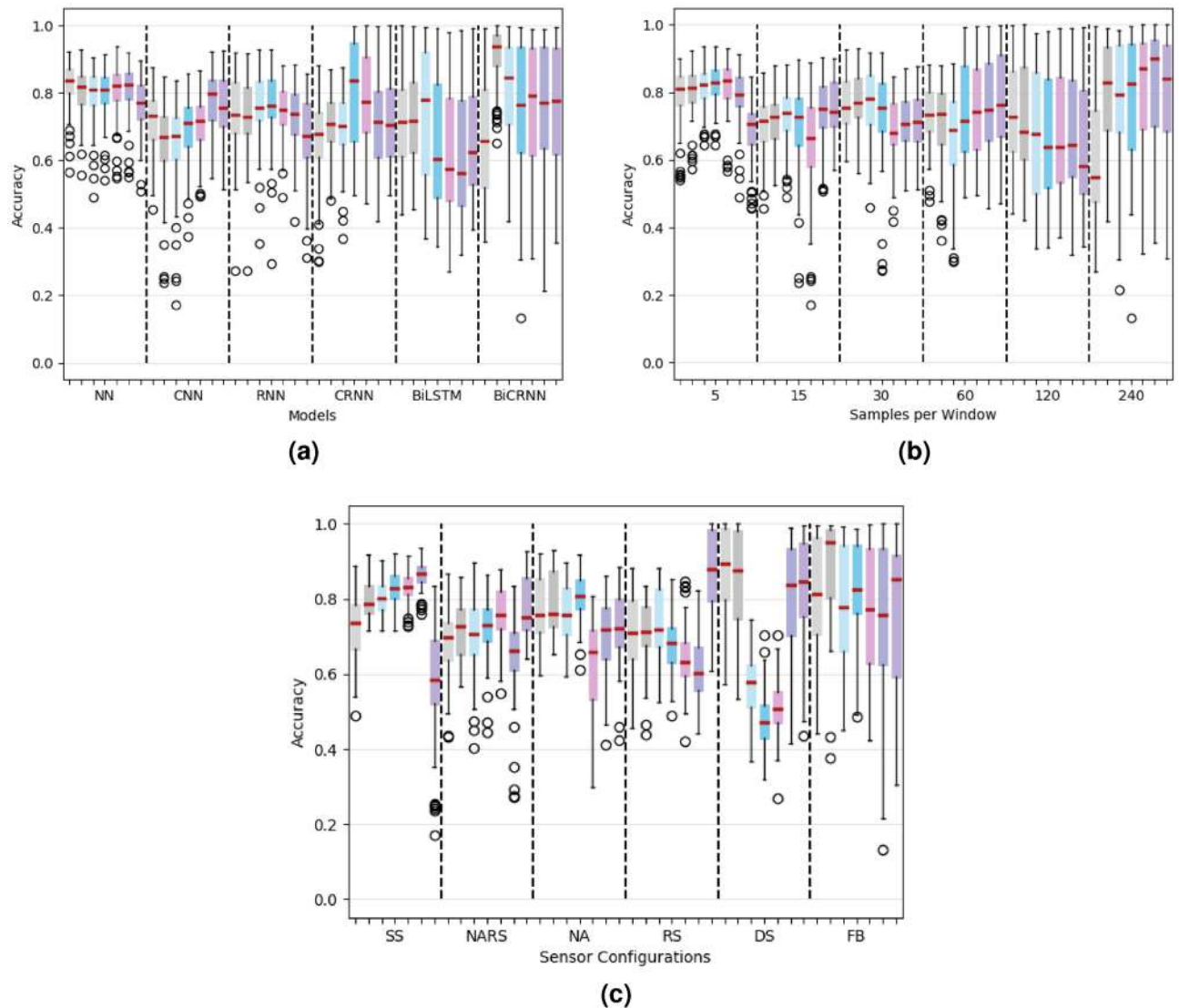


Fig. 4. Boxplots for each individual class considering (a) classification models, (b) window sizes, and (c) sensors configuration. From left to right, the classes are symmetric squat, asymmetric squat, symmetric stoop, asymmetric stoop, walking, carrying, and sitting.

Accuracy	F1-score	Model	Time window (samples)	Sensor configuration	Inputs
0.9336	0.9341	CNN	120	FB	AVJ
0.9331	0.9323	NN	30	FB	AVJ
0.9300	0.9291	NN	60	FB	AVJ
0.9290	0.9287	NN	60	RS	AVJ
0.9289	0.9285	NN	240	FB	AVJ

Table 3. The 5 configurations with the highest accuracy and F1-score in decreasing order.

