LATENCY-AWARE CONSTRUCTION ACTIVITY RECOGNITION USING KINEMATIC AND PHYSIOLOGICAL SENSORS

Umutcan Kılıç and Semra Çomu
Boğaziçi University, Türkiye

Abstract

Wearable sensors are increasingly being used to monitor worker activity and safety. While Human Activity Recognition (HAR) studies predominantly rely on kinematic data, physiological signals remain mostly underexplored due to the delayed nature of physiological responses. This study investigates the fusion of physiological and kinematic modalities by systematically evaluating temporal offsets to compensate for physiological latency. Results show that classification performance is sensitive to delay magnitude, with optimal performance achieved using an 8-second window and a 4-second time lag. Based on these findings, a latency-aware fusion protocol is proposed to enable robust multimodal HAR in wearable-based worker monitoring systems.

Introduction

The construction industry is characterized by dynamic and hazardous working environments, making occupational safety and health (OSH) a significant concern. To mitigate OSH-related risks, various methods have been proposed to monitor the physical and physiological status of the workforce (Antonaci et al., 2024). Enabling continuous and non-intrusive worker monitoring, Human Activity Recognition (HAR), which refers to automatic identification of human activities and movements using machine learning and deep learning algorithms, has emerged as a prominent approach for data-driven safety assessments and productivity tracking (Sedaghati et al., 2025).

HAR systems can be categorized into two types based on data modalities: visual and non-visual (Sun et al., 2022). Compared to vision-based systems, non-visual HAR solutions are cost-effective, capable of fine-grained recognition, less vulnerable to environmental conditions, and better at protecting worker privacy, making them attractive options for construction settings (Sun et al., 2022).

Non-visual HAR systems are predominantly based on kinematic data derived from Inertial Measurement Units (IMUs), such as accelerometers and gyroscopes (Shin et al., 2025). While effective for classifying physical movements such as walking, bending, or standing, kinematic sensors are inherently limited to describing motion-related patterns. They provide minimal insight into the physiological effects of construction tasks on workers' bodies, such as physical exertion, fatigue, or physiological stress, which are critical metrics for comprehensive safety monitoring (Kazar & Comu, 2021). On the other hand, physiological data, including Heart Rate (HR), Blood Volume Pulse (BVP), Electrodermal Activity (EDA), and Skin Temperature (ST), offer a promising complementary data stream to fill this gap. However, despite its potential to provide a holistic view of worker well-being, physiological data remains mostly unexplored in the context of activity recognition models (Demrozi et al., 2020).

One of the primary barriers to effectively integrating physiological data into HAR models is temporal misalignment. Unlike kinematic signals, which change instantaneously with body movement, physiological responses exhibit biological latency. For instance, while individual variability exists, a cardiovascular response to physical exertion typically occurs a few seconds after the onset of the stressor (Javorka et al., 2003). Similarly, EDA responses typically exhibit a latency of 2 to 6 seconds following a sudden physiological stressor (Yu et al., 2024). Consequently, standard sensor fusion approaches that align data streams strictly based on timestamps result in a mismatch between the physical action and the biological reaction, introducing noise into the dataset rather than enhancing it.

This study aims to address the synchronization challenge by investigating different temporal alignment strategies for multimodal HAR. We utilize machine learning (ML) to test the performance of proposed data fusion strategies and compare different data modalities. In particular, we systematically analyze the impact of different window sizes and the introduction of artificial time-lags to the physiological data stream, aiming to bridge the gap between physical action and physiological reaction. Based on this analysis, the study proposes a latency-aware fusion protocol designed to enable robust multimodal worker monitoring.

Methodology

The experimental study was conducted in a controlled environment at the Structure Laboratory of Boğaziçi University, spanning approximately 660 m² with a ceiling height of 12 meters. The experimental dataset was collected from five healthy male subjects, with no record of an injury or condition that might affect their postures or physiological responses, during independent sessions.

Before each session, the participants were briefed regarding the experiment and provided their informed consents. Table 1 summarizes the demographic details of participants, including age, height, and weight.

Table 1: Participant details

ID	Age (years)	Weight (kg)	Height (cm)
1	26	91	193
2	25	105	187
3	57	98	180
4	58	59	170
5	26	63.5	174

