

Article

Neural Network-Based Granular Activity Recognition from Accelerometers: Assessing Generalizability Across Diverse Mobility Profiles

Metin Bicer ^{1,2}, James Pope ³, Lynn Rochester ^{1,2,*}, Silvia Del Din ^{1,2,*} and Lisa Alcock ^{1,2}

¹ Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK; metin.bicer@newcastle.ac.uk (M.B.); lynn.rochester@newcastle.ac.uk (L.R.); lisa.alcock@newcastle.ac.uk (L.A.)

² NIHR Newcastle Biomedical Research Centre, Newcastle University and the Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, NE4 5PL, UK

³ School of Engineering Mathematics and Technology, University of Bristol, Bristol, BS8 1QU, UK; jp16127@bristol.ac.uk

* Correspondence: silvia.del-din@newcastle.ac.uk; Tel.: +44-(0)1912081131

Highlights

What are the main findings?

- Outperformed fixed sliding-window methods for frame-level activity recognition.
- Length of data window did not affect the performance.
- Satisfactory cross-cohort performance.

What are the implications of the main findings?

- Granular recognition for capturing transitions.
- Handling of long and arbitrary length real-world wearable sensor data.

Abstract

Human activity recognition (HAR) lies at the core of digital healthcare applications that monitor different types of physical activity. Traditional HAR methods often struggle to adapt to variable-length, real-world activity data and to generalise across cohorts (e.g., from young to old cohorts). Thus, the aim of this study was to investigate HAR using wearable sensor data, with a particular focus on cross-cohort evaluation. Each dataset included two accelerometers (right thigh and lower back) sampling at 50 Hz, capturing a range of daily-life activities that were annotated using video recordings from chest-mounted cameras synchronised with the accelerometers. Neural networks were trained on young cohorts' data and tested on old cohorts' data. The effects of network architecture, sampling frequency and sensor location on classification performance were investigated. Network performance was evaluated using accuracy, recall, precision, F1-score and confusion matrices. The gated recurrent unit architecture achieved the best performance when trained solely on young cohorts' data, with weighted F1-score of 0.95 ± 0.05 and 0.93 ± 0.05 for young and old cohorts, respectively, resulting in a highly generalizable method. Classification performance across multiple sampling frequencies was comparable. The thigh-mounted sensor consistently achieved higher performance than the lower back sensor across activities except lying. Furthermore, combining datasets significantly improved performance on the old cohort (weighted F1-score: 0.97 ± 0.02) due to increased variability in the training data. This study highlights the importance of network architecture and dataset composition in HAR and demonstrates the potential of neural networks

Academic Editor: Lorenzo Scalise

Received: 7 January 2026

Revised: 11 February 2026

Accepted: 13 February 2026

Published: 18 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

for robust, real-world activity recognition across age-defined cohorts, specifically between young and old cohorts.

Keywords: human activity recognition; deep learning; wearable sensor

1. Introduction

Improving quality of life drives advances across domains, from smart homes [1] and industry [2] to healthcare [3], where physical mobility, a key indicator of health [4], is routinely evaluated through mobility assessment. Traditional mobility assessments lack ecological validity, but wearable sensors enable long-term monitoring in real-world environments [3]. This shift allows research to move from lab-based tests to daily-life monitoring, which rely on human activity recognition (HAR), including walking and sedentary behaviours [5–7]. Traditional HAR approaches often lack generalizability to old populations due to validation on small, homogeneous cohorts (typically 10–20 healthy adults) [7,8]. These methods relied on fixed thresholds and rule-based trees [9,10], which assume specific mobility characteristics, reducing applicability to diverse populations and increasing computational complexity. In contrast, data-driven HAR approaches detect activities directly from wearable signals, either with or without feature engineering. Feature-based methods extract time- and frequency-domain measures (e.g., mean, variance, entropy) [11], but lack established rules for selecting optimal features, particularly for activities with similar characteristics [12]. To overcome this challenge, researchers increasingly adopted neural networks [13]. Convolutional (CNNs) and recurrent (RNNs) neural networks have shown strong potential for HAR [14]. CNNs, adapted for time-series analysis, use 1D convolutional filters to extract spatial features [15–17], while RNNs capture temporal dependencies [18,19]. Hybrid CNN–RNN networks exploit both spatial and temporal information in sensor data [20,21]. More recently, architectures that model longer-range dependencies, such as temporal convolutional networks, self-attention mechanisms and transformer-based architectures, have also been applied to solve this problem [22–24].

Previous studies predicted a single activity label per 2–3 s window [25,26], which reduces granularity and can mislead training when multiple activities occur within a window [27]. HAR at a more granular level (i.e., dense prediction across the time series, where activity transitions are reflected by changes in the predicted label window rather than by assigning a single label to a multi-second window) remains underexplored [27,28]. Furthermore, existing HAR datasets differ in sensor modalities and placements, which complicates comparisons between studies [29,30]. Healthy old adults are also underrepresented in commonly used datasets [31–33], meaning that cross-cohort evaluation is not routinely performed, although it has been explored in recent work [34]. Sensor placement and the number of sensors influence both performance and usability: while multi-sensor configurations can improve classification accuracy, extensive setups may restrict natural movement and reduce wearability, particularly for old adults and healthcare applications. Sampling frequency presents a similar trade-off, as higher frequencies capture finer motion detail but increase power consumption and storage requirements, limiting feasibility for long-term monitoring in real-world applications. Consequently, optimising sensor placement and sampling frequency is essential to balance accuracy, usability, and real-world applicability.

This study aimed to develop and evaluate neural networks for recognising daily activities using publicly available datasets from young and old cohorts. We trained models on data from the young cohort and evaluated their performance on the old cohort to assess

cross-cohort generalisability. The objectives were to (1) develop neural networks for HAR in the young cohort, (2) evaluate their performance in the old cohort, and (3) examine whether combining datasets from the young and old cohort improves model performance. To support these objectives, we investigated the influence of network architecture, sampling frequency, and sensor location. By focusing on frame-level recognition (i.e., predicting at each time step) and cross-cohort evaluation between the young and old cohorts, this work supports the development of practical HAR methods for real-world healthcare applications.

2. Materials and Methods

2.1. Datasets

This study utilised two publicly available HAR datasets [34,35], both containing tri-axial accelerometer data, sampled at 50 Hz, recorded from the lower back and the right thigh in ambulatory settings. Sensor orientation was consistent between studies, and no additional axis re-alignment was required. The accelerometer data were used in their raw form for training and evaluation, with no gravity removal or coordinate normalisation applied in this study. Both datasets used in this study provided activity labels for each frame, derived from synchronised video recordings, using the annotation protocols described in their respective publications [34,35]. In this work, we used these labels as provided and did not apply any additional post-processing methods. This was a deliberate choice to preserve reproducibility and to evaluate the robustness models under realistic annotation conditions, where some degree of ambiguity around activity transitions is unavoidable. Additionally, both datasets exhibit substantial class imbalance, with postural and ambulatory activities (e.g., sitting, standing, walking) dominating the data, and more challenging activities such as stair ascent, stair descent and cycling occurring relatively infrequently. No additional resampling, class balancing or label modification was applied in this study. Detailed quantitative distributions for each activity class are provided in the corresponding publications [34,35].

