 doi:10.20944/preprints202107.0505.v1

Article

Deep Learning for Classifying Physical Activities from Accelerometer Data

Vimala Nunavath [1,†,‡], Sahand Johansen [2,‡], Tommy Sandtorv Johannessen [2,‡], Lei Jiao [2,‡], Bjørge Herman Hansen [3,*], Sveinung Berntsen [3,*], and Morten Goodwin [2,‡]

[1] Affiliation 1; vimala.nunavath@usn.no;
[2] Affiliation 2; Sahand.johansen@gmail.com; tommy.s.johannessen@gmail.com; lei.jiao@uia.no; morten.goodwin@uia.no
[3] Affiliation 3; bjorge.h.hansen@uia.no; sveinung.berntsen@uia.no
* Correspondence: vimala.nunavath@usn.no
† Current address: University of South-Eastern Norway, Hasbergsvei 36, Krona, 3616 Kongsberg
‡ These authors contributed equally to this work.

Abstract: Physical inactivity increases the risk of many adverse health conditions, including the world's major non-communicable diseases, such as coronary heart disease, type 2 diabetes, and breast and colon cancers, shortening life expectancy. There are minimal medical care and personal trainers' methods to monitor a patient's actual physical activity types. To improve activity monitoring, we propose an artificial-intelligence-based approach to classify the physical movement activity patterns. In more detail, we employ two deep learning (DL) methods, namely a deep feed-forward neural network (DNN) and a deep recurrent neural network (RNN) for this purpose. We evaluate the proposed models on two physical movement datasets collected from several volunteers who carried tri-axial accelerometer sensors. The first dataset is from the UCI machine learning repository, which contains 14 different activities-of-daily-life (ADL) and is collected from 16 volunteers who carried a single wrist-worn tri-axial accelerometer. The second dataset includes ten other ADLs and is gathered from 8 volunteers who placed the sensors on their hips. Our experiment results show that the RNN model provides the accuracy performance compared to the state-of-the-art methods in classifying the fundamental movement patterns with an overall accuracy of 84.89% and an overall F1-score of 82.56%. Our results indicate that the proposed method will provide the medical doctors and trainers a promising way to precisely track and understand a patient's physical activities for better treatment.

Keywords: Classification; Deep Learning; Health; Machine Learning; Accelerometer data; Sensors; Physical activity)

1. Introduction

Physical activity (PA) is defined as "any bodily movement produced by skeletal muscles that result in energy expenditure above resting level" [1]. Physical activity recognition has been required by several real-life applications, such as monitoring older people, lifelong systems for monitoring energy expenditure and supporting weight-loss programs, and digital assistants for weightlifting exercises, etc. [2]. It is well known that physical inactivity is a risk factor for a variety of chronic diseases, particularly diabetes, cardiovascular disease, obesity, and depression [3–6]. There are minimal methods for medical care and personal trainers to monitor a patient's actual physical activity types and training diaries where they commonly use logs. However, there are questions around the accuracy and credibility of these diaries, as they may be subject intentionally or unintentionally to social desirability and recall bias [3]. Today's de facto method is personal diaries, of which the accuracy and credibility can be put into question, as these could be either intentionally or unintentionally subject to social desirability and recall bias. Therefore, a reliable and objective movement registration method is vital.

In recent years, much research has been done to recognize the PA based on inertial sensor data from body-worn and smartphone sensors accelerometer, and gyroscope [2]. This paper focuses on the use of body-worn tri-axial accelerometers to collect activities and movement patterns where the raw data is converted into activity count variables, which is further used to classify physical activity intensity and energy expenditure [3].

 doi:10.20944/preprints202107.0505.v1

A challenge in physical activity monitoring is automatic detection of which activity is taking place from inertial sensor measurements, such as to detect walking, lying, and standing up which can be done by using various deep learning algorithms namely deep feed-forward neural networks (DNNs) and deep recurrent neural networks (RNNs). This works because the deep learning is shown to be well suited for classification of complex data patterns. Thus, the main motivation of this paper is to explore the use of a deep learning approach for classifying physical activity movements using body-worn accelerometer data.

In this study, two datasets are used: the first one is the publicly available dataset which contains labeled accelerometer data recordings acquired from UCI Machine learning Repository [7]. It is a dataset for activities-of-daily-life (ADL) collected through wrist-worn accelerometers. The dataset has five categories and 13 daily living activities, namely brushing teeth, combing hair, climbing stairs, descend stairs, walking, drinking from a glass, pouring water into a glass, eating with a knife and fork, eating with a spoon, using telephone, getting out of the bed, lying down in the bed, standing up from a chair, sitting down on the chair. Note that some of the activities in this data are highly related to physical activities (e.g., walking), while others are not (e.g., brushing teeth). The second dataset is collected ourselves from eight voluntary participants wearing hip-worn accelerometers, which contains labeled accelerometer data recordings of ten movement patterns of ADL, where some of the movement patterns are different variations of a movement, such as different speeds of walking. The collected movement patterns including cycling, jogging, laying still, sitting, sitting in a vehicle, sitting relaxed, walking stairs, standing, walking fast, and walking normally.

The paper is organized as follows. Section 2 provides the existing literature in the field of human activity recognition from the body-worn accelerometer data. The considered research methodology, including data acquisition process, use of different classifiers' network architecture, and the descriptions of the publicly available dataset and the acquired dataset are presented in Section 3. Section 4 presents and discusses the experimental results obtained using two datasets. Finally, a conclusion and future research developments are given in Section 5.

2. Related Work

In the literature many researchers have applied deep learning in health care for classification [8–12], prediction [13,14] and diagnosis [15]. If we consider the literature on classification, recognizing, and predicting physical activities from body-worn accelerometer data, many attempts have been made over the past years. So, this section presents the existing literature on the use of deep learning for classifying physical activity using wearable sensor-based accelerometer data.

In [16], the researchers proposed an activity recognition scheme for classifying older people physical activity using machine learning algorithms namely Support Vector Machine (SVM) and Convolutional Neural Networks (CNN) on recordings from wearable sensors. The study results showed the classification accuracy of SVM model was 81.7% and CNN model was 82.47%.

In [17], the researchers' main objective was to design algorithms suitable for physical activities (PA) monitoring using a wrist-worn accelerometry device. To achieve the objective, the researchers used sixty participants who completed an ordered series of 10–12 semi-structured activities in the laboratory and outdoor environment and various machine learning algorithms, namely Support Vector Machine (SVM), decision trees, neural network, naive Bayes, logistic regression, and hidden Markov models for pattern recognition to classify PA. Before training and testing the model, all the features were standardized, and the SVM feature evaluation method was applied to select the feature set best for the classification. The metrics that were used to assess and compare the performance of the algorithms with respect to the correct classification of the PA were precision, recall, ROC, and F1-score. In their study, the results showed the classification accuracy of the decision tree model for right wrist GENEA data was 96.97% the left wrist GENEA data was 95.93%, and waist GENEA data was 99.14%. The classification accuracy of the naive Bayes model for right wrist GENEA data was 95.26% the left wrist GENEA data was 95.33%, and waist GENEA data was 98.12%. The classification accuracy of the logistic regression model for right wrist GENEA data was 95.65% the left wrist GENEA data was 95.73%, and waist GENEA data was 99.40%. The classification accuracy of the SVM model for right wrist GENEA data was 96.76% the left wrist GENEA data was 96.40%, and waist GENEA data was 99.30%. The classification accuracy of the neural network

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

model for right wrist GENEA data was 96.76% the left wrist GENEA data was 95.93%, and waist GENEA data was 99.60%.

In [18], the researchers' main objective was to design algorithms suitable for predicting physical activity type from wrist-worn accelerometer data. To achieve the objective, the researchers used a regularized logistic regression model and collected data from 52 children and adolescents who completed 12 activity trials that were categorized into 7 activity classes: lying down, sitting, standing, walking, running, basketball, and dancing.. In their study, features were extracted from 10-s windows and fed to the model. The study results showed the classification accuracy for the hip and wrist was 91.0% $\pm$ 3.1% and 88.4% $\pm$ 3.0%, respectively. The hip model exhibited excellent classification accuracy for sitting (91.3%), standing (95.8%), walking (95.8%), and running (96.8%); acceptable classification accuracy for lying down (88.3%) and basketball (81.9%); and modest accuracy for dance (64.1%). The wrist model exhibited excellent classification accuracy for sitting (93.0%), standing (91.7%), and walking (95.8%); acceptable classification accuracy for basketball (86.0%); and modest accuracy for running (78.8%), lying down (74.6%) and dance (69.4%).

In [19], the researchers worked on classifying physical activities such as sitting, walking, lying, stair climbing, standing running, cycling using on-body accelerometer data using single-frame classification algorithms: Naïve Bayesian (NB), SVM, Binary decision tree (C4.5), Gaussian Mixture Model (GMM), Logistic classifier, Nearest mean (NM), k-NN, parzen classifier, and ANN (multi-layer perceptron). The study results showed the classification accuracy for the NB model was 97.4%, GMM model was 92.2%, Logistic model was 94.0%, Parzen model was 92.7%, SVM model was 97.8% , NM model was 98.5%, k-NN model was 98.3%, ANN model was93.0%, C4.5 model was 93.0%.

In [20], the researchers developed and tested machine learning models for predicting activity type in preschool-aged children. Overall recognition accuracy for the standard feed-forward artificial neural network was 69.7%. Recognition accuracy for sedentary activities, light activities and games, moderate-to-vigorous activities, walking, and running was 82%, 79%, 64%, 36%, and 46%, respectively. In comparison, the overall recognition accuracy for the Deep Learning Ensemble Network was 82.6%. For sedentary activities, light activities and games, moderate-to-vigorous activities, walking, and running, recognition accuracy was 84%, 91%, 79%, 73%, and 73%, respectively. The lower accuracy for the classification of sedentary activities is to be expected because of the free-living nature of included PA types.

