Human Interaction and Emerging Technologies (IHIET-FS 2025), Vol. 196, 2025, 238–248
https://doi.org/10.54941/ahfe1005972

Lightweight Transformer for Robust Human Activity Recognition Using Smartphone IMU Data

Hossein Shahverdi[1] and Seyed Ali Ghorashi[2]

[1]Independent Researcher
[2]School of Architecture, Computing and Engineering, University of East London, UK

ABSTRACT

Human-activity recognition (HAR) underpins a wide range of m-health, smart-home, and context-aware services, yet conventional approaches frequently struggle with overfitting, class imbalance, and limited capacity to capture long-range temporal dependencies. In this study we introduce a lightweight, end-to-end Transformer pipeline that learns directly from raw smartphone inertial signals, eliminating the need for manually engineered features. We evaluate the approach on MotionDetection, a 12-channel dataset collected from 24 volunteers who performed a scripted series of everyday movements while carrying a Samsung Galaxy Note 20 Ultra. After windowing and minimal preprocessing, the Transformer attains 98% validation accuracy with no discernible overfitting. Relative to a strong CNN-BiLSTM baseline, it improves the macro F1-score by 3.6 percentage points while employing a smaller parameter budget, underscoring its computational efficiency. These findings indicate that Transformer architecture can provide a robust, scalable foundation for real-world HAR on commodity mobile devices, paving the way for battery-friendly, on-device activity monitoring in health and ambient-assisted applications

Keywords: Human activity recognition, Transformers

INTRODUCTION

Human Activity Recognition (HAR) using wearable and smartphone-based sensors has emerged as a critical component of modern healthcare, fitness tracking, and ambient intelligence systems. Accurately detecting activities such as walking, jogging, sitting, or climbing stairs enables a wide range of real-world applications, from fall detection in eldercare to personalized fitness coaching and context-aware mobile services (Lara and Labrador, 2013; Avci et al., 2010).

Traditional HAR approaches typically relied on hand-crafted statistical features extracted from accelerometer and gyroscope signals, followed by classical machine learning models such as support vector machines (SVMs) or hidden Markov models (HMMs) (Kwapisz et al., 2011; Wang et al., 2019). Although effective under constrained conditions, these methods often struggle in real-world settings due to variability in sensor noise, subject diversity, and subtle inter-class similarities, particularly among dynamic

locomotion activities. With the advent of deep learning, convolutional neural networks (CNNs) and recurrent neural networks (RNNs), including Long Short-Term Memory (LSTM) architectures, have significantly advanced HAR by enabling end-to-end feature extraction directly from raw sensor signals (Hammerla et al., 2016; Ordóñez and Roggen, 2016; Hochreiter and Schmidhuber, 1997). However, these models still encounter notable limitations: CNNs are effective for local pattern detection but inefficient at modeling long-range dependencies, while RNNs, though capable of sequence modeling, suffer from vanishing gradients and training inefficiencies over extended time windows.

More recently, attention-based architectures—particularly transformers—have demonstrated remarkable success across sequential domains such as natural language processing and time-series forecasting (Vaswani et al., 2017; Zhou et al., 2021). Transformers leverage self-attention mechanisms to capture both short-term patterns and long-range dependencies without relying on recurrence, offering efficient parallel computation and improved scalability.

In this work, we explore the application of transformer encoders for smartphone-based human activity recognition. We develop a compact and lightweight transformer architecture specifically designed for raw inertial measurement unit (IMU) signals, aiming to capture the complex temporal dynamics of human motion without the need for manual feature engineering. Using the MotionDetection dataset, collected under real-world conditions, we instructed participants to perform activities naturally while maintaining consistent device placement and activity labeling across trials. The proposed model demonstrates superior recognition accuracy over baseline CNN-BiLSTM architectures while remaining computationally efficient for potential deployment on mobile platforms.

The main contributions of this work are as follows:

- We introduce a new dataset, called MotionDetection, collected from 24 participants under realistic and semi-controlled conditions using only a standard smartphone placed in the front pocket. The protocol captured six activities with attention to minimizing bias while preserving naturalistic behavior.
- We propose lightweight transformer architecture specifically designed for raw smartphone IMU data, enabling efficient end-to-end learning without reliance on handcrafted features.
- We implement an effective windowing and class-balancing strategy to mitigate dataset imbalance, particularly improving recognition performance for challenging activities such as stair traversal.