The first dataset, referred to as the “Young Cohort Dataset” [35], comprised data from 22 young participants (8 female, 14 male; age: 38.6 ± 14 years; height: 177.3 ± 8.3 cm; weight: 72.9 ± 10.6 kg; BMI: 23.1 ± 2.3 kg/m²). The per-person average duration of data collected from this cohort was 97.90 ± 32.95 min. This dataset included nine distinct activities: walking, running, standing, stair ascent, stair descent, sitting, lying, cycle-sit and cycle-stand. The second dataset, referred to as the “Old Cohort Dataset” [34], included data from 18 old adult participants (9 female, 8 male; age: 79.6 ± 7.6 years; height: 173.0 ± 7.8 cm; weight: 80.0 ± 9.3 kg; BMI: 26.8 ± 2.7 kg/m²). The per-person average duration of data for this cohort was 41.84 ± 5.77 min. This dataset contained a subset of activities more common in an old population: walking, stairs ascending, stairs descending, standing, sitting and lying. Further details about the datasets can be found in their respective articles.

For both datasets, each timepoint included six sensor channels: three axes for the lower back accelerometer and three axes for the right thigh accelerometer. A single training or testing example window was a $6 \times T$ matrix, where T denotes the number of timepoints (i.e., the window length) for that specific example. Each time window had corresponding labels (ground-truth) for every timepoint such that a single example of a label window is a vector of length T . The window length for training was fixed to $T = 200$ frames.

2.2. Neural Network Architectures

A critical design principle for all investigated architectures was their ability to process variable-length input windows and to predict an activity at each individual timepoint

within the given window. This approach was chosen to ensure high temporal granularity in activity detection while maintaining computational efficiency by eliminating use of data windowing.

All investigated architectures were explicitly designed to process variable-length input windows and to generate predictions at each timepoint. No zero padding was applied at the input level to equalise window lengths, as the networks operate natively on windows of arbitrary length. Hybrid architectures were evaluated to assess whether combining local feature extraction (CNN) with long-range temporal modelling (RNN) offers advantages over purely recurrent models for dense activity recognition.

2.2.1. One-Dimensional Convolutional Neural Network

One-dimensional CNNs are a type of neural network architecture designed for analysing sequential data, where the temporal or spatial order of the input is meaningful. These networks operate by applying filters (also known as kernels) that slide over the input, enabling the network to automatically learn spatially localised patterns, such as characteristic features in time series data.

In our implementation, the input was a window of sensor data denoted as $\mathbf{X} = [x_1, x_2, \dots, x_T]$ where T is the length of the window and each $x_t \in \mathbb{R}^n$ was composed of n sensor channel readings at time t . A convolutional layer consisted of a set of kernels, each defined by a weight vector $\mathbf{W} = [w_1, w_2, \dots, w_k]$ with kernel size k . Kernels also had an associated bias term b . The convolution operation between the kernel $\mathbf{W}$ and a data window of $\mathbf{X}$, followed by a non-linear activation function (Rectified Linear Unit, *ReLU*), produced an output feature map o according to $o_t = \text{ReLU}(\sum_{i=1}^k w_i \cdot x_{t+i-1} + b)$ at timepoint t . The output calculation was repeated across the entire window and for multiple kernels, enabling the network to learn diverse feature representations. The sliding movement of the kernels across the input window was governed by a parameter called the stride (i.e., the step size by which the kernel moves).

The implemented CNN architecture consisted of a series of 1D convolutional layers, each followed by a dropout and batch normalisation layer. The number of layers (1–5), number of kernels per layer (32–1024), kernel and stride size (1–11), dropout rate (0.3, 0.6) and learning rate (0.001–0.1) were optimised through grid search as described in Section 2.3.

2.2.2. Gated Recurrent Unit Network

RNNs are a class of neural architectures designed to model sequential data by maintaining a memory of previous inputs through internal hidden states. Traditional RNNs, however, are limited in modelling long-range dependencies due to issues such as vanishing or exploding gradients during training. To address these limitations, gated recurrent architectures such as Long Short-Term Memory (LSTM) networks and gated recurrent units (GRUs), as well as attention-augmented recurrent variants, have been proposed to improve temporal modelling. In preliminary experiments, we evaluated vanilla RNN, LSTM, and GRU architectures, including attention-augmented variants. As the standard GRU architecture provided improved performance with more stable training behaviour and lower computational cost, it was selected as the recurrent architecture for subsequent experiments. GRUs regulate information flow through gating mechanisms that dynamically control the retention and updating of temporal information.

In our implementation, at each time step t , a GRU cell received the current input x_t and the previous hidden state h_{t-1} . It then computed the new hidden state h_t using two gates: the update gate $z_t = A(\mathbf{W}_z x_t + \mathbf{U}_z h_{t-1} + b_z)$ and the reset gate $r_t = A(\mathbf{W}_r x_t + \mathbf{U}_r h_{t-1} + b_r)$ where A denotes the activation function, $\mathbf{W}_*$ and $\mathbf{U}_*$ are learnable weight matrices and b_* are bias terms. The update gate z_t controls the extent to which the

hidden state is updated with new information, while the reset gate r_t determines how much of the past information to forget. Using these gates, the GRU computes a candidate hidden state $\tilde{h}_t = \tanh[W_h x_t + U_h(r_t \cdot h_{t-1}) + b_h]$. The final hidden state is then updated by interpolating the previous hidden state and the candidate state, $h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t$. This gating mechanism allows the GRU to capture dependencies over longer windows without suffering from gradient degradation.

The implemented GRU architecture consisted of a series of recurrent layers, each followed by a dropout and batch normalisation layer. The number of layers (1–3), number of units per layer (32–256), dropout rate (0.3, 0.6) and learning rate (0.001–0.1) were optimised through grid search as described in Section 2.3.

2.2.3. Hybrid Convolutional–Recurrent Networks

To leverage the strengths of both convolutional and recurrent neural networks in modelling sequential data, we implemented a hybrid architecture combining 1D CNNs with GRUs. One of the two architectures processes input window and its output is fed to the next architecture to output final predictions. We investigated two alternatives. The first hybrid architecture, called ConvRec, consisted of one or more layers of 1D CNNs processing the raw multivariate input window to extract local features. If strides greater than one were used, we applied padding to these features to retain input window length. Because the GRUs processed the full window while maintaining the input window length, the network retains the ability to generate a prediction for each individual time step. The second alternative, called ReConv, was to learn temporal features at the beginning with the GRU layers that was then followed by 1D convolutional layers to learn local features among the temporal features.

The hybrid networks were constructed by combining the CNN and GRU architectures as described in the previous section. The same hyperparameter search space was applied to the respective CNN and GRU components.

2.3. Training and Evaluation Strategy

Training of network requires optimising the kernel weights and biases. The optimisation problem was minimising the cross-entropy loss function quantifying the difference between the predicted probability distribution of the network and the true distribution of the data. The optimisation was carried out in PyTorch (V2.4.1) using a gradient-based optimisation method called adaptive moment estimation.

The Young Cohort Dataset was used for training neural networks (Objective 1) and initially divided into training, validation and test folds, following the approach described in [35], to tune hyperparameters. Hyperparameter tuning involved a random grid search with 50 configurations on several parameters (learning rate, number of convolutional or GRU layers, units, kernel and stride size, dropout rate) to optimise performance on the validation set, preventing overfitting to the training data. After the hyperparameter tuning, a Leave-One-Participant-Out Cross-Validation (LOPO-CV) strategy was employed to comprehensively assess the networks' generalizability within the young cohort. In this strategy, for each fold, data from one participant was held out as the test set, while the remaining participants' data were used for training. This process was repeated 22 times, ensuring that each participant's data served as the unseen test set once. The predictions and true labels for all 22 individuals were then concatenated into single arrays to compute overall evaluation metrics for the young cohort. This approach provides a robust measure of generalizability at the individual participant level.

