International Journal of Innovative Information Systems & Technology Research 14(3):254-270, July-Sept, 2026

© SEAHI PUBLICATIONS, 2026 www.seahipublications.org ISSN: 2467-8562

DOI:10.5281/zenodo.22135609

A Hybrid 1D-CNN and Transformer Architecture for Diverse Human Activity Recognition

[1]Oluwole A. Ayegbusi, [2]Moses O. Onyesolu . [3]Ednah O. Aliyu. and [4]Olatunde D. Akinrolabu

[1]Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria
[2]Department of Software Engineering, Nnamdi Azikiwe University, Awka, Nigeria
[3]Department of Computer Science, Adekunle Ajasin University, Akungba-Akoko, Nigeria
[4]Department of Data Science. Adekunle Ajasin University, Akungba-Akoko, Nigeria

Corresponding Author: woleayegbusi@gmail.com

ABSTRACT

Human Activity Recognition (HAR) is crucial for healthcare, rehabilitation, and smart environments, but traditional approaches struggle to capture both local dependencies and long-range temporal relationships in sensor data. This research develops a hybrid deep learning architecture combining 1D Convolutional Neural Networks (CNNs) and Transformer-based encoders for temporal feature learning in HAR. Implemented using Python with TensorFlow/Keras, the system was trained on wearable sensor datasets comprising 18 distinct activities following a Data Science development lifecycle. The architecture uses CNN layers to extract localized spatial-temporal patterns, while a Transformer encoder with positional embedding models' global sequential dependencies. Batch normalization and dropout ensure stable convergence and prevent overfitting. Experimental evaluation shows the proposed CNN-Transformer hybrid achieved 100% and 99% accuracy in Class 4 and Class 10 activities, respectively, and over 96% overall accuracy with consistently high precision, recall, and F1-scores across activity classes, outperforming Transformer-only baseline. The model excels at both static activities (standing, sitting, walking) through effective local feature extraction and complex transition-based activities (walking upstairs/downstairs) by capturing long-range temporal dependencies. These findings demonstrate that the CNN-Transformer hybrid provides a robust, computationally efficient solution for real-time HAR with significant potential for healthcare monitoring, assistive technology, and intelligent context-aware systems.

Keywords: Human Activity Recognition, Wearable Sensors, Transformer, 1D Convolutional Neural Network, Temporal Feature Learning, Deep Learning

1. INTRODUCTION

Human Activity Recognition (HAR) is an interdisciplinary research field that draws on Artificial Intelligence (AI), Machine Learning (ML), pattern recognition, signal processing, and ubiquitous computing. It concerns the automatic detection and classification of human activities or behaviours from data acquired through sensing technologies. In sensor-based HAR, measurable signals associated with human movement and, where applicable, physiological or contextual information are processed and mapped to predefined activity classes. The overall objective is to transform low-level sensor observations into meaningful representations of human activities, enabling systems to recognize what a person is doing

and, in some applications, when and how an activity is performed (Lara & Labrador, 2013; Bulling et al., 2014; Kaur et al., 2024).

The fundamental principle of HAR can be understood as a sensing-and-inference process. First, sensors capture signals generated by human movement or interaction with the surrounding environment. These signals may originate from wearable and embedded sensors, such as accelerometers, gyroscopes, magnetometers, and physiological sensors, as well as from ambient and vision-based sensing technologies. The acquired signals are then preprocessed to reduce noise and other unwanted variations and are typically segmented into meaningful temporal windows. Relevant representations are subsequently obtained either through handcrafted feature extraction or automatically learned using deep learning. Finally, a recognition model analyses the resulting representations and assigns an activity label, such as walking, sitting, standing, running, or falling. This general processing pipeline is commonly described in HAR literature as an activity-recognition chain involving sensing, preprocessing, segmentation, feature representation, and classification, followed by evaluation of recognition performance (Bulling et al., 2014; Lara & Labrador, 2013).

An important concept in HAR is that human activities are temporal and hierarchical in nature. Simple movements or gestures can form actions, while sequences of actions can constitute more complex activities or interactions. Consequently, HAR is not limited to recognizing isolated movements; it can also involve the recognition of activities of daily living, human-object interactions, behavioural states, and complex sequences of actions. The sensing modality determines what aspect of human behaviour can be observed. For example, inertial sensors primarily capture kinematic changes associated with body motion, physiological sensors capture changes in bodily state, cameras capture visual appearance and motion, while ambient sensors can capture interactions between individuals and their surrounding environment (Bulling et al., 2014; Demrozi et al., 2020; Dang et al., 2020; Bian et al., 2022).

Based on the sensing mechanism, HAR systems are broadly categorized into vision-based and sensor-based approaches. Vision-based HAR uses cameras or related visual sensing technologies to infer activities from images or video sequences and can provide rich information about body posture, motion, and human interactions. However, continuous visual monitoring raises privacy concerns and may be affected by factors such as illumination, occlusion, camera viewpoint, and environmental conditions. Sensor-based HAR, in contrast, relies on wearable, smartphone, ambient, physiological, and other non-visual sensing technologies. Wearable inertial sensors, particularly accelerometers and gyroscopes, are widely used because they can continuously capture motion-related signals while offering relatively unobtrusive monitoring and, in many applications, greater privacy than camera-based approaches (Aggarwal & Ryoo, 2011; Chaquet et al., 2013; Lara & Labrador, 2013; Dang et al., 2020; Demrozi et al., 2020).