In [21], the researchers explored deep, convolutional, and recurrent neural network machine learning approaches across three representative datasets that contain movement data captured with wearable sensors for the identification of physical activities. The researchers used f1-score as a performance metric to classify the class distribution. The F1-score of the DNN model was 0.904 on the PAMAP2 dataset, 0.633 on the Daphnet Gait dataset (DG), 0.575 on the Opportunity (Opp) dataset. The F1-score of the CNN model was 0.937 on the PAMAP2 dataset, 0.684 on the Daphnet Gait dataset (DG), 0.591 on the Opportunity (Opp) dataset. The F1-score of their Long Short-Term Memory (LSTM) model was 0.882 on the PAMAP2 dataset, 0.760 on the Daphnet Gait dataset (DG), 0.698 on the Opportunity (Opp) dataset.

In [22], the researchers proposed the use of LSTM-based deep RNNs (DRNNs) to build human activity recognition models for classifying activities mapped from variable-length input sequences and developed architectures based on deep layers of unidirectional and bidirectional RNNs, independently, as well as a cascaded architecture progressing from bidirectional to unidirectional RNNs. These models were then tested on the benchmark datasets UCI-HAD, USC-HAD, Opportunity, Daphne FOG, Skoda to validate their performance and generalizability for a large range of activity recognition tasks. The study results on the UCI-HAD dataset showed that the unidirectional DRNN model achieved the 96.7% classification accuracy, CNN got 95.2%, SVM got 96.0%, sequential extreme learning machine got 93.3%. On the USC-HAD dataset, the researchers obtained classification accuracy with the unidirectional DRNN model was 97.8%, CNN got 97.0%, least squares-SVM got 95.6%, random forest got 90.7%. On the Opportunity dataset, the researchers obtained f1-score with the unidirectional DRNN model was 0.92, CNN and unidirectional RNN got 0.915, CNN got 0.883, SVM got 0.847, and deep belief networks (DBNs) got 0.745 scores. On the Daphne FOG dataset, the researchers obtained an f1-score of 0.93 with DRNNs, 0.89 with CNN, and 0.83 with

 doi:10.20944/preprints202107.0505.v1

k-NN on the Skoda dataset, the researchers obtained classification accuracy of 92.6% with DRNNs, 91.7% with CNN, 89.4% with DBNs, and 86.0% with HMMs.

In [23], the authors worked on comparing different the classification techniques k-nearest neighbors, Gaussian naive Bayes, linear discriminant analysis, stochastic gradient descent, support vector machines with several kernel functions, decision trees, random forest, extra randomized trees, adaptive boosting, and deep neural networks for automatic cross-person activity recognition and used the data from the PAMAP2 (Physical Activity Monitoring in the Aging Population) dataset collected from eight participants performing 12 activities with wearable devices. The experimental results show that with large training sets, they obtained very high average accuracies (e.g., 96% using extra randomized trees) with the best classifier. However, when the data volume was drastically reduced (where available data are only 0.001% of the continuous data), deep neural networks performed the best, achieving 60% in overall prediction accuracy.

In [24], the researchers presented an analysis of deep and shallow feature representations for accelerometer data on human activity recognition (HAR) with data collected from the wrist and thigh. They considered the three types of representations hand-crafted, frequency transform, and deep features, including two-hybrid approaches for HAR, and used CNN and introduced CNN hybrid approaches, i.e., CNN-SVM CNN-kNN. The classification performance of the models was measured using the F1-score. The results obtained show that the classification performance with data collected from the wrist got the highest F1-scores at 0.850 with CNN-SVM and 0.845 with CNN-kNN, respectively, and 0.839 with CNN. Whereas, the best F1-score results achieved for the data collected from the thigh by Discreet Cosine Transform (DCT) was 0.967, which was slightly better than achieved by the deep learning features of CNN, CNN-SVM, and CNN-kNN with scores of 0.959, 0.957, and 0.949, respectively.

In [25], the researchers' goal was to identify physical activities using a CNN accurately. The used large-scale exercise motion data were collected from a forearm-worn wearable sensor. In their study, the authors formatted the time-series data consisting of accelerometer and orientation measurements into images in order to allow the CNN to extract discriminative features automatically. Their experimental results show that the best performing configuration classifies 50 gym exercises with 92.1% accuracy.

In [26], the researchers' goal was to accurately and automatically recognize and classify physical activity through wearable sensors by utilizing a Deep Belief Network (DBN) model. The authors used a publicly available dataset from the UCI machine learning repository for the experiments. For the comparison purpose, they used the SVM model. The experimental results show that the DBN model achieved an accuracy of 97.5% with 40 hidden units. At the same time, SVM achieved an accuracy of 94.12%.

In [27], the researchers worked on a dataset of non-disabled participants, which was recorded using three custom wireless motion sensors for classifying the physical activities. To classify, the researchers used Support Vector Machine with Radius Basis Function Kernel (RBF-SVM) model, and the experimental results showed that they achieved a motion classification accuracy of 97.35% between 8 body motions.

In [28], the researchers proposed and deployed deep neural networks for human activity recognition (HAR) in the context of activities of daily living using multi-channel time-series. These time-series data were acquired from body-worn devices, which were composed of different types of sensors. The deep architectures process these measurements to find basic and complex features in human corporal movements and classify them into a set of human actions. The experiments were carried out on publicly available three different datasets. The first one is the Opportunity dataset, the second one is the Pamap2 dataset, and the third one is the industrial dataset. The experimental results showed that the classification accuracy and the weighted F1-score for the CNN architecture on the Opportunity-Gestures dataset were 92.24% and 0.92. On the Opportunity-Locomotion dataset, baseline CNN performed well with 90.07% accuracy performance and F1-score of 0.903, and on the PAMAP2 dataset, CNN with 93.68% and F1-score with 0.937.

In [29], the objective of the researchers was to investigate the performance of different artificial neural networks (ANN) for classifying the physical activities using two public databases, wearable action recognition database (WARD) and HAPT, which got downloaded from the UCI machine

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

learning repository. Their experiment results showed that the highest recognition rate of ANN with raw data was only 91.8% on the WARD while it was 96.8 on the HAPT dataset. When hyper-parameters were tuned, and with entire feature sets, the ANN model achieved approximately 88% on WARD and 80% on the HAPT dataset. Also, the researchers examined the effect of feature selection on the classification results; with the original feature set, the network obtained 97% accuracy on HAPT and 99.4% accuracy on WARD.

In [30], the authors presented an LSTM deep recurrent neural network for the classification of six daily life activities from accelerometer and gyroscope data which was acquired from a waist-mounted inertial sensor recorded at 50 Hz sampling frequency. Their experimental results show that the LSTM model achieved 92% average accuracy in a multi-class scenario.

In [31], the researchers used convolutional neural networks to classify physical activities by using raw data obtained from a set of inertial sensors. To evaluate the chosen classifier, the authors use precision and recall scores for every combination of sensors and activities and then computed the F-scores. Their experimental results show that for the walk group activities, two majority classes, Bawk and Sdwk, obtained an F1-score of 0.96 and 0.97, respectively, than the other activities.

Unlike the above works, our work in this paper classifies PAs using two different deep learning algorithms, i.e., DNN and RNN. In addition, when compared to the results mentioned above, we divided the raw data into three other sliding windows, such as three, five, and ten seconds windows. Further, in [22], the researchers used the RNN algorithm for the classification. However, the datasets they used are different from the datasets we used in this paper. So, our contribution in this paper is not comparable.

3. Materials and Methods

In this section, we present the proposed methodology including data collection and the used classifiers' network architecture.

3.1. Data collection

In this study, the classification of movement patterns is done on two datasets. The first is a publicly available dataset which contains labelled accelerometer data recordings acquired from UCI Machine learning Repository [7]. It is a dataset for activities-of-daily-life (ADL) collected through wrist-worn accelerometers including 5 categories and 13 daily living activities namely brushing teeth, combing hair, climbing stairs, descend stairs, walking, drinking from glass, pouring water into glass, eating with knife and fork, eating with spoon, using telephone, getting out of the bed, lying down in the bed, standing up from chair, sitting down on chair. This data set is here by referred to as the wrist-worn dataset.

The second dataset is collected ourselves through hip-worn accelerometers from eight voluntary participants which contains labelled accelerometer data recordings of ten movement patterns of free-living ADL, where some of the movement patterns are different variations of a movement, such as different speeds of walking. The collected movement patterns are cycling, jogging, laying still, sitting, sitting in a vehicle, sitting relaxed, walking stairs, standing, walking fast, and walking normal. Before using the accelerometers, the participant were given instructions to perform activities in their own way without specific constraints. The second dataset is here by referred to as the hip-worn dataset.

3.2. Data pre-processing

Data pre-processing is one of the most important steps in the data mining process. It consists of filtering data, replacing the missing and outlier's values and extracting/selecting features. To extract features from raw data, windowing techniques are generally used, which consist of dividing sensor signals into small time segments. Segmentation and classification algorithms are then applied respectively to each window.

Three types of windowing techniques are usually used: (i) sliding window where signals are divided into fixed-length windows; (ii) event-defined windows, where pre-processing is necessary to locate specific events, which are further used to define successive data partitioning and (iii) activity-defined windows where data partitioning is based on the detection of activity changes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

The sliding window approach is well-suited to real-time applications since it does not require any pre-processing treatments [32].

In this study, we used the sliding window technique to extract features from raw data. The raw data is divided into three types of sliding windows of three-, five-, and ten second windows respectively. More details on the sliding window process is described in subsections 3.2.1 and 3.2.2.