The Old Cohort Dataset was reserved for final testing to assess generalizability to an unseen population (Objective 2). Data from each of the 18 participants in the Older Cohort Dataset was passed through all 22 trained networks from the LOPO-CV. This resulted in

22 individual activity predictions at each timepoint for every participant. To obtain a final prediction for each timepoint, we applied a majority voting scheme: the most frequently predicted activity among the 22 networks was selected as the final activity label. In the event of a tie, the class with the highest summed prediction probability across networks was selected. These final predictions and their corresponding true labels were combined into single arrays to calculate overall evaluation metrics for the old cohort. This rigorous testing methodology ensured an unbiased and comprehensive assessment of the real-world generalizability.

We run additional experiments on sampling rate and sensor locations. First, a set of experiments was conducted in which the dataset was independently downsampled to 25, 20, 15, and 10 Hz, and model performance was evaluated at each sampling frequency to assess the effect of temporal resolution. Next, effects of sensor location were investigated by training networks using thigh-only, lower-back-only and combined sensor configurations. These experiments were carried out using the same strategy outlined above for the architecture selection.

We also developed a general neural network, based on the best-performing architecture, sampling frequency and sensor location, using a combined dataset by pooling all samples from the two cohorts (Objective 3). This neural network was evaluated using the same LOPO-CV strategy. To assess the impact of dataset composition, performance was evaluated separately for young and old participants. This allowed us to quantify how combining datasets influenced performance across the two cohorts.

We evaluated the performance of the trained models under different window lengths, specifically comparing 200-frame windows (4 s) with extended 3000-frame windows (1 min).

2.4. Performance Metrics

The performance of the trained networks was quantitatively assessed using several common classification metrics: precision, recall, accuracy and F1-score. These metrics were calculated both for each individual activity class (per-activity) and as macro averages across all activities. To account for imbalances in the number of samples per activity type, weighted averages of these metrics were also computed, using the proportion of samples for each activity as weights. Weighted F1-score served as the criteria for hyperparameter tuning. Also, this metric was used to compare the performance of models based on architecture, sampling frequency and sensor location. Performance differences between experimental conditions were assessed using paired t-tests on F1-scores, with statistical significance defined at $p < 0.05$.

A detailed breakdown of the networks' performance was provided through the analysis of confusion matrices. A confusion matrix visually summarises the performance of a classification algorithm. The diagonal elements of the matrix represent the number (or percentage) of correct predictions for each activity class, where the predicted label perfectly matches the actual ground truth label. The off-diagonal elements indicate misclassifications, showing instances where one activity class was incorrectly predicted as another. Analysing the confusion matrix is crucial for identifying specific patterns of confusion between different classes and for understanding a network's strengths and weaknesses for each activity.

3. Results

3.1. Neural Network Architecture Validation on the Young Cohort Dataset and Generalizability to the Old Cohort Dataset

The recurrent architecture, GRU, achieved the highest average (0.89 ± 0.10) and weighted (0.95 ± 0.05) F1-score among the architectures (Figures 1 and 2), with the differences being statistically significant ($p < 0.05$). The best-performing architecture identified through hyperparameter optimisation consisted of a two-block bidirectional GRU network: the first block contains two stacked layers with 128 hidden units, followed by a second layer with also 128 neurons, with dropout (0.3) applied between recurrent layers. On the other hand, CNN resulted in the worst predictive performance (average F1-score: 0.79 ± 0.20 , weighted F1-score: 0.92 ± 0.09). Combining RNNs with CNNs did not significantly affect the performance of the recurrent architecture. Showing results that were slightly better than others, ConvRec's performance was the closest to the RNN. While all architectures performed equally well for the common activity types (walking, running, standing, sitting, lying) with an F1-score over 0.85, activities involving stairs and cycling while standing up achieved the lowest scores. RNN still outperformed others with the lowest F1-score of 0.70 for the stair descent.

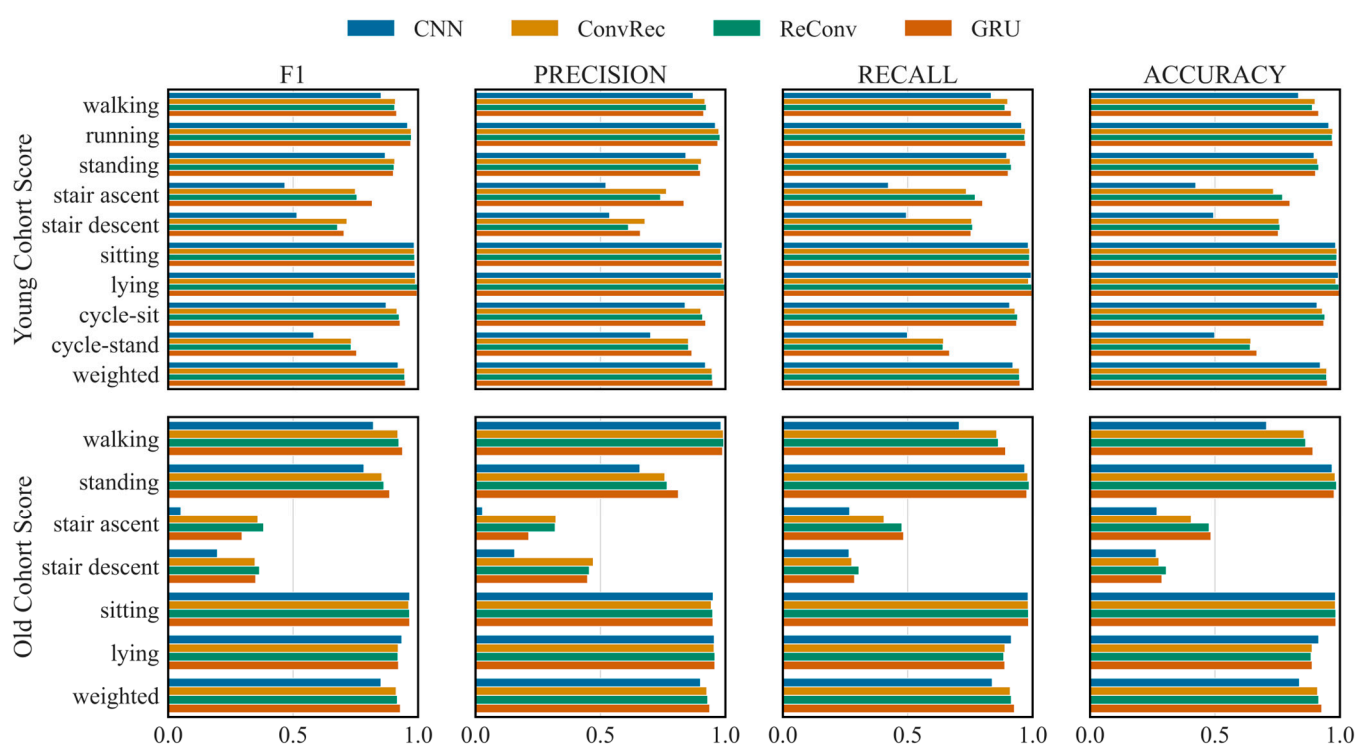


Figure 1. F1-score, precision, recall and accuracy on the Young and Old Cohort Datasets for different architectures. CNN: convolutional neural network. GRU: gated recurrent unit. ConvRec: CNN followed by GRU. ReConv: GRU followed by CNN. The metrics per activity and their average weighted by their frequency (“weighted”) are presented.

