Where to Sense Before What to Classify: A Full-Body Tracking Pilot Study on Real-Time HAR

Ruben Schlonsak[1,2], Marco Gabrecht[1,2], Jierui Li[1,3], and Denys J.C. Matthies[1,2]

[1] Technical University of Applied Sciences Lübeck, Lübeck, Germany
ruben.schlonsak@th-luebeck.de, denys.matthies@th-luebeck.de
[2] Fraunhofer IMTE, Lübeck, Germany
[3] East China University of Science and Technology, Shanghai, China

Abstract. Human Activity Recognition (HAR) research increasingly emphasizes complex deep models, while real-time wearable deployments are also constrained by latency and by *where* the body is sensed. In this controlled 3 participant pilot study, we use a full-body motion tracking suit with 17 IMUs to explore how 1D-CNN and LSTM classifiers behave across six daily activities, six window sizes, single coordinate axes, and twelve individual body segments. Three observations emerge from this setup: (1) LSTMs achieve higher average accuracy (M=0.95 vs. M=0.85), whereas CNNs classify 3.6× faster, indicating a relevant accuracy–latency trade-off for interactive use; (2) window size affects the two model families differently (CNN: r=+0.22; LSTM: r=−0.47), suggesting that window length should be tuned with respect to the chosen architecture; and (3) toe and foot segments achieve 96–97% accuracy in this controlled activity set, approaching the performance of full-body input. While these findings require validation with more participants and deployable low-cost sensors, they indicate that sensor placement and windowing can substantially shape real-time HAR performance. We derive preliminary design implications for body-worn HAR systems and discuss the limitations of transferring suit-based results to practical wearable deployments.

Keywords: Human Activity Recognition · Wearable Sensors · Motion Capture · Sensor Placement · Real-Time Classification.

1 Introduction

Human Activity Recognition (HAR) has shifted from handcrafted features to deep learning. Despite strong benchmark results, real-world deployments continue to face brittleness, latency, and power constraints. Recent surveys argue that further progress depends not only on network depth, but also on *how and where* movement is sensed [22]. CNNs, LSTMs, and Transformers dominate current research, yet their complexity increases compute and memory requirements [11,5]. At the same time, sensor placement can strongly affect recognition

performance [10,32], and edge deployments require explicit attention to window size and latency [9,16].
These deployment-centered factors are often studied in isolation. Model comparisons typically fix the sensor configuration and window length, while placement studies often fix the model architecture. As a result, practitioners building interactive wearable systems receive limited guidance on the joint design question: given a latency budget, which body location, window length, and architecture should be combined? Full-body motion capture provides a controlled way to explore this question, because all body segments are observed simultaneously and can therefore be compared on the same movement instances rather than across separate sensor recordings.
This paper examines these factors in a controlled pilot study using a full-body motion tracking suit. We evaluate two widely used architecture families for sensor-based HAR, a 1D convolutional neural network and an LSTM, across six daily activities. We systematically vary the temporal window, the coordinate representation, and the body segment used as input.
Our contributions are:

- We conduct a controlled pilot study with three participants using a full-body tracking suit to compare 1D-CNN and LSTM classifiers across six activities, six window sizes, coordinate representations, and twelve body segments.
- We quantify architecture-dependent accuracy–latency trade-offs: LSTMs achieve higher average accuracy (M=0.95 vs. M=0.85), while CNNs classify 3.6× faster; window size affects the two architectures differently (CNN accuracy: r=+0.22; LSTM accuracy: r=−0.47).
- We provide preliminary evidence that distal lower-limb segments preserve much of the discriminative information in the studied locomotion and posture activities, with suit-derived toe and foot segments reaching 96–97% accuracy, and derive practical implications for sensor placement, windowing, and coordinate handling in real-time body-worn HAR.

The remainder of this paper is organized as follows. Section 2 reviews deep learning for HAR, sensor placement, and real-time constraints. Section 3 describes the apparatus, dataset, preprocessing, and evaluation protocol. Section 4 reports results on model comparison, window size, body segments, and coordinate representation. Section 5 discusses the findings, limitations, and resulting design implications. Sections 6 and 7 outline future work and conclude.

2 Related Work

2.1 Deep Learning in HAR

Early HAR relied on hand-engineered features with classifiers such as SVMs and Random Forests [2,29,33]. While effective in controlled settings, these approaches required domain expertise and were sensitive to user variability. Deep learning enabled end-to-end learning from raw sensor streams [17], reducing

manual feature design and improving robustness. Zeng et al. pioneered CNNs for accelerometer data [40]; Ordoñez and Roggen introduced DeepConvLSTM, combining convolutional and recurrent modeling [26].

Subsequent studies converted sensor readings into signal images and applied 2D CNNs [18], or stacked 1D convolutions across multichannel wearable data [37,38], and by the late 2010s deep models had become the de facto standard for HAR [15,13]. Since vanilla CNNs have a limited temporal receptive field, many approaches integrate recurrent layers: Dua et al. fused spatial and temporal features in a multi-layer CNN–GRU network [12], and Yin et al. combined a multiscale CNN with a bidirectional LSTM, outperforming standalone CNNs or LSTMs on complex activity sequences [39].