To capture a comprehensive set of kinematic and physiological data, the participants were instrumented with two wearable devices. An Apple iPhone 16 Pro Max smartphone (Apple Inc., Cupertino, CA, USA) was positioned at the lumbar region using an adjustable cloth belt to record gross body movements via its Inertial Measurement Unit (IMU). Because of its proximity to the human body's center of mass, this placement enables effective analysis of overall movement patterns (Wlodzimierz, 2018). The "Physics Toolbox Sensor Suite" application was utilized to log tri-axial gyroscope and accelerometer data (both at 100 Hz), storing the files locally in Comma Separated Values (CSV) format. Simultaneously, an Empatica EmbracePlus smartwatch (Empatica Inc., Boston, MA, USA) was worn on the non-dominant wrist to acquire kinematic data associated with hand movements and physiological signals. Although the smartwatch hardware includes a gyroscope, only tri-axial accelerometer data was recorded (at 64 Hz) due to access restrictions imposed by the "Academic & Basic Research" plan of Empatica. Physiological data included Blood Volume Pulse (BVP) recorded at 64 Hz, Systolic Peaks, EDA recorded at 4 Hz, and ST recorded at 1 Hz. The data collected by the smartwatch was retrieved via the manufacturer's dedicated data access platform as Avro files and subsequently converted to CSV format for further processing. Based on raw Systolic Peaks data, instantaneous HR values were derived by computing inter-beat intervals (IBI) and converting them to beats per minute. HR values outside the physiologically plausible range of 40–200 bpm were discarded as artifacts. The resulting irregularly sampled HR measurements were transformed into a continuous signal using linear interpolation and resampled at 10 Hz to obtain a uniformly spaced time series suitable for temporal analysis and multimodal synchronization. In addition, ST signals, originally recorded at 1 Hz, were upsampled via linear interpolation to a uniform sampling frequency of 4 Hz to prevent data loss at sliding window boundaries.

To capture physiological responses under varying levels of physical exertion, participants followed a physically demanding experimental protocol lasting approximately 30 minutes, designed to simulate continuous construction work. Each session began with a 5-minute seated rest baseline measurement phase (no movement or talking) to capture resting levels of EDA, ST, and cardiovascular signals derived from BVP, including systolic peaks and IBI (Posada-Quintero et al., 2018; Damoun et al., 2024). For analysis purposes, the first and last 30 seconds of the baseline measurement phase were omitted to account for participant settling time. Following the initial omitted period, the subsequent 240 seconds were labeled as the sitting baseline. Any remaining data from the seated baseline measurement was excluded from the analysis to prevent data leakage.

The experiment sessions were structured into four identical activity blocks and were performed across four custom-built stations designed to replicate common construction-related tasks. The first station involved six repetitions of walking on a narrow wooden plank (measuring 2.95 m in length and 0.15 m in width, elevated 25 cm above ground level) to challenge balance and stability. For safety purposes, buffer planks with a width of 20 cm were installed on both sides; however, participants were instructed to walk strictly on the central narrow plank. The second station consisted of two repetitions of accessing a mezzanine platform (measuring 6.3 m by 2.0 m at a height of 2.8 m) via a spiral staircase. Upon reaching the platform, participants walked to its end before descending back to ground level. The third station required participants to climb a standard scaffold (measuring 2.7 m by 0.8 m), featuring two working levels at heights of 0.60 m and 2.55 m. Participants ascended the scaffold using a ladder, reached the upper working level, and descended. The fourth station focused on material handling and involved transferring four separate loads (one 10 kg metal plate, one 5 kg metal plate, and two 1 kg wooden plates) from a platform at 30 cm height onto a pallet truck bed positioned at 10 cm height. Participants then pulled the loaded pallet truck over a distance of approximately 25 meters before returning to the starting position to unload the weights.

Each activity block began with a 60-meter regular walking segment, followed by the execution of all four stations in sequence, and concluded with an additional visit to the first station. After completing the first two blocks, participants took a one-minute standing rest period to emulate a brief work break. After this short recovery phase, the remaining two blocks were performed. The experiments were concluded with a one-minute seated rest period to capture post-activity physiological recovery dynamics. Figure 1 illustrates the content of an experimental session.

To establish precise ground truth for activity labeling, the experimental sessions were recorded on video at 60 fps using an Apple iPhone 16e (Apple Inc., Cupertino, CA, USA) and a GoPro Hero 4 (GoPro Inc., San Mateo, CA, USA). The smartphone captured the third-person view, while the GoPro recorded the first-person view of a given participant. Third-person videos were embedded with real-time timestamps with millisecond precision using the "Timestamp Camera Basic" application. Subsequently, first- and third-person recordings were synchronized using visual cues via CapCut software.