RELATED WORKS

Human Activity Recognition (HAR) has long been a central area of research, with early methods relying on handcrafted statistical features combined with traditional machine learning classifiers such as support vector machines (SVMs) and hidden Markov models (HMMs). While these approaches proved effective under controlled settings, they struggled to generalize across

varying users and environments due to limitations in feature design and model adaptability.

With the growth of deep learning, more powerful end-to-end models have emerged, capable of learning complex spatiotemporal patterns directly from raw sensor data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), particularly long short-term memory (LSTM) architectures, were among the first to significantly advance HAR performance. Raj and Kos (2023) demonstrated that a carefully optimized CNN could achieve up to 97.2% accuracy on the WISDM dataset. Mekruksavanich and Jitpattanakul (2023) extended this by incorporating channel attention mechanisms into multi-channel CNNs, reaching accuracies as high as 98.9% across various benchmark datasets. Recurrent models have also been widely explored. Zhang et al. (2023) proposed an attention-based residual BiLSTM network, integrating residual connections and layer normalization to enhance sequence modeling. Their model achieved 97–99% accuracy on datasets like UCI-HAR and WISDM, demonstrating the value of enhancing RNN-based models with attention and architectural refinements (Zhang et al., 2023).

More recently, attention-based architecture, particularly Transformers, have gained momentum in HAR research. Xing et al. (2023) introduced SVFormer, a Vision Transformer-based framework for semi-supervised video HAR, showing strong performance on large-scale datasets such as Kinetics-400. While most Transformer-based work has focused on video or multimodal HAR, Pramanik et al. (2023) explored their utility in sensor-based HAR, proposing a deep reverse attention model that dynamically emphasizes relevant signal segments over time. Their approach achieved state-of-the-art results on five wearable HAR datasets. Hybrid models that combine CNNs with Transformers have also shown promise. Djenouri and Belbachir (2023) introduced the Convolutional Visual Transformer Network (CVTN), in which CNNs first extract spatial features, and Transformers model temporal dependencies. Similarly, Chen et al. (2025) proposed a two-stream GCN-Transformer framework for skeleton-based HAR, demonstrating strong improvements on various benchmark datasets.

Our work is directly inspired by these advances. However, unlike prior efforts that primarily target video-based or multimodal HAR, we develop a compact and efficient Transformer architecture tailored specifically to raw smartphone IMU data. By eliminating recurrence and relying solely on self-attention, our model captures both short- and long-range dependencies effectively, while maintaining a low parameter count and high computational efficiency. This enables fast convergence and robust performance across diverse activity classes, setting it apart from conventional CNN or RNN-based HAR pipelines.

METHODS

Dataset

In this study, we introduce MotionDetection, a smartphone-based dataset collected to support the development and evaluation of deep learning models

for human activity recognition. Data was recorded from 24 volunteer participants representing diverse demographics in terms of age, gender, height, and weight. Participants were equipped with a standard smartphone (Samsung Galaxy Note 20 Ultra) placed naturally in the front pocket of their trousers, mimicking common real-world smartphone usage without additional attachments or specialized sensor positioning. All participants wore flat shoes and were asked to perform six activities in a semi-controlled environment to ensure consistency while preserving natural motion patterns. The activities included walking, jogging, walking upstairs, walking downstairs, sitting, standing.

Each activity was repeated across 15 trials per participant, comprising both short-duration (≈30–60 seconds) and longer-duration (≈2–3 minutes) sessions to capture variability in movement speed and context. Participants initiated and terminated recording manually using the smartphone interface, ensuring that the device placement remained consistent throughout each activity session. The MotionDetection dataset was collected on a university campus, utilizing predefined routes for dynamic activities (walking, jogging, stairs) and designated seating/standing zones for static activities. Data acquisition employed the SensingKit framework, capturing 12-channel motion signals at a sampling rate of 50 Hz. The recorded features include a) attitude: roll, pitch, yaw, b) gravity: x, y, z axes, c) user acceleration: x, y, z axes, and d) rotation rate: x, y, z axes.

In total, the dataset comprises over 9,000 segmented windows after preprocessing, with slight natural imbalance among activities due to differences in recording durations (e.g., stair activities were inherently shorter than level walking or jogging sessions). This realistic distribution reflects practical activity patterns in daily life while maintaining sufficient coverage for each class. To address minor class imbalance, we later applied synthetic minority oversampling at the window level during training (described in Section 3.2).

Data Preprocessing

To prepare the raw MotionDetection signals for model training, we applied a systematic segmentation and labeling process followed by balancing techniques to address class distribution.

