Real-Time Temporal Convolution based Worker Motion Recognition with Three IMU Sensor Combinations*

Minjun Chang[1], Sai Machiraju[2], Francis Baek[3] and Yong K. Cho[†]

[1]School of Civil and Environmental Engineering, Georgia Institute of Technology, United States of America
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, United States of America
[3]Assistant Professor, School of Civil and Environmental Engineering, Georgia Institute of Technology, United States of America
[†]Professor, School of Civil and Environmental Engineering, Georgia Institute of Technology, United States of America

minjunchang@gatech.edu, saimachi@gatech.edu,francis.baek@ce.gatech.edu,yong.cho@ce.gatech.edu

Abstract -

On active jobsites, unobserved worker motions drive injuries and near-misses, unsafe habits evade periodic audits, so reliable activity monitoring is essential to prevent delays and costly rework. In recent years, Inertial-Measurement-Unit(IMU) based motion recognition using Deep Learning approach is used widely in many domains. While the method has been proven to effectively recognize human motion using 5 to 21 IMUs, it is impractical to attach that many IMUs to construction workers. Minimal-sensing approaches using a single sensor are also proposed for limited motion patterns. To address these limitations, this paper presents a Temporal Convolutional Network (TCN) based real-time human motion recognition system that uses only three IMU sensors. TCN was chosen because it replaces recurrent states with small 1-D convolutions and dilations, keeping parameters and memory modest while inference runs in parallel on short windows, making it suitable for on-board deployment. Three IMUs are attached to the neck, back, and right wrist (NBW) that deliver 6-axis data at 50Hz, which are windowed for per-window z-normalized inference using a lightweight TCN with dilated, causal 1D convolutions and residual blocks. The model predicts 19 basic motion classes, classified by anatomical terms for body movements. This paper evaluated a lab-made 15-participant dataset, achieving the highest average prediction accuracy of 83.26% on the NBW combination in a single-split configuration, demonstrating the capability to achieve reasonable accuracy for field deployment. It also compares training results across diverse combinations of the three sensors to demonstrate the possibility of reducing sensor counts.

Keywords -

Inertial-Measurement-Unit, Worker Motion Recognition, 1D-CNN, Construction Safety Management

*This work was supported in part by the National Science Foundation (NSF) under Award No. 2515786 and by the Georgia Research Alliance (GRA). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or GRA

1 Introduction

Construction sites face frequent injuries, near-misses, and costly rework. Unsafe motions often occur between periodic audits, when supervisors cannot observe early warning signs[1]. Manual observation cannot scale across multiple crews or shifts. Paper based reports lag behind events, and fixed surveillance cameras struggle in dynamic construction environments[2]. Current practice relies on periodic visual audits, manual incident reports, and fixed cameras that lack continuous coverage across multiple crews. Unlike attendance-based check-in systems, the proposed approach requires no manual input from workers; the PPU devices automatically timestamp and stream sensor data continuously, enabling passive and unobtrusive monitoring. Real-time IMU-based motion recognition fills this gap by flagging hazardous postures before injuries occur, supporting ergonomic risk assessment through cumulative high-strain motion tracking, and quantifying task productivity without dependence on line-of-sight or observer presence.

Field-deployable human-motion recognition for construction workers remains constrained by sensor burden and latency, making many systems impractical on active jobsites due to donning time, discomfort, cost, battery swaps, RF congestion, and synchronization drift. Commercial motion-capture suits typically mount a large number of IMUs across the body[3] to achieve stable pose tracking, trading usability for coverage and drift robustness. Academic "sparse-IMU" efforts reduce sensor count but often depend on non-causal or offline temporal models that leverage future frames[4], additional sensing modalities[5], or heavyweight optimization pipelines compromising strict real-time operation on embedded hardware. The resulting gap is a method that delivers reliable, real-time recognition of a rich motion set with minimal sensor attachment and short, causal windows suitable for noisy and time-sensitive construction environments.