3.2.1. Wrist-worn Dataset

The wrist-worn dataset consists of 14 different motion primitives which were performed by 16 volunteers. To test and examine the performance of the proposed algorithms, these motion primitives are divided into broader, less complex, ADL categories. The correlation between these ADL categories and the motion primitives are shown in Table 1.

Table 1: Correlation between ADL categories and Motion primitives

ADL	Motion Primitives
Personal hygiene	Brush teeth
	Comb Hair
Mobility	Climb Stairs
	Descend Stairs
	Walk
Feeding	Drink from glass
	Pour water into glass
	Eat with knife and fork
	Eat with spoon
Communication	Use telephone
Functional transfers	Get out of bed
	Lie down in bed
	Stand up from chair
	Sit down on chair

The distribution of the data-points for these ADL's and motion primitives are shown in Figure 1, where the ADL distribution is on the left and the motion primitives on the right. When we looked at the ADL distribution, it can be seen that it is not evenly distributed as three categories; mobility (1), feeding (2) and functional transfers (4) which consists of above 100000 data-points each. While personal hygiene (0) and communication (3) only consists of about 50 000 and 20 000 data-points respectively. The overall distribution of the movement primitives can be considered somewhat even, with the exceptions of walking, eating with spoon, and lying down, as most of them are within the range of 20 000 to 40 000 data-points.

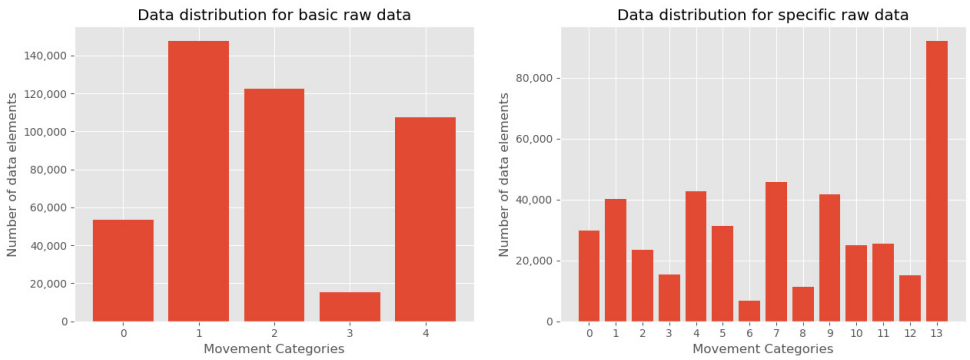

Figure 1. Raw data distribution for the wrist-worn dataset

As mentioned above, the raw data is divided into sliding windows i.e., a three second, a five second- and a ten second- sliding window in order to be used for training the algorithms. Whenever the accelerometers gather the data, it collects 32 data points each second.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

In the sliding windows, the three second sliding window differs from the other two sliding windows, as the shift for each window also is set to three seconds. It is created by collecting three seconds of data (which is a total of 96 data-points) then sliding three seconds to create the next sliding window. Thus, there is no overlapping of data points within this type of sliding window. Dividing the dataset into these three second sliding windows yields a dataset consisting of approximately 4000 sliding windows distributed.When generating the five second sliding windows, the shift for each window is set to one second. Thus, each sliding window contains an overlap of four seconds to the previous window. This yields a dataset of approximately 10 000 sliding windows.

The final type of sliding window is ten seconds of data, also sliding by one second for each window. This creates a dataset consisting of approximately 6000 sliding windows.

3.2.2. Hip-worn dataset

The hip-worn dataset consists of ten movement patterns performed by eight volunteers, where some of the movement patterns are different variations of a movement including different speeds of walking. The collected movement patterns are cycling, jogging, laying still, sitting, sitting in a vehicle, sitting relaxed, walking stairs, standing, walking fast, and walking normal. This dataset consists of 1.6 million data-points distributed among the categories as shown in Figure 2.

Similar to the wrist-worn dataset, the hip-worn dataset is also divided into three types of sliding windows i.e., three-, five-, and ten second windows respectively. The hip-worn dataset is unevenly distributed and most of the data-points are in the relaxing categories lying down, and sitting. This pattern in the distribution is carried over to the sliding window datasets. Generating the dataset for three second sliding windows is done same as the wrist-worn, three seconds of data shifting is done with three seconds. However, the accelerometers used on this dataset collects 30 data-points each second, giving the three second sliding window dataset around 15000 sliding windows with the distribution.

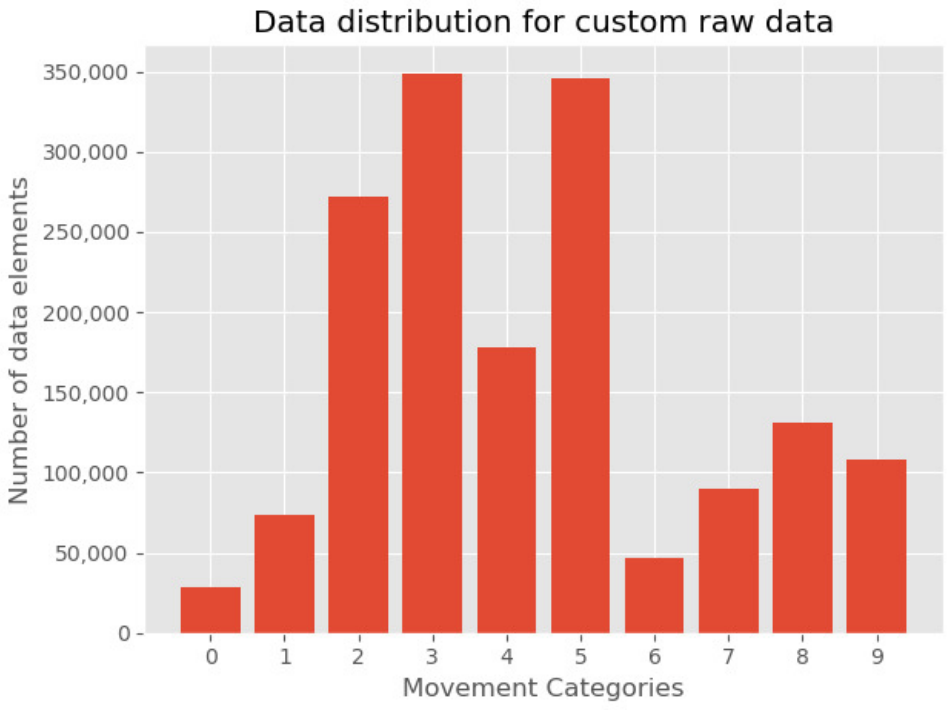

Figure 2. Raw data distribution for the hip-worn dataset

The dataset is created with five- and ten- second sliding windows with a shift of one second, the same as for the wrist-worn dataset. Both types of sliding window creates a dataset with approximately 54 000 sliding windows. After the data pre-processing, the data is fed to the deep learning algorithms. The description of the used deep learning networks architecture can be seen in the below subsection 3.3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

3.3. Deep learning networks architecture

In this paper, two deep learning algorithms DNNs and RNNs are used for classifying the activity movement types. The network structure of both used DNNs and RNNs is described below.

The first proposed network structure is the deep feed-forward network which is shown in Figure 3a. The model consists of five layers i.e, an input and output layer, and three hidden layers. The three hidden layers have a constant number of cells, more specifically, the first hidden layer has 512 cells, the second hidden layer has 258 cells, and the last hidden layer has 128 cells. While the input and the output layer adapt to the dataset. The input layer will always fit the length of the samples, while the output layer will always have n cells, where n is the number of classes that need to be classified.

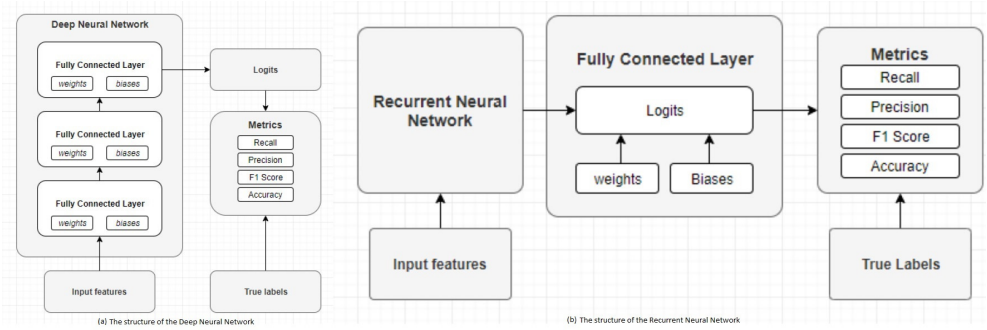

Figure 3. The structure of the Deep Neural Network and Recurrent Neural Network models

During the training phase, the input data is split up into smaller batches of 100 elements. The model will predict a label for each element and the weights and biases are adjusted from the sum of the errors between the predicted labels and true labels. The testing dataset is not split up into smaller batches. Thus, the entire testing dataset is used for calculating the performance metrics.

The second proposed network structure is the recurrent neural network which consists of two components, the recurrent element of the network and a fully connected network. The first component, the recurrent element, consists of cells which in our case are Gated Recurrent Unit's (GRU). Each of these cells are connected to each other and they all have a fixed number of states. The number of cells, and the number of states each cell has is provided as a parameter to the network. The recurrent element of the network returns two values the current state of the recurrent network S_t, and the output of each state O_t. The output of each state will be an array of length t, where t is the number of items in the time-series that is sent into the network.

The last element of the outputs is used and sent forward to the fully connected network. The fully connected network is much smaller than the previous model and only consists of two layers. The input layer, which takes the input from the recurrent network and an output layer that is used to predict the class. A simplified illustration of the recurrent neural network model is shown in Figure 3b.