The continuous 12-channel time-series data were segmented into overlapping windows using a fixed-length sliding window approach. Each window contained 50 consecutive samples, corresponding to approximately one second of recorded motion data given the 50 Hz sampling rate. A step size of 30 samples (i.e., 40% overlap between consecutive windows) was used to ensure sufficient temporal coverage while minimizing redundancy. Each segmented window was assigned a label based on the activity being performed at the start of the window. This strategy is consistent with best practices in sensor-based activity recognition, ensuring that window labels accurately reflect the dominant activity present within each segment.

Formally, for a given time-series $\{X_t, y_t\}$ where X_t represents the 12-channel sensor readings at time t and y_t the corresponding activity label,

the *k-th* window W_k is defined as:

$$W_k = \{X_t : t \in [k.step_{size},\ k.step_{size} + window_{size}]\} \tag{1}$$

Before model training, all input features were standardized by subtracting the mean and dividing by the standard deviation computed across the entire training set. This normalization step ensures that all sensor channels contribute equally during model optimization and mitigate bias due to differing signal scales. Despite a relatively even trial design, natural variations in activity durations resulted in slight class imbalance, particularly for stair-related activities, which inherently involved shorter recording sessions compared to level walking or jogging. To mitigate the risk of biased training outcomes, we applied the Synthetic Minority Oversampling Technique (SMOTE) at the window level. SMOTE synthesizes new training examples for underrepresented classes by interpolating existing samples, thereby enhancing the model's exposure to minority class patterns without simple duplication. The application of SMOTE ensured that each activity class contributed roughly equally during model training, improving generalization across all activities. After segmentation and balancing, the dataset comprised approximately 9,400 labeled windows evenly distributed across the six activity classes. These windows were subsequently partitioned into training and validation sets using an 80%/20% stratified split, maintaining class proportions in both subsets.

Model Architecture

The proposed model is based on a transformer encoder architecture tailored for multivariate time-series classification using raw smartphone IMU signals. The model processes each 50-sample input window as a sequence of sensor readings, learning both local motion patterns and long-range temporal dependencies without relying on recurrence. Each input segment is a matrix of shape 50×12, where 50 denotes the number of timesteps and 12 corresponds to the IMU features (attitude, gravity, rotation rate, and user acceleration along three axes each). The raw input is first passed through a linear embedding layer that projects each 12-dimensional sensor reading into a 512-dimensional latent space:

$$z_t = Linear\,(X_t)\,,\ \forall t \in [1, \ldots, 50] \tag{2}$$

resulting in an embedded sequence of shape 50×512.

To inject temporal order information into the model, fixed sinusoidal positional encodings are added to the embedded sequence. This enables the transformer to distinguish between timesteps, a necessary feature given its permutation-invariant self-attention mechanism. The resulting position-aware sequence is denoted by:

$$\acute{Z} = Z + PE \tag{3}$$

where PE is the positional encoding matrix.

The position-encoded sequence is passed through a stack of six transformer encoder layers, each consisting of: a) multi-head self-attention with 8

attention heads and model dimension of 512, b) feed-forward network (FFN) with hidden size 2048 and ReLU, activation, c) Layer normalization, residual connections, and dropout (0.5) applied to both attention and FFN sub-blocks.

This stack enables the model to attend to dependencies across all 50 timesteps, capturing temporal correlations across the entire duration of motion. The final output of the transformer encoder is a sequence of shape 50×512. To reduce this to a fixed-size representation, we apply temporal mean pooling across the sequence dimension:

$$h = \frac{1}{T} \sum_{t=1}^{T} Z_t, \quad T = 50 \tag{4}$$

The pooled vector $h \in \mathbb{R}^{512}$ is then goes through a dropout layer (rate = 0.5) and a fully connected classification layer:

$$\hat{y} = Softmax(W.h + b) \tag{5}$$

where $W \in \mathbb{R}^{6 \times 512}$ and $\hat{y} \in \mathbb{R}^{6}$ contains the class probabilities for the six activity types. The proposed model is summarized in Fig. 1.