To address the limitations of basic CNN/RNN architectures, recent work explores attention mechanisms and Transformers. Mahmud et al. introduced sensor-level, temporal, and self-attention modules for adaptive sensor fusion [23]; others integrated Squeeze-and-Excitation channel attention into CNNs to improve sensitivity to informative signal axes and sensor locations [14,19]. Luptáková et al. adapted a Transformer to smartphone motion data and achieved over 99% accuracy on UCI-HAR, exceeding an LSTM baseline by nearly 10% [11]; Cao et al. proposed sparse-attention Transformers that reduce computation while maintaining accuracy [5]. However, models with 50–500K parameters challenge deployment on microcontrollers with 256 KB to 1 MB of RAM, raising concerns about memory, latency, and energy use in edge settings. Notably, most of this literature reports accuracy as the primary or only metric, while the prediction time of the trained model, which directly bounds the responsiveness of an interactive system, is reported far less consistently. The present study treats both quantities as equally important outcomes.

2.2 Sensor Placement

Sensor location fundamentally constrains recognition performance by determining which body dynamics are observable. For locomotion-related activities, lower-limb placements consistently outperform upper-body locations [1,27], as they capture primary gait dynamics. Davoudi et al. systematically evaluated five standard placements (wrist, hip, ankle, upper arm, thigh) and found that a single hip-mounted accelerometer yielded the highest balanced accuracy, while using all five devices improved accuracy only marginally [10], indicating that placement often outweighs sensor count. Such results echo earlier work identifying the waist/hip, thigh, and ankle as high-value sites that capture the gross movement of legs and torso [7,28,31]. Similarly, in fall detection, waist- and thigh-mounted sensors yielded the highest sensitivity, whereas arm-mounted sensors were less informative [34].

Ankle-mounted sensors excel for gait analysis [24,3], capturing foot-strike and stance patterns that are weakly expressed elsewhere. Recent studies additionally highlight the benefits of multimodal sensor fusion for recognition accuracy [4,6,25]. More recently, foot-mounted IMUs have been shown to reconstruct lower-limb motion and detect terrain characteristics [36], suggesting that the

distal lower limb carries considerably richer information than its traditional role as a mere step counter implies.

A limitation of most placement studies is that each candidate location requires its own device and often its own recording session, so placements are compared across slightly different movement instances. Full-body motion capture sidesteps this problem: because all segments are tracked simultaneously, every placement can be evaluated on identical movements, isolating the effect of location from the effect of movement variability. We use this property here, and we term the resulting perspective *embodiment-first design*: recognition performance is frequently shaped by sensor placement as much as by model architecture.

2.3 Real-Time Constraints

For HAR systems deployed on wearable or mobile devices, "real time" typically means low end-to-end latency: the delay between an activity happening and the system's response should usually stay well under a second for interactive applications such as fitness coaching or safety monitoring [20]. Achieving this requires careful design of the entire pipeline, from sensor sampling and windowing to model inference [8]. Larger windows provide more context and often higher accuracy, but incur longer delays since the system must wait for the window to fill, an inherent trade-off explored in depth [41,31,35].

Window size thus directly impacts recognition delay and responsiveness. Czekaj et al. showed that reducing window lengths from 5 s to 1 s cuts latency by a factor of five, often with limited accuracy loss [9]. Shorter windows therefore improve interactivity but reduce temporal context, creating an inherent accuracy–latency trade-off. Importantly, this trade-off has two distinct components that are easily conflated: the *accumulation delay*, i.e., the time required to gather enough samples to fill a window, and the *prediction time* of the model itself. The former is determined solely by window length and sampling rate, the latter by architecture and implementation. Our results show that the two components favor different design choices, which is why we report them separately.

TinyML enables neural networks on microcontrollers through quantization and efficient architectures [42,21,16], making on-device inference feasible under tight resource constraints. However, most work emphasizes compressing deep models rather than asking whether informative sensor placement and architecture-aware windowing may already be sufficient for real-time HAR.

3 Method

3.1 Apparatus and Data Collection

We used the Rokoko Smartsuit Pro II [30], a full-body motion tracking suit equipped with 17 IMUs, each combining an accelerometer, a gyroscope, and a magnetometer, as shown in Fig. 1. The suit streams data wirelessly to the Rokoko Studio software, which fuses the raw inertial signals into absolute 3D

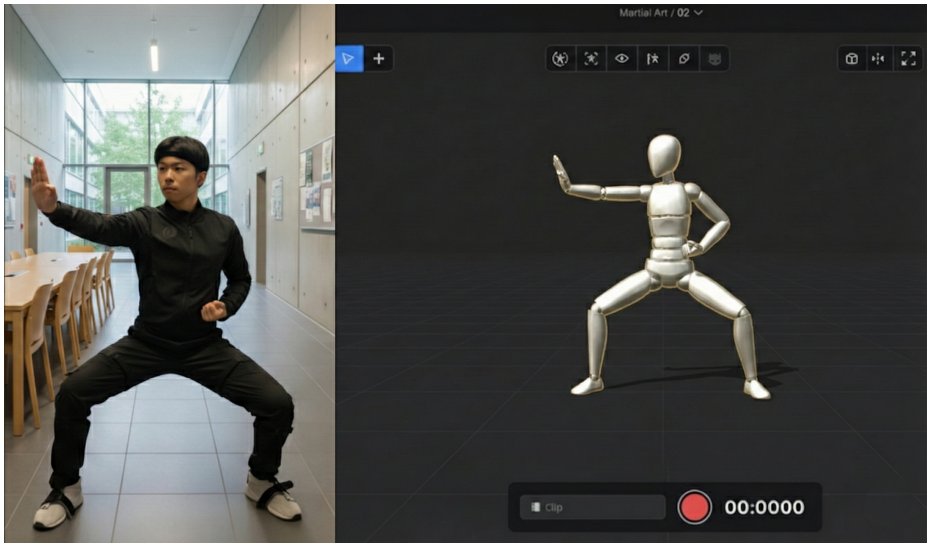

Fig. 1. The equipped Rokoko Smartsuit Pro II (left) and the visualization in Rokoko Studio (right).

positions of 21 body segments (head, neck, hips, and, for both body sides, shoulder, arm, forearm, hand, thigh, shin, foot, toe, and toe tip) at a sampling rate of 30 Hz. Working on fused positional data rather than raw inertial signals has two consequences that matter for interpreting our results: it removes sensor-level noise and drift handling from the learning problem, and it expresses all body segments in a common spatial reference frame, which is precisely what enables the per-segment comparisons below.