4. Results

The results of our proposed models are presented and discussed in this section. The results are compared to other state-of-the-art methods. In this work, we have done several types of experiments by using four different learning rates i.e., *0.1*, *0.01*, *0.001* and *0.0001*. These learning rates are tested in combination with three well known optimizers i.e., Adagrad, Adam and Stochastic Gradient Descent (SGD). Furthermore, these combinations of optimizers and learning rates are tested for all three types of sliding windows i.e., three, five and ten seconds, both for the basic and specific movement categories. Considering the number of data-points available in the datasets, all experiments are stopped after 2000 training steps. Thus, reducing the chance of over-fitting through repeatedly training on the same data.

4.1. Deep Artificial Neural Network

Throughout this section, a detailed explanation of the best performing results with regards to the overall F1-score is given. For each type of sliding window (three-, five- and ten- seconds), the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

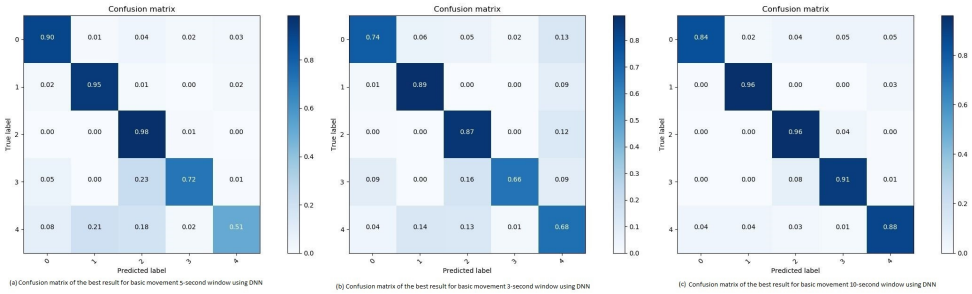

Figure 4. Confusion matrix of the best result for basic movement 3, 5 and 10-second window using *DNN*

best performing combination of optimizer and learning rate will be given for both the basic and specific movement categories.

Low learning rates are shown to be more suited for movement recognition, for both the Adagrad- and the Adam- optimizer. When using a high learning rate of 0.1, both the Adagrad and Adam optimizer get a low F1-score. The recall of these results shows that the algorithms guesses "randomly". The probable cause of this might be the high learning rate. Thus, the algorithm diverges from the global minimum, which prevents it from learning the patterns of the movements. Furthermore, the result of the SGD optimizer is to some extent unexpected. However, the result can be explained through the proposed structure of the DNN. The model uses Rectified Linear Unit (ReLU) as its activation function. Using ReLU with SGD often leads to vanishing gradients, which stagnates the learning of the algorithm.

4.1.1. Experiments with Wrist-Worn Dataset- Basic Movement

Three second sliding window: Testing the different hyper-parameter combinations on the three second sliding window shows the highest achieved overall accuracy and F1-score are 80.94% and 78.46% respectively. This result is reached when using a low learning rate 0.0001 with the Adam optimizer.

Figure 4b shows a confusion matrix for the results of each movement type. This matrix is created to evaluate the algorithm, and interpret why the algorithm achieved its results and where the incorrect classifications occurs. The recall(R), precision (P), and F1-score (F1) percentages about the classification of the movement types with 3-second window are shown in Table 2.

The classification inaccuracies for three seconds sliding windows are understandable considering the data distribution.The most distinct pattern in the matrix is that each category is to some extent inaccurately recognized as a functional transfer, even without it being the category with the largest amount of data-points. However, the probable cause of this pattern is the fact that the data is collected through a wrist-worn accelerometer. Each of the 14 participants might have different personal traits when moving, especially hand movements. This could affect the classification. Thus, giving functional transfers a lower precision score shown in Table 2.

The ADL category which is inaccurately classified the most is the communication category, which contains the least amount of data-points. This is in line with what we expect, given the nature of the activities and placement of the sensor. Furthermore, when it is incorrectly classified, it is often interpreted as similar movement patterns, either hygiene or feeding. Functional transfers is the category with the second highest percentage of inaccurate classification. Again, some of these classification might be affected by traits of the participants, thus classifying it as feeding. On the other hand, functional transfers are often classified as mobility, which is a more comparable category with regards to the movement types in those categories.

Five second sliding window: Using the Adam optimizer combined with a low learning rate of 0.0001, on the five second sliding windows, the proposed DNN achieves an overall accuracy of 86.01% and an overall F1-score of 82.37% and can be seen in Table 5 . There are clear patterns in the inaccurate classifications of the ADL categories shown in the confusion matrix in Figure 4a. The recall(R), precision (P), and F1-score (F1) percentages about the classification of the movement types with 5-second window are shown in Table 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

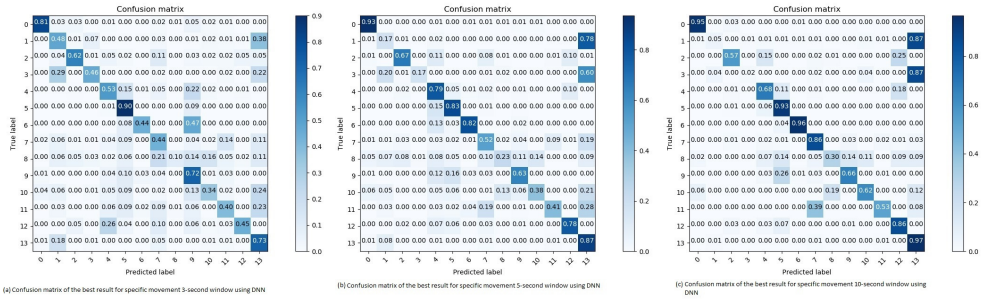

Figure 5. Confusion matrix of the best result for specific movement 3,5,10- seconds window using *DNN*

The DNN model incorrectly classifies the communication data as feeding (20% of the time), similar to the three second sliding window. Considering the resemblance in the movement patterns when using a telephone and eating, this confusion is understandable, especially when seeing the difference in the data distribution.

Explaining the miss-classification of the functional transfer category is partially based on assumptions. The two most commonly classified classes are mobility and feeding. As mentioned for the three second sliding window, the specific movement types in the mobility class and the functional transfer class (see Table 1) are comparable. It is understandable that getting up from a chair or the bed is interpreted as one of the mobility movements. However, looking at possible reasons for the DNN to interpret functional transfers as feeding, the most reasonable explanation would, as mentioned, be the placement of the accelerometer.

Ten second sliding window: The Adam optimizer and a low learning rate of 0.0001 is again the best performing combination when testing it on ten seconds sliding windows. Achieving an impressive overall accuracy of 92.4% and an overall F1-score of 89.43% (see Table 5). Looking at the confusion matrix in Figure 4c, it is clear to see that a ten second sliding window allows the DNN model to observe and distinguish the differences between these basic movement patterns. The probable cause of the lower F1-score of communication is its low number of data-points, which influences its precision score. Thus, predicting 5% of the hygiene data-points as communication heavily affects the precision score. The recall(R), precision (P), and F1-score (F1) percentages about the classification of the movement types with 10-second window are shown in Table 2.

4.1.2. Experiments with Wrist-Worn Dataset - Specific Movement

As the number of movement types are increased in the specific movement distribution of the UCI dataset, the complexity of recognizing them also increases. An important point given that we ultimately want to predict all types of free-living (or at least the predominant types) activities with acceptable precision. In addition, the distribution of data-points is lower for each categories as they are no longer combined as for the basic distribution.

Three second sliding window: The best result achieved for three second sliding windows for specific movement is an accuracy of 59.93% and a F1-score of 56.59% (see Table 5). This result is achieved by the Adagrad optimizer with a learning rate of 0.01. Analyzing the confusion matrix in Figure 5a, the most difficult movement to recognize is lying down in bed. The recall(R), precision (P) and F1-score (F1) percentages about the classification of the specific movement types with 3-second window are shown in Table 2. Considering that it is the movement with the second lowest amount of sliding windows.Thus, it is to some extent expected.

The other reasonable miss-classified movement patterns are full body movements. Both climbing and descending stairs are often wrongly interpreted as walking, which is understandable as walking is the largest category in terms of data-points and the movements are comparable, which again is not surprising since, from a bio-mechanical view, both sensor signals are very similar. In addition, the two functional transfers, sitting down and standing up, are also interpreted as walking. Furthermore, descending stairs is either wrongly recognized as either walking or descending stairs. The accelerometer registers the g-forces, and as it is placed on the wrist. Thus, the patterns on climbing and descending stairs are relatively equal as the axes changes when rotating the hand.