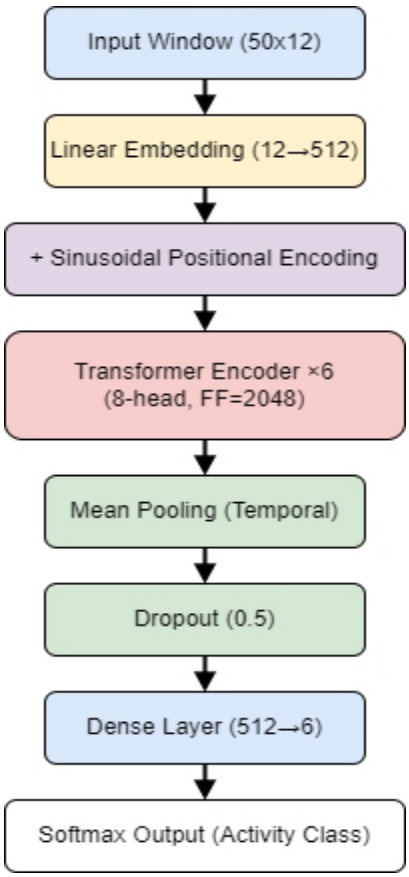

Figure 1: The proposed model.

Training Configuration

The proposed transformer model was trained end-to-end using supervised learning. The dataset was divided into training (80%) and validation (20%) sets using a stratified split to preserve class distribution. Each windowed sequence of 50 IMU samples (with 40% overlap) was treated as an independent instance and labeled according to the activity present at the start of the window. Labels were encoded into integer class indices using a standard label encoding strategy. The model was trained using the Adam optimizer with a learning rate of $1 \times 10\_4$. The loss function was categorical cross-entropy, suitable for multi-class classification tasks. Dropout (0.5) was applied after the transformer encoder and before the classification head to improve generalization. No additional regularization techniques (e.g., weight decay) were needed, as convergence was achieved with minimal overfitting.

Each training run was performed on a free-tier Kaggle GPU instance, which provided access to a single NVIDIA Tesla P100 with 16 GB of memory. The batch size was set to 32, and the model was trained for 40 epochs. Training typically completed within 10–12 minutes per run, with the validation loss stabilizing well before the final epoch.

Evaluation was conducted on the held-out validation set, using overall accuracy, precision, recall, F1-score (both macro- and weighted-average), and the confusion matrix. To mitigate the effect of natural class imbalance (especially in stair-related activities), we applied SMOTE oversampling to the training set only. Validation data remained untouched to ensure fair assessment of generalization.

RESULTS

The transformer model demonstrated strong classification performance across all six activity classes in the MotionDetection dataset. Training was stable and converged rapidly, with validation loss and accuracy curves indicating minimal overfitting. The model consistently reached its optimal performance within the first 10 epochs of the 40-epoch training schedule. On the held-out validation set, the model achieved an overall classification accuracy of 98.0%. The macro-averaged F1-score was 0.96, with a weighted F1-score of 0.98, confirming robust performance across both majority and minority classes. The precision, recall, and F1 metrics for each activity class are summarized in Table 1.

Table 1: Per-class precision, recall, and F1-score for the six activities in the MotionDetection dataset, based on predictions from the transformer model.

Activity	Precision	Recall	F1-Score
Downstairs	0.94	0.89	0.92
Upstairs	0.94	0.99	0.96
Walking	0.97	0.99	0.98
Jogging	0.99	1.00	1.00
Sitting	1.00	1.00	1.00
Standing	0.99	1.00	1.00

Minor misclassifications occurred primarily between downstairs and upstairs classes—an expected outcome due to the kinematic similarity between these two types of stair-based locomotion. In contrast, static and level-plane activities such as sitting, standing, walking, and jogging were recognized with near-perfect consistency. Training and validation loss curves remained closely aligned, with the validation loss stabilizing at approximately 0.11 and accuracy reaching a plateau near 98%. These results suggest the model effectively captured temporal and spatial patterns within the motion signals without overfitting.

Figures 2 and 3 illustrate the training curves and confusion matrix respectively.