Recordings were exported as BVH files and converted to CSV using the `bvh-converter` library, yielding one timestamp column, 63 position channels (21 segments $\times$ 3 axes), and an auxiliary hip-height channel, for a total of 64 feature columns; Fig. 2 shows the overall pipeline.

Three participants took part in the recording, all male, aged 22–23 years, with body heights between 170 and 174 cm. Each participant performed six daily activities indoors: jogging, jumping, lying, sitting, standing, and walking. The activity set deliberately spans three qualitatively different signal regimes: periodic locomotion, namely jogging and walking, impulsive movement, namely jumping, and quasi-static postures, namely lying, sitting, and standing. For each participant and activity, approximately 20 s were recorded for training and, in a separate recording, approximately 9 s for testing, so that training and test windows never originate from the same recording session. Each activity was stored in a separate CSV file and annotated by hand. Because the data stream leaving Rokoko Studio is already fused and gap-free, no additional filtering or missing-value treatment was required.

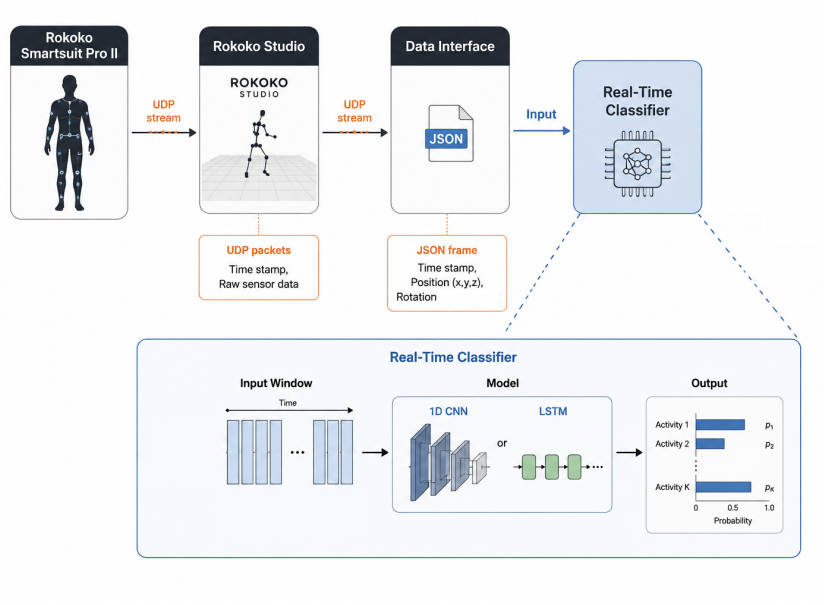

Fig. 2. Overall workflow for recording motion data from the Rokoko suit and classifying activities.

3.2 Preprocessing and Segmentation

Because Rokoko Studio exports absolute coordinates, identical activities performed at different locations or with different headings produce different raw signals: a person walking eastward and the same person walking northward generate distinct trajectories even though the activity is the same. We therefore converted the horizontal trajectories into relative movement via first-order differencing (`pandas.DataFrame.diff()`), which encodes frame-to-frame displacement instead of absolute position and thereby removes the dependence on starting point and heading. The gravity-aligned vertical axis was additionally examined in unmodified form, since absolute height carries posture information that differencing would destroy (Sect. 4.4).

Activity labels were encoded numerically, and the streams were segmented with a sliding window (stride = 2 samples) using 6 sizes $W \in \{8, 16, 32, 64, 128, 256\}$ frames, corresponding to ≈0.27–8.53 s at 30 Hz. Note that on recordings of fixed length, the number of extractable windows shrinks as the window grows: at W=8 the training material yields roughly 5300 windows, whereas at W=256 only about 180 remain. This sample-count effect becomes relevant when interpreting the results for very large windows. Unless stated otherwise, results refer to W=32 (≈1 s).

3.3 Models

We compared two widely used architectures for sensor-based HAR. The first is a single-headed 1D-CNN: two convolutional layers with ReLU activations perform feature learning along the time axis, a 1D max-pooling layer condenses the learned feature maps, a dropout layer with rate 0.5 counteracts overfitting, and a softmax output layer produces the class distribution. The second is an LSTM network, whose gated recurrent state is designed to capture temporal dependencies across the window. Both models were implemented with the TensorFlow library and trained for 50 epochs with a batch size of 64, using the Adam optimizer and categorical cross-entropy loss. We intentionally kept both architectures small and standard: the goal of this study is not to maximize accuracy through architecture search, but to isolate the effects of windowing, coordinate representation, and sensor placement under models that are realistic candidates for embedded deployment.

3.4 Evaluation Protocol

The evaluation addresses four hypotheses, which structure the results section:

- **H1:** The CNN is better suited than the LSTM for constructing a real-time classifier.
- **H2:** Larger windows make classification more accurate.
- **H3:** Using a single coordinate axis yields higher accuracy than using all three.
- **H4:** Data from specific body segments suffices for accurate recognition.

Because neural network training is stochastic, every configuration (model $\times$ window size, as well as the axis and body-segment ablations) was trained and evaluated 10 times; we report means (M) and standard deviations (SD) across runs, and confusion matrices for the reference configuration (W=32). Accuracy is complemented by the measured prediction time per window, since both jointly determine real-time suitability.