We addressed this gap by targeting three IMUs placed on the rear neck, lower back, and right wrist and en-

forcing strict causality with short windows (≈ 2 s at 50 Hz). Instead of relying on look-ahead or offline smoothing, our model is a lightweight Temporal Convolutional Network (TCN)[3] with dilated, causal 1D convolutions and residual blocks, providing a large effective receptive field without future context. The streaming pipeline performs per-window z-normalization and runs inference under low-memory conditions, targeting wearability and deployability on constrained edge hardware. Our working hypothesis is that the temporal structure of human motion can be captured by dilated causal convolutions over short windows when the three IMUs are chosen to provide complementary distal and proximal information. We compare several sensor configurations: neck+back+wrist (NBW), neck+back (NB), back+wrist (BW), neck+wrist (NW), and single sensor setups under matched computational budgets. We evaluate two training methods that reflect practical deployment: leave-one-subject-out (LOSO)[6] for cross-worker generalization and a single fixed split for rapid development. We measure accuracy, macro-F1, latency, and throughput, and perform ablation studies to characterize the trade-offs among different sensor sets and training methods. The selected sensor–training configuration is determined through a multi-objective criterion that jointly accounts for recognition reliability and field deployability, including performance, inference latency, model footprint, and sensor donning burden. Overall, our contributions are (i) a causal, edge-ready TCN pipeline for real-time worker motion recognition with only three IMUs, and (ii) a systematic ablation across seven sensor combinations and two training techniques that offer accuracy–practicality trade-offs for job site deployment.

2 Background

2.1 IMU based Human Motion Recognition

Wearable IMUs are widely used for human activity and pose estimation because they are inexpensive and robust to lighting or occlusion. Early and influential human activity recognition(HAR) systems paired multiple body-worn IMUs with deep neural networks and achieved high accuracy on curated datasets, but typically at the cost of many sensors and offline processing. Ordoñez et al. [7] proposed DeepConvLSTM, which demonstrated strong performance on multimodal IMU streams using a convolutional front-end and a recurrent back-end, establishing a deep-learning baseline for wearable HAR. Sparse-IMU pose estimators such as Deep Inertial Poser (DIP)[4] reconstruct full-body kinematics from six IMUs but rely on bidirectional temporal models and sliding windows with look-ahead, which limit strictly causal, real-time use on embedded devices. In manufacturing and construction-adjacent settings, multimodal systems often combine an IMU armband with video to improve accuracy, again trading off deployment simplicity for additional sensors and compute[8]. Prior construction-focused studies have shown that decomposing construction tasks into basic motion primitives enables detailed safety and productivity analysis, but these studies are often conducted in controlled settings, highlighting the gap between laboratory validation and field deployment[9]. Reviews and case studies report growing interest in using wearables for worker activity monitoring and safety, but note persistent barriers, such as the burden of multiple sensors, annotation costs, and synchronization challenges in dynamic construction environments. These factors point toward methods that minimize sensor count, remain causal, and operate with tight latency budgets while retaining recognition fidelity[10]. Recent construction focused IMU work by Park et al.[11] introduces the Scaffolding Worker IMU Time-series(SWIT) dataset for deep-learning recognition of hazardous and routine behaviors on scaffolds. The paper motivates IMU-based monitoring to overcome occlusion limits in camera systems and addresses the shortage of public, construction-specific datasets by releasing labeled 100 Hz time-series across ten behavior categories. State-of-the-art wearable HAR often assumes dense IMU coverage, which inflates cost and increases donning time and maintenance overhead[12]. On active construction sites many body-worn units hinder movement and PPE use and degrade reliability through frequent battery swaps, RF congestion, and synchronization drift. The gap is a method that uses few IMUs, runs causally in real time, and remains practical for noisy construction sites.

2.2 Deep Learning Methods in Motion Recognition

Given the constraints of existing wearable systems, deep learning has been widely explored to extract meaningful motion patterns from IMU signals. Architectures for IMU time series generally fall into three families. CNN–RNN hybrids[7] extract local temporal features with 1-D convolutions and model longer dependencies with recurrent units, but recurrent state adds memory overhead and can complicate low-latency deployment. Pure convolutional approaches, including Temporal Convolutional Networks (TCNs), replace recurrence with dilated causal convolutions and residual connections as they parallelize efficiently, support long receptive fields with short windows, and are well-suited to streaming inference on edge hardware[13]. Temporal convolutional stacks deliver strong accuracy with efficient inference on short windows[14] for action segmentation. TCNs have shown competitive or superior accuracy to recurrent models while preserving strict causality, motivating their use for real-time HAR from IMUs. Some approaches use multimodal fusion, pairing IMUs with cameras or audio to

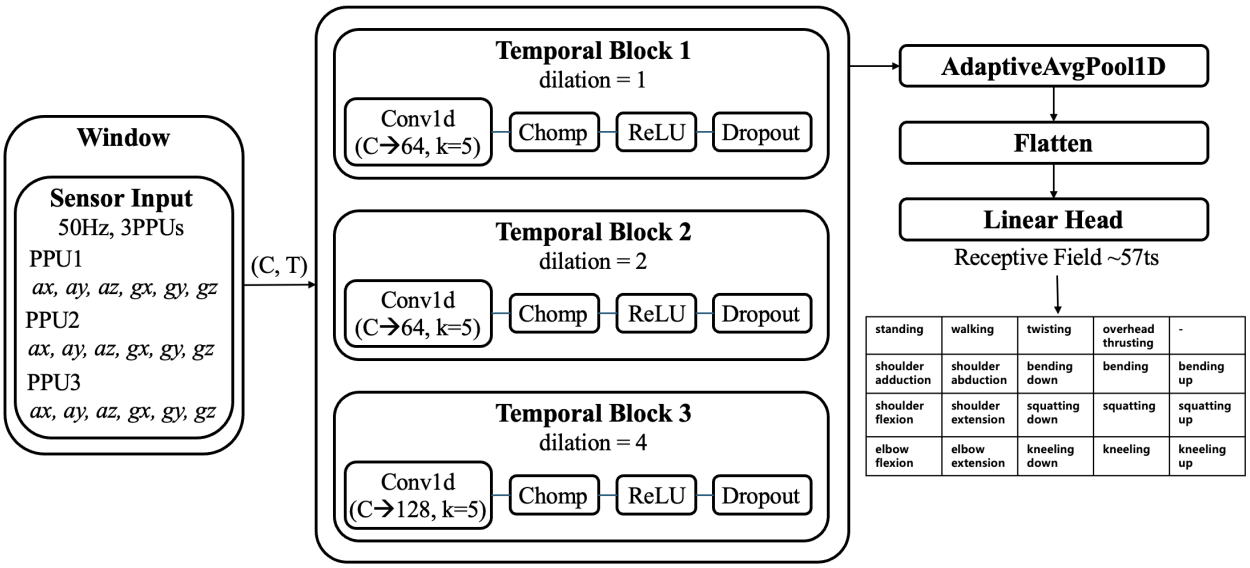

Figure 1. TCN Network Structure

improve recognition in controlled environments[6]. Another method explores optimization-based pose estimation from sparse IMUs, achieving high fidelity [9], but depends on non-causal inference or heavy priors that are hard to sustain on embedded systems. This leads to delayed monitoring time and extended computational cost. As a result, many existing approaches trade deployment practicality for accuracy or model complexity, assuming long temporal contexts or computationally intensive inference pipelines. These limitations underscore a gap in prior work for causal, lightweight temporal models that operate on short windows and support real-time inference with minimal sensor burden.

3 Methodology

3.1 Network Architecture

Our objective is to design and validate a light, edge-deployable system that recognizes 19 anatomically defined motion classes in real-time using three IMUs, while systematically identifying the most effective sensor combination and training protocol for downstream device design. We adopted TCN because it is causal, low-latency, and efficient on short windows typical of embedded streaming. Unlike bidirectional RNNs or Transformers, which require future context, a causal TCN operates on past data only and avoids recurrent state, making it easier to deploy in nonstationary jobsite conditions.

Our model described in Fig. 1 ingests a (C, T) IMU window (T=100≈2s at 50 Hz; C depends on the sensor combination). The backbone has three dilated residual blocks (channels 64/64/128, kernel 5) with shared dilations 1→2→4, ReLU+dropout, and 1×1 skips when needed. An AdaptiveAvgPool1d and a linear head produce logits for 19 motions. Right-padding and truncation ensure strict causality. This topology yields an effective receptive field of 57 steps (1.14 s), large enough to capture brief transients and sub-second pose evolution, helping separate look-alike actions like bend-down and kneel-down. We apply per-window, channel-wise z-normalization and a majority-vote smoother over recent outputs. Majority-vote smoother is a simple method of temporal post-processing where instead of displaying each prediction from time window independently, the smoother covers a short sequence(≈2s) of neighboring window predictions to select the most frequent label as output.

Computation scales modestly with kernel and channel counts, enabling real-time throughput on edge processors as memory is bounded by the fixed window. This balance of dilated temporal coverage, causal inference, and compactness suits field deployment with just three IMUs and tight latency under power budgets.

3.2 Sensors

Each ICM-20948 IMU is integrated into a wearable Personal Protection Unit (PPU) enclosure. We deploy three PPUs at the neck, lower back, and right wrist (Fig. 2). Upper-body placements such as the wrist, neck, and lower back have been shown to remain compatible with protective gear and routine work activities while maintaining user comfort[11, 15]. Prior work by Kim and Cho [12] systematically evaluated IMU quantity and placement on the body for deep learning-based worker motion recognition, identifying upper-body nodes as providing the best balance

ID	Motion	ID	Motion	ID	Motion	ID	Motion	ID	Motion
1	standing	2	walking	3	twisting	4	overhead thrusting	5	shoulder adduction
6	shoulder abduction	7	bending down	8	bending	9	bending up	10	shoulder flexion
11	shoulder extension	12	squatting down	13	squatting	14	squatting up	15	elbow flexion
16	elbow extension	17	kneeling down	18	kneeling	19	kneeling up		

Table 1. Anatomical terms for movement classification

of complementary motion coverage and minimal donning burden. Alternative locations such as the thigh and ankle were excluded because they would interfere with kneeling activities, footwear, and lower-body PPE common on construction sites. The neck unit sits at the base of the neck on the dorsal side to capture head-torso coupling while avoiding helmet interference. The back unit is centered over L5 with a belt clip to track trunk orientation and large sagittal motions. The wrist unit is strapped proximal to the ulnar styloid to capture hand activity with minimal glove obstruction. All mounts use rigid, anti-slip interfaces with consistent arrow markings to control sensor orientation.

After mounting, a brief functional check confirms sensor placement and synchronization. Subjects then perform a standardized calibration motion consisting of slow trunk flexion, an arm raise, and a return to neutral. This sequence verifies orientation consistency across the neck, back, and wrist channels and helps detect any strap slippage. All PPUs time-stamp packets on-device, and a lightweight offset correction is applied at the receiver to keep inter-PPU clock skew within 100 ms. To support deployment on remote sites with limited or absent network connectivity, each PPU stores raw sensor data locally on an onboard SD card.

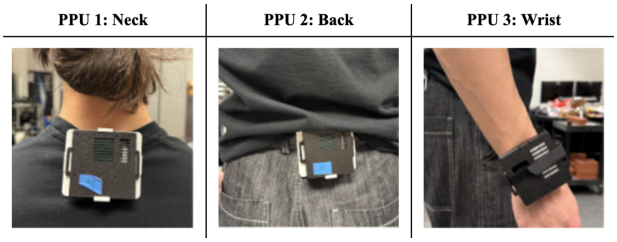

Figure 2. Neck, Back, Wrist PPUs Placement

3.3 Data Collection

Motion data was collected from 15 subjects in a height range of 153 cm to 185 cm and weight range of 44kg to 83kg. Subjects were directed to perform 13 classified motions(19 including -up, -down breakdowns) in a lab environment. Sensors streamed nine-axis IMU signals at 50 Hz to a local server; all devices were timestamp-synchronized with time tolerance of 100ms. Trials were executed in pre-defined and separate orders of motions for corresponding ground truth labels.

To address motions, construction related activities such as hammering, lifting, placing, and carrying were broken down into basic motions as described in Table 1. Motions were defined using standard anatomical terms (e.g., adduction/abduction and flexion/extension), and we split them into direction-specific phases (e.g., . . . -down and . . . -up) to disambiguate symmetric transitions where applicable. The specific classification of the motions are displayed on Table 1.

3.4 Training Configuration

We trained exclusively on six-axis IMU features (three-axis accelerometer and three-axis gyroscope) and excluded magnetometer channels, as magnetic interference from steel members and powered equipment can destabilize headings on active jobsites. Two evaluation regimes were used: (i) a single split with 80% train and 20% test (equivalently, 8:2 train:test) fixed across all experiments for ablation consistency, and (ii) leave-one-subject-out(LOSO), which withholds each participant in turn to assess subject-level generalization. Windowing and inference settings matched deployment (short, causal windows), and all models were trained with identical optimization hyperparameters so differences reflect only sensor combinations and split protocols.

Table 2. Sensor Combinations used for Training

combination	selected IMUs	combination	selected IMUs
1	Neck+Back+Wrist(NBW)	5	Neck(N)
2	Neck+Back(NB)	6	Back(B)
3	Neck+Wrist(NW)	7	Wrist(W)
4	Back+Wrist(BW)		

3.5 Model Inference

3.5.1 Combination of Sensors

To identify the minimal instrumentation that preserves recognition accuracy, we trained the TCN on seven sensor configurations as described in Table 2: three single-IMU baselines (wrist, neck, back), three dual-IMU pairs (wrist+neck, wrist+back, neck+back), and one tri-IMU set (wrist+neck+back). This factorial ablation isolates the marginal value of each body location and their complementarities for dynamic and quasi-static motions. For comparability, all runs used identical windowing, optimization settings, and preprocessing. Only the input channel count varied with the chosen combination (6 channels per IMU, excluding the magnetometer, as specified in 3.4). The tri-IMU model serves as an upper-bound reference, while dual and single configurations quantify the trade-off between wearability and performance needed for field deployment.

Table 3. Training results across combinations

Method	Comb.	Acc.	Bal. Acc.	Macro F1	Wtd. F1
single split	NBW	0.9670	0.9396	0.9402	0.9671
single split	NB	0.9526	0.9343	0.9249	0.9531
single split	**NW**	**0.9748**	**0.9451**	**0.9485**	**0.9743**
single split	BW	0.9644	0.9176	0.9253	0.9636
single split	N	0.9570	0.9276	0.9246	0.9570
single split	B	0.8874	0.8227	0.8227	0.8893
single split	W	0.9470	0.8783	0.8893	0.9448
LOSO	**NBW**	**0.9607**	**0.9315**	**0.9302**	**0.9608**
LOSO	NB	0.9266	0.8828	0.8845	0.9263
LOSO	NW	0.9522	0.9137	0.9155	0.9517
LOSO	BW	0.9494	0.9013	0.9032	0.9491
LOSO	N	0.9160	0.8653	0.8673	0.9159
LOSO	B	0.8714	0.8029	0.8042	0.8702
LOSO	W	0.9263	0.8565	0.8593	0.9247

3.5.2 Training Output

Table 3 summarizes single-split performance across sensor combinations. The results are evaluated in terms of Accuracy and Balanced Accuracy (weighted by the number of samples per class), along with F1 scores. The F1-score as defined in Equation (1) is a main machine learning evaluation metric that balances precision and recall into a single score to represent the harmonic mean. We selected the Macro F1-score, which is the unweighted average of the F1-scores computed independently for each class, and the Weighted F1-score which averages the F1-scores per class, weighted by the number of samples per class.

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c} \tag{1}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c \tag{2}$$

Table 4. Activity–motion lists used in the evaluation.

No	Activity	Motion
1	walking	walking squatting-down
2	brick lifting	squatting squatting-up elbow-flexion
3	carrying	walking
4	brick placing	bending-down bending-up
5	grab hammer in closet	shoulder-abduction thrusting shoulder-abduction
6	hammering	elbow-flexion elbow-extension shoulder-flexion shoulder-extension
7	grab heavy toolbox	kneeling-down
8	place toolbox on top	kneeling-down thrusting
9	stretching	twisting shoulder-extension shoulder-flexion

$$\text{Weighted-F1} = \frac{\sum_{c=1}^{C} N_c \cdot \mathrm{F1}_c}{\sum_{c=1}^{C} N_c} \tag{3}$$

Neck–wrist achieves the highest accuracy 0.9748 and the highest balanced accuracy of 0.9451. It also delivers the best Macro-F1 of 0.9485 and weighted F1 of 0.9743. The tri-sensor NBW configuration achieves a competitive accuracy of 0.9670 and a balanced accuracy of 0.9396. Neck–back and back–wrist trail by small margins across all metrics. Single-sensor settings degrade, with back only at 0.8874 accuracy and wrist at 0.9470. These results indicate that the wrist provides strong discriminative cues and that the neck complements it well. The limited gains of NBW over neck–wrist suggest a two-sensor design is sufficient for field deployment

3.5.3 Real-time Inference

In-lab performance test was conducted with three participants who did not participate in training data collection process. Performance test consists of a series of activities commonly observed in construction fields, and each activity is a combination of 1 to 4 basic motions defined for data collection. The test sequence of activities and their motion composition are described in Table 4.

4 Result

4.1 Result of Single Split

We outline an 8:2 ratio single-split TCN training using 7 sensor combinations to verify its accuracy on the inference test. Results in Table 5. show that the model achieves an average accuracy of 77.64%. The optimal combination is NBW, achieving an average accuracy of 83.26%,

Table 5. Single-split inference results

Combination	Run#1	Run#2	Run#3	Run#4	Run#5	Run#6	Run#7	Run#8	Run#9	AVG
NBW	**0.8182**	**0.8608**	**0.8356**	**0.8033**	**0.8194**	**0.8209**	**0.8361**	**0.8358**	0.8636	**0.8326**
NB	0.7848	0.7595	0.7848	0.7237	0.7722	0.7975	0.7910	0.8058	0.7879	0.7785
NW	0.8228	0.8354	0.8481	0.8033	0.8472	0.8060	**0.8710**	0.8507	0.8333	0.8353
BW	0.7975	0.8101	0.8101	0.7105	0.7595	0.7848	0.8217	0.8358	0.8030	0.7926
N	0.7595	0.7470	0.7397	0.7377	0.7778	0.7848	0.8230	0.8060	0.7727	0.7720
B	0.6835	0.6962	0.7089	0.6579	0.6330	0.6962	0.7164	0.7015	0.6970	0.6878
W	0.7342	0.7215	0.7260	0.6842	0.7342	0.7468	0.7761	0.7463	0.7576	0.7363

Table 6. LOSO inference results

Combination	Run#1	Run#2	Run#3	Run#4	Run#5	Run#6	Run#7	Run#8	Run#9	AVG
NBW	**0.7468**	**0.8075**	**0.7733**	**0.8226**	**0.8194**	**0.7971**	**0.7903**	**0.8060**	**0.8182**	**0.7979**
NB	0.6582	0.6076	0.6027	0.5672	0.5890	0.6076	0.6000	0.6119	0.5909	0.6039
NW	0.7179	0.7722	0.7534	0.7143	0.7361	0.7612	0.7869	0.7463	0.7576	0.7495
BW	0.7215	0.6582	0.6579	0.6212	0.6489	0.6418	0.6515	0.6716	0.6061	0.6532
N	0.6962	0.6709	0.6962	0.6557	0.6667	0.7206	0.6825	0.6866	0.7273	0.6892
B	0.5897	0.5949	0.5890	0.5385	0.5833	0.6119	0.6615	0.6567	0.6218	0.6053
W	0.6456	0.6582	0.6712	0.6190	0.6667	0.6418	0.6769	0.7313	0.6970	0.6675

confirming that three sensors provide robust performance. BW and NB form a middle tier with averages of 79.26% and 77.85%, indicating moderate loss when one sensor is removed. Among single-sensor setups, Neck performs best at 77.20%, while Wrist reaches 73.63% and Back remains lowest at 68.78%. Including the wrist consistently improves performance across all multi-sensor configurations. These results suggest that single-split training better captures richer temporal patterns, and that the NW configuration offers the best balance between accuracy and sensor reduction.

4.2 Result of LOSO

We present accuracy results for 7 sensor combinations and 19 motions for the TCN model trained using the LOSO method. Results show the model predicts with an average of 68.09% accuracy total. The optimal combination is NBW, with an average accuracy of 79.79%. Numerical results of accuracy are shown in Table 6. NW is the next strongest at 74.95%, which indicates a viable two-sensor option. Removing the wrist most adversely affects accuracy, with NB averaging 60.39% and BW dropping to 65.32%. For single sensors, the ranking is Neck at 68.92%, Wrist at 66.75%, and Back at 60.53%. Adding the wrist to the neck raises accuracy by 6.03 points, while adding the wrist to the back raises it by 4.79 points.

4.3 Sensor Combination x Training Method

Both single-split and LOSO training methods achieve the highest accuracy with the NBW combination, indicating that using all three sensor points is superior to leaving out one or two sensors during training. Single split training performed better across all 7 combinations than LOSO training in the current dataset, achieving an average accuracy of 83.26% in the NBW combination.

5 Discussion

The causal TCN is benchmarked across seven sensor configurations under simple single-split and leave-one-subject-out regimes. The LOSO method was applied to enhance person-to-person generalization. Although the LOSO method was expected to perform better due to its generalization to different human samples, the results indicate better inference performance on the simple 8:2 split. This result may have occurred due to a small number of training samples that were not large enough to ensure cross-person generalization with the LOSO method.

The results show that Neck+Back+Wrist(NBW) combination of sensors is optimal for worker motion monitoring with an average of 83.26%(8:2 split) and 79.79%(LOSO) accuracy. This indicates adding sensors to major nodes of body movement will increment motion prediction accuracy but not significantly. However, the hazardous nature of the construction field makes it difficult to add more sensors especially on body nodes since devices on movement nodes can interfere with natural movements. Excluding the NBW combination, Neck+Wrist(NW) combination also displayed reasonably high accuracy compared to the other 5 combinations, showing the possibility of reducing sensors while maintaining high prediction accuracy.

The results in both training methods reveal similar error sources: (i) predicting a single motion when the activity contains multiple motions, (ii) using a two-second window that can blur short transitional states, and (iii) misclassifying coupled motions such as elbow extension occurring with shoulder extension. The first and second error sources

share similar ground which means solving one is likely to solve the other together. The first issue can be addressed by scaling up the prediction to the activity level, where an activity is a combination of multiple consecutive motions. This will still predict motions, but inaccurate predictions may occur due to the order of predictions, or transition motions may be ignored on a larger scale. The third issue should be solved by weighting wrist gyro and acceleration data during training, as most of the confusion arises from elbow and shoulder actions that rely heavily on wrist PPU motion data.