The experimental set-up design, as presented in the previous section, has allowed to collect about 2 hours of recording of different activities. For the purpose of this paper, the authors’ analysis was focused on 7 activities, but future works could take advantage of the presented dataset by combining, or further classifying, the 7 presented activities.

The results presented above can be used as a tool to understand, according to the specific application, which is the set of optimal hyper-parameters to be chosen. Here in the following, the authors present some interpretations that can be used as a tool to read the outcomes.

The accuracies of the presented classifiers, in some cases, can be considered well-promising. Considering the model selection, several existing studies^{16–22,28–30} have utilized machine learning and deep learning algorithms, obtaining an accuracy above 90%. These studies have focused mainly on daily life activities and simple tasks that are rarely performed in occupational scenarios. Nonetheless, the accuracies obtained in this study are comparable to the accuracies of HAR algorithms utilized in literature, suggesting that the algorithms used in this work can potentially be applied to MMH activities. Additionally, there are other studies^{23–26,42} that emphasized MMH and applied different HAR algorithms, i.e., CNN, NN, LDA, BiLSTM, etc., for identifying activities achieving an accuracy of 76 - 97%. This indicates that our developed algorithm truly carries the potential to accurately identify MMH activities.

However, in some of the cases the accuracies are quite limited. The limited accuracy results are related to the nature of the classification problem under analysis. Indeed, lifting activities are impossible, also for a human observer, to be correctly identified in the beginning and the end of the activity. This statement becomes more intuitive if one imagines the movement of a person that is upright and starts to bend with the intent to lift something. Until “late” in the activity, there is no obvious way to identify the type of lifting activity the person is about to perform. Generally, this is because there is little unique information, in the initial phases of the movement, to reveal the exact type of bending intent. Consequently, a time window in the beginning of a lifting activity, could be mapped to any of the 4 considered lifting activities. Thus, by the nature of the problem it is expected that no mapping exists that can accurately predict all lifting activities, throughout the whole duration of the activity. Linear Discriminant Analysis (LDA) can be used to show that there is indeed no clear distinction at some cases of the 4 lifting activities. Two major components are used for the analysis and 4 increasing time windows, namely 10, 30, 60 and 120 sample windows, respectively. As it can be clearly seen from the scatter plots in Fig. 5, there are many overlapping data points. The overlapping is reduced as the window size increases, which is intuitively correct, since larger windows allow for more information and thus more indications of intent. Also, it is possible to note how the overlapping of activities of the same lifting style but with different symmetry is more significant than the overlapping between stooping and squatting, suggesting an even harder task for the classifier.

Therefore, according to the specific HAR application, algorithms could be designed to simply distinguish between lifting or non-lifting activities or, when the lifting style is important, also the tuning of other hyper-parameters besides the window size should be considered.

From the boxplots, it can be seen that models such as CNN, RNN or CRNN have less variance with respect to NN proving that, if there are sufficient inputs, there is better performance/generalization. However, comparable, or even poorer accuracy results with respect to NN, arise due to increasing size, the so called curse of dimensionality. While the number of samples in the dataset is comparable to other commonly used HAR datasets, it is still questionable if they are sufficient to provide generalization capabilities to models with many parameters. Moreover, large windows significantly reduce the number of data points available for training, which makes it difficult for the deep models to generalize. It would be interesting, in future works, to repeat the same analysis with a more diverse and larger dataset.

Also, it emerged that very large window sizes (e.g., 240 samples corresponding to 4 seconds) are not efficient (only 17 classifiers in the top 100) when the movements under analysis have high frequencies (namely walking or bending, as opposed to sitting or standing still). On the other hand, as it emerged from the LDA, too small window sizes (e.g., 5 samples corresponding to about a 1/10 of second) are too fast to recognize a movement that is in its initial phase.