Table 2: Best result of basic and specific movement of wrist-worn dataset 3,5,10-second window using *DNN*

Category	Basic movement with DNN								
	3-W[1]			5-W[2]			10-W[3]		
	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]
Hygiene	73.7%	83.61%	78.35%	89.69%	84.33%	86.93%	84.1%	96.49%	89.87%
Mobility	89.33%	88.7%	89.01%	94.53%	89.38%	91.88%	95.95%	97.51%	96.72%
Feeding	87.13%	84.44%	85.76%	98.37%	83.69%	90.43%	95.87%	93.67%	94.75%
Communication	66.23%	79.69%	72.34%	71.72%	72.82%	72.26%	91.3%	70.59%	79.62%
F-Transfer	67.57%	64.6%	66.05%	50.54%	88.89%	64.44%	87.74%	81.4%	84.45%
	Specific movement with DNN								
	3-W[1]			5-W[2]			10-W[3]		
	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]
Brush teeth	81.29%	90.0%	85.42%	93.49%	91.56%	92.52%	94.77%	98.82%	96.75%
Climb stairs	48.07%	39.55%	43.39%	16.95%	32.72%	22.33%	4.9%	35.0%	8.59%
Comb hair	61.82%	79.07%	69.39%	17.29%	79.31%	28.4%	57.26%	84.81%	68.37%
Descend stairs	46.15%	57.69%	51.28%	17.29%	79.31%	28.4%	2.17%	33.33%	4.08%
Drink	52.74%	64.63%	58.08%	79.06%	59.97%	68.2%	68.28%	60.77%	64.3%
Eat w/knife&fork	90.0%	60.71%	72.51%	83.3%	74.58%	78.7%	93.01%	82.06%	87.19%
Eat w/ spoon	44.44%	59.26%	50.79%	44.44%	59.26%	50.79%	95.79%	93.81%	94.79%
Get out bed	43.75%	49.46%	46.43%	51.93%	68.18%	58.96%	85.66%	83.27%	84.44%
Lie down bed	9.52%	66.67%	16.67%	23.42%	37.68%	28.89%	29.55%	61.9%	40.0%
Pour water	72.14%	52.73%	60.92%	63.38%	80.12%	70.77%	66.48%	85.21%	74.69%
Sit down	34.12%	46.03%	39.19%	38.06%	68.6%	48.96%	62.5%	60.61%	61.54%
Stand up	39.58%	39.58%	39.58%	40.67%	49.19%	44.53%	52.78%	55.88%	54.29%
Telephone	44.87%	77.78%	56.91%	78.26%	66.12%	71.68%	85.54%	58.44%	69.44%
Walk	72.75%	67.41%	69.98%	87.47%	64.0%	73.91%	96.55%	83.3%	89.44%

[1] Three-second window
[2] Five-second window
[3] Ten-second window
[a] Recall
[b] Precision
[c] F1-score

doi:10.20944/preprints202107.0505.v1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

Another conspicuous miss-interpretation is when the algorithm recognizes eating with spoon as pouring water into a glass. The assumption here is that the collected data of pouring water is from a mug. Thus, the resemblance in hand movements affects the algorithm. As an example, when eating soup one slowly moves the spoon from the plate, up towards the mouth before one tilts the spoon inside the mouth. The same pattern goes for pouring water from a mug into a glass. First the mug is lifted, then tilted to pour the water.

Five second sliding window: Testing the DNN on the five second sliding windows, the highest performing hyper-parameter combination is the Adagrad optimizer and a learning rate of 0.01. Achieving an overall accuracy of 67.83% and an overall F1-score of 62.11% (see Table 5). The results of the five second sliding windows are similar to the results of the three second sliding windows, where most of the movement categories have a slight increase in their recall score (see Table 2). As mentioned in subsection 3.2.1 3.2.1, each five second sliding window consists of some overlapping data. Allowing the algorithm to more easily recognize the patterns of the movements. However, the movements which were miss-interpreted as walking are more frequently interpreted incorrect. This, pattern has a correlation to the distribution, as these movement categories, climbing and descending stairs, have fewer sliding windows for five seconds than for three seconds. Thus, the algorithm has fewer samples to train on, which affects the result and this can be seen in the confusion matrix results in Figure 5b.

Ten second sliding window: For ten second sliding windows, Adagrad combined with a learning rate of 0.01 gives an overall accuracy of 81.29% and an F1-score of 66.75% (see Table 5). The recall(R), precision (P) and F1-score (F1) percentages about the classification of the specific movement types with 10-second window are shown in Table 2.

Looking at the results shown in the confusion matrix in Figure 5c, it follows the same patterns as the five second sliding window. The recall score averagely increases, due to the increase of the window size. As the window size increase, the amount of sliding widows decrease for the climbing- and descending- stairs movements. Thus, further decreasing their recall score, as they are more consistently interpreted as walking. The most distinguish pattern in the ten second sliding window confusion matrix is the miss-interpretation of standing up. It is approximately 40% of the time interpreted as getting out of bed. Indicating that the similarity of getting out of the bed, and standing up from a chair is quiet high when monitoring ten seconds of those movements.

4.2. Recurrent Neural Network

In this section, we discuss the best results of the Recurrent Neural Network (RNN) experiments. As for the DNN, a detailed explanation of the best combination of hyper-parameters is given for each sliding window, for both the basic and specific movement categories. In addition, the results for the secondary dataset, and the hip-worn data are explained.

The hyper-parameters used for the RNN are same as the DNN. However, two additional parameters are tested i.e., number of cells in the RNN, and the size of the cells. The RNN is run with four and eight cells for each of the learning rates. Furthermore, these were both tested with two different sizes for the cells i.e., 16 and 32 respectively.

Looking at the result overviews, lower learning rates often perform better than higher. However, Adagrad performs surprisingly bad with the lowest learning rate. The assumption here is that the number of training steps are to low for the Adagrad optimizer to learn the patterns with such a low learning rate. The same pattern for the SGD optimizer are shown for the RNN; SGD does not learn with the ReLU cells, except it is able to achieve relatively impressive results with the lowest learning rate.

4.2.1. Experiments with Wrist-Worn Dataset - Basic Movement

The best performing results with regards to the overall F1-score for the basic movements for each sliding window type are explained below.

Three second sliding window: Running a RNN, consisting eight cells of size 32, with the Adam optimizer and a low learning rate of 0.001, an overall accuracy of 84.89% and an overall F1-score of 82.56% is achieved. When these results are compared to the DNN results (see Table 2,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

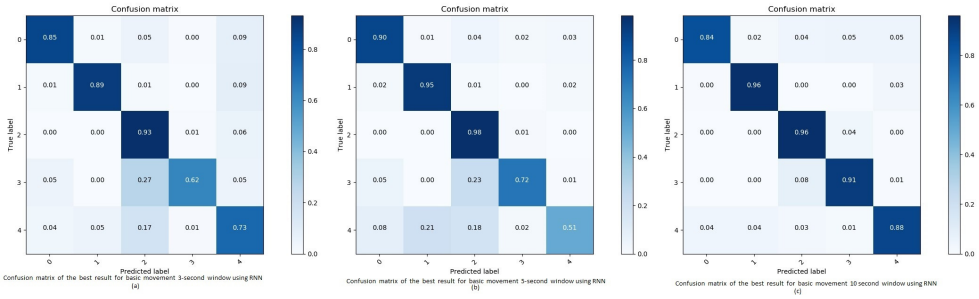

Figure 6. Confusion matrix of the best result for basic movement 3,5,10- second window using *RNN*

there is an increase of approximately 4%.

Looking at the confusion matrix in Figure 6a, the results again shows that whenever a sliding window is miss-interpreted it is often guessed as a hand movement. As an example, the category with the lowest amount of sliding windows, communication, is often guessed as feeding. This is understandable because of the similarity in movements, and also due to the fact that feeding is the category with the second most sliding windows.

The other noticeable result is the interpretation of functional transfers (see Table 3). As discussed in previous sections, the placement of the accelerometer can be used as an explanation. Placing it at the wrist is probably affecting the movement pattern, as a small hand gesture can have an influence on the algorithms. Thus, the assumption is that functional transfers is classified as feeding due to the accelerometer placement. The recall(R), precision (P), and F1-score (F1) percentages about the classification of the basic movement types using RNN for 3-second window are shown in Table 3.

Five second sliding window: Using a learning rate of 0.01 and the Adagrad optimizer with the five second sliding windows achieves an impressive 94.65% overall accuracy and a 93.03% overall F1-score which is tabulated in Table 5.The recall(R), precision (P) and F1-score (F1) percentages about the classification of the basic movement types using RNN for 5-second window are shown in Table 3. The only result standing out in the confusion matrix in Figure 6b is the communication category. However, the "low" accuracy is explained by checking at the distribution.which consisting of the fewest sliding windows, communication is expected to be the hardest class to predict for the RNN.

Ten second sliding window: The highest performing combination of hyper-parameters for ten second sliding windows, is also the highest performing result for basic movements in general. Combining the Adam optimizer with a learning rate of 0.001, in a RNN with 4 cells of size 32, an overall accuracy of 98.75% and an overall F1-score of 98.06% is achieved (see Table 5). The recall(R), precision (P), and F1-score (F1) percentages about the classification of the basic movement types using RNN for 10-second window are shown in Table 3.

Considering the overlapping of data-point in the sliding windows, and the fact that RNN uses prior knowledge to improve. These results are expected. The few miss-interpreted sliding windows are assumed to be caused by the "noise" from the placement of the accelerometer. Figure 6c shows the confusion matrix of the best result for basic movement of 10-second window using RNN.

4.2.2. Experiments with Wrist-Worn Dataset - Specific Movement

Throughout this section we explain and discuss the results of the best performing hyper-parameters for all three types of sliding windows for specific movement.