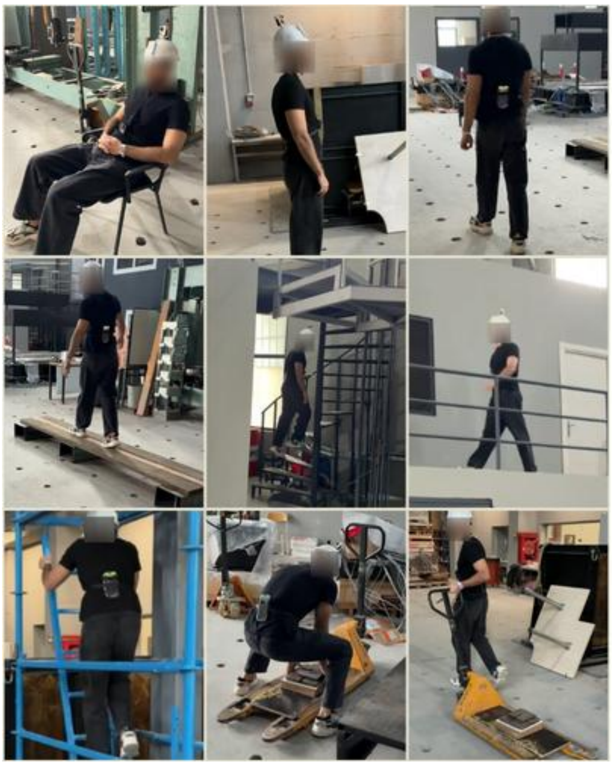


Figure 1: Typical content of an experimental session

Ground truth labeling was performed manually using ELAN software, where activity start and end points were annotated relative to the video timeline (ELAN, 2025). Based on the video footage, various activity types were labeled as shown in Table 2. All annotations were performed by trained annotators following a predefined labeling protocol, which was designed to minimize inter-rater variance by establishing rules that define exact start and end moments of activities. Interruptions during the experiment or other unclassified movements were labeled as junk and excluded from the analysis.

Table 2: Activity Types and Annotation Labels

Activity	Types	Label
Sitting	-	Sitting
Standing	-	Standing
Walking	Normal, Plank, Mezzanine Platform, Scaffolding	Walking
Climbing Stairs	-	N_Stairs
Climbing Ladder	-	V_Stairs
Bending	Hip Hinge, Squat	Bending
Cart	Pulling, Pushing	Cart
Junk	-	Junk

To mitigate temporal discrepancies caused by the Variable Frame Rate (VFR) inherent in smartphone video encoding while converting video-relative timestamps to Unix timestamps for feature extraction purposes, we implemented a piecewise linear interpolation-based conversion algorithm. To serve as anchor points, we extracted the first and the last frames of a given video, as well as a reference frame every 3600 frames (approximately 60 seconds) using FFmpeg (Version 6.0). For each extracted frame, we recorded the exact start time of the frame relative to the video start and the Unix timestamp embedded within the frame, both with millisecond precision. The algorithm was then applied to convert the manually annotated video-relative activity start and end timestamps into sensor-compatible Unix timestamps using (1),

$$T_{unix}(t) = U_i + \frac{(U_{i+1} - U_i)}{(V_{i+1} - V_i)} \cdot (t - V_i) \qquad (1)$$

where t represents the annotated video-relative timestamp, and (V_i, U_i) and (V_{i+1}, U_{i+1}) denote the video-relative and Unix timestamps of the two consecutive anchor frames enclosing t, respectively. This approach effectively compensated for the non-linear time drift associated with VFR, ensuring millisecond-level alignment between the visual ground truth and the multimodal sensor data.

Following the conversion of all ground truth annotations to Unix timestamps, the synchronized multimodal sensor data were segmented using a sliding window approach. Sliding window segmentation is commonly employed in time-series machine learning to compute window-level feature representations, where each window contains a fixed-length segment of data and adjacent windows overlap by a predefined percentage to preserve temporal continuity and increase the number of training samples. Based on the selected window lengths and overlap ratios, sliding windows were generated over the entire duration of each experimental session independently for each participant.

Since experiment sessions followed a continuous flow, a consensus-based labeling strategy was employed to assign activity labels at the window level. Rather than isolating different activities, we generated windows with selected lengths and overlap percentages from the start of a session, marked with the start Unix timestamp of the sitting baseline measurement. For each window, ground truth labels corresponding to the annotated activities were examined, and a window was assigned a class label only if at least 75% of its duration was occupied by a single activity. Windows that did not meet this criterion were considered mixed and excluded from the analysis. This threshold was empirically selected as a trade-off between label purity and data retention, aiming to reduce noise introduced by transitional movements while preserving a sufficient number of valid samples.

For each window, statistical and temporal features were extracted independently from each sensor channel. For BVP, HR, EDA, and ST signals, extracted features included mean, standard deviation, maximum, range, energy (2), slope (defined as the difference between the last and first sample within each window, normalized by the window duration), and delta (defined as the difference between consecutive windows).

$$E = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{2}$$

Baseline-based z-score normalization was applied to level-dependent physiological features (mean, standard deviation, maximum, and energy) using (3) to reduce inter-subject variability and ensure comparability of the data across participants, whereas change-based features (range, slope, and delta) were computed in the original signal scale to preserve absolute variation within each window.

$$z_i = \frac{(x_i - \mu)}{\sigma} \tag{3}$$

where x_i denotes the physiological feature for the i[th] window, and μ and σ are the mean and standard deviation of the feature calculated over the baseline measurement period, respectively.