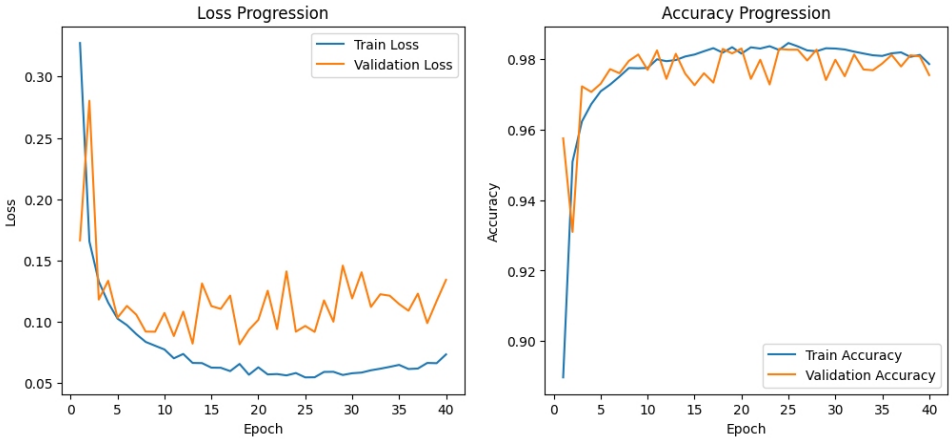

Figure 2: Training and validation loss and accuracy curves over 40 epochs. The model converges rapidly, with minimal overfitting, and validation accuracy stabilizing around 98%.

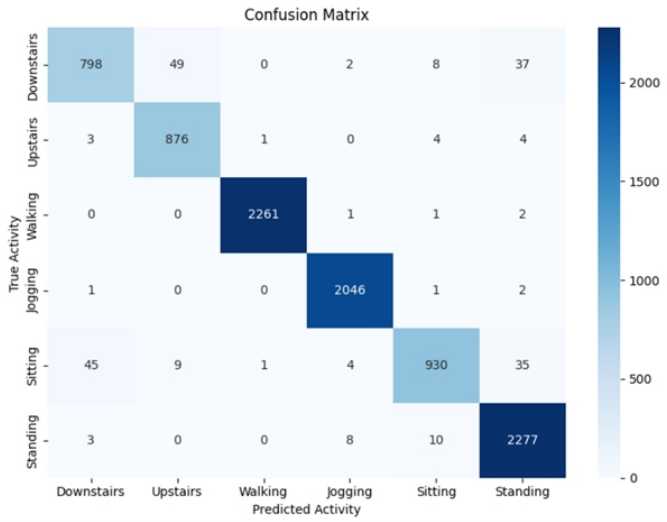

Figure 3: Confusion matrix of the transformer model on the validation set. Rows represent true activity classes and columns predicted classes.

DISCUSSION

The primary objective of this study was to investigate the effectiveness of transformer-based neural networks for smartphone-based human activity recognition. Our proposed transformer model demonstrated superior overall accuracy compared to two commonly used baseline architectures, CNN-BiLSTM and DeepConvLSTM, as clearly illustrated in the validation accuracy progression shown in Figure X. The progression plot highlights that the transformer model consistently outperformed the baseline methods across the key training epochs (10, 20, 30, and 40), maintaining the highest accuracy at every milestone. Although DeepConvLSTM exhibited a rapid improvement, closely approaching the transformer's performance at later epochs, it still lagged slightly behind, reflecting the stronger capacity of the transformer architecture to model temporal dependencies in multivariate IMU data. The CNN-BiLSTM showed steady improvement yet consistently underperformed compared to the transformer and DeepConvLSTM across all epochs.

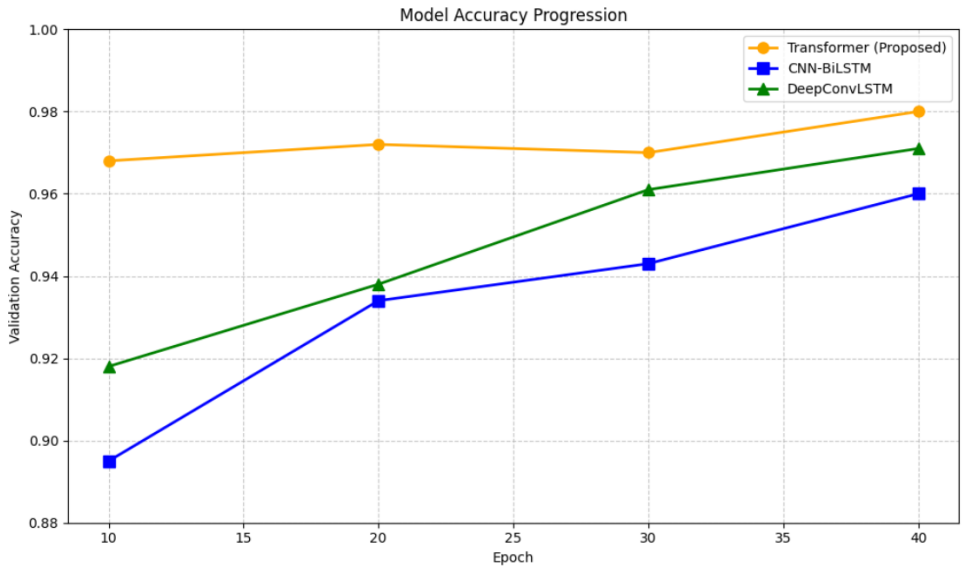

Figure 4: Comparison of validation accuracy progression at key epochs (10, 20, 30, 40) between the proposed transformer-based model and two baseline methods (CNN-BiLSTM and DeepConvLSTM).