Three second sliding window: The results of the RNN for three second sliding window, are noticeably better than for the DNN. A RNN with four cells, with size 32, a learning rate of 0.001 and the Adam optimizer, gives an overall accuracy of 70.39% and an overall F1-score of 65.58% (see Table 5). One of the main differences between the results of the RNN and the DNN are both descending and climbing the stairs. The recall(R), precision (P) and F1-score (F1) percentages about the classification of the specific movement types using RNN for 3-second window are shown in Table 3. Figure 7a shows that the RNN has reduced the number of miss-interpretations of stair

Table 3: Best result of basic and specific movement of wrist-worn dataset 3,5,10-second window using *RNN*

Category	Basic movement with DNN								
	3-W[1]			5-W[2]			10-W[3]		
	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]
Hygiene	84.81%	88.42%	86.58%	96.95%	93.73%	95.31%	99.85%	98.79%	99.32%
Mobility	88.62%	95.61%	91.98%	94.86%	98.36%	96.58%	99.58%	99.08%	99.33%
Feeding	93.12%	80.86%	86.56%	97.16%	95.66%	96.41%	99.66%	98.35%	99.0%
Communication	62.34%	85.71%	72.18%	83.33%	91.67%	87.3%	92.93%	99.42%	96.07%
F-Transfer	73.22%	73.53%	73.38%	90.85%	87.79%	89.29%	94.71%	98.27%	96.45%
	Specific movement with RNN								
	3-W[1]			5-W[2]			10-W[3]		
	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]
Brush teeth	83.87%	86.09%	84.97%	96.01%	96.21%	96.11%	99.32%	99.77%	99.54%
Climb stairs	72.38%	71.2%	71.78%	88.78%	75.15%	81.4%	92.31%	85.71%	88.89%
Comb hair	69.09%	83.52%	75.62%	92.54%	92.23%	92.39%	98.72%	98.72%	98.72%
Descend stairs	75.38%	87.5%	80.99%	81.2%	81.82%	81.51%	91.3%	84.0%	87.5%
Drink	56.72%	73.08%	63.87%	90.42%	81.36%	85.65%	95.16%	94.65%	94.91%
Eat w/knife&fork	92.35%	73.02%	81.56%	92.9%	87.77%	90.26%	98.94%	99.36%	99.15%
Eat w/ spoon	63.89%	50.0%	56.1%	94.68%	89.9%	92.23%	100%	95.0%	97.44%
Get out bed	44.23%	57.86%	50.14%	79.23%	77.34%	78.27%	90.16%	94.02%	92.05%
Lie down bed	11.11%	36.84%	17.07%	52.25%	63.04%	57.14%	68.18%	85.71%	75.95%
Pour water	74.13%	65.35%	69.46%	88.03%	84.84%	86.41%	97.25%	95.68%	96.46%
Sit down	48.24%	45.56%	46.86%	72.9%	68.48%	70.62%	100%	88.89%	94.12%
Stand up	57.29%	37.67%	45.45%	57.33%	69.92%	63.0%	94.44%	85.0%	89.47%
Telephone	83.33%	66.33%	73.86%	54.11%	95.73%	69.14%	93.98%	93.98%	93.98%
Walk	86.15%	84.3%	85.22%	89.14%	93.84%	91.43%	97.49%	98.42%	97.95%

[1] Three-second window
[2] Five-second window
[3] Ten-second window
[a] Recall
[b] Precision
[c] F1-score

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

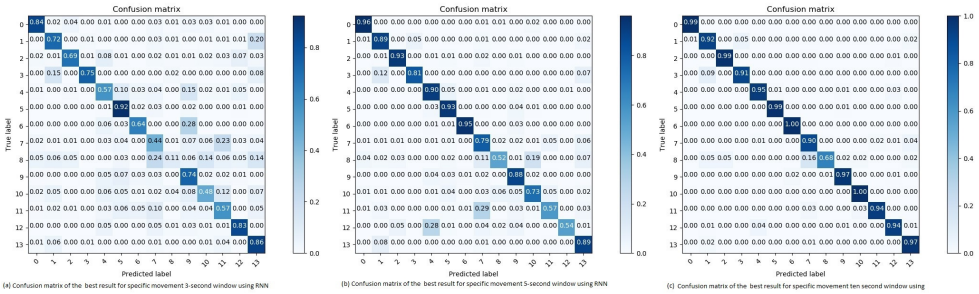

Figure 7. Confusion matrix of the best result for specific movement 3,5 and 10- second window using *RNN*

movements as walking. Thus, we assume that the "memory" of the RNN are able to remember the small differences between walking and moving upwards or downwards.

Many of the miss-interpreted movements are movements of similar types, mostly different hand movements. Thus, some sliding windows are interpreted incorrectly as another hand-movement. Examples of such miss-interpretations are drinking which may be interpreted as pouring water, Eating with a spoon which may be interpreted as pouring water. These movements are all hand-gestures which understandably can be confused with each other, and not a mis-classification of high relevance since it is of little importance to physical activity epidemiology.

The bad results of body movements from lying down, sitting down, and standing up has a low precision score due to their low data distribution.Considering the low amount of sliding windows, these classes are misinterpreted as a few different categories. However, they are mostly interpreted as similar movements. Lying down as either getting out of bed, sitting down or walking, sitting down as standing up, and standing up as getting out of bed.

Five second sliding window: An overall accuracy of 85.6% and an overall F1-score of 81.67% is the highest performing result of five second sliding windows with RNN (see Table 5). The RNN used to accomplish these results is a network consisting on 4 cells of size 32. This network uses Adam as its optimizer and a learning rate of 0.01.

The result of the five second sliding windows are overall increased compared to the three second sliding windows. This is probably due to the overlap in the sliding windows, which further allows the network to recognize the differences in the movement patterns. Again, the worst performing categories are the ones with the lowest distribution of sliding windows. Thus, they are interpreted as movements with similar patterns to its own. The recall(R), precision (P), and F1-score (F1) percentages about the classification of the specific movement types using RNN for 5-second window are shown in Table 3. Figure 7b shows the confusion matrix of the best result for specific movement five second window using RNN.

Ten second sliding window: Ten second sliding windows are again the highest performing distribution of the data-points. Achieving an overall accuracy of 96.52% and an overall F1-score of 93.43%, when using Adam as the optimizer, a learning rate of 0.001 on a RNN consisting of 8 cells of size 32 and this is seen in Table 5. The recall(R), precision (P), and F1-score (F1) percentages about the classification of the specific movement types using RNN for 10-second window are shown in Table 3.

Figure 7c shows the confusion matrix of the best result for specific movement 10-second window using RNN. Most miss-interpretations are eliminated with the exception of lying down, the category with the fewest sliding windows. There are some concerns to these results as the small shift of one second for each ten seconds sliding window might cause the algorithms to over-fit during training. Thus, achieving such impressive results. The problem however is the size of the dataset, increasing the shift of the sliding window drastically reduces the number of sliding windows in the dataset. Leading to poor training as the size of the dataset would be low.

4.2.3. Experiments with Hip-Worn Dataset

In this section, we discuss the highest performing results for the RNN on the dataset which is collected ourselves from voluntary participants. This dataset is also tested for each type of sliding window, which are separately discussed throughout this section. The hip-worn dataset is tested using

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

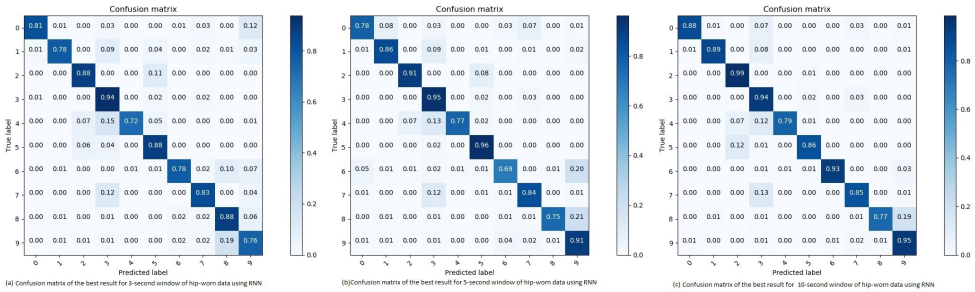

Figure 8. Confusion matrix of the best result for 3,5,10-second window of hip-worn data using *RNN*

only the proposed RNN model, with the different hyper-parameter combinations discussed above. The decision to not run this dataset through the proposed DNN model is the fact that the results of the wrist-worn dataset shows that the RNN model consistently outperforms the DNN for this type of time series movement patterns.

Three second sliding window: The highest performing result for the three second sliding window gets an overall accuracy of 85.5%, and a F1-score of 84.04% (see Table 5). Examining the results shown in the confusion matrix in Figure 8a, the wrongly interpreted sliding windows are reasonable. Examples are laying down which is interpreted as sitting relaxed, which in some cases might be a person almost lying in a sofa. Walking is sometimes confused with walking fast, which might be explained by differences in walking speed between participants. One persons normal walking speed might be the same speed as another persons speed when walking fast. Thus, may be confusing the algorithm.

Looking at the categories which are shown in Table 4, and their individual performances which are shown in Figure 8a, the lower performing category is sitting in a vehicle. As there are multiple vehicle options, it can be hard to predict this category. For instance, if one of the participants where sitting in a bus, then the three second sliding window have a possibility to be when the bus is at a stop. Thus, making the algorithm believe it is a person who is sitting still. The results explained above are achieved through a RNN with 4 cells, with cell sizes of 32, a learning rate of 0.1 and Adagrad as its optimizer. The recall(R), precision (P), and F1-score (F1) percentages about the classification of the specific movement types using RNN for 10-second window are shown in Table 3.

Five second sliding window: Using Adagrad as the optimizer for a RNN with four cells, of size 32, combined with a 0.1 learning rate, an accuracy of 88.48% and a F1-score of 85.29% is obtained (see Table 5). Again the category of sitting in a vehicle is among the lowest performing categories, as different vehicles have different driving patterns which might confuse it with other categories, and vibration from the vehicle and ground may introduce noise and since the logs are self-reported, we cannot rule out that some reporting is unreliable. However, looking at the results of both the category for walking stair and walking fast, shown in the confusion matrix in Figure 8b, their recall score decreased compared to the three second sliding window. The recall(R), precision (P) and F1-score (F1) percentages about the classification of the hip-worn data using RNN for 5-second window are shown in Table 4. We assume that the patterns between walking stairs and walking in normal speed are easier to distinguish when using three seconds of data compared to five. Thus, the results gets worse for five second sliding window within this category.

Ten second sliding window: A clear pattern in the results is that ten second sliding windows performs better compared to the three- and five- second sliding windows. Testing the RNN with the hip-worn dataset is no exception. When combining a learning rate of 0.01 with a RNN with four cells, with a size of 32, and using Adam as the optimizer, an accuracy of 89.31% with a F1-score of 89.36% is achieved (see Table 5).