For kinematic data (accelerometer and gyroscope), axis-wise (x, y, z) features (mean, standard deviation, minimum, maximum, and range) and magnitude-based features including signal magnitude area (SMA), vector magnitude–based features (mean and standard deviation), and linear or angular jerk (approximated using first-order differences of acceleration and angular velocity signals, respectively) were extracted.

Table 3 and Table 4 summarize the feature sets generated from the multimodal sensor data.

Table 3: Physiological feature set and applied normalization methods

Sensor	Features	Normalization
BVP	Energy, Mean, Standard Deviation	z-Score
BVP	Range	None
HR	Maximum, Mean	z-Score
HR	Range, Slope	None
EDA	Energy, Maximum, Mean, Standard Deviation	z-Score
EDA	Range, Slope	None
ST	Mean	z-Score
ST	Delta	None

To explicitly investigate the effect of biological response latency on activity classification, we utilized a systematic parameter search strategy. While the sensor data were segmented using multiple window lengths (1, 2, 4, and 8 seconds) combined with overlap ratios of 50% and 75%, we simultaneously introduced temporal offsets ranging from 0 seconds (no delay) to 8 seconds in 1-second increments. This resulted in a total of 72 distinct window configurations. In this protocol, kinematic features were extracted synchronously with the ground truth activity labels to capture immediate physical motion. In contrast, physiological features were extracted from a time window shifted forward by the delay amount, aiming to align the delayed biological response with the causing physical activity. This approach allows for a controlled assessment of how explicit temporal alignment influences classification performance, independent of window size.

To evaluate the contribution of physiological signals to HAR models, we compared the performance of models trained on six different sensor combinations. This approach resulted in 240 distinct model configurations: for kinematic-only models, 3 sensor combinations with 8 window configurations each; and for models involving physiological data, 3 sensor combinations with 72 window configurations each. Based on the extracted feature vectors for each model configuration, activity classification models were trained using a Random Forest (RF) classifier with 100 estimators (trees), which is widely adopted as a strong baseline in machine learning (ML)–based HAR studies (Demrozi et al., 2020). To avoid data leakage and ensure robust, subject-independent evaluation, a Leave-One-Subject-Out (LOSO) cross-validation scheme was employed. In each fold, data from one participant were reserved for testing, while data from all remaining participants were used for training. Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, macro and weighted F1-scores, as well as confusion matrices, enabling both class-wise and overall performance assessment. This procedure was iterated for all 240 model configurations, enabling a comparative assessment to identify the optimal temporal alignment strategy and to see the performance differences between different sensor configurations. Since all five experimental sessions had similar durations, F1-scores were reported as the average performance across all LOSO folds, unless otherwise specified.

Table 4: Kinematic feature set

Sensor Type	Feature Type	Features
Accelerometer	Axis-wise (x, y, z)	Maximum, Mean, Minimum, Range, Standard Deviation
Accelerometer	Magnitude-based	Linear Jerk, Mean Vector Magnitude, Signal Magnitude Area, Standard Deviation of Vector Magnitude
Gyroscope	Axis-wise (x, y, z)	Maximum, Mean, Minimum, Range, Standard Deviation
Gyroscope	Magnitude-based	Angular Jerk, Mean Vector Magnitude, Signal Magnitude Area, Standard Deviation of Vector Magnitude

Results and discussion

First, we analyzed the distribution of ground truth activity labels to characterize the dataset. Based on the video annotations, we calculated the combined total duration of the five experiment sessions to be 9756.833 seconds. As summarized in Table 5, "Walk" constituted the largest portion of the data with 4277.145 seconds (43.84%), followed by "Junk" segments and "N_Stairs".

Following the exclusion of invalid segments (i.e., Junk) and the removal of transitional periods where 75% label consensus was not met, valid windows were retained for model input. Table 6 summarizes the resulting window counts across different activity types for various window lengths at 50% overlap. On the other hand, setting the overlap to 75% resulted in approximately twice the number of analysis windows compared to the values given in Table 6.

The resulting class distribution reveals a natural imbalance often encountered in HAR literature (Lim et al., 2021). For instance, the dataset used by Sikder & Nahid (2021) exhibits an imbalance ratio of 8.41:1 (max:min), which is comparable to the 11.76:1 ratio observed in our dataset. “Walking” emerged as the dominant class, accounting for approximately 51% of the total samples. In contrast, short-duration or less frequent activities such as “Bending” and “V_Stairs” represented approximately 5% and 6% of the samples, respectively. This observed imbalance underscores the necessity of utilizing macro-averaged metrics alongside weighted metrics to ensure a fair assessment of the model's capability to recognize less frequent activities.