Analysing the sensor configurations, it can be seen that, considering all the classes, the best performance is obtained when there is at least one sensor in the arm (see Fig. 3f). In addition, from the analysis of Fig. 4c it is possible to note how in the RS, DS and FB configurations there are classifiers for which the sitting accuracy reaches around 100%, while this does not happen in the other configurations. Hence, it is possible to conclude that a sensor in the arm is extremely useful to properly classify sitting activities. While the FB configuration can be considered as a gold standard, due to the excessive number of sensors to be used, it is hard to imagine how it could be applied in occupational scenarios. So which is the best proxy of this configuration? Considering sitting and carrying, it appears that the performance of the DS configuration are almost equivalent to the FB. Interestingly, the same does not apply to the DS configuration, where the carrying classification is quite

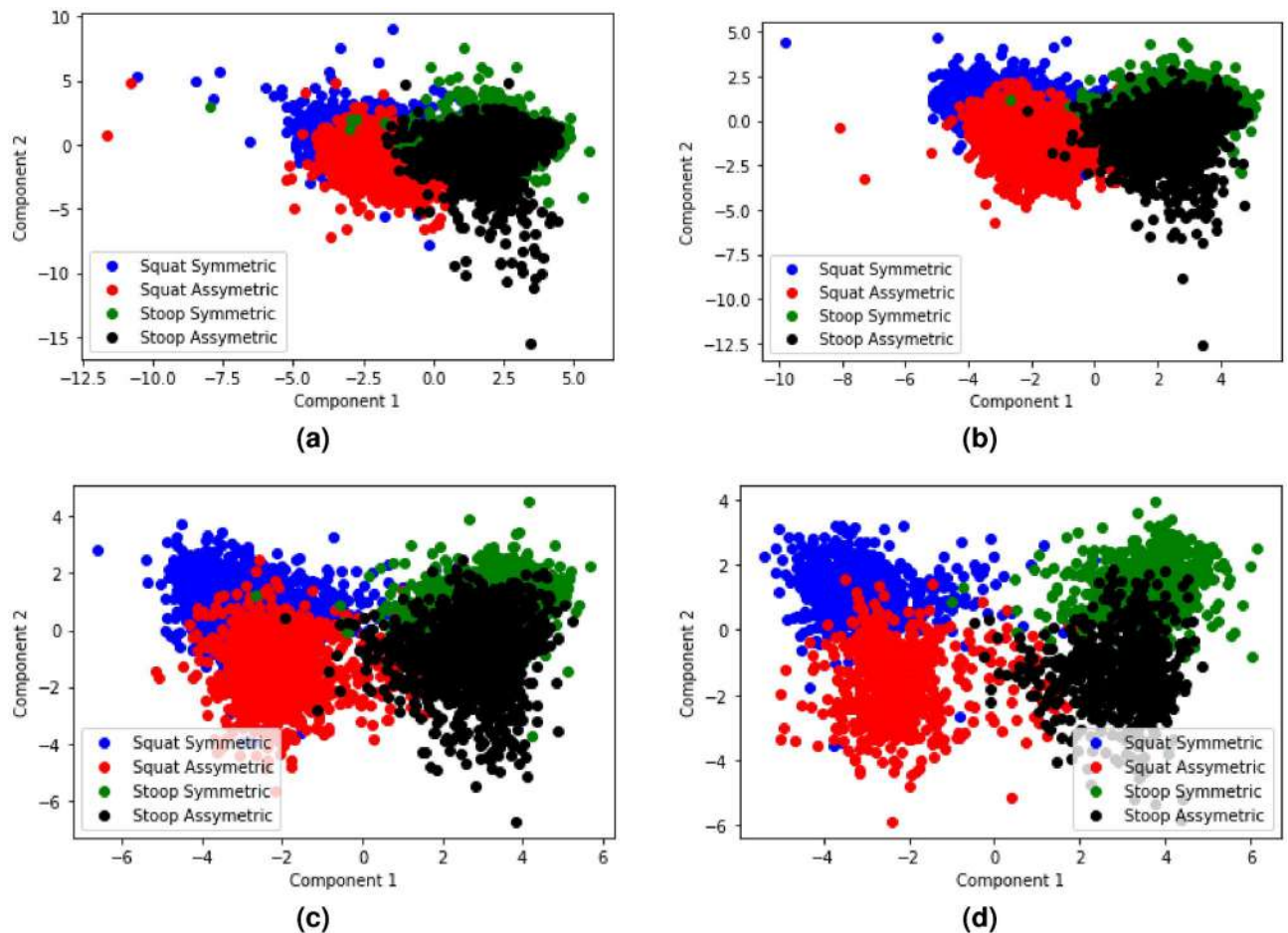


Figure 5. These figures show the Linear Discriminant Analysis (LDA) outcome when considering symmetric squatting (blue), asymmetric squatting (red), symmetric stooping (green), and asymmetric stooping (black). The figures show the results obtained when considering (a) 4 samples, (b) 10 samples, (c) 30 samples, and (d) 120 samples.