The results shown in the confusion matrix in Figure 8c are impressively high. Most categories are interpreted correctly more than 85% of the time with the exception of sitting in a vehicle and walking fast. Other incorrectly interpreted sliding windows are confused with related categories, such as standing interpreted as sitting, and sitting relaxed as laying still.The recall(R), precision (P),

Table 4: Best result of 3,5,10-second window of hip-worn data using *RNN*

Category	Movements in hip-worn data with RNN								
	3-W[1]			5-W[2]			10-W[3]		
	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]	R[a]	P[b]	F1[c]
Cycling	81.33%	82.99%	82.15%	78.08%	78.74%	78.41%	87.55%	89.67%	88.6%
Jogging	77.67%	94.28%	85.17%	85.82%	89.39%	87.57%	89.41%	96.54%	92.84%
Laying still	87.8%	86.82%	87.31%	90.61%	94.86%	92.69%	98.52%	82.88%	90.02%
Sitting	93.73%	83.56%	88.36%	94.84%	86.27%	90.35%	94.27%	87.44%	90.73%
Sitting in a vehicle	72.09%	98.53%	83.26%	76.81%	97.51%	85.93%	79.01%	98.36%	87.63%
Sitting relaxed	88.19%	85.55%	86.85%	96.19%	90.63%	93.33%	85.97%	96.12%	90.76%
Walking stairs	77.96%	83.04%	80.42%	68.61%	78.83%	73.37%	93.18%	94.41%	93.79%
Standing	83.27%	82.45%	82.85%	84.24%	80.89%	82.53%	85.09%	81.83%	83.43%
Walking fast	87.97%	81.27%	84.49%	74.51%	98.56%	84.86%	77.3%	98.55%	86.64%
Walking normal	75.63%	77.2%	76.41%	90.78%	70.02%	79.06%	94.69%	76.51%	84.63%

[1] Three-second window
[2] Five-second window
[3] Ten-second window
[a] Recall
[b] Precision
[c] F1-score

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2021 doi:10.20944/preprints202107.0505.v1

Table 5: Summary of the *DNN* and *RNN* results

DNN			
Dataset	**Sliding Window**	**Accuracy**	**F1 Score**
Wrist Worn: Basic Movement	3 Second	80.94%	78.46%
	5 Second	86.01%	82.37%
	10 Second	**92.4%**	**89.43%**
Wrist-Worn: Specific Movement	3 Second	59.93%	56.59%
	5 Second	67.83%	62.11%
	10 Second	**81.29%**	**66.75%**
RNN			
Wrist Worn: Basic Movement	3 Second	84.89%	82.56%
	5 Second	94.65%	93.03%
	10 Second	**98.75%**	**98.06%**
Wrist-Worn: Specific Movement	3 Second	70.39%	65.58%
	5 Second	85.6%	81.67%
	10 Second	**96.52%**	**93.43%**
Hip-Worn	3 Second	85.5%	84.04%
	5 Second	88.48%	85.29%
	10 Second	**89.31%**	**89.36%**

and F1-score (F1) percentages about the classification of the hip-worn data movement types using RNN for 10-second window are shown in Table 4.

5. Discussion

This paper has proposed and presented two deep learning models for the classification of physical movement patterns from on-body accelerometer sensors. Our models were trained on two on-body accelerometer sensor datasets.

For recognizing ADLs with DNNs, few categories are accomplished with good success. The DNN achieve accuracies between 80-93% and F1-scores between 80-90% for basic movements (see Table 5). When increasing the complexity of the dataset by using specific movement categories, the accuracies significantly decrease. This is expected as there are more categories to learn and recognize. The achieved accuracy for the specific movement types, using DNN, is between 60-80%, while the F1-scores are between 55-65% (see Table 5). Considering each experiment trained the model for 2000 training steps, then providing it with unseen data to classify, the DNN is able to classify "unknow" data.

When different combinations of hyper-parameters are used, the performance is depending on which categorization of the movement patterns was used. For the broad categories, ADLs, low learning rates, and the Adam optimizer achieve the highest F1 scores, i.e., 78.5% for three seconds, 82.4% for five seconds, and 89.4% for ten seconds sliding windows. Recognizing the specific movements, Adagrad, with a learning rate of 0.01 gets the highest results, i.e., 56.6% for three seconds, 62.1% for five seconds, and 67.8% for ten seconds sliding windows.

Let's compare whether the proposed DNN model performs better than the state-of-the-art algorithm. The state-of-the-art algorithm performs better at some categories such as drinking from a glass, climbing stairs, pouring water into a glass, and standing up from a chair. In contrast, the DNN model performs better at the others getting out of bed, sitting down on a chair, and walking. Additionally, the DNN model is trained to recognize all of the categories in the wrist-worn dataset, not just a selection.

Analyzing Table 6, the proposed DNN model achieves relatively good results considering the complexity of the classification. Comparing it to the state-of-the-art, which classifies seven categories, the DNN achieves comparable results using all 14 categories. Thus, we argue that the proposed DNN model, at the very least, matches the performance of the state-of-the-art algorithm.

Compared to our proposed DNN model, the accuracies are consistently higher for the used RNN model. They are achieving accuracies between 85-99% and F1-scores between 83-98% for the

Table 6: Comparison of our results with existing state-of-the-art

	DNN results			State-of-the-art		RNN results		
Category	***Recall***	***Precision***	***F1 score***	***True positives***	***True negatives***	***Recall***	***Precision***	***F1 score***
Brushing teeth	94.77%	98.82%	96.75%	-	-	99.32%	99.77%	99.54%
Climbing stairs	4.9%	35.0%	8.59%	20%	93.34%	92.31%	85.71%	88.89%
Comb hair	57.26%	84.81%	68.37%	-	-	98.72%	98.72%	98.72%
Descend stairs	2.17%	33.33%	4.08%	-	-	91.3%	84.0%	87.5%
Drinking	68.28%	60.77%	64.3%	100%	83.34%	95.16%	94.65%	94.91%
Eat w/ fork and knife	93.01%	82.06%	87.19%	-	-	98.94%	99.36%	99.15%
Eat w/spoon	95.79%	93.81%	94.79%	-	-	100%	95.0%	97.44%
Getting out of bed	85.66%	83.27%	84.44%	60%	66.67%	90.16%	94.02%	92.05%
Lying down on the bed	29.55%	61.9%	40.0%	-	-	68.18%	85.71%	75.95%
Pour water into glass	66.48%	85.21%	74.69%	100%	80%	97.25%	95.68%	96.46%
Sitting down on a chair	62.5%	60.61%	61.54%	0%	93.34%	100%	88.89%	94.12%
Standing up from a chair	52.78%	55.88%	54.29%	60%	83.34%	94.44%	85.0%	89.47%
Using the telephone	85.54%	58.44%	69.44%	-	-	93.98%	93.98%	93.98%
Walking	96.55%	83.3%	89.44%	40%	70%	97.49%	98.42%	97.95%

doi:10.20944/preprints202107.0505.v1

basic movement types. For the specific movement, the accuracies are between 70-97%, while the F1-scores are between 65-94% (see Table 5).

In addition to the ADLs and specific movement types for the wrist-worn dataset, we also tested the hip-worn dataset with the RNN. The results of this dataset are impressive and the most important finding in this study from an epidemiological perspective, considering that it consists of movements of similar type, with different intensities in a free-living setting. They are achieving accuracies between 85-90% and F1-scores of 84-90%. Thus, showing that the RNN model is able to perform at a high level on new datasets. Table 6 shows a comparison of the results of the two proposed deep learning models and the state-of-the-art algorithms. The comparable percentages are the true positives of the state-of-the-art and the recall score of the deep learning models.

Further, when we compare our results with the existing literature (see section 2), for classifying the physical movement activities, our results show higher accuracy when RNN is used, but compared with all other literature, they have a larger and different dataset, making the results incomparable.

6. Conclusion and Future Work

This paper proposes and presents two distinct deep learning approaches, built upon DNN and RNN, to classify physical movement patterns from body-worn tri-axial accelerometer data. We perform numerous experiments with different combinations of hyper-parameters, optimizers, and learning rates for the classification of the movement patterns. Both the models train on two separate accelerometer datasets. The first dataset, from the UCI machine learning repository, contains 14 various activities of daily life (ADL) and collected from 16 volunteers who carried a single wrist-worn tri-axial accelerometer. The second dataset, collected by us, has ten different ADLs from eight volunteers who placed the sensors on their hip and carried them out during their daily activities. Both the DNN and RNN models are evaluated under several performance metrics such as precision, recall, F1-score, and accuracy. The experimental results showed that the proposed DNN model is able to recognize “unlabeled” data with an acceptable overall recall percentage of 64%, using 14 categories, which is a 10% increase compared to the state-of-the-art algorithm, which has only a 54% overall recall score for seven types. Whereas, the proposed RNN model performs significantly better on the time-series data, reaching an overall recall score of 94%, which is a 40% increase compared to the DNN. The RNN model results surpass the percentages for most categories compared to the state-of-the-art, even when classifying all classes in the dataset. Furthermore, the RNN model is able to recognize different movement patterns from a new dataset consisting of movement types of varying intensities. Thus, our results show that the proposed RNN model is the best-suited algorithm of the discussed algorithms for movement pattern recognition. As future work, we intend to extend this work to experiment with restructuring the proposed deep learning models first to recognize these broad ADL categories, then classify which specific movement type within the ADL it is and test different optimizers in combination with varying rates of learning.