To evaluate the contribution of physiological signals to HAR models, we first compared the performance of models trained on different data modalities, fused using a simple fusion strategy (no temporal offsets), with window lengths of 1, 2, 4, and 8 seconds, and 50% and 75% overlap. The average classification performances of the resulting models are summarized in Figure 2.

As anticipated, scenarios relying solely on physiological data yielded the lowest results (Average Macro F1: 16.29%). Notably, a theoretically uniform random classifier would achieve an F1 score of 14.29%, which is only 12.28% worse than the physiological-only model. This marginal difference confirms that physiological signals alone are insufficient for distinguishing complex locomotor activities. In contrast, the model utilizing only the smartphone IMU data from the lumbar region, a standard configuration in HAR literature, achieved an average Macro F1 score of 69.61%, establishing a strong baseline. For instance, the RF-based LOSO model trained by Chen and Shen (2017) using a similar sensor configuration and placement achieved an average Macro F1 score of 67.21% while classifying 5 activities, indicating that our model performed comparably to previously reported results. The highest performance was observed in the configuration that combined kinematic data from both the smartphone and the smartwatch (Average Macro F1: 79.17%). On the other hand, the simple fusion of physiological signals into this optimal kinematic model resulted in a slight performance decline (Average Macro F1: 77.54%). These findings support the claim that the simple fusion of inherently latent physiological data may introduce noise or redundant information, thereby degrading the classifier's predictive capability rather than enhancing it.

Subsequently, we investigated the interaction between model performance and artificial temporal delays. Figure 3 illustrates the fusion model's sensitivity to different temporal delays across various window sizes and overlap ratios. Based on the results, we identified two key trends.

First, larger time windows consistently yielded higher performance. Specifically, the 8-second window achieved the highest Macro F1 scores across nearly all delay settings, with a few exceptions marginally favoring the 4-second window. This superiority is likely attributable to the ability of longer windows to simultaneously capture slow-varying physiological signals and sudden kinematic changes associated with an activity within the same window.

Secondly, the introduction of artificial delays revealed a surprising but expected concave relationship with performance. While simple fusion (0s delay) provided satisfactory results, particularly for longer windows, introducing a temporal shift to the physiological data significantly enhanced Macro F1 scores. For instance, the highest performance observed in this study was obtained at a 4-second delay with an 8-second window and 75% overlap (Macro F1: 81.22% ± 7.64%), highlighting the potential performance gains of integrating physiological sensors under a latency-aware fusion framework. In comparison, the simple fusion counterpart without temporal delay (0-second delay with an 8-second window and 75% overlap) achieved a lower Macro F1 of 80.12% with a higher standard deviation (± 8.14%), indicating that compensating for physiological latency not only improves overall accuracy but also enhances the consistency of the model across different subjects. However, further increasing the delay almost always led to a degradation in classification performance. This finding aligns with established physiological literature, which suggests that physiological responses typically lag behind the onset of a physical stressor by approximately 2–6 seconds (Javorka et al., 2003; Yu et al., 2024).

Lastly, we compared the confusion matrices and class-wise performance metrics for the overall best-performing model (latency-aware sensor fusion model utilizing an 8-second time window, 75% overlap, and a 4-second temporal shift) and its kinematic-only counterpart (Smartphone + Smartwatch). As illustrated in Figure 4 and Figure 5, both models demonstrated strong diagonal dominance. However, the inclusion of temporally aligned physiological data yielded noticeable improvements in specific activities. A closer inspection of the class-wise performance reveals that the most significant gains were observed in the “Cart” and “Walking” activities. These two activities present a critical challenge for kinematic-only models: the body posture similarity between the activities. As a result, kinematic sensors alone fail to

Table 5: Descriptive statistics of annotated activty classes

Activity	Count	Total Duration (s)	Avg. Duration (s)	Min. Duration (s)	Max. Duration (s)	Distribution (%)
Walking	488	4277.145	8.765	0.435	103.649	43.84%
Junk	486	1683.380	3.464	0.068	94.000	17.25%
N_Stairs	33	1078.334	32.677	3.333	117.434	11.05%
Standing	84	751.653	8.948	5.900	11.800	7.70%
Cart	32	666.601	20.831	3.500	31.800	6.83%
Bending	105	585.852	5.580	0.834	23.470	6.00%
V_Stairs	40	390.132	9.753	5.559	14.951	4.00%
Sitting	5	323.736	64.747	54.423	71.011	3.32%

capture the physical effort required to maneuver an approximately 110 kg pallet truck. In contrast, the latency-aware fusion model integrates the physiological markers of this high-exertion task, allowing the classifier to distinguish it from walking, despite the ambiguous kinematic cues. As a result, the number of "Walking" activities falsely classified as "Cart" decreased from 50 to 8, increasing the F1-score for the "Cart" activity from 60.83% to 67.14%. This suggests that while kinematic sensors effectively capture the distinct motion trajectories of the activities, they may struggle to differentiate the subtle intensity variations associated with them. The physiological signals, acting as an indicator of effort, provide the necessary complementary information to resolve these ambiguities.

Table 6: Total number of activity samples for different window lengths at 50% overlap

Window (s)	1	2	4	8
Walking	8065	3790	1629	650
Standing	2121	1046	505	239
N_Stairs	1416	664	294	102
Cart	1301	637	302	135
Bending	1073	480	180	67
V_Stairs	742	351	156	54
Sitting	593	293	142	67
Total Samples	15311	7261	3208	1314

These class-specific improvements highlight the mechanism behind the performance boost. The 4-second temporal offset effectively synchronizes the delayed physiological response with the immediate kinematic changes. Without this alignment (as observed in the simple fusion case discussed earlier), the physiological marker of a demanding activity registers after the activity label has ceased, creating noise. By shifting the windows, we ensure that the classifier receives both the motion pattern of an activity and the corresponding metabolic effect simultaneously. Consequently, the fusion model does not merely rely on how the subject moves, but also how much effort the movement entails, leading to a more robust and context-aware recognition system.

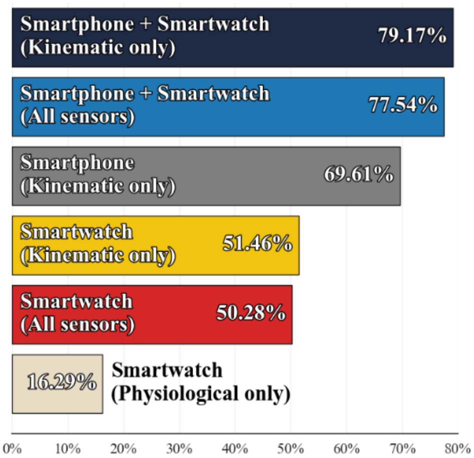


Figure 2: Average Macro F1-scores across various sensor combinations and their simple fusion

As a final step, we measured the average inference time of the best-performing model to evaluate the feasibility of the proposed approach for real-world applications. Using a standard desktop CPU (2.50 GHz, 16 threads), the model classified a single data window in approximately 0.06 milliseconds on average. This low computational cost indicates that the proposed latency-aware fusion model is lightweight enough to be deployed on edge devices or microprocessors available in modern wearable devices with negligible delay, satisfying the critical requirement for real-time safety monitoring in construction environments.

Ultimately, this study demonstrates that bridging the temporal gap between instantaneous motion and delayed physiological response is essential for developing context-aware HAR systems capable of distinguishing

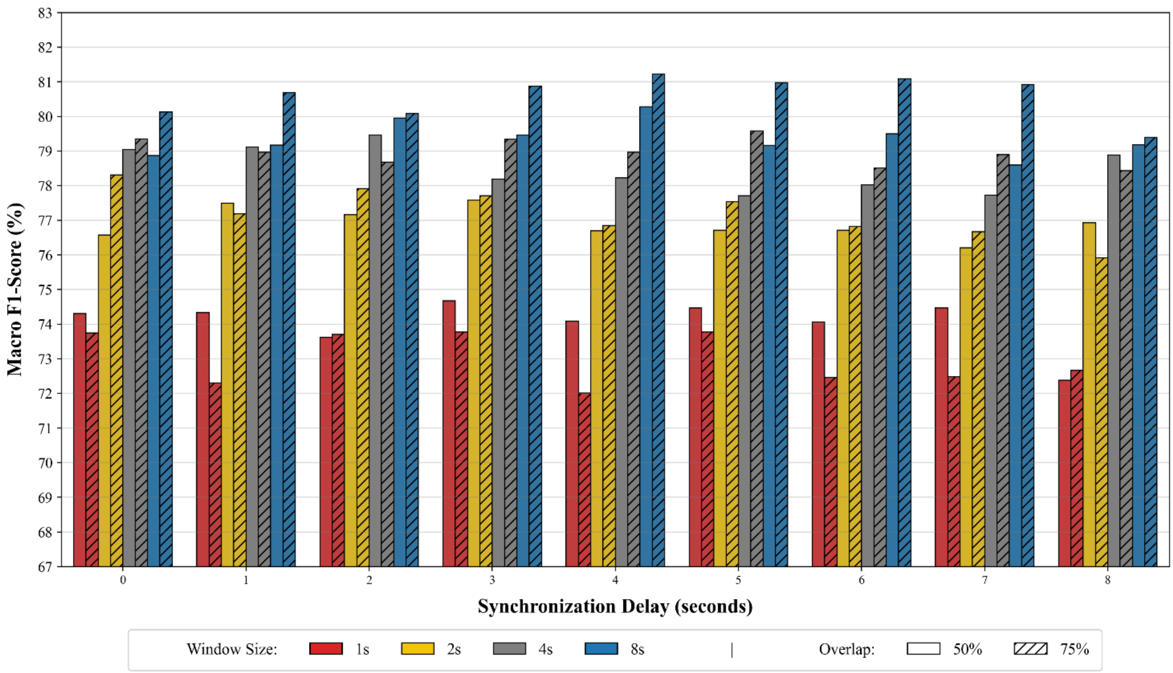
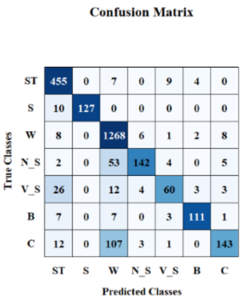


Figure 3: Performance variations of the fusion model as a function of artificial temporal synchronization dealys for different windows sizes and overlap ratios

activities defined not just by how we move, but by how much effort is exerted. The proposed latency-aware fusion model achieved a Macro F1 score of 81.22% and a Weighted F1 score of 86.79% (Pooled Macro F1 score of 82.78% and Pooled Weighted F1 score of 87.42%), underscoring its potential to enhance construction site safety through robust worker monitoring.

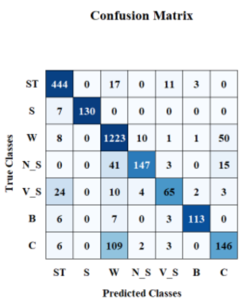


	ST	S	W	N_S	V_S	B	C
ST	455	0	7	0	9	4	0
S	10	127	0	0	0	0	0
W	8	0	1268	6	1	2	8
N_S	2	0	53	142	4	0	5
V_S	26	0	12	4	60	3	3
B	7	0	7	0	3	111	1
C	12	0	107	3	1	0	143

Class-wise Performance Metrics

Activity	Precision	Recall	F1-Score
Standing	87.50%	95.79%	91.46%
Sitting	100.00%	92.70%	96.21%
Walk	87.21%	98.07%	92.32%
N_Stairs	91.61%	68.93%	78.67%
V_Stairs	76.92%	55.56%	64.52%
Bending	92.50%	86.05%	89.16%
Cart	89.38%	53.76%	67.14%
Macro Avg	89.30%	78.69%	82.78%
Weighted Avg	88.34%	88.22%	87.42%

Figure 4: (a) Confusion matrix and (b) class-wise performance metrics for the latency-aware sensor fusion model using an 8-second time window, 75% overlap, and a 4-second temporal delay.



	ST	S	W	N_S	V_S	B	C
ST	444	0	17	0	11	3	0
S	7	130	0	0	0	0	0
W	8	0	1223	10	1	1	50
N_S	0	0	41	147	3	0	15
V_S	24	0	10	4	65	2	3
B	6	0	7	0	3	113	0
C	6	0	109	2	3	0	146

Class-wise Performance Metrics

Activity	Precision	Recall	F1-Score
Standing	89.70%	93.47%	91.55%
Sitting	100.00%	94.89%	97.38%
Walk	86.92%	94.59%	90.59%
N_Stairs	90.18%	71.36%	79.67%
V_Stairs	75.58%	60.19%	67.01%
Bending	94.96%	87.60%	91.13%
Cart	68.22%	54.89%	60.83%
Macro Avg	86.51%	79.57%	82.59%
Weighted Avg	86.39%	86.76%	86.29%

Figure 5: (a) Confusion matrix and (b) class-wise performance metrics for the kinematic-only model using an 8-second time window and 75% overlap

Conclusion

This study investigated the impact of temporal alignment strategies on the performance of multimodal Human Activity Recognition (HAR) systems for construction worker monitoring, focusing on the fusion of kinematic and physiological data. By systematically evaluating different window sizes and artificial temporal delays, the study aimed to address the synchronization challenges caused by the latency of physiological responses.

In total, we trained and evaluated 240 distinct model configurations to systematically investigate the impact of temporal alignment. The results demonstrated that the temporal alignment of data streams significantly affects the classification performance of multimodal models. The latency-aware fusion model, utilizing an 8-second time window with 75% overlap and a 4-second temporal delay, achieved the highest performance with a Pooled Macro F1-score of 82.78% and a Pooled Weighted F1-score of 87.42%, significantly outperforming both the kinematic-only counterpart and the simple fusion approach. These improvements are largely attributed to the effective compensation of physiological latency, which ensures that the delayed metabolic markers of physical exertion are correctly aligned with the corresponding motion data.