In addition, a real-time classifier was implemented to verify that the pipeline operates on live data. The program loads a trained model, connects to Rokoko Studio, and receives motion data as UDP packets at 30 Hz, each containing a JSON-encoded frame of timestamped position values. Frames are buffered until a window is filled, then parsed into a data frame and passed to the model for inference. The decision interval of this system is therefore lower-bounded by the window accumulation time $\Delta t_{\min} \geq W/f_s$ plus the model's prediction time, which makes the window size a first-class latency parameter rather than a mere preprocessing detail.

4 Results

4.1 CNN vs. LSTM: Accuracy and Prediction Time

Table 1 summarizes the aggregate model performance across all tested window sizes, with each configuration evaluated over 10 stochastic runs. Overall, the

Table 1. Model comparison across window sizes (mean and SD over 10 runs per configuration).

Model	Accuracy		Prediction time [s]	
	M	SD	M	SD
1D-CNN	0.85	0.0064	**0.141**	0.0002
LSTM	**0.95**	**0.0005**	0.509	0.0046

LSTM achieves higher accuracy and substantially lower run-to-run variation (M=0.95, SD=0.0005) than the 1D-CNN (M=0.85, SD=0.0064), indicating more stable classification performance. This gain in accuracy comes at a clear computational cost: the 1D-CNN requires only 0.141 s per prediction on average, whereas the LSTM requires 0.509 s. Thus, the CNN provides a 3.6× faster inference time, making it more attractive when responsiveness is constrained by the model-computation part of the real-time HAR pipeline.

At the reference window of W=32, both models exceed 90% overall accuracy. The confusion matrices show that errors are not uniformly distributed but concentrate on semantically adjacent pairs: jogging is occasionally confused with walking for both models, since the two share periodic gait structure and differ mainly in intensity, and the CNN additionally confuses standing with sitting, two postures whose relative-movement signals are both close to zero. Jumping and lying are recognized nearly perfectly by both models, as their signals are the most distinctive in the set.

With respect to **H1**, the picture is deliberately two-sided: the LSTM wins on accuracy and stability, the CNN on prediction time. For a real-time classifier whose responsiveness is bounded by compute, the 3.6× faster CNN is the more suitable backbone, which supports H1 in its deployment-oriented reading.

4.2 Window Size: Opposite Effects per Architecture

Window length exerts model-specific effects. Over 8–128 frames, CNN accuracy correlates positively with window size (r=+0.22): longer windows expose more repetitions of the underlying movement cycle for the convolutional filters to integrate, which benefits a model whose receptive field is local. LSTM accuracy, in contrast, correlates negatively with window size (r=−0.47): the recurrent state already extracts the salient temporal cues from short sequences, so longer sequences mainly add redundancy and optimization burden. This directly contradicts the common intuition captured in **H2**; the hypothesis holds only for the CNN and must be rejected for the LSTM.

Prediction time grows with window size for both models (CNN: r=0.61; LSTM: r=0.90), and in an online setting the earliest possible decision is additionally lower-bounded by the accumulation time W/f_s, spanning ≈0.27 s (8 frames) to ≈8.53 s (256 frames) at 30 Hz. Both latency components therefore

push in the same direction: shorter windows respond faster, both because less data must be gathered and because less data must be processed.

At W=256, deviations from these monotone trends appear, and the CNN's prediction time at W=128 does not exceed that at W=64 as expected. We attribute these irregularities to the sample-count effect described above: with a stride of 2 on fixed-length recordings, very large windows drastically reduce the number of training samples, leaving too few examples for stable learning and timing estimates. For interactive applications, short windows ($\leq$1 s) combined with an LSTM, or with a CNN when prediction time is critical, offer the best trade-off.

4.3 Sensor Placement: Single Body Segments

To test **H4**, we trained both models (W=32) on each body segment's channels in isolation, using identical movements for every placement since all segments were tracked simultaneously. Figure 3 shows the resulting accuracy distributions over 10 runs per segment. Across segments, LSTMs (M=0.92, SD=0.0021) outperform CNNs (M=0.89, SD=0.0090) on average, mirroring the full-body ranking.

The single best segments are located at the distal lower limb: toe data yields the highest CNN accuracy (M=0.97, SD=0.0003), and foot data the highest LSTM accuracy (M=0.96, SD=0.0008), both close to the accuracy obtained with full-body input. The weakest segments for the CNN are the hand (M=0.74, SD=0.0019), forearm (M=0.81, SD=0.0024), and shin (M=0.87, SD=0.0057); for the LSTM, the shin (M=0.86, SD=0.0006) and hips (M=0.87, SD=0.0019). Beyond their high means, toe and foot exhibit the tightest distributions across runs, whereas several other segments produce visibly wider spreads and occasional outlier runs. Within this controlled activity set, H4 is supported: a single well-chosen distal lower-limb segment preserved most of the discriminative information.

4.4 Coordinate Representation

Converting the horizontal (x/z) trajectories to relative movement removes the dependence on starting position and heading and is a prerequisite for location-independent recognition. To test **H3**, we additionally fed the models a single coordinate axis at a time (W=32).

The gravity-aligned vertical axis is by far the most informative single axis: an LSTM using only y-data reaches M=0.98 (SD=0.0001), its best result overall, and y is also the best single axis for the CNN (M=0.85, SD=0.0026). The horizontal axes perform substantially worse in isolation: z-only input yields the CNN's weakest result (M=0.75, SD=0.0042), and x-only input the LSTM's weakest (M=0.81, SD=0.0008). This asymmetry is biomechanically plausible: the six activities differ strongly in vertical signature (body height while lying vs. sitting vs. standing; vertical oscillation while jogging and jumping), while their horizontal signatures overlap.