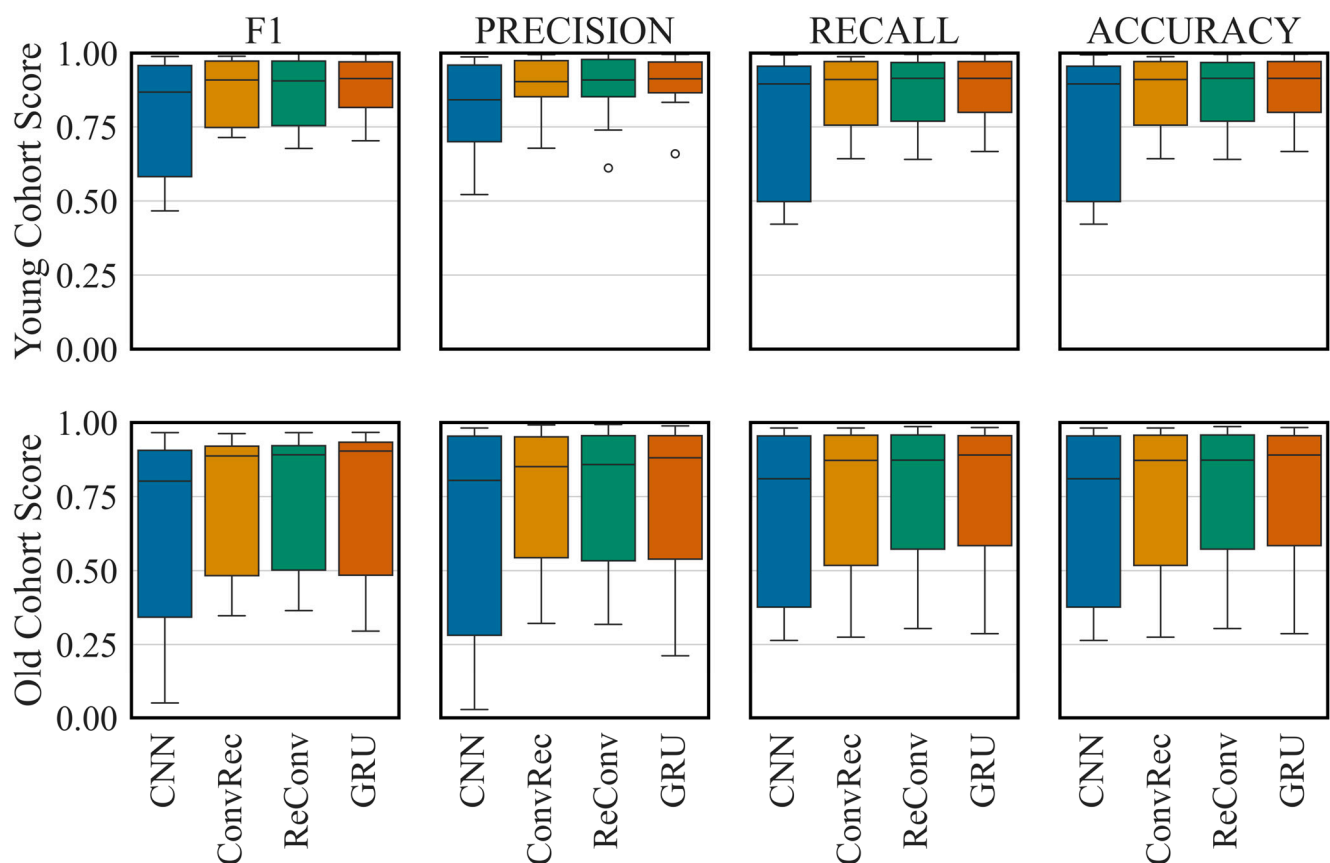


Figure 2. The distributions of F1-score, precision, recall and accuracy on the Young and Old Cohort Datasets for different architectures. CNN: convolutional neural network. GRU: gated recurrent unit. ConvRec: CNN followed by GRU. ReConv: GRU followed by CNN. Points shown beyond the whiskers represent statistical outliers inherent to the boxplot visualisation.

Overall, the models did not exhibit strong signs of global overfitting, as they maintained good performance on the Old Cohort Dataset despite being trained solely on the Young Cohort Dataset. RNN had the best performance (F1-score, average: 0.73 ± 0.29 , weighted: 0.93 ± 0.05) followed by the ReConv (F1-score, average: 0.74 ± 0.26 , weighted: 0.92 ± 0.05). While the common activities achieved an F1-score over 0.90, the stair ascent and descent tasks could be best predicted by the ReConv (weighted F1-score: 0.38 and 0.37, respectively) improving the performance of RNN for these tasks (0.30 and 0.35, respectively).

3.2. Sampling Frequency, Sensor Locations and Window Length

RNN still outperformed other architectures with downsampled data and in different sensor locations (Figures 3 and 4). Downsampling across sampling frequencies (50, 25, 20, 15, and 10 Hz) had no statistically significant effect on overall performance ($p > 0.05$). Weighted F1-scores in the LOPO-CV evaluation remained close to 0.95 across all frequencies (Figures S1 and S2), and generalisation performance on the old cohort was similarly stable, with weighted F1-scores ranging from 0.94 ± 0.05 (15 Hz) to 0.93 ± 0.05 (all other frequencies). The activities most affected by changes in sampling frequency were stair ascent and descent. For example, reducing the sampling frequency from 50 Hz to 25 Hz degraded stair ascent performance by 4%, while stair descent performance improved by 10%. In contrast, when downsampling from 50 Hz to 10 Hz, recognition performance on the old cohort for stair ascent and descent decreased by 16% and 72%, respectively.



Figure 3. F1-score, precision, recall and accuracy on the Young and Old Cohort Datasets for different experiments (sampling frequency and sensor location). Neural network architecture and hyperparameters were optimised for each experiment. The metrics per activity and their average weighted by their frequency (“weighted”) are presented.