poor (max accuracy slightly above 80%). Also, for both RS and DS configurations, the walking classification is not satisfactory. This result might indicate that when considering many inputs (as in the FB, RS, and DS configurations) the information provided by both arms is extremely valuable. However, as the number of inputs gets reduced (and so the training complexity), simpler configurations show promising capabilities in classifying walking and carrying even without arm sensors. This is the case of the SS, NA, and NARS configurations (note however that the sitting classification is very poor, as expected). The analysis of these 3 latter configurations shows an interesting trend: while it could be expected that information on the sternum and on the lower limb joints might be necessary to properly classify the symmetry of a lift and its style, the little differences between the SS configuration and the NA and NARS configuration suggest the opposite.

This first analysis also suggests that for a proper sensor configuration choice, it is important to understand if certain classes are of more importance than other. In this case, the average accuracy should not be trusted since there could be poor underlying performance in some classes. For example, in an instance where sitting detection is not needed, the SS configuration could offer the best cost-to-accuracy ratio with a single sensor.

A macroscopic observation of Fig. 4, shows that as complexity and information are increased the accuracy of all the individual classes can increase. However, that comes with an increase in variance, which means that the sensitivity in the rest of the hyperparameters is increased, requiring better hyperparameter tuning and more computational power. It is interesting, but also expected, that the effect is reverse when the average accuracy is observed, according to Fig. 3. As indicated by the heatmaps, increased processing and information in the input signals have a big impact in the classification accuracy. In particular, classifiers that utilize joint angle information outperform those that use accelerations, velocities and orientations or accelerations and velocities alone. This observation is consistent with intuition, since joint angles are the product of specific, to this problem, processing which gives insight on the HAR problem. However, this processing poses a need for extra computational effort and a complete set of sensors. Consequently, although accuracy is significantly improved in most cases, robustness and computational efficiency are being compromised and extra overhead is added.

Furthermore, it can be seen in Table 3, the top 5 scoring configurations use all the available information as inputs, namely the accelerations, the velocities and the joint angles. Most of them (4 out of 5) have a FB sensor

configuration and time windows more than 0.5 seconds, meaning that most of the available information is used. Moreover, all classifiers except the first one, are based in the NN model. This is possibly because the abundance of information results in a simple mapping to be learnt, meaning that models with many parameters are prone to overfitting.

Although the detailed analysis we conducted in this work using different HAR algorithms for MMH can be helpful to the research community, there are some limitations that need to be addressed in the future. First, data from actual work environments with a wider range of activities should be used to validate the algorithms. Second, the algorithms' accuracy may be further increased by fine-tuning the hyperparameters of the models. Thirdly, to assess each component's unique contribution to the HAR algorithms, an ablation investigation has to be conducted which can be done by assessing the significance of each attribute of HAR algorithm in a step-by-step manner and reporting the accuracy of models at each step. Finally, it is important to examine and contrast the computational complexity of various HAR methods and can be achieved by examining the training and testing duration of individual HAR algorithms.

Conclusions

While there are many studies that have focused on automatic recognition of daily tasks, there is a lack of datasets and classifiers on manual material handling (MMH) related activities. In this study, the authors provide a guide on how to approach the problem of Human Activity Recognition (HAR) when considering MMH activities. Particular focus was given to the selection of the classification hyper-parameters, namely the classification models, the sensors' configuration and the data extracted from them and, lastly, the choice of the time series segmentation. A dataset of 10 subjects performing a complex set of MMH activities was collected and used to train a variety of classifiers. The results have been discussed highlighting how particular hyper-parameters choices can influence the classification outcome and how such results can be used to improve HAR performance in applications that utilize wearable devices such as exoskeletons.

Although this work aims to address as many factors impacting the HAR problem as possible, several of them are left uncovered. Therefore, future works, can help not only validating the current findings with a more diverse dataset, but also discovering or highlighting new factors that affect the classification. Moreover, the complexity of the classification problem to be solved is dependent on the model inputs, the sensor configurations and the time windows, however no network hyperparameter tuning takes place in the current work. Thus, the analysis performed in future works instead of testing different configurations on the same networks, could focus on a hyperparameter tuning algorithm in order to achieve the best possible results. Furthermore, an ablation study of HAR algorithms and computational complexity can be analyzed as well.

Data availability

The datasets generated during the current study are available from the corresponding author on reasonable request.