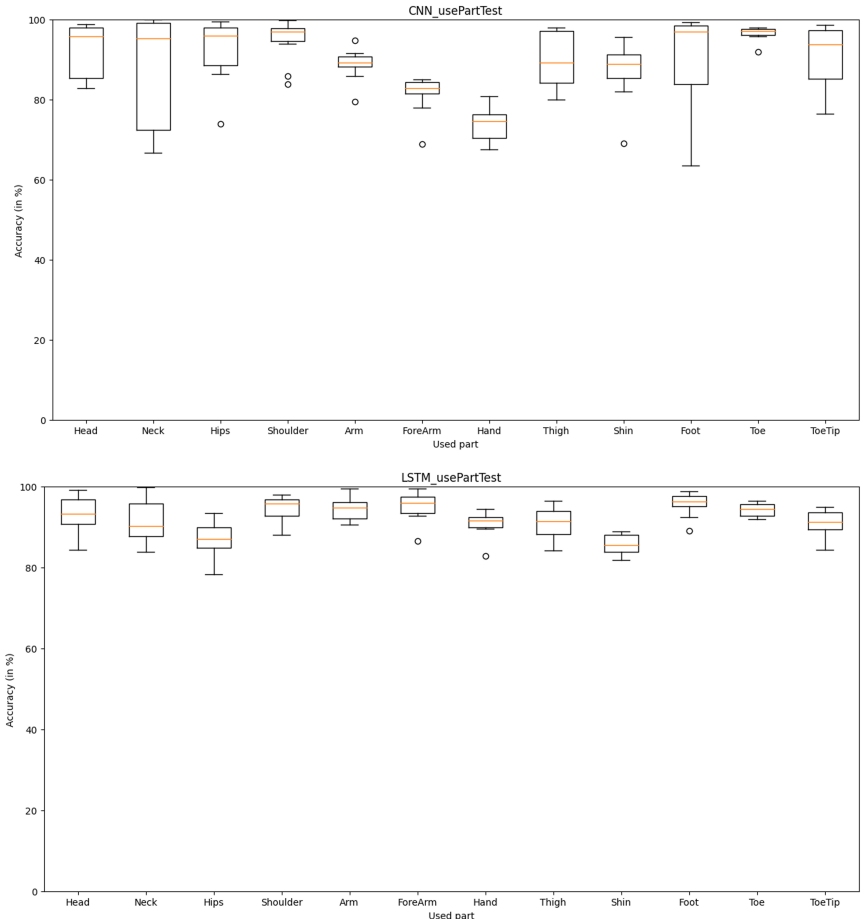

Fig. 3. Classification accuracy by body segment for CNN (top) and LSTM (bottom) models over 10 runs each (W=32). Toe and foot achieve the highest and most stable accuracy.

Single-axis input is, however, not uniformly beneficial, so H3 must be rejected as a general rule: the CNN with all three axes (M=0.88, SD=0.0021) outperforms its y-only variant, and discarding horizontal movement makes classes that differ mainly in horizontal displacement indistinguishable in principle, even if the present activity set does not fully expose this failure mode. The practical conclusion is to preserve the gravity-aligned axis unmodified and to represent horizontal axes relatively, rather than to discard them.

5 Discussion

Across all evaluations, two deployment-centered factors, namely where the body is sensed and how the signal is windowed, influenced recognition performance at least as strongly as the choice between the two deep architectures. A single toe or foot sensor recovered nearly the full-body accuracy, and choosing the window size appropriate to the architecture mattered more than adding temporal context indiscriminately.

5.1 Why Placement and Windowing Dominate

Distal lower-limb signals are maximally discriminative. For the locomotion- and posture-centric activity set studied here, the feet and toes carry the strongest signatures: step impacts, stance phases, and vertical excursions separate jogging, jumping, and walking, while their near-absence separates the static postures. The distal position at the end of the kinematic chain also amplifies movement amplitudes, so the same gait cycle produces larger, cleaner signals at the toe than at the hip. Upper-limb segments, by contrast, move more idiosyncratically and overlap across classes, which explains the low accuracy of hand and forearm data for the CNN. This aligns with prior placement studies favoring the lower limb for locomotion [1,10] and extends them by showing that even the most distal segments, toe and toe tip, are excellent sites, a region conventional placement studies rarely instrument.

Architectures consume context differently. The opposite window-size correlations indicate that "more context" is not universally better. Convolutions integrate local patterns and benefit from seeing several movement cycles per window, whereas the LSTM's recurrent state extracts sufficient temporal structure from short sequences and degrades when sequences grow long relative to the available training data. Window size should therefore be treated as an architecture-coupled hyperparameter with latency consequences, not as a fixed preprocessing convention inherited from prior work.

Gravity is the reference frame that matters. The vertical axis alone supports up to 98% accuracy, confirming that posture height and vertical dynamics are the dominant cues for this activity set. This also cautions against normalization schemes that remove absolute vertical information during preprocessing: a pipeline that blindly standardizes all channels would destroy exactly the signal that separates lying from sitting from standing.

Accuracy and responsiveness are separable goals. The LSTM is more accurate; the CNN responds faster. Which matters more depends on the application: a retrospective activity diary can afford the LSTM's prediction time, while an interactive coaching system that must react within a fraction of a second cannot. Reporting only accuracy, as much of the literature does, hides this decision entirely.

5.2 Limitations

Data were collected indoors from three participants, which is the central limitation of this study. The dataset is also small in absolute terms (roughly 60 s of training material per activity), which particularly affects the large-window conditions.