The development of HAR has progressed from handcrafted feature-based approaches to data-driven deep learning methods. Early sensor-based HAR systems generally extracted statistical, temporal, and frequency-domain features from sensor signals and employed conventional classifiers such as Support Vector Machines, K-Nearest Neighbors, Decision Trees, and Random Forests (Bao & Intille, 2004; Kwapisz et al., 2011; Lara & Labrador, 2013). Although effective for many well-defined activities, these approaches depend substantially on the quality of handcrafted features and may be limited in their ability to represent complex temporal dependencies. The emergence of deep learning enabled models to learn hierarchical representations directly from raw or minimally processed sensor signals. Convolutional Neural Networks (CNNs) are effective at learning local patterns, whereas Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are designed to model sequential dependencies (Ordóñez & Roggen, 2016; Hammerla et al., 2016). More recently, Transformer-based architectures have introduced self-attention mechanisms that enable models to capture relationships between representations at different temporal positions, leading to increasing interest in their application to HAR (Luptáková et al., 2022; Seenath & Dharmaraj, 2024).

The widespread availability of smartphones, smartwatches, fitness trackers, and other wearable devices equipped with embedded inertial sensors has further expanded the practical application of HAR. These

technologies enable the acquisition of motion-related signals for applications including healthcare and rehabilitation, elderly care, sports monitoring, smart environments, surveillance, and human–computer interaction (Lara & Labrador, 2013; Dentamaro et al., 2024).

Despite substantial progress, reliable HAR in real-world environments remains challenging. Recognition performance can be influenced by inter-subject variability, differences in sensor placement and orientation, activity similarity, temporal variability, sensor noise, class imbalance, limited labelled data, device heterogeneity, and limited cross-population generalization (Bulling et al., 2014; Dang et al., 2020). These challenges become particularly important when systems are expected to operate outside controlled laboratory conditions, where users perform activities naturally and sensor configurations and environmental conditions may vary. Recent research therefore increasingly focuses on robust representation learning, multimodal sensing, adaptive and personalized models, computationally efficient architectures, and approaches capable of generalizing across users, devices, and environments**.**

In this context, understanding the principles of sensing, data acquisition, representation, and recognition is essential for evaluating the evolution of HAR systems. In particular, the choice of sensor modality directly determines the type and quality of information available to a recognition model and consequently affects accuracy, robustness, privacy, computational requirements, and deployability. This motivates a systematic examination of HAR sensing technologies and their associated data acquisition strategies, ranging from wearable inertial sensors to physiological, ambient, acoustic, radio-frequency, and vision-based modalities.

2. Literature Review

2.1 Conceptual Review

The fundamental principle of HAR is to transform continuous observations of human movement or physiological behaviour into meaningful activity labels using computational learning algorithms. Depending on the sensing modality, HAR systems may exploit information from wearable inertial sensors, physiological sensors, cameras, ambient sensors, or radio-frequency devices. In wearable-sensor-based HAR, accelerometers and gyroscopes are particularly common because they capture changes in linear acceleration and angular velocity associated with body movement. The resulting multichannel signals provide temporal patterns that can be analysed to distinguish different human activities.

A typical HAR system consists of several interconnected stages: data acquisition, preprocessing, segmentation, feature representation or extraction, activity classification**,** and performance evaluation. During data acquisition, sensor signals are continuously collected from one or more sensing devices attached to or carried by a subject. Preprocessing is then applied to reduce noise, remove artefacts, normalize signal ranges, and prepare the data for subsequent analysis. Because human activities occur over time, continuous sensor streams are generally divided into temporal windows or segments, each of which is associated with an activity label. The segmented signals are subsequently transformed into informative representations from which a learning model can discriminate among activity classes. In conventional systems, this representation is obtained through manually designed features, whereas deep learning approaches can learn discriminative representations directly from sensor data.

The recognition process is further influenced by the nature and temporal characteristics of the activities being considered. Human activities can broadly be categorized as static, dynamic, and transitional activities. Static activities, such as standing and sitting, involve relatively limited body movement over a period of time. Dynamic activities, such as walking, running, or climbing stairs, produce more pronounced and repetitive motion patterns. Transitional activities, such as sit-to-stand, stand-to-sit, and lie-to-stand, represent changes between body states and are characterized by relatively short, variable, and non-stationary temporal patterns. Consequently, transitional activities can be more difficult to distinguish from neighbouring activities because their sensor signatures may contain overlapping characteristics and their durations can vary substantially across individuals.

The effectiveness of HAR therefore depends not only on the choice of classification algorithm but also on the ability of the learning model to capture the spatial, temporal, and contextual characteristics embedded

in sensor signals. Differences in users, sensor placement, movement execution, activity duration, and environmental conditions can produce substantial variations in the observed signals. These variations create challenges for developing HAR models that are accurate, generalizable, and computationally efficient.