Received: 28 December 2023; Accepted: 26 November 2024

Published online: 31 March 2025

References

1. Amyotte, P. R., MacDonald, D. K. & Khan, F. I. An analysis of CSB investigation reports concerning the hierarchy of controls. *Process Saf. Prog.* **30**, 261–265 (2011).
2. Amyotte, P. R. & Eckhoff, R. K. Dust explosion causation, prevention and mitigation: An overview. *J. Chem. Health Saf.* **17**, 15–28 (2010).
3. Bell, J. L. et al. Evaluation of a comprehensive slip, trip and fall prevention programme for hospital employees. *Ergonomics* **51**, 1906–1925 (2008).
4. Floyd, H. L. A practical guide for applying the hierarchy of controls to electrical hazards. In *2015 IEEE IAS Electrical Safety Workshop*, 1–4 (IEEE, 2015).
5. Toxiri, S. et al. Back-support exoskeletons for occupational use: An overview of technological advances and trends. *IIEE Trans. Occup. Ergon. Human Factors* **7**, 237–249 (2019).
6. Kermavnar, T., de Vries, A. W., de Looze, M. P. & O'Sullivan, L. W. Effects of industrial back-support exoskeletons on body loading and user experience: An updated systematic review. *Ergonomics* **64**, 685–711 (2021).
7. Crea, S. et al. Occupational exoskeletons: A roadmap toward large-scale adoption. methodology and challenges of bringing exoskeletons to workplaces. *Wearable Technol.* **2** (2021).
8. Di Natali, C. et al. From the idea to the user: A pragmatic multifaceted approach to testing occupational exoskeletons. *Wearable Technol.* **6**, e5 (2025).
9. Di Natali, C. et al. Smart tools for railway inspection and maintenance work, performance and safety improvement. *Trans. Res. Procedia* **72**, 3070–3077 (2023).
10. Poliero, T. et al. Applicability of an active back-support exoskeleton to carrying activities. *Front. Robot. AI* **7**, 579963 (2020).
11. Baltrusch, S., Van Dieën, J., Van Bennekom, C. & Houdijk, H. The effect of a passive trunk exoskeleton on functional performance in healthy individuals. *Appl. Ergon.* **72**, 94–106 (2018).
12. Poliero, T. et al. Versatile and non-versatile occupational back-support exoskeletons: A comparison in laboratory and field studies. *Wearable Technol.* **2**, e12 (2021).
13. Anguita, D., Ghio, A., Oneto, L., Parra Perez, X. & Reyes Ortiz, J. L. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*, 437–442 (2013).
14. Weiss, G. M., Yoneda, K. & Hayajneh, T. Smartphone and smartwatch-based biometrics using activities of daily living. *IEEE Access* **7**, 133190–133202 (2019).
15. Sikder, N. & Nahid, A.-A. Ku-har: An open dataset for heterogeneous human activity recognition. *Pattern Recogn. Lett.* **146**, 46–54 (2021).
16. Thakur, D., Guzzo, A. & Fortino, G. Intelligent adaptive real-time monitoring and recognition system for human activities. *IEEE Trans. Indus. Inf.* (2024).