The activity set comprises six coarse-grained activities; fine-grained or upper-body-centric tasks such as manipulation, gestures, or household work may favor entirely different placements, and the dominance of the feet reported here should not be extrapolated to such tasks. The motion capture suit provides fused, high-fidelity position estimates rather than raw IMU signals, and is itself not suitable for daily deployment; results may therefore overestimate what individual low-cost wearables can achieve, although the body-segment ablation suggests that foot- and toe-mounted sensing retains most of the discriminative information. Finally, prediction times were measured on a host system, not on a microcontroller; absolute values will shift on embedded hardware even if the relative CNN/LSTM ordering is expected to persist.

5.3 Threats to Validity

Task composition. The activity set is dominated by locomotion and posture classes, which naturally favors lower-limb sensing. If the task distribution overweighted manipulative activities, the placement ranking would plausibly invert. Conclusions about placement should therefore always be read relative to the task set.

Representation pipeline. BVH→CSV conversion, the suit's internal sensor fusion, and coordinate handling all influence separability. The finding that the unmodified vertical axis helps is contingent on reliable height estimation; drift or inconsistent calibration would attenuate this effect. Likewise, first-order differencing suppresses slow horizontal drift but also attenuates low-frequency movement components that other tasks might need.

Hardware and sampling-rate limits. Results are tied to the Rokoko suit, its segment set, and its 30 Hz output. Performance with raw IMUs, other sampling rates, magnetic disturbances, or outdoor surfaces may differ. The 30 Hz rate also caps the temporal resolution available to both models, which may compress the difference between architectures relative to higher-rate setups.

Statistical inference. We report means and standard deviations over 10 runs per configuration but did not perform formal hypothesis tests across configurations. The placement gaps are large (toe/foot $\approx$0.96–0.97 vs. hand $\approx$0.74 for the CNN) and run-to-run variance is small, so we treat them as strong design signals rather than definitive statistical claims.

5.4 Design Rules

From the empirical findings, we derive the following guidelines for real-time, body-worn HAR targeting locomotion and posture activities:

1. **Match sensor placement to the target activity dynamics.** For the locomotion- and posture-oriented activities studied here, foot and toe segments reached 96–97% accuracy, close to full-body input, and were among the most stable placements across runs. This supports prioritizing distal lower-limb sensing when activities are dominated by gait, impacts, and body-height changes. For upper-body, gesture, or object-interaction tasks, placement should be re-evaluated rather than transferred directly.
2. **Choose the window per architecture, not per convention.** LSTMs perform best with short windows, CNNs improve with longer ones. Around 1 s (32 frames at 30 Hz) both exceed 90% accuracy while keeping the accumulation delay interactive. Never adopt a window length from prior work without checking it against the chosen architecture.
3. **Preserve the gravity-aligned vertical axis; relativize the horizontal axes.** Vertical position dynamics are the single most informative cue; horizontal trajectories should be converted to relative movement to remove location and heading dependence, but not discarded.
4. **Prefer CNNs when prediction time is critical.** CNNs classify 3.6× faster than LSTMs at accuracy sufficient for many applications; LSTMs are preferable when accuracy and run-to-run stability dominate the requirements and the latency budget tolerates them.
5. **Budget both latency components explicitly.** End-to-end responsiveness is bounded by the accumulation delay W/f_s *plus* the model's prediction time. Optimizing one while ignoring the other, e.g., compressing the model but keeping a 5 s window, wastes the effort.
6. **Report deployment metrics.** Beyond accuracy, studies should report prediction time and the window accumulation delay, as both bound the responsiveness of real-time systems and are essential for fair comparisons between architectures.

5.5 Reproducibility and Reporting

To strengthen comparability across HAR studies, we recommend: (i) releasing preprocessing code, including coordinate handling and differencing choices; (ii) documenting how train and test recordings were separated, since window-level leakage under sliding-window segmentation silently inflates results; (iii) publishing ablation sheets that jointly report accuracy, prediction time, and window size; and (iv) including confusion matrices to reveal interaction-relevant errors such as jogging/walking confusion, which aggregate accuracy conceals.

6 Future Work

Population diversity and cross-subject validation. The three participant design is the most pressing limitation. We will extend data collection to multiple participants of varying age, height, and body composition, enabling leave-one-subject-out evaluation and testing whether the placement and windowing effects replicate across users. Gait patterns vary substantially with age and mobility, so populations such as older adults or users of assistive devices are of particular interest.

Minimalist hardware validation. The body-segment ablation suggests that foot-only systems are viable, but this requires validation beyond motion capture suits. We will evaluate commercial insole IMUs and foot-mounted low-cost sensors against the full-body baseline, asking whether reduced sensor fidelity and the absence of studio-grade sensor fusion degrade accuracy, how shoe type affects signal quality, and whether pressure sensing can complement inertial data for stance detection.

Adaptive inference. Fixed windows impose unnecessary latency for activities with natural boundaries. Gait provides heel-strike and toe-off events; transitions between activities create detectable discontinuities. Event-triggered segmentation could decouple decision timing from window length and reduce the accumulation delay that currently lower-bounds responsiveness. Confidence-based early exit may further reduce computation for unambiguous activities such as lying.

Privacy–utility trade-off. Full-body kinematics enable person re-identification from gait signatures, presenting a privacy risk for continuous monitoring. Future work should quantify this risk and explore representations that preserve activity discriminability while obscuring identity, e.g., body-normalized coordinates or adversarial training against identification. The finding that a single foot sensor suffices is encouraging here as well: less captured kinematics means less identifying information in the first place.