The distinct advantage of the transformer in handling sequential IMU data is likely attributed to its self-attention mechanism, allowing simultaneous consideration of both short-range and long-range temporal dependencies without the drawbacks of recurrent architecture, such as gradient vanishing or explosion. Consequently, the transformer model maintained more stable and robust accuracy trends, converging rapidly and reliably, as depicted clearly by the higher and steadier accuracy curve. The observed accuracy plateau beyond epoch 30 suggests the transformer's quick convergence and effective learning capacity. Moreover, the slight narrowing of the accuracy

gap at epoch 40 between transformers and DeepConvLSTM indicates that deeper CNN-LSTM hybrid models can also achieve high performance, albeit with potentially higher computational complexity or slower convergence rates.

Overall, these results confirm the superiority of transformer architectures for HAR applications. Their inherent computational efficiency, as indicated by quicker convergence, and their robustness against overfitting, make transformers especially suited for mobile and resource-constrained environments.

CONCLUSION

In this study, we proposed a lightweight transformer-based neural network specifically tailored for smartphone-based human activity recognition (HAR) using raw inertial sensor signals. Our results indicate that transformer architecture significantly outperforms conventional deep learning approaches such as CNN-BiLSTM and DeepConvLSTM, achieving superior accuracy, faster convergence, and enhanced robustness against overfitting. The introduced MotionDetection dataset, collected under realistic conditions, provided a robust testbed for evaluating our model. With carefully implemented windowing and class-balancing techniques, our transformer consistently demonstrated reliable classification performance, particularly excelling in differentiating challenging activities like stair climbing.

Future research directions include exploring transformer generalization across diverse subject populations, investigating interpretability through attention visualization, and optimizing the model further for real-time, energy-efficient deployment on mobile devices. Overall, the findings of this study confirm the potential of transformer architectures as a powerful, scalable, and computationally efficient solution for practical HAR applications.

REFERENCES

Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R., & Havinga, P. (2010). "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," *Proceedings of the 23rd International Conference on Architecture of Computing Systems*.

Chen, W., Liu, C., & Shen, J. (2025). Two-Stream Graph Convolutional Transformer for Skeleton-Based Human Activity Recognition. *Scientific Reports*.

Djenouri, Y., & Belbachir, A. (2023). Convolutional Visual Transformer Network for Human Activity Recognition. *ICCV Workshops (ICCVW)*.

Hammerla, N. Y., Halloran, S., & Ploetz, T. (2016). "Deep, convolutional, and recurrent models for human activity recognition using wearables," *IJCAI*.

Hochreiter, S., & Schmidhuber, J. (1997). "Long short-term memory," *Neural Computation*.

Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*.

Lara, O. D., & Labrador, M. A. (2013). "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Communications Surveys & Tutorials*.

Mekruksavanich, S., & Jitpattanakul, A. (2023). Multi-channel CNN with Channel Attention for Sensor-Based Human Activity Recognition. *Scientific Reports*.

Ordóñez, F. J., & Roggen, D. (2016). "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*.

Pramanik, S., Dutta, S., & Tripathy, A. (2023). Transformer-Based Deep Reverse Attention Network for Sensor-Based Human Activity Recognition. *Engineering Applications of Artificial Intelligence*.

Raj, A., & Kos, A. (2023). An Improved CNN-Based Method for Human Activity Recognition Using Wearable Sensors. *Scientific Reports*.

Vaswani, A., et al. (2017). "Attention is All You Need," *NeurIPS*.

Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*.

Xing, X., Xu, C., Liu, X., et al. (2023). SVFormer: Semi-supervised Video Transformer for Human Activity Recognition. *Proceedings of CVPR 2023*.

Zhang, Y., Wu, H., Chen, L., & Liu, J. (2023). Attention-Based Residual BiLSTM Network for Human Activity Recognition. *IEEE Access*.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., & Xiong, H. (2021). "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *AAAI*.