Author Contributions: Conceptualization, V.N, L.J, M.G, B.H.H, and S.B; methodology, V.N, L.J, M.G, B.H.H, and S.B; software, S.J and T.S.H; validation, S.J and T.S.H ; formal analysis, S.J and T.S.H; investigation, S.J and T.S.H; resources, S.J and T.S.H; data curation, S.J and T.S.H; writing—original draft preparation, V.N and M.G; writing—review and editing, V.N and M.G; visualization, S.J and T.S.H; supervision, V.N, L.J, M.G, B.H.H, and S.B; project administration, V.N, L.J, M.G, B.H.H, and S.B; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data collected through research presented in the paper are available on request from the corresponding authors.

Acknowledgments: We would like to thank Faculty of Health and Sport Sciences, University of Agder, Norway for providing the human activity data to us and CAIR for proposing the project and allowing us to do the research in this topic.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Not applicable.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep learning
DNN	Deep feed-forward neural network
RNN	Recurrent neural network
ADL	activities-of-daily-life
PA	Physical activity
SVM	Support Vector Machine
CNN	Convolutional Neural Networks
GMM	Gaussian Mixture Model
NB	Naïve Bayesian
LSTM	Long Short-Term Memory
DBN	Deep belief networks
HAR	Human activity recognition
ANN	Artificial neural networks
GRU	Gated Recurrent Unit
SGD	Stochastic Gradient Descent
ReLU	Rectified Linear Unit

References

1. Caspersen, C.J.; Powell, K.E.; Christenson, G.M. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public health reports* **1985**, *100*, 1–126.
2. Kawaguchi, N.; Nishio, N.; Roggen, D.; Inoue, S.; Pirttikangas, S.; Van Laerhoven, K. *Human activity sensing: corpus and applications*, 1st ed.; Springer Nature, 2019; pp. 1–250.
3. Procter, D.S.; Page, A.; Cooper, A.R.; Nightingale, C.M.; Ram, B.; Rudnicka, A.R.; Whincup, P.; Clary, C.; Lewis, D.J.; Cummins, S.; Ellaway, A.; Giles-Corti, B.; Cook, D.G.; Owen, C.G. An open-source tool to identify active travel from hip-worn accelerometer, GPS and GIS data. The international journal of behavioral nutrition and physical activity, 2018, pp. 1–10.
4. Warburton, D.E.; Nicol, C.W.; Bredin, S.S. Health benefits of physical activity: the evidence. *CMAJ* **2006**, *174*, 801–809, [http://www.cmaj.ca/content/174/6/801.full.pdf]. doi:10.1503/cmaj.051351.
5. Stuij, M. Physical activity, that's a tricky subject- Experiences of health care professionals with physical activity in type 2 diabetes care. *BMC health services research* **2018**, *18*, 1–13. doi:10.1186/s12913-018-3102-1.
6. Bredahl, T.V.; Puggaard, L.; Roessler, K.K. Exercise on Prescription. Effect of attendance on participants' psychological factors in a Danish version of Exercise on Prescription: A Study Protocol. *BMC health services research* **2008**, *8*, 1–8. doi:10.1186/1472-6963-8-139.
7. UCI Machine Learning Repository: Dataset for ADL Recognition with Wrist-worn Accelerometer Data Set. https://archive.ics.uci.edu/ml/datasets/Dataset+for+ADL+Recognition+with+Wrist-worn+Accelerometer#. Accessed: 2019-09-15.
8. Talo, M. Automated classification of histopathology images using transfer learning. *Artificial Intelligence in Medicine* **2019**, *101*, 1–16. doi:10.1016/j.artmed.2019.101743.
9. Bardou, D.; Zhang, K.; Ahmad, S.M. Lung sounds classification using convolutional neural networks. *Artificial intelligence in medicine* **2018**, *88*, 58–69. doi:10.1016/j.artmed.2018.04.008.
10. He, B.; Guan, Y.; Dai, R. Classifying medical relations in clinical text via convolutional neural networks. *Artificial intelligence in medicine* **2019**, *93*, 43–49. doi:10.1016/j.artmed.2018.05.001.
11. Banerjee, I.; Ling, Y.; Chen, M.C.; Hasan, S.A.; Langlotz, C.P.; Moradzadeh, N.; Chapman, B.; Amrhein, T.; Mong, D.; Rubin, D.L.; others. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artificial intelligence in medicine* **2019**, *97*, 79–88. doi:10.1016/j.artmed.2018.11.004.
12. Ting, H.W.; Chung, S.L.; Chen, C.F.; Chiu, H.Y.; Hsieh, Y.W. A drug identification model developed using deep learning technologies: experience of a medical center in Taiwan. *BMC Health Services Research* **2020**, *20*, 1–9. doi:10.1186/s12913-020-05166-w.
13. Jiménez, F.; Palma, J.; Sánchez, G.; Marín, D.; Palacios, F.; López, L. Feature Selection based Multivariate Time Series Forecasting: An Application to Antibiotic Resistance Outbreaks Prediction. *Artificial Intelligence in Medicine* **2020**, pp. 1–16. doi:10.1016/j.artmed.2020.101818.
14. Miled, Z.B.; Haas, K.; Black, C.M.; Khandker, R.K.; Chandrasekaran, V.; Lipton, R.; Boustani, M.A. Predicting dementia with routine care EMR data. *Artificial Intelligence in Medicine* **2020**, *102*, 1–8. doi:10.1016/j.artmed.2019.101771.
15. Zhang, N.; Cai, Y.X.; Wang, Y.Y.; Tian, Y.T.; Wang, X.L.; Badami, B. Skin cancer diagnosis based on optimized convolutional neural network. *Artificial Intelligence in Medicine* **2020**, *102*, 1–8. doi:10.1016/j.artmed.2019.101756.
16. Papagiannaki, A.; Zacharaki, E.I.; Kalouris, G.; Kalogiannis, S.; Deltouzos, K.; Ellul, J.; Megalooikonomou, V. Recognizing Physical Activity of Older People from Wearable Sensors and Inconsistent Data. *Sensors* **2019**, *19*, 1–8. doi:10.3390/s19040880.
17. Zhang, S.; Rowlands, A.; Murray, P.; Hurst, T. Physical activity classification using the GENEA wrist-worn accelerometer. *Medicine and science in sports and exercise* **2012**, *44*, 1–7. doi:10.1249/MSS.0b013e31823bf95c.
18. Trost, S.G.; Zheng, Y.; Wong, W.K. Machine learning for activity recognition: hip versus wrist data. *Physiological measurement* **2014**, *35*, 1–8. doi:10.1088/0967-3334/35/11/2183.

19. Mannini, A.; Sabatini, A.M. Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors* **2010**, *10*, 1154–1175. doi:10.3390/s100201154.
20. Hagenbuchner, M.; Cliff, D.P.; Trost, S.G.; Van Tuc, N.; Peoples, G.E. Prediction of activity type in preschool children using machine learning techniques. *Journal of Science and Medicine in Sport* **2015**, *18*, 426–431. doi:10.1016/j.jsams.2014.06.003.
21. Hammerla, N.Y.; Halloran, S.; Plötz, T. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables **2016**. p. 1533–1540. doi:10.5555/3060832.3060835.
22. Murad, A.; Pyun, J.Y. Deep recurrent neural networks for human activity recognition. *Sensors* **2017**, *17*, 25–56. doi:10.3390/s17112556.
23. Saez, Y.; Baldominos, A.; Isasi, P. A comparison study of classifier algorithms for cross-person physical activity recognition. *Sensors* **2017**, *17*, 1–26. doi:10.3390/s17010066.
24. Sani, S.; Massie, S.; Wiratunga, N.; Cooper, K. Learning deep and shallow features for human activity recognition. International Conference on Knowledge Science, Engineering and Management. Springer, 2017, pp. 469–482. doi:10.1007/978-3-319-63558-3_40.
25. Um, T.T.; Babakeshizadeh, V.; Kulić, D. Exercise motion classification from large-scale wearable sensor data using convolutional neural networks. International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017, pp. 2385–2390. doi:10.1109/IROS.2017.8206051.
26. Hassan, M.M.; Huda, S.; Uddin, M.Z.; Almogren, A.; Alrubaian, M. Human activity recognition from body sensor data using deep learning. *Journal of medical systems* **2018**, *42*, 1–8. doi:10.1007/s10916-018-0948-z.
27. Mascret, Q.; Bielmann, M.; Fall, C.L.; Bouyer, L.J.; Gosselin, B. Real-Time Human Physical Activity Recognition with Low Latency Prediction Feedback Using Raw IMU Data. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018, pp. 239–242. doi:10.1109/EMBC.2018.8512252.
28. Moya Rueda, F.; Grzeszick, R.; Fink, G.; Feldhorst, S.; ten Hompel, M. Convolutional neural networks for human activity recognition using body-worn sensors. Informatics. Multidisciplinary Digital Publishing Institute, 2018, Vol. 5, pp. 1–26. doi:10.3390/informatics5020026.
29. Suto, J.; Oniga, S. Efficiency investigation of artificial neural networks in human activity recognition. *Journal of Ambient Intelligence and Humanized Computing* **2018**, *9*, 1049–1060. doi:10.1007/s12652-017-0513-5.
30. Welhenge, A.M.; Taparugssanagorn, A. Human activity classification using long short-term memory network. *Signal, Image and Video Processing* **2019**, *13*, 651–656. doi:10.1007/s11760-018-1393-7.
31. Bevilacqua, A.; MacDonald, K.; Rangarej, A.; Widjaya, V.; Caulfield, B.; Kechadi, T. Human Activity Recognition with Convolutional Neural Networks. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2018, pp. 541–552.
32. Preece, S.J.; Goulermas, J.Y.; Kenney, L.P.; Howard, D.; Meijer, K.; Crompton, R. Activity identification using body-mounted sensors—a review of classification techniques. *Physiological measurement* **2009**, *30*, 1–64. doi:10.1088/0967-3334/30/4/R01/meta.