2.2 Traditional Machine Learning Approaches

Traditional HAR systems generally follow a pipeline consisting of sensor data acquisition, signal preprocessing, handcrafted feature extraction, feature selection, and activity classification (Contoli et al. 2024; da silva et al.,2026). Popular machine learning algorithms such as Decision Trees, Support Vector Machines, Random Forests, Naïve Bayes, and Hidden Markov Models have achieved satisfactory performance for recognizing simple activities(Rabbi et al.,2021; Lohia et al.,2022.). However, these methods depend heavily on manually engineered features extracted from time-domain and frequency-domain representations of sensor signals. The effectiveness of such systems is therefore closely tied to domain expertise, making them less robust when applied to diverse users, sensor placements, and complex activity patterns (Trabelsi et al.,2022; Ciortuz et al.,2025).

2.2 Deep Learniung Learning Approach

Deep learning has revolutionized Human Activity Recognition by enabling automatic feature extraction from raw sensor data. Convolutional Neural Networks (CNNs) have been very effective in extracting local spatial representations from multivariate sensor data (Ige & Sibiya, 2024), while Recurrent Neural Networks (RNNs), in particular Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, have been widely used to model. (Ławryńczuk & Zarzycki, 2025).temporal dependencies in sequential activity data (Ławryńczuk & Zarzycki, 2025).

Hybrid CNN-RNN architectures have further improved recognition accuracy by combining spatial and temporal feature learning within a unified framework(Wang et al.,2020). Nevertheless, vanilla recurrent architectures process sequential information iteratively and often suffer from gradient degradation, increased computational cost, and limited ability to model long-range dependencies in lengthy activity sequences(Hochreiter & Schmidhuber, 1997 ; Pascanu et al., 2013). Likewise, CNNs are inherently limited in capturing global temporal context because of their localized receptive fields(Chibuike & Yang, 2024).

Transformer architectures have recently emerged as a promising alternative for sequence modelling. Originally introduced for natural language processing, Transformers employ self-attention mechanisms that allow every element in a sequence to directly interact with every other element regardless of temporal distance (Vaswani et al.,2017). This capability enables efficient modelling of long-range dependencies while supporting highly parallel computation during training(Shaikh et al., 2025). These characteristics make Transformers particularly suitable for sensor-based HAR, where activities often involve complex temporal relationships extending over long observation windows. Recent studies have reported encouraging results using Transformer models for activity recognition(Shavit & Klein, 2021 ; Lamaakal et al, 2025); however, many existing approaches employ computationally intensive architectures that are difficult to deploy on wearable and resource-constrained edge devices. Furthermore, relatively few studies have explored lightweight Transformer-guided frameworks capable of simultaneously capturing local motion characteristics and long-term temporal dependencies, particularly for transitional activities such as sit-to-stand, stand-to-sit, and lie-to-stand.

Transitional activities are difficult to handle due to their low frequency, short duration, and possess temporal properties that are remarkably different from static and dynamic activities (Li et al., 2019). These factors often lead to class imbalance and reduced recognition accuracy. Furthermore, many state-of-the-art deep learning frameworks include millions of trainable parameters, which makes them computationally expensive to deploy on wearable devices having limited processing, memory, and energy resources (Incel & Bursa, 2024). Consequently, there remains a need for lightweight architectures capable of achieving high recognition performance while maintaining computational efficiency.

Hence, this study proposes a Hybrid 1D CNN and Transformer Architecture for diverse Human Activity Recognition that combines the complementary strengths of one-dimensional Convolutional Neural

Networks and a lightweight Transformer encoder. The 1D-CNN extracts discriminative local features from multichannel inertial sensor signals, while the Transformer encoder captures long-range contextual dependencies using self-attention. The resulting feature representations are subsequently used for effective classification of static, dynamic, and transitional human activities. The proposed architecture is intentionally lightweight such that it can be easily deployed on wearable sensing platforms and edge computing environments without sacrificing recognition accuracy.

The main contributions of this work can be summarized in three folds. First, we propose a lightweight deep learning hybrid architecture by utilizing both a 1D-CNN layer and a Transformer encoder, which learns local and global representations simultaneously from wearable sensors' data. Secondly, we propose an approach for recognizing various types of human activities, i.e., static, dynamic, and transition activities, which have been a long-standing issue in HAR. Finally, the efficacy and robustness of our proposed approach will be tested with conventional evaluation metrics and compared to other deep learning methods to demonstrate its effectiveness, robustness, and computational efficiency. The proposed framework contributes toward the development of intelligent HAR systems suitable for next-generation healthcare monitoring, fitness tracking, rehabilitation support, and context-aware ubiquitous computing applications.

3. MATERIALS AND METHODS

3.1 Research Design

This study adopts an experimental research design to develop and evaluate a lightweight Transformer-guided deep learning framework for Human Activity Recognition (HAR) using wearable sensor data. The proposed framework combines the local feature extraction capability of a one-dimensional Convolutional Neural Network (1D-CNN) with the global temporal modeