CardioHARNet: A Lightweight Hybrid Deep Model Using Raw IMU Signals for Human Activity Recognition

Morsheda Akter
School of Computer Science and Engineering
Pusan National University
Busan, South Korea
anniislam1108@gmail.com

Junyoung Son
School of Computer Science and Engineering
Pusan National University
Busan, South Korea
jyson@pusan.ac.kr

***Abstract*—In this paper, we introduce CardioHARNet, a lightweight hybrid deep neural architecture that integrates Convolutional Neural Networks (CNNs) for local feature extraction, Convolutional Block Attention Modules (CBAM) for adaptive channel–temporal weighting, and bidirectional LSTM (BiLSTM) for long-term dependency modeling. The model operates directly on raw IMU time-domain signals from accelerometer and gyroscope sensors without any handcrafted feature engineering. Experiments on the KU-HAR dataset, which contains 18 daily activities with 1,945 original samples and 20,750 windowed segments, demonstrate that CardioHARNet achieves 95.57% test accuracy. It outperforms the re-implemented 1D CNN baseline by 7.7 percentage points while using significantly fewer parameters (174K vs. 566K). The results show that CBAM enhances discriminative feature learning from raw signals without additional preprocessing. The proposed model is promising for real-time wearable IoT and early movement-risk management.**

***Index Terms*—Human Action Recognition, Convolutional Neural Networks, Attention Mechanism, IMU sensors, KU-HAR dataset, Deep Learning**

I. Introduction

Human Activity Recognition (HAR) has become essential for healthcare monitoring, rehabilitation, sports analytics, and intelligent IoT systems [1]–[3] based on wearable and smartphone inertial sensors. Initial HAR systems were highly dependent on hand-crafted features or time-frequency features [4], [5], which did not generalize well across subjects or activities and required domain knowledge. Several works transform IMU signals into images or spectrograms for use with 2D CNNs [6]. However, preprocessing increases computational complexity, making these approaches unsuitable for wearable IoT devices in real-time.

In recent years, deep neural networks have significantly improved HAR performance. Convolutional Neural Networks capture local temporal patterns from raw sensor signals [7], while recurrent networks such as LSTM are useful for capturing long-term dependencies in human motion [8]. Nevertheless, CNN-based models lack temporal memory, whereas LSTM-based models struggle with fine-grained local feature extraction. Moreover, standard CNN-LSTM hybrids do not explicitly focus on the most informative sensor channels or time segments. Recent attention-based frameworks [9], [10] highlight the importance of guiding the network to address salient movement cues, but most require frequency-domain transformations or multimodal inputs. Although several public datasets, such as Opportunity, WISDM, and UCI-HAR, have accelerated benchmarking of deep learning approaches, they do not explicitly emphasize informative sensor channels or time segments. These characteristics make them unsuitable for real-time deployment on resource-constrained wearable and IoT devices, where memory footprint, inference latency, and power consumption are critical. In this research, we focus on the KU-HAR dataset, released in 2021 by Nahid et al. [11]. The dataset contains 18 daily activities and fitness activities collected from 90 participants using smartphone accelerometer and gyroscope. The original DeepConvLSTM baseline achieved only 90.87% accuracy [11], which leaves considerable room for future improvement. While some studies have reported up to 99% using heavier ensembles [12], but at the cost of increased complexity. To address these limitations, we propose CardioHARNet, a lightweight hybrid architecture combining CNNs, Convolutional Block Attention modules, and BiLSTM to learn both local patterns, long-term temporal structure, and channel-temporal importance. Our model processes raw IMU time-domain signals without any hand-crafted features or spectrogram generation, which is suitable for low-power wearable devices. After combining the last hidden state with mean-pooled temporal features, the model effectively captures both local dynamics and global patterns. Extensive experiments on KU-HAR demonstrate that CardioHARNet achieves a competitive accuracy of **95.57%** with fewer parameters, making it more suitable for wearable devices.