To address the suggested problems while maintaining the project's real-time prediction concept, we plan to scale from motion to activity to align with practical field display. This is to define an activity as a combination of multiple basic motions; for example, brick lifting as a combination of squatting down, elbow flexion, and squatting up. Our data collection and inference were conducted in a controlled laboratory setting with scripted motion sequences and fixed timing. This environment is less dynamic and more predictable than active construction sites; thus, real-world variability, interruptions, and tool interactions may reduce performance. We will validate on live construction sites in future work. Our next step is to extend the TCN to predict real-world activities as time-ordered combinations of motions, and to embed a pre-trained model on a microcontroller for live construction-site deployment. On the hardware side, we are actively working to improve the wearability of the PPU system. Current efforts focus on reducing the wrist unit to approximately 2 inches in diameter through improved circuit design and power management, bringing it closer to the form factor of a standard wristband. This size reduction is a stepping stone toward embedding the sensing units into existing PPE such as hard hat liners and safety vests, which would eliminate separate donning steps and lower the adoption barrier on active jobsites.

6 Conclusion

This work presented a causal TCN for real-time worker motion recognition using three IMUs. The system uses short windows and per-window normalization, and it runs at 50 Hz with modest compute. Across seven sensor configurations, the NBW setup achieved the best accuracy. Single-split averaged 83.26% for NBW, and LOSO averaged 79.79%, which shows good cross-person generalization. NW was the strongest two-sensor choice and suggests a path to lower burden.

We will expand subjects and environments, add latency and power benchmarks, and deploy a compact model on a PPU. We will also study placement variants and class-weighted training to reduce upper-limb confusion. Overall, the results show that a small, causal model with three IMUs can support practical jobsite monitoring.

References

[1] Jochen Teizer, Brian S. Allread, Craig E. Fullerton, and Jimmie Hinze. Autonomous pro-active real-time construction worker and equipment operator proximity safety alert system. *Automation in Construction*, 19(5):630–640, 2010.

[2] J. Gong and C.H. Caldas. Computer vision-based video interpretation model for automated productivity analysis of construction operations. *Journal of Computing in Civil Engineering*, 23(3):151–160, 2009.

[3] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[4] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics*, 2018. SIGGRAPH Asia.

[5] Y. Ren et al. Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *arXiv preprint arXiv:2205.15410*, 2022.

[6] Davoud Gholamiangonabadi, Nikita Kiselov, and Katarina Grolinger. Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection. *IEEE Access*, 8:133982–133994, 2020.

[7] F. Ordoñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 2016.

[8] W. Tao, M. C. Leu, and Z. Yin. Multi-modal recognition of worker activity for human-centered intelligent manufacturing. *Engineering Applications of Artificial Intelligence*, 2019.

[9] Kinam Kim and Yong Cho. Automatic recognition of workers' motions in highway construction by using motion sensors and long short-term memory (lstm) networks. *Journal of Construction Engineering and Management*, 147, 10 2020.

[10] S. Chatterjee et al. Accoustate: Auto-annotation of imu-generated activity signatures under smart infrastructure. *arXiv preprint arXiv:2112.06651*, 2021.

[11] Minsoo Park, Seongwoo Son, Yuntae Jeon, Dongyoung Ko, Mingeon Cho, and Seunghee Park. Scaffolding worker imu time-series dataset for deep learning-based construction site behavior recognition. *Advanced Engineering Informatics*, 65:103232, 2025.

[12] Kinam Kim and Yong K. Cho. Effective inertial sensor quantity and locations on a body for deep learning-based worker's motion recognition. *Automation in Construction*, 113:103126, 2020.

[13] Jinwoo Kim, Kyeongsuk Lee, and JungHo Jeon. Systematic literature review of wearable devices and data analytics for construction safety and health. *Expert Systems with Applications*, 257:125038, 2024.

[14] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[15] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. Physical activity recognition with mobile phones: challenges, methods, and applications. *IEEE Signal Processing Magazine*, 32(4):52–59, 2015.