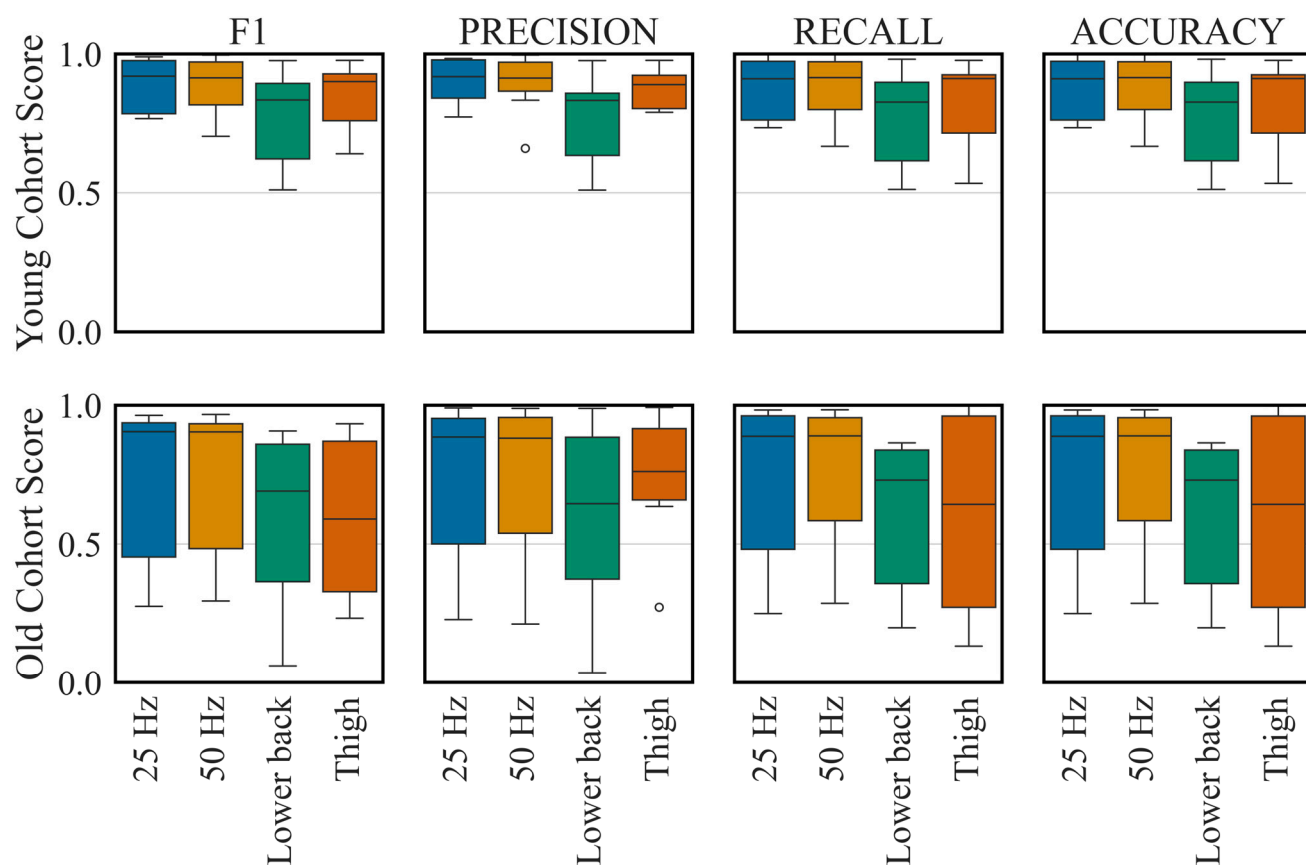


Figure 4. The distributions of F1-score, precision, recall and accuracy on the Young and Old Cohort Datasets for different experiments (sampling frequency and sensor location). Neural network architecture and hyperparameters were optimised for each experiment. Points shown beyond the whiskers represent statistical outliers inherent to the boxplot visualisation.

While using both sensors yielded the best performance in the LOPO-CV evaluation (Figures 3 and 4), the thigh sensor alone achieved significantly higher performance than the lower back sensor alone (weighted F1-score: 0.91 ± 0.08 vs. 0.80 ± 0.14 , $p < 0.05$). Standing, stair ascent and descent predictions were highly dependent on the thigh accelerometer. Thigh accelerometer or both accelerometers predicted the standing activity with F1-score of 0.90 unlike the lower back sensor alone which had only F1-score of 0.51. Sitting and lying predictions were improved by having the two sensors together. The network's performance on the old cohort was remarkably improved by the two accelerometers compared to the single sensor configurations (weighted F1-score: 0.81 ± 0.12 , 0.84 ± 0.20 and 0.93 ± 0.05 for lower back, thigh and combined). Prediction of sitting by the combined sensors improved to 0.97 F1-score from 0.84 for thigh and 0.75 for lower back sensor.

The effect of window length was negligible, without compromising performance. The performance of the RNN processing both sensors at 50 Hz using the 200-frames and 3000-frames of windows were practically the same (weighted F1-score: 0.94 ± 0.05 in both cases, $p > 0.05$).

3.3. A Generalizable Neural Network by Combining Datasets

The RNN with tuned hyperparameters was trained using both sensors sampled at 50 Hz (Figures 5, S3 and S4). When trained and tested on the young cohort (Figure 5A), performance was high across activities, with recall values of 0.91 (walking), 0.97 (running), 0.90 (standing), 0.80 (stair ascent) and 0.75 (stair descent). Applying this model to the old cohort (Figure 5B) revealed reduced performance for stair ascent and descent (recall 0.48 and 0.29), with frequent confusion with standing (20% and 36%) and walking (13% and 20%), while other activities remained robust (e.g., walking 0.89, sitting 0.98). Training on the combined dataset preserved performance on the young cohort (Figure 5C; weighted F1-score: 0.94 ± 0.04 , $p > 0.05$) and substantially improved performance on the old cohort (Figure 5D; weighted F1-score: 0.97 ± 0.02 , $p < 0.05$), with recall for stair ascent and descent increasing to 0.80 and 0.62, respectively, and reduced misclassification into walking (19% and 23%).

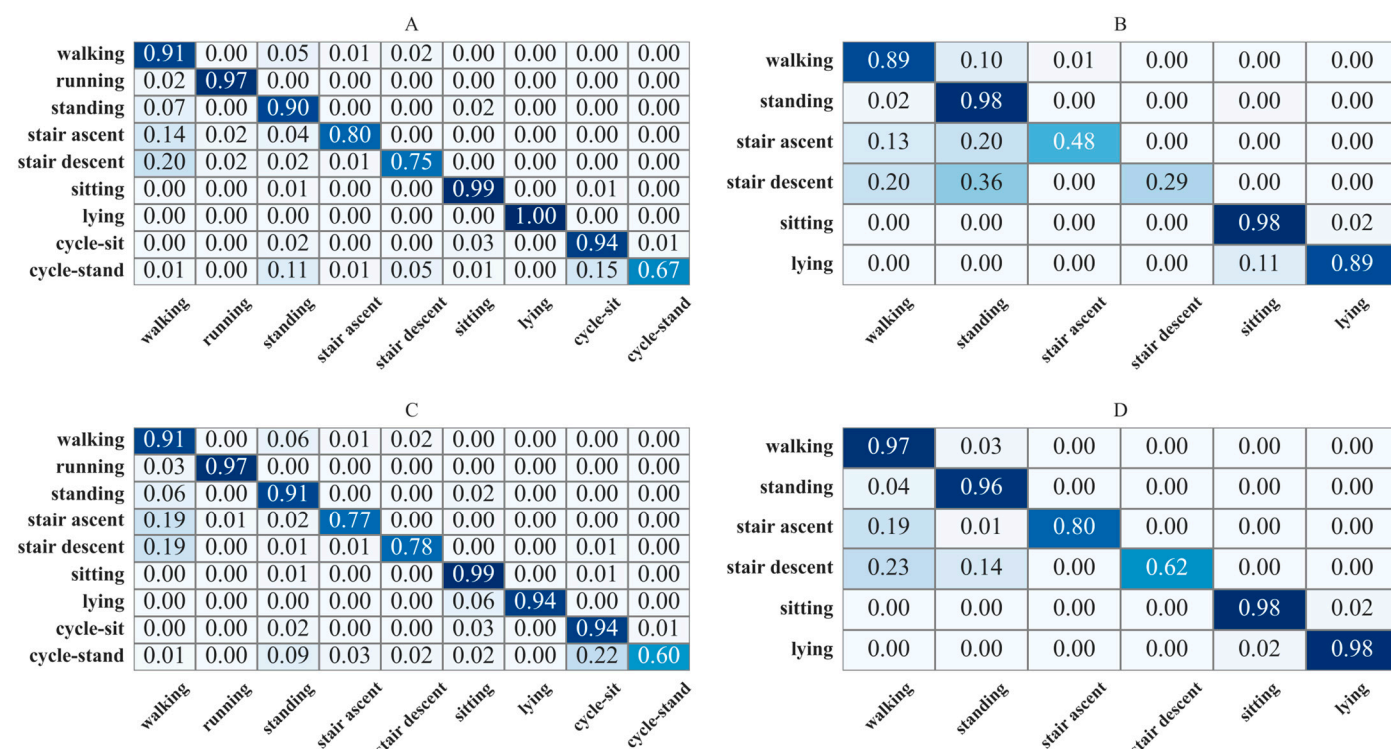


Figure 5. Confusion matrix for predictions on datasets, where darker shades indicate higher values. (A): Trained on the Young, tested on the Young Cohort Dataset. (B): Trained on the Young, tested on the Old Cohort Dataset. (C): Trained on the combined, tested on the Young Cohort Dataset. (D): Trained on the combined, tested on the Old Cohort Dataset.