II. Related Work

Human Activity Recognition (HAR) has rapidly changed from early hand-crafted feature engineering to modern deep neural architectures. This section summarizes the field in three main directions: (1) feature-based approaches in traditional methods, (2) deep learning techniques of IMU-based HAR, and (3) attention-based and hybrid frameworks for understanding the evaluation.

A. Feature-based approaches in traditional methods

Traditional Human Activity Recognition relied heavily on hand-crafted statistical, temporal, and frequency-domain features extracted from accelerometer and gyroscope signals. In early works, accelerometer or gyroscope data were divided into fixed segments and a set of predefined descriptors, including mean, variance, energy, correlation, and frequency domain coefficients. Classical machine learning models such as decision trees, SVMs, HMMs, and k-NN used these features as input. Several works, including Bao and Intille's, showed that the acceleration-based annotation dataset and the feature extraction pipeline [5] could achieve reasonable performance for daily activities, and biometric gait-based wearable authentication was used for feature engineering in Casale et al. [4].

Although these methods performed well in some contexts, they suffered from poor generalization and struggled with complex or transitional movements. Additionally, the feature-engineering process itself is time-consuming and limits the scalability of these approaches. To learn discriminative representations directly from raw signals, these challenges motivated the shift toward deep-learning-based HAR models.

B. Deep learning techniques of IMU-based HAR

Deep neural networks have become the dominant approach for IMU-based activity recognition due to the growing availability of wearable sensors and larger annotated datasets. Convolutional Neural Networks (CNNs) were the first architectures to demonstrate strong performance on raw inertial signals, mainly because of their ability to capture local temporal patterns and short-term motion characteristics. Chen et al. [6] proposed a CNN-based model that defined a strong capability for capturing local temporal-spatial patterns from IMU time-series. Recurrent networks such as Long Short-Term Memory (LSTM) models [8] have also been widely adopted in HAR because human activities naturally exhibit sequential structure. Long-range temporal dependencies can be captured by LSTM-based models and are effective for activities that span multiple time windows or involve smooth transitions.

Hybrid architectures that integrate CNNs and LSTMs further improved performance by combining CNNs' local feature extraction from raw sensor channels, while LSTMs learn the progression of movements over time. Ordóñez and Roggen's DeepConvLSTM [7] is one of the most widely used baselines in many HAR benchmarks. However, these models often lack limited attention to informative channels. As a result, they do not emphasize the most informative segments of the signal, especially when activities are visually or kinematically similar(e.g., walking vs walking-backward). Researchers are encouraged to explore attention mechanisms to enhance deep neural models for HAR because of these limitations. Several non-deep learning approaches have also been explored for KU-HAR, such as metaheuristic-driven feature selection with XGBoost classifiers [13], achieving optimized high accuracy through efficient feature engineering. However, these methods offer low computational complexity that causes underperforming deep learning hybrids in capturing complex temporal patterns, motivating our focus on lightweight deep learning architectures.

C. Attention-based and hybrid frameworks

Recent studies have shown that HAR models can be improved by adding attention mechanisms to interpret sensor signals. In traditional pipelines, CNN-LSTM often treats all sensor channels and all time steps uniformly. Attention modules suppress irrelevant noise by focusing on informative regions to help the network. The Convolutional Block Attention Module (CBAM) proposed by Woo et al. [9] introduced channel attention followed by spatial (or temporal) attention, allowing the model to reinforce important feature maps and temporal patterns cautiously. Beyond CBAM, AttnSense [10] is a multi-level attention framework that highlights essential motion segments across different sensor modalities. By capturing fine-grained temporal relevance, these architectures achieve strong performance at multiple stages of the network pipeline. Several works have pushed accuracies higher on KU-HAR using attention-enhanced hybrids, such as ResLSTM variants reaching 97.05% with 386K parameters [14]. However, these often require more computational resources, making them unsuitable for real-time wearable applications.

Our work aims to integrate attention mechanisms into a lightweight, raw-signal deep model to solve these limitations. CardioHARNet integrates CNNs, CBAM modules, and a bidirectional LSTM to learn both local features and long-term temporal structure, as well as channel–temporal importance directly from raw IMU data. The model does not rely on handcrafted features or expensive transformations and balances high accuracy (95.57%) with low overhead for wearable devices.

III. Methodology

A. Dataset Description

In our research, we used a publicly available dataset published in 2021. The KU-HAR dataset contains raw accelerometer and gyroscope recordings from 90 participants who performed 18 daily activities. A total of 1,945 raw activity samples and 20,750 subsamples were collected from the participants. Each trial is captured using a smartphone IMU and stored as time-domain CSV files.

B. Dataset Processing and Window Generation

1) Feature Organization: Each sample in the KU-HAR dataset contains six raw IMU channels. They are obtained from the smartphone's accelerometer and gyroscope as:

- **Accelerometer:** a_x, a_y, a_z
- **Gyroscope:** g_x, g_y, g_z

No handcrafted features, statistical descriptors, or frequency-domain transformations were applied in our study. To preserve the natural temporal structure of human movement, our models operate directly on the raw time-domain signals.

2) Sliding Window Segmentation: We implemented a fixed-length sliding window for sensor-based HAR with the following parameters.

$$W = 128, \quad \text{stride} = 64.$$

This processing generates overlapping segments that capture short-term motion dynamics while also increasing the number of training samples. For each activity recording, windowed sequences are formed as:

$$X_i = [s_i, \, s_{i+1}, \, \ldots, \, s_{i+W-1}],$$

where s_i represents the i-th timestamped IMU sample. The label assigned to each window corresponds to the activity of its source sequence.

Through this segmentation, the dataset expands from 1,945 raw recordings to approximately 20,750 windowed samples, enabling stable training of deep neural networks.

3) Normalization: Channel-wise normalization was applied to every IMU signal using the formula:

$$x' = \frac{x - \mu}{\sigma + 10^{-8}},$$

where μ and σ are calculated entirely from the training set to avoid data leakage.

4) Train–Test Split: We divided the processed samples into training and testing sets using an 80/20 stratified split to ensure a balanced representation of all 18 activity classes. We used the PyTorch Dataset and DataLoader modules to manage data loading and batching for efficient training.

C. CardioHARNet Architecture

We propose a CardioHARNet model that integrates three complementary components to learn discriminative motion patterns directly from raw IMU time-series data.

1) CNN Feature Extractor: We used three sequential 1D convolutional blocks to increase the depth of the feature (32, 64, and 128). Batch Normalization, ReLU activation, and MaxPooling are included with Conv1D in each block. With the help of these layers, we extracted short-term temporal features and reduced the sequence length from $128 \rightarrow 64 \rightarrow 32 \rightarrow 16$ hierarchically. After Block 2 and Block 3, we applied CBAM attention modules to refine important sensor channels and relevant temporal regions.

2) Bidirectional LSTM Layer: The output is reshaped to (B, T, C) for the CNN feature extractor and passed through a bidirectional LSTM with a hidden size of 64 per direction.

$$BiLSTMoutput : (B, T, 128)$$

To improve temporal summarization, we concatenated the final hidden state with a mean-pooled feature.

3) Classification Head: A fully connected lightweight classifier consisting of dropout, a ReLU layer, and a final dense layer that computes probabilities over 18 activity classes through a softmax function.

$$\hat{y} = \text{Softmax}(Wh + b)$$

TABLE I: CardioHARNet Architecture

Layer	Filters/Units	Kernel	Output Shape
Input	6 ch	–	(128, 6)
Conv Block 1	32	7	(64, 32)
BN + ReLU + MaxPool	–	–	–
Conv Block 2	64	5	(32, 64)
BN + ReLU + MaxPool	–	–	–
CBAM (Block 2)	–	–	(32, 64)
Conv Block 3	128	3	(16, 128)
BN + ReLU + MaxPool	–	–	–
CBAM (Block 3)	–	–	(16, 128)
BiLSTM (Bi, 1 layer)	64×2	–	(16, 128)
Feature Fusion	–	–	(256)
FC Layer	128	–	(128)
Output (Softmax)	18	–	(18)

IV. Experiments and Results

A. Performance Comparison

We compare CardioHARNet with a re-implemented 1D CNN baseline model to evaluate the performance of the proposed model. Both models were trained on the KU-HAR dataset under the same experimental settings using the Adam optimizer, batch size 64, and StepLR scheduler.

The proposed CardioHARNet, the baseline 1D CNN, and recent state-of-the-art models on KU-HAR are summarized in Table II. CardioHARNet achieves 95.57% test accuracy with only 174K parameters, outperforming the re-implemented 1D CNN baseline (87.87%, 566K parameters) by 7.70 percentage points. Although some recent models achieve higher accuracies, they typically require significantly more parameters. CardioHARNet offers a competitive balance between accuracy and efficiency, highlighting the benefits of integrating CBAM attention and BiLSTM-based temporal modeling with CNN feature extraction for resource-constrained wearable and IoT devices.

TABLE II: Performance Comparison of CardioHARNet with Baseline 1D CNN and Recent SOTA on KU-HAR

Model	Test Accuracy (%)	Params	Features
Baseline 1D CNN (re-implemented)	**87.87**	**566K**	**simple 1D CNN**
DeepConvLSTM [11]	**90.87**	**$\sim$1M**	**Original baseline**
Deep-HAR Ensemble [12]	**$\sim$99**	**> 500K**	**Ensemble DL**
Metaheuristic XGBoost [13]	**$\sim$94–96**	**Low (non-DL)**	**Feature sel. + meta**
ResLSTM [14]	**97.05**	**386K**	**Residual LSTM**
CardioHARNet (Ours)	**95.57**	**174K**	**CNN + CBAM + BiLSTM**

B. Confusion Matrix Analysis

1) Baseline CNN: The confusion matrix of the baseline CNN model is presented in Figure 1. Several activities with similar movement patterns are difficult to classify, such as *sit vs. stand*, *walk vs. walk-circle*, *stair-up vs. stair-down*, and *Talk-Sit vs. Talk-Stand*. These misclassifications arise from the lack of explicit temporal modeling and insufficient attention to discriminative channels. The model achieves only 87.87% in the test set due to these misclassifications.

2) Proposed CardioHARNet: The confusion matrix of the proposed CardioHARNet is shown in Figure 2. It is shown that the matrix is strongly diagonal, indicating accurate per-class prediction performance across all 18 activities. The

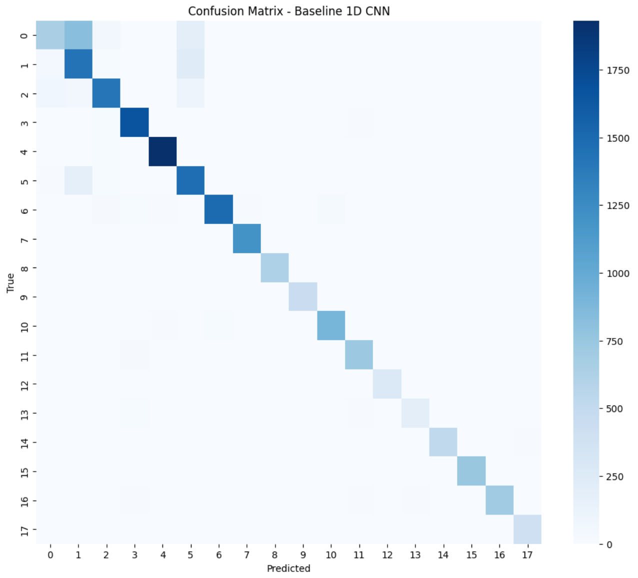

Fig. 1: Confusion matrix of Baseline 1D CNN.

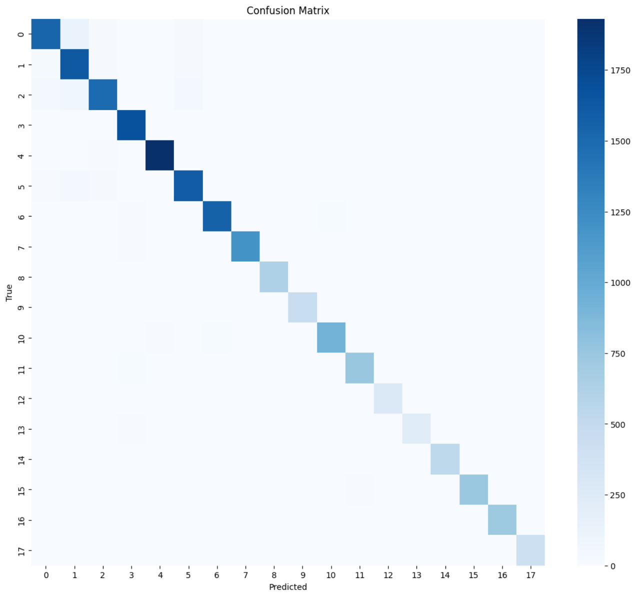

Fig. 2: Confusion matrix of CardioHARNet.

CBAM modules help to highlight the most informative IMU channels, while the BiLSTM captures long-term temporal dependencies. Together, they reduce ambiguity between similar motion classes. Some misclassifications remain between *walk vs. walk-circle*, and *Upstairs vs. Downstairs*. Overall, CardioHARNet achieves 95.57% accuracy, demonstrating the benefits of combining CNN and CBAM with BiLSTM to capture both fine-grained and global temporal dynamics.

C. Training and Testing Accuracy Curve

The training and testing accuracy curves are illustrated in Figure 3 for CardioHARNet. Within the first 10 to 12 epochs, most of the improvements are achieved, and the model reaches high accuracy rapidly. The testing accuracy curve closely follows the training accuracy curve throughout the training, indicating that the model generalizes well and does not overfit.

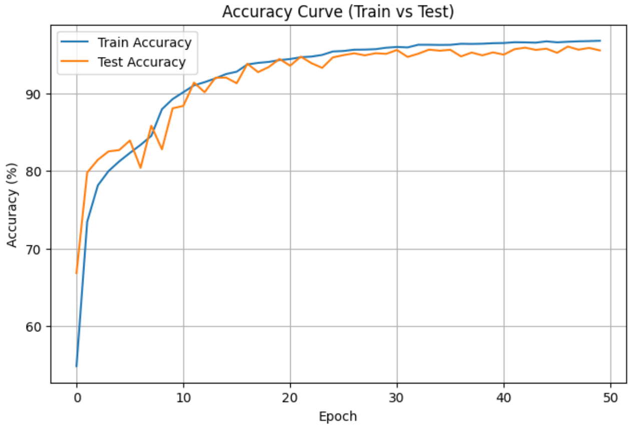

Fig. 3: Training and testing accuracy curves for CardioHARNet on the KU-HAR dataset.

V. Discussion

CardioHARNet provides a clear improvement over the baseline 1D CNN model in experimental results, particularly for activities that depend on subtle temporal signals or fine-grained limb coordination. By integrating CBAM, the model significantly improves in emphasizing informative sensor channels and suppressing irrelevant movements. This effect is visible in the reduction of confusion matrices, where similar activity pairs such as *walk* vs. *walk-backward* or *sit-up* vs. *push-up* are handled more reliably by CardioHARNet, compared to the baseline network.

To enhance performance, an additional BiLSTM layer is added to the model. Although convolutional filters are used to capture short-range transitions, they cannot model long-range temporal dependencies. The BiLSTM models activity sequences over time to address the problem, enabling the model to retain both short-term motion transitions and broader activity context. Since user movements are often continuous and non-uniform, this is essential for practical HAR environments.

Another important strength of the proposed model is that CardioHARNet learns directly from raw IMU time series without relying on spectrograms or engineered features. This architecture avoids the need for expensive preprocessing steps and makes the model more suitable for low-power wearable devices. The model remains compact by combining lightweight convolutions, attention, and a single-layer BiLSTM to achieve high accuracy.

However, some limitations remain. Activities with inherently overlapping motion patterns are still difficult to separate properly, and variation in how users execute the same activity with different intensities and styles can affect performance. Moreover, though the model is lighter than many spectrogram-based architectures, additional optimization or compression would be required for ultra-low-power microcontroller deployment.

In summary, our results demonstrate that integrating attention into a CNN–BiLSTM pipeline is an effective strategy for raw-signal human activity recognition. The model strikes a balance between accuracy, efficiency, and practical deployability, making it a promising solution for real-world wearable and IoT applications.

VI. Conclusion

This paper presents CardioHARNet, a lightweight hybrid model based on raw IMU signals for accurate and efficient human activity recognition. CardioHARNet integrates 1D convolutional layers and CBAM attention module with a bidirectional LSTM model. Bidirectional LSTM plays a critical role in capturing long-term temporal dependencies that cannot be fully modeled by convolutional layers alone. Although CNN blocks learn motion patterns and CBAM enhances channel-temporal to emphasize informative sensor signals, the BiLSTM aggregates temporal context across extended time windows. This combination enables CardioHARNet to simultaneously capture fine-grained local dynamics and global temporal structures, leading to improved recognition performance, particularly for activities with subtle or overlapping motion patterns. Another key strength is that the model operates directly on time-domain data, which eliminates the preprocessing cost and improves its applicability to wearable and IoT gadgets, where previous HAR systems only depend on handcrafted features or spectrogram transformations.

After performing the experiments in the KU-HAR dataset, our model achieved 95.57% test accuracy and performed better than the baseline 1D CNN model. In addition, confusion matrix analysis is added to highlight the strength of our model to recognize activities that exhibit subtle or overlapping movement characteristics. These results validate the effectiveness of integrating channel-temporal attention with sequence modeling for challenging HAR tasks.

In future work, we plan to explore microcontroller-level optimization, cross-dataset generalization, and real-time deployment on embedded wearable platforms. Overall, CardioHARNet is a promising step towards reliable and practical HAR systems that are both low-cost and deployable for health care, fitness, and human-oriented IoT applications.

VII. Acknowledgment

This work was supported by Korea Planning & Evaluation Institute of Industrial Technology(KEIT) grant funded by the Korea government(MOTIE) (RS-2025-04003004, IS^{2}D (Intelligent Self-evolving Security Dome))

References

[1] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys*, vol. 46, no. 3, pp. 1–33, 2014.

[2] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[3] O. Lara and M. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[4] P. Casale, O. Pujol, and P. Radeva, "Personalization and user verification in wearable systems using biometric walking patterns," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 796–804, 2010.

[5] L. Bao and S. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing*, 2004, pp. 1–17.

[6] Y. Chen *et al.*, "Human activity recognition using a hybrid deep learning approach," *IEEE Access*, vol. 8, pp. 139–152, 2020.

[7] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[10] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "Attnsense: Multi-level attention mechanism for multimodal human activity recognition," in *IJCAI*, 2019, pp. 3109–3115.

[11] A.-A. Nahid, N. Sikder, and I. Rahim, "Ku-har: An open dataset for human activity recognition," Feb. 2021, version 5.

[12] P. Kumar and S. Suresh, "Deep-har: An ensemble deep learning model for recognizing simple, complex, and heterogeneous human activities," *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 30 435–30 462, Feb 2023.

[13] P. Sarker, J.-J. Tiang, and A.-A. Nahid, "Metaheuristic-driven feature selection for human activity recognition on ku-har dataset using xgboost classifier," *Sensors*, vol. 25, no. 17, p. 5303, Aug 2025.

[14] S. AlMuhaideb, L. AlAbdulkarim, D. M. AlShahrani, H. AlDhubaib, and D. E. AlSadoun, "Achieving more with less: A lightweight deep learning solution for advanced human activity recognition (har)," *Sensors*, vol. 24, no. 16, p. 5436, Aug 2024.