Furthermore, the analysis revealed that the inclusion of temporally aligned physiological data is particularly critical for distinguishing activities that share similar kinematic patterns but differ in physical intensity. For instance, the latency-aware model successfully differentiated between "Walking" and high-exertion "Cart" pulling tasks, reducing misclassification rates where kinematic-only models failed due to postural similarities. In contrast, the simple fusion of physiological signals without temporal correction resulted in a slight performance degradation compared to kinematic-only models, suggesting that misaligned physiological data acts as noise rather than an informative feature.

While these findings highlight the strong potential of latency-aware multimodal sensor fusion for robust worker monitoring, certain limitations should be acknowledged. First, the experiments were conducted in a controlled laboratory environment with a limited sample size of five healthy male participants (N=5). Although sufficient for a proof-of-concept temporal alignment analysis using a Leave-One-Subject-Out validation scheme, this small sample size inherently restricts the statistical power required for formal significance testing (e.g., paired t-tests) between alternative model configurations. For the generalizability of the results and robust validation of the optimal physiological temporal delay, future research should utilize a more diverse population varying significantly in age, baseline fitness levels, body mass index (BMI) and construction experience. Secondly, although the activities were designed to simulate realistic construction tasks, the controlled nature of the study does not fully replicate the complex, unstructured, and dynamic conditions of an active construction site, where factors such as environmental conditions and worker interactions may influence kinematic and physiological sensor readings.

By addressing these challenges, multimodal HAR systems can be further refined to provide a more reliable and comprehensive solution for occupational safety and health management in the construction industry.

References

Antonaci, F. G., Olivetti, E. C., Marcolin, F., Castiblanco Jimenez, I. A., Eynard, B., Vezzetti, E., & Moos, S. (2024). Workplace well-being in industry 5.0: A worker-centered systematic review. Sensors, 24(17), 5473. https://doi.org/10.3390/s24175473

Chen, Y., & Shen, C. (2017). Performance analysis of smartphone-sensor behavior for human activity recognition. IEEE Access, 5, 3095–3110. https://doi.org/10.1109/access.2017.2676168

Damoun, N., Amekran, Y., Taiek, N., & El Hangouche, A. J. (2024). Heart rate variability measurement and influencing factors: Towards the standardization of methodology. Global Cardiology Science and Practice, 2024(4). https://doi.org/10.21542/gcsp.2024.35

Demrozi, F., Pravadelli, G., Bihorac, A., & Rashidi, P. (2020). Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. IEEE Access, 8, 210816–210836. https://doi.org/10.1109/access.2020.3037715

Javorka, M., Žila, I., Balhárek, T., & Javorka, K. (2003). On- and off-responses of heart rate to exercise – relations to heart rate variability. Clinical Physiology and Functional Imaging, 23(1), 1–8. https://doi.org/10.1046/j.1475-097x.2003.00460.x

Kazar, G., & Comu, S. (2021). Exploring the relations between the physiological factors and the likelihood of accidents on construction sites. Engineering, Construction and Architectural Management, 29(1), 456–475. https://doi.org/10.1108/ecam-11-2020-0958

Lim, X. Y., Gan, K. B., & Abd Aziz, N. A. (2021). Deep CONVLSTM network with dataset resampling for Upper Body Activity Recognition using minimal number of IMU sensors. Applied Sciences, 11(8), 3543. https://doi.org/10.3390/app11083543

ELAN (Version 7.0) [Computer software]. (2025). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Posada-Quintero, H. F., Reljin, N., Mills, C., Mills, I., Florian, J. P., VanHeest, J. L., & Chon, K. H. (2018). Time-varying analysis of electrodermal activity during exercise. PLOS ONE, 13(6). https://doi.org/10.1371/journal.pone.0198328

Sedaghati, N., Ardebili, S., & Ghaffari, A. (2025). Application of human activity/action recognition: A Review. Multimedia Tools and Applications, 84(28), 33475–33504. https://doi.org/10.1007/s11042-024-20576-2

Shin, J., Hassan, N., Miah, A. S., & Nishimura, S. (2025). A comprehensive methodological survey of human activity recognition across diverse data modalities. Sensors, 25(13), 4028. https://doi.org/10.3390/s25134028

Sikder, N., & Nahid, A.-A. (2021). Ku-Har: An open dataset for heterogeneous human activity recognition. Pattern Recognition Letters, 146, 46–54. https://doi.org/10.1016/j.patrec.2021.02.024

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–20. https://doi.org/10.1109/tpami.2022.3183112

Wlodzimierz, E. S. (2018). Center of mass of the human body helps in analysis of balance and movement. MOJ Applied Bionics and Biomechanics, 2(2). https://doi.org/10.15406/mojabb.2018.02.00057

Yu, X., Lu, J., Liu, W., Cheng, Z., & Xiao, G. (2024). Exploring physiological stress response evoked by passive translational acceleration in healthy adults: A pilot study utilizing electrodermal activity and heart rate variability measurements. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-61656-5