4. Discussion

Our findings showed that the recurrent architecture achieved the best overall performance; however, the hybrid architectures (ReConv and ConvRec) also performed competitively and achieved comparable performance for several activities. Sampling frequency did not practically affect prediction performance. In contrast, sensor location influenced performance, with the thigh sensor outperforming the lower back sensor, and the combined configuration yielding the best results. The inclusion of data from the old cohort in the training phase improved the performance.

4.1. Neural Network Architecture Validation on the Young Cohort Dataset and Generalizability to Old Cohort Dataset

The neural network architecture emerged as the most important factor in the predictive performance. The recurrent network, GRU, effectively learned hidden features to predict activities at each time step and outperformed prior work on the same dataset using LOPO-CV [35] with higher accuracy (0.81 ± 0.18 vs. 0.89 ± 0.10). However, some activities, including stair ascent and descent, remained a persistent challenge for the current modeling approach. On the old cohort, our model also outperformed Ustad, Logacjov [34], who merged stair ascent and descent into walking, whereas our approach treated them separately. The discrepancy between the two evaluation schemas was expected, given mobility differences between cohorts, with some old participants using walking aids. Unlike their approach, which predicted a single activity for each 5 s window (250 frames), our method captured more fine-grained activity changes. Furthermore, the impact of data window length in model performance evaluation was minimal. The networks trained on 200-frame windows maintained comparable performance even when tested with much longer windows. For instance, testing with a 10 min window (vs. 4 s training window) revealed only

minor trade-offs between activities (e.g., improvements in stair tasks but slight decline in cycling tasks), while overall metrics remained stable, with improved computational efficiency at longer windows. Despite GRU's superior performance, combining it with CNNs (tested in different configurations, including CNN followed by RNN, RNN followed by CNN and multi-head CNNs) did not improve results. GRU remained the best architecture for stair ascent and descent in LOPO-CV, while in the old cohort these activities were predicted slightly better by ReConv.

Analysing per-activity performance, the GRU confusion matrix showed explainable and reasonable patterns. First, the lowest accuracy in LOPO-CV occurred for cycle-stand, which was almost equally misclassified as standing and cycle-sit. This confusion was reasonable given the overlap in postures as the activity names imply. Second, stair ascent and descent were misclassified as walking (LOPO-CV: 13% and Old Cohort: 20%). In the old cohort, 20% of stair ascent and 36% of descent were misclassified as standing, likely due to the fact that old people tend to perform these tasks significantly slower than young people due to stair negotiation [36,37]. Similarly, 10% of walking in old participants was classified as standing, likely because very slow walking approximates a standing posture, consistent with the age-related decline in gait speed [38].

A small number of predictions did not belong to the old cohort's label set, which is inevitable given the predefined activity classes. These cases were rare, with only 0.003% and 0.08% of the dataset predicted as running and cycle-sit, respectively.

4.2. Sampling Frequency and Sensor Locations

Downsampling did not affect overall network performance on either dataset with only a reduction observed at 10 Hz for stair ascent and descent. These findings are consistent with prior work suggesting that sampling frequencies around 20 Hz are sufficient for wearable activity recognition [39] and that higher frequencies do not improve accuracy beyond this threshold [40]. However, more aggressive downsampling reduces the temporal resolution of the signal, which may limit the ability to capture rapid transitions and fine-grained temporal dynamics. In the context of frame-level activity recognition, this effectively shifts the problem closer to a window-based prediction regime. Therefore, while sampling frequencies of approximately 15–20 Hz appear sufficient for maintaining overall performance, higher frequencies may still be preferable when fine temporal resolution is required. Additionally, using both sensors as inputs yielded better results than using a single sensor, as expected due to increased information content for neural networks to exploit. Between the single-sensor models, the thigh sensor outperformed the lower back sensor overall, though performance varied by activity, consistent with previous work [41]. This difference is biomechanically plausible. The thigh undergoes pronounced, activity-specific movement patterns during locomotor tasks such as walking, running, and stair ascent/descent, resulting in clearer and more discriminative acceleration signatures. In contrast, acceleration signals measured at the lower back predominantly reflect trunk motion, which is more attenuated and often more similar across different activities. The largest discrepancies were observed for standing when using only the lower back sensor and for lying when using only the thigh sensor. This can be explained by the fact that the lower back sensor maintains a similar orientation during standing and sitting, whereas the thigh sensor exhibits similar orientations during sitting and lying. Consequently, the reduced signal contrast at each location for these activity pairs limits class separability and increases confusion when only a single sensor is used. This confusion in single-sensor models could potentially be reduced by labelling transitional movements, as transitions contain information about both the preceding and following activities that the network can exploit, even though recognising transitions remains inherently challenging [42].

4.3. A Generalizable Neural Network by Combining Datasets

Combining the two datasets resulted in the best predictive performance on the old cohort. This suggested that inclusion of data from people with different mobility characteristics is required to achieve satisfactory recognition performance [34]. The networks, serving as general or population-wide recognition models, can be personalised for each individual [43] through fine-tuning, which involves additional training using the individual's data.

4.4. Limitations

This study has several limitations. First, the data collection procedure involved pre-defined activity protocols, which may have reduced the ecological validity of natural behaviour. Second, the range of neural architectures evaluated in this study was limited, and other architectures (e.g., 2D CNNs, transformers) were not investigated, which may have influenced the observed performance. Furthermore, there was a tendency of the networks to predict more frequent activities, which is a common problem in imbalanced datasets [44]. Considering the confusion matrices, our networks were not strongly affected; however, this issue could be further mitigated by oversampling less frequent labels while adding slight perturbations to the duplicated sensor data to avoid overfitting. This issue could also be handled by data augmentation techniques ranging from time-based numerical approaches [45] to more sophisticated methods such as generative adversarial networks and diffusion models [46–48].