17. Thakur, D. & Biswas, S. Permutation importance based modified guided regularized random forest in human activity recognition with smartphone. *Eng. Appl. Artif. Intell.* **129**, 107681 (2024).
18. Thakur, D. & Biswas, S. Online change point detection in application with transition-aware activity recognition. *IEEE Trans. Human-Mach. Syst.* **52**, 1176–1185 (2022).
19. Uddin, M. Z. & Soylu, A. Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning. *Sci. Rep.* **11**, 1–15 (2021).
20. Tang, Y., Teng, Q., Zhang, L., Min, F. & He, J. Layer-wise training convolutional neural networks with smaller filters for human activity recognition using wearable sensors. *IEEE Sens. J.* **21**, 581–592 (2020).
21. Yin, X., Liu, Z., Liu, D. & Ren, X. A novel cnn-based bi-lstm parallel model with attention mechanism for human activity recognition with noisy data. *Sci. Rep.* **12**, 1–11 (2022).
22. Ashraf, H., Bröls, O., Schwartz, C. & Boutaayamou, M. Comparison of machine learning algorithms for human activity recognition. In *BIO SIGNALS, 2023* (SciTePress, Lisbon, Portugal, 2023).
23. Bastani, K., Kim, S., Kong, Z., Nussbaum, M. A. & Huang, W. Online classification and sensor selection optimization with applications to human material handling tasks using wearable sensing technologies. *IEEE Trans. Human-Mach. Syst.* **46**, 485–497 (2016).
24. Grzeszick, R. et al. Deep neural network based human activity recognition for the order picking process. In *Proceedings of the 4th international Workshop on Sensor-based Activity Recognition and Interaction*, 1–6 (2017).
25. Kim, S. & Nussbaum, M. A. An evaluation of classification algorithms for manual material handling tasks based on data obtained using wearable technologies. *Ergonomics* **57**, 1040–1051 (2014).
26. Porta, M., Kim, S., Pau, M. & Nussbaum, M. A. Classifying diverse manual material handling tasks using a single wearable sensor. *Appl. Ergon.* **93**, 103386 (2021).
27. Pesenti, M. et al. Imu-based human activity recognition and payload classification for low-back exoskeletons. *Sci. Rep.* **13**, 1184 (2023).
28. Chen, B., Grazi, L., Lanotte, F., Vitiello, N. & Crea, S. A real-time lift detection strategy for a hip exoskeleton. *Front. Neurobot.* **12**, 17 (2018).
29. Poliero, T., Mancini, L., Caldwell, D. G. & Ortiz, J. Enhancing back-support exoskeleton versatility based on human activity recognition. In *2019 Wearable Robotics Association Conference (WearRAcon)*, 86–91 (IEEE, 2019).
30. Jamssek, M., Petrić, T. & Babić, J. Gaussian mixture models for control of quasi-passive spinal exoskeletons. *Sensors* **20**, 2705 (2020).
31. Di Natali, C. et al. Equivalent weight: Connecting exoskeleton effectiveness with ergonomic risk during manual material handling. *Int. J. Environ. Res. Public Health* **18**, 2677 (2021).
32. Zelik, K. E. et al. An ergonomic assessment tool for evaluating the effect of back exoskeletons on injury risk. *Appl. Ergon.* **99**, 103619 (2022).
33. Burgess-Limerick, R. Squat, stoop, or something in between?. *Int. J. Ind. Ergon.* **31**, 143–148 (2003).
34. Haladjian, J. The wearables development toolkit: An integrated development environment for activity recognition applications. *Proc. ACM Interact., Mob., Wearable Ubiquitous Technol.* **3**, 1–26 (2019).
35. Banos, O., Galvez, J.-M., Damas, M., Pomares, H. & Rojas, I. Window size impact in human activity recognition. *Sensors* **14**, 6474–6499 (2014).
36. Huynh, T. & Schiele, B. 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*, 159–163 (2005).
37. Wang, J., Chen, Y., Hao, S., Peng, X. & Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recogn. Lett.* **119**, 3–11 (2019).
38. Sadek, S., Al-Hamad, A., Michaelis, B. & Sayed, U. A fast statistical approach for human activity recognition. *Int. J. Comput. Inf. Syst. Indus. Manag. Appl.* **4**, 7–7 (2012).
39. Attal, F. et al. Physical human activity recognition using wearable sensors. *Sensors* **15**, 31314–31338 (2015).
40. Banos, O., Damas, M., Pomares, H., Prieto, A. & Rojas, I. Daily living activity recognition based on statistical feature quality group selection. *Expert Syst. Appl.* **39**, 8013–8021 (2012).
41. Abadi, M. et al. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–283 (2016).
42. Barazandeh, B. et al. Robust sparse representation-based classification using online sensor data for monitoring manual material handling tasks. *IEEE Trans. Autom. Sci. Eng.* **15**, 1573–1584 (2017).

Author contributions

Conceptualization: A.S, T.P. and C.D.N.; Methodology: A.S, and T.P.; Software: A.S.; Formal analysis: A.S., and T.P.; Investigation: A.S., and T.P.; Data curation: A.S., and T.P.; Writing- original draft preparation: A.S., and T.P.; Writing-review and editing: T.P., J.A., D.G.C. and C.D.N.; Visualization: A.S, T.P. and C.D.N.; Supervision: D.G.C. and C.D.N.; Project administration: C.D.N.; Funding acquisition: C.D.N.; All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the STREAM project funded by Shift2Rail Joint Undertaking, established under the European Unions Horizon 2020 framework programme for research and innovation, under grant agreement No 101015418. This work was also supported by the BEEYONDERS project funded by the European Union's Horizon Europe research and innovation programme under grant agreement N° 101058548. Responsibility for the information and views expressed in the paper/article lies entirely with the authors.

Declaration

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.D.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025