7 Conclusion

This paper examined real-time HAR with a full-body motion tracking suit, comparing 1D-CNN and LSTM classifiers across window sizes, coordinate axes, and individual body segments on six daily activities. Three results carry practical weight. First, LSTMs are more accurate and more stable, but CNNs classify 3.6× faster, and prediction time, not accuracy, is the binding constraint for interactive systems. Second, window size helps CNNs yet hurts LSTMs, so it must be tuned per architecture, and it simultaneously sets the accumulation delay that bounds responsiveness. Third, single toe or foot sensors reach 96–97% accuracy, rivaling full-body sensing, and do so with the lowest run-to-run variance of all placements.

These results support an *embodiment-first* perspective: before scaling model complexity, practitioners should ask where to sense and how to window. The design rule is simple: start with the body, not the model. Future validation with more participants and deployable foot-mounted sensors is needed to test how far these suit-derived design rules transfer to practical wearable HAR systems.

References

1. Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., Amirat, Y.: Physical human activity recognition using wearable sensors. Sensors **15**(12), 31314–31338 (2015). https://doi.org/10.3390/s151229858
2. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) Pervasive Computing. pp. 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24646-6_1
3. Beaufils, B., Chazal, F., Grelet, M., Michel, B.: Robust stride detector from ankle-mounted inertial sensors for pedestrian navigation and activity recognition with machine learning approaches. Sensors **19**(20) (2019). https://doi.org/10.3390/s19204491
4. Benos, L., Tsaopoulos, D., Tagarakis, A.C., Kateris, D., Bochtis, D.: Optimal sensor placement and multimodal fusion for human activity recognition in agricultural tasks. Applied Sciences **14**(18) (2024). https://doi.org/10.3390/app14188520
5. Cao, K., et al.: Human behavior recognition based on sparse transformer for wearable sensors. Scientific Reports **13**, 15789 (2023). https://doi.org/10.1038/s41598-023-43075-6
6. Cao, M., Wan, J., Gu, X.: Clear: Multimodal human activity recognition via contrastive learning based feature extraction refinement. Sensors **25**(3) (2025). https://doi.org/10.3390/s25030896
7. Casale, P., Pujol, O., Radeva, P.: Human activity recognition from accelerometer data using a wearable device. In: Vitrià, J., Sanches, J.M., Hernández, M. (eds.) Pattern Recognition and Image Analysis. pp. 289–296. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21257-4_36
8. Cero Dinarević, E., Baraković Husić, J., Baraković, S.: Step by step towards effective human activity recognition: A balance between energy consumption and latency in health and wellbeing applications. Sensors **19**(23) (2019). https://doi.org/10.3390/s19235206
9. Czekaj, Ł., Kowalewski, M., Domaszewicz, J., Kitłowski, R., Szwoch, M., Duch, W.: Real-time sensor-based human activity recognition for efitness and ehealth platforms. Sensors **24**(12), 3891 (2024). https://doi.org/10.3390/s24123891
10. Davoudi, A., Mardini, M.T., Nelson, D., Albinali, F., Ranka, S., Rashidi, P., Manini, T.M.: The effect of sensor placement and number on physical activity recognition and energy expenditure estimation in older adults: Validation study. JMIR Mhealth Uhealth **9**(5), e23681 (May 2021). https://doi.org/10.2196/23681
11. Dirgová Luptáková, I., Kubovčík, M., Pospíchal, J.: Wearable sensor-based human activity recognition with transformer model. Sensors **22**(5) (2022). https://doi.org/10.3390/s22051911
12. Dua, N., Singh, S.N., Semwal, V.B.: Multi-input cnn-gru based human activity recognition using wearable sensors. Computing **103**(7), 1461–1478 (Jul 2021). https://doi.org/10.1007/s00607-021-00928-8

13. Gu, F., Chung, M.H., Chignell, M., Valaee, S., Zhou, B., Liu, X.: A survey on deep learning for human activity recognition. ACM Comput. Surv. **54**(8) (Oct 2021). https://doi.org/10.1145/3472290
14. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7132–7141 (2017), https://api.semanticscholar.org/CorpusID:140309863
15. Ignatov, A.: Real-time human activity recognition from accelerometer data using convolutional neural networks. Applied Soft Computing **62**, 915–922 (2018). https://doi.org/10.1016/j.asoc.2017.09.027
16. Jaramillo, I.E., Jeong, J.G., Lopez, P.R., Lee, C.H., Kang, D.Y., Ha, T.J., Oh, J.H., Jung, H., Lee, J.H., Lee, W.H., Kim, T.S.: Real-time human activity recognition with imu and encoder sensors in wearable exoskeleton robot via deep learning networks. Sensors **22**(24) (2022). https://doi.org/10.3390/s22249690, https://www.mdpi.com/1424-8220/22/24/9690
17. Jeyakumar, J.V., Sarker, A., Garcia, L.A., Srivastava, M.: X-CHAR: A concept-based explainable complex human activity recognition model. Proc ACM Interact Mob Wearable Ubiquitous Technol **7**(1) (Mar 2023). https://doi.org/10.1145/3580804
18. Jiang, W., Yin, Z.: Human activity recognition using wearable sensors by deep convolutional neural networks. In: Proceedings of the 23rd ACM International Conference on Multimedia. p. 1307–1310. MM '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2733373.2806333
19. Khan, Z.N., Ahmad, J.: Attention induced multi-head convolutional neural network for human activity recognition. Applied Soft Computing **110**, 107671 (2021). https://doi.org/10.1016/j.asoc.2021.107671
20. Lane, N.D., Georgiev, P.: Can deep learning revolutionize mobile sensing? In: Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications. p. 117–122. HotMobile '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2699343.2699349
21. Lin, J., Zhu, L., Chen, W.M., Wang, W.C., Han, S.: Tiny machine learning: Progress and futures [feature]. IEEE Circuits and Systems Magazine **23**(3), 8–34 (2023). https://doi.org/10.1109/mcas.2023.3302182
22. Liu, H., Gamboa, H., Schultz, T.: Sensor-based human activity and behavior research: Where advanced sensing and recognition technologies meet. Sensors **23**(1) (2023). https://doi.org/10.3390/s23010125, https://www.mdpi.com/1424-8220/23/1/125
23. Mahmud, S., Tonmoy, M.T.H., Bhaumik, K.K., Rahman, A.K.M.M., Amin, M.A., Shoyaib, M., Khan, M.A.H., Ali, A.A.: Human activity recognition from wearable sensor data using self-attention. In: ECAI 2020: 24th European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 1332–1339. IOS Press (2020). https://doi.org/10.3233/FAIA200236
24. Mannini, A., Rosenberger, M., Haskell, W.L., Sabatini, A.M., Intille, S.S.: Activity recognition in youth using single accelerometer placed at wrist or ankle. Medicine & Science in Sports & Exercise **49**(4), 801–812 (2017). https://doi.org/10.1249/MSS.0000000000001144
25. Oh, Y., Kim, S.: Multi-modal lifelog data fusion for improved human activity recognition: A hybrid approach. Information Fusion **110**, 102464 (2024). https://doi.org/10.1016/j.inffus.2024.102464
26. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1) (2016). https://doi.org/10.3390/s16010115

27. Özdemir, A.T., Barshan, B.: Detecting falls with wearable sensors using machine learning techniques. Sensors **14**(6), 10691–10708 (2014). https://doi.org/10.3390/s140610691
28. Quesada, F.J., Moya, F., Espinilla, M., Martínez, L., Nugent, C.D.: Feature sub-set selection for activity recognition. In: Inclusive Smart Cities and e-Health (ICOST 2015). Lecture Notes in Computer Science, vol. 9102, pp. 307–312. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19312-0_27
29. Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3. p. 1541–1546. IAAI'05, AAAI Press (2005)
30. Rokoko Electronics ApS: Smartsuit pro ii — inertial motion capture suit (2021), https://www.rokoko.com/products/smartsuit-pro, product page
31. Stisen, A., Blunck, H., Bhattacharya, S., Prentow, T.S., Kjærgaard, M.B., Dey, A., Sonne, T., Jensen, M.M.: Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys '15). pp. 127–140. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2809695.2809718
32. Tao, W., Chen, H., Moniruzzaman, M., Leu, M.C., Yi, Z., Qin, R.: Attention-based sensor fusion for human activity recognition using imu signals. vol. abs/2112.11224 (2021), https://api.semanticscholar.org/CorpusID:245353860
33. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) Pervasive Computing. pp. 158–175. Springer Berlin Heidelberg, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24646-6_10
34. Teng, S., Kim, J.Y., Jeon, S., Gil, H.W., Lyu, J., Chung, E.H., Kim, K.S., Nam, Y.: Analyzing optimal wearable motion sensor placement for accurate classification of fall directions. Sensors **24**(19) (2024). https://doi.org/10.3390/s24196432
35. Wang, G., Li, Q., Wang, L., Wang, W., Wu, M., Liu, T.: Impact of sliding window length in indoor human motion modes and pose pattern recognition based on smartphone sensors. Sensors **18**(6), 1965 (2018). https://doi.org/10.3390/s18061965
36. Wang, Y., Fehr, K.H., Adamczyk, P.G.: Impact-aware foot motion reconstruction and ramp/stair detection using one foot-mounted inertial measurement unit. Sensors **24**(5) (2024). https://doi.org/10.3390/s24051480
37. Yang, J.B., Nguyen, M.N., San, P.P., Li, X.L., Krishnaswamy, S.: Deep convolutional neural networks on multichannel time series for human activity recognition. In: Proceedings of the 24th International Conference on Artificial Intelligence. p. 3995–4001. IJCAI'15, AAAI Press (2015)
38. Yao, S., Hu, S., Zhao, Y., Zhang, A., Abdelzaher, T.: Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In: Proceedings of the 26th International Conference on World Wide Web. p. 351–360. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). https://doi.org/10.1145/3038912.3052577
39. Yin, X., Liu, Z., Liu, D., Ren, X.: A novel cnn-based bi-lstm parallel model with attention mechanism for human activity recognition with noisy data. Scientific Reports **12**(1), 7878 (May 2022). https://doi.org/10.1038/s41598-022-11880-8
40. Zeng, M., Nguyen, L.T., Yu, B., Mengshoel, O.J., Zhu, J., Wu, P., Zhang, J.: Convolutional neural networks for human activity recognition using mobile sensors.

In: 6th International Conference on Mobile Computing, Applications and Services. pp. 197–205 (2014). https://doi.org/10.4108/icst.mobicase.2014.257786
41. Zhao, Y., Yang, R., Chevalier, G., Xu, X., Zhang, Z.: Deep residual bidir-lstm for human activity recognition using wearable sensors. Mathematical Problems in Engineering **2018**, 1–13 (2018). https://doi.org/10.1155/2018/7316954
42. Zhou, Y., Zhao, H., Huang, Y., Riedel, T., Hefenbrock, M., Beigl, M.: Tinyhar: A lightweight deep learning model designed for human activity recognition. In: Proceedings of the 2022 ACM International Symposium on Wearable Computers. p. 89–93. ISWC '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3544794.3558467