5. Conclusions

Overall, this study demonstrated that RNNs are well-suited for HAR using accelerometers in the context of granular activity recognition, likely thanks to their ability to model longer-range temporal dependencies. The trained networks provided accurate, fine-grained predictions at the frame-level, capturing subtle transitions between activities while remaining robust to variations in window length. Including data from the old cohort during training improved generalizability by capturing greater population diversity. From a practical perspective, these findings are important for real-world deployment. The observation that performance was maintained at lower frequencies, and that acceptable accuracy can be achieved using a single thigh-mounted sensor, supports the feasibility of low-power, minimally intrusive wearable systems using the relatively lightweight RNN architectures evaluated in this study, particularly when compared with the higher computational demands typically associated with recent architectures. This has direct implications for long-term adherence and acceptability, particularly in old cohorts and healthcare populations. Overall, the presented approach is a promising step to continuous health monitoring using low-cost, lightweight wearable sensors and neural networks to process their data.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/s26041320/s1>. Figure S1. F1-score, precision, recall and accuracy on the Young and Old Cohort Datasets for multiple sampling frequencies. Neural network architecture and hyperparameters were optimised for each experiment. The metrics per activity and their average weighted by their frequency (“weighted”) are presented. Figure S2. The distributions of F1-score, precision, recall and accuracy on the Young and Old Cohort Datasets for multiple sampling frequencies. Neural network architecture and hyperparameters were optimised for each experiment. Points shown beyond the whiskers represent statistical outliers inherent to the boxplot visualisation. Figure S3. F1-score, precision, recall and accuracy on the Young and Old Cohort Datasets for neural networks trained solely on the Young Cohort Dataset and the

Combined Dataset. The metrics per activity and their average weighted by their frequency (“weighted”) are presented. Figure S4. The distributions of F1-score, precision, recall and accuracy on the Young and Old Cohort Datasets for neural networks trained solely on the Young Cohort Dataset and the Com-bined Dataset. The metrics per activity and their average weighted by their frequency (“weighted”) are presented. Points shown beyond the whiskers represent statistical outliers inherent to the boxplot visualisation.

Author Contributions: Conceptualization, M.B. and L.A.; methodology, M.B. and L.A.; software, M.B.; validation, M.B.; formal analysis, M.B.; investigation, M.B.; resources, L.A.; data curation, M.B.; writing—original draft preparation, M.B.; writing—review and editing, M.B., J.P., L.R., S.D.D. and L.A.; visualisation, M.B.; supervision, L.A. and J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Institute for Health Research (NIHR) Newcastle Biomedical Research Centre (BRC) based at the Newcastle upon Tyne Hospital NHS Foundation Trust, Newcastle University and the Cumbria, Northumberland and Tyne and Wear (CNTW) NHS Foundation Trust. This study was also supported by the UK Research and Innovation (UKRI) Engineering and Physical Sciences Research Council (EPSRC) (TORUS, Grant Ref: EP/X036146/1). LR was awarded by the NIHR senior investigator award. SDD and LR were supported by the IDEA-FAST project that has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 853981. This JU receives support from the European Union’s Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations (EFPIA). SDD, LR and LA were also supported by the NIHR/Wellcome Trust Clinical Research Facility (CRF) infrastructure at Newcastle upon Tyne Hospitals NHS Foundation Trust. SDD was supported by the UKRI EPSRC (Grant Ref: EP/X031012/1).

Data Availability Statement: The datasets used in the study are openly available in GitHub at <https://github.com/ntnu-ai-lab/harth-ml-experiments> (accessed on 12 January 2026).

Conflicts of Interest: S.D.D. and L.R. are shareholders in Enoda Health Ltd.; this shareholding had no relationship to the work under consideration. All opinions are those of the authors and not the funders. The content in this publication reflects the authors’ view, and neither IMI nor the European Union, EFPIA, NHS, NIHR, UKRI or any associated partners are responsible for any use that may be made of the information contained herein.

Abbreviations

The following abbreviations are used in this manuscript:

HAR	Human activity recognition
CNN	Convolutional neural network
RNN	Recurrent neural network
GRU	Gated recurrent unit
ConvRec	Convolutional–Recurrent network
ReConv	Recurrent–Convolutional network
LOPO	Leave-one-participant-out

References

1. Tapia, E.M.; Intille, S.S.; Larson, K. Activity recognition in the home using simple and ubiquitous sensors. In Proceedings of the International Conference on Pervasive Computing, Vienna, Austria, 21–23 April 2004; Springer: Berlin/Heidelberg, Germany, 2004.
2. Tezcan, Ö.; Akcay, C.; Sari, M.; Cavus, M. Sensor-Based Automatic Recognition of Construction Worker Activities Using Deep Learning Network. *Sensors* **2025**, *25*, 3988.

3. Rochester, L.; Mazzà, C.; Mueller, A.; Caulfield, B.; McCarthy, M.; Becker, C.; Miller, R.; Piraino, P.; Viceconti, M.; Dartee, W.P., et al. A roadmap to inform development, validation and approval of digital mobility outcomes: The Mobilise-D approach. *Digit. Biomark.* **2020**, *4*, 13–27.
4. Viceconti, M.; Tome, M.; Dartee, W.; Knezevic, I.; Hernandez Penna, S.; Mazzà, C.; Caulfield, B.; Garcia-Aymerich, J.; Becker, C.; Maetzler, W. On the use of wearable sensors as mobility biomarkers in the marketing authorization of new drugs: A regulatory perspective. *Front. Med.* **2022**, *9*, 996903.
5. Micó-Amigo, M.E.; Bonci, T.; Paraschiv-Ionescu, A.; Ullrich, M.; Kirk, C.; Soltani, A.; Küderle, A.; Gazit, E.; Salis, F.; Alcock, L.; et al. Assessing real-world gait with digital technology? Validation, insights and recommendations from the Mobilise-D consortium. *J. Neuroeng. Rehabil.* **2023**, *20*, 78.
6. Paraschiv-Ionescu, A.; Newman, C.J.; Carcreff, L.; Gerber, C.N.; Armand, S.; Aminian, K. Locomotion and cadence detection using a single trunk-fixed accelerometer: Validity for children with cerebral palsy in daily life-like conditions. *J. Neuroeng. Rehabil.* **2019**, *16*, 1–11.
7. Abdul Jabbar, K.; Sarvestan, J.; Zia Ur Rehman, R.; Lord, S.; Kerse, N.; Teh, R.; Del Din, S. Validation of an Algorithm for Measurement of Sedentary Behaviour in Community-Dwelling Older Adults. *Sensors* **2023**, *23*, 4605.
8. Nazarahari, M.; Rouhani, H. Detection of daily postures and walking modalities using a single chest-mounted tri-axial accelerometer. *Med. Eng. Phys.* **2018**, *57*, 75–81.
9. Hickey, A.; Galna, B.; Mathers, J.C.; Rochester, L.; Godfrey, A. A multi-resolution investigation for postural transition detection and quantification using a single wearable. *Gait Posture* **2016**, *49*, 411–417.
10. Godfrey, A.; Bourke, A.; O'laighin, G.; Van De Ven, P.; Nelson, J. Activity classification using a single chest mounted tri-axial accelerometer. *Med. Eng. Phys.* **2011**, *33*, 1127–1135.
11. Figo, D.; Diniz, P.C.; Ferreira, D.R.; Cardoso, J.M. Preprocessing techniques for context recognition from accelerometer data. *Pers. Ubiquitous Comput.* **2010**, *14*, 645–662.
12. Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; Liu, Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–40.
13. Foumani, N.M.; Tan, C.W.; Webb, G.I.; Salehi, M. Improving position encoding of transformers for multivariate time series classification. *Data Min. Knowl. Discov.* **2023**, *38*, 22–48.
14. Ordóñez, F.J.; Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115.
15. Zhang, H.; Xiao, Z.; Wang, J.; Li, F.; Szczerbicki, E. A Novel IoT-Perceptive Human Activity Recognition (HAR) Approach Using Multihead Convolutional Attention. *IEEE Internet Things J.* **2020**, *7*, 1072–1080.
16. Nafea, O.; Abdul, W.; Muhammad, G.; Alsulaiman, M. Sensor-Based Human Activity Recognition with Spatio-Temporal Deep Learning. *Sensors* **2021**, *21*, 2141.
17. Xu, S.; Zhang, L.; Huang, W.; Wu, H.; Song, A. Deformable Convolutional Networks for Multimodal Human Activity Recognition Using Wearable Sensors. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14.
18. Mekruksavanich, S.; Jitpattanakul, A. LSTM Networks Using Smartphone Data for Sensor-Based Human Activity Recognition in Smart Homes. *Sensors* **2021**, *21*, 1636.
19. Zhao, Y.; Yang, R.; Chevalier, G.; Xu, X.; Zhang, Z. Deep residual bidir-LSTM for human activity recognition using wearable sensors. *Math. Probl. Eng.* **2018**, *2018*, 7316954.
20. Xu, C.; Chai, D.; He, J.; Zhang, X.; Duan, S. InnoHAR: A Deep Neural Network for Complex Human Activity Recognition. *IEEE Access* **2019**, *7*, 9893–9902.
21. Mekruksavanich, S.; Phaphan, W.; Jitpattanakul, A. Harnessing Deep Learning for Activity Recognition in Seniors' Daily Routines with Wearable Sensors. In Proceedings of the 2024 47th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 10–12 July 2024; pp. 164–167.
22. Nair, N.; Thomas, C.; Jayagopi, D.B. Human activity recognition using temporal convolutional network. In Proceedings of the Proceedings of the 5th International Workshop on Sensor-Based Activity Recognition and Interaction, Berlin, Germany, 20–21 September 2018.
23. Khan, Z.N.; Ahmad, J. Attention induced multi-head convolutional neural network for human activity recognition. *Appl. Soft Comput.* **2021**, *110*, 107671.
24. Han, H.; Zeng, H.; Kuang, L.; Han, X.; Xue, H. A human activity recognition method based on Vision Transformer. *Sci. Rep.* **2024**, *14*, 15310.

25. Yi, M.-K.; Lee, W.-K.; Hwang, S.O. A human activity recognition method based on lightweight feature extraction combined with pruned and quantized CNN for wearable device. *IEEE Trans. Consum. Electron.* **2023**, *69*, 657–670.
26. Xia, K.; Huang, J.; Wang, H. LSTM-CNN architecture for human activity recognition. *IEEE Access* **2020**, *8*, 56855–56866.
27. Yao, R.; Lin, G.; Shi, Q.; Ranasinghe, D.C. Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. *Pattern Recognit.* **2018**, *78*, 252–266.
28. Thu, N.T.H.; Han, D.S. Handle Dense Labeling in Human Activity Recognition Using Self Attention and BiLSTM. In Proceedings of the 2024 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 5–8 January 2024.
29. Koşar, E.; Barshan, B. A new CNN-LSTM architecture for activity recognition employing wearable motion sensor data: Enabling diverse feature extraction. *Eng. Appl. Artif. Intell.* **2023**, *124*, 106529.
30. Roggen, D.; Calatroni, A.; Rossi, M.; Holleczeck, T.; Förster, K.; Tröster, G.; Lukowicz, P.; Bannach, D.; Pirkel, G.; Ferscha, A. Collecting complex activity datasets in highly rich networked sensor environments. In Proceedings of the 2010 Seventh International Conference on Networked Sensing Systems (INSS), Kassel, Germany, 15–18 June 2010.
31. Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J.L. A public domain dataset for human activity recognition using smartphones. In Proceedings of the ESANN, Bruges, Belgium, 24–26 April 2013; p. 3.
32. Reiss, A.; Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In Proceedings of the 2012 16th International Symposium on Wearable Computers, Newcastle, UK, 18–22 June 2012.
33. Shoaib, M.; Bosch, S.; Incel, O.D.; Scholten, H.; Havinga, P.J. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors* **2016**, *16*, 426.
34. Ustad, A.; Logacjov, A.; Trollebø, S.Ø.; Thingstad, P.; Vereijken, B.; Bach, K.; Maroni, N.S. Validation of an activity type recognition model classifying daily physical behavior in older adults: The HAR70+ model. *Sensors* **2023**, *23*, 2368.
35. Logacjov, A.; Bach, K.; Kongsvold, A.; Bårdstu, H.B.; Mork, P.J. HARTH: A human activity recognition dataset for machine learning. *Sensors* **2021**, *21*, 7853.
36. Lee, H.-J.; Chou, L.-S. Balance control during stair negotiation in older adults. *J. Biomech.* **2007**, *40*, 2530–2536.
37. Hamel, K.A.; Cavanagh, P.R. Stair performance in people aged 75 and older. *J. Am. Geriatr. Soc.* **2004**, *52*, 563–567.
38. Elshehabi, M.; Del Din, S.; Hobert, M.A.; Warmerdam, E.; Sünkel, U.; Schmitz-Hübsch, T.; Behncke, L.M.; Heinzl, S.; Brockmann, K.; Metzger, F.G. Walking parameters of older adults from a lower back inertial measurement unit, a 6-year longitudinal observational study. *Front. Aging Neurosci.* **2022**, *14*, 789220.
39. Junker, H.; Lukowicz, P.; Troster, G. Sampling frequency, signal resolution and the accuracy of wearable context recognition systems. In Proceedings of the Eighth International Symposium on Wearable Computers, Arlington, VA, USA, 31 October–3 November 2004.
40. Yang, X.; Dinh, A.; Chen, L. Implementation of a wearable real-time system for physical activity recognition based on Naive Bayes classifier. In Proceedings of the 2010 International Conference on Bioinformatics and Biomedical Technology, Chengdu, China, 16–18 April 2010.
41. Atallah, L.; Lo, B.; King, R.; Yang, G.-Z. Sensor positioning for activity recognition using wearable accelerometers. *IEEE Trans. Biomed. Circuits Syst.* **2011**, *5*, 320–329.
42. Li, J.-H.; Tian, L.; Wang, H.; An, Y.; Wang, K.; Yu, L. Segmentation and recognition of basic and transitional activities for continuous physical human activity. *IEEE Access* **2019**, *7*, 42565–42576.
43. Hong, J.-H.; Ramos, J.; Dey, A.K. Toward personalized activity recognition systems with a semipopulation approach. *IEEE Trans. Hum. Mach. Syst.* **2015**, *46*, 101–112.
44. Sun, Z.; Song, Q.; Zhu, X.; Sun, H.; Xu, B.; Zhou, Y. A novel ensemble method for classifying imbalanced data. *Pattern Recognit.* **2015**, *48*, 1623–1637.
45. El Zein, H.; Mourad-Chehade, F.; Amoud, H. CSI-based human activity recognition via lightweight CNN model and data augmentation. *IEEE Sens. J.* **2024**, *24*, 25060–25069.
46. Bicer, M.; Phillips, A.T.; Melis, A.; McGregor, A.H.; Modenese, L. Generative deep learning applied to biomechanics: A new augmentation technique for motion capture datasets. *J. Biomech.* **2022**, *144*, 111301.

47. Bicer, M.; Phillips, A.T.; Melis, A.; McGregor, A.H.; Modenese, L. Generative adversarial networks to create synthetic motion capture datasets including subject and gait characteristics. *J. Biomech.* **2024**, *177*, 112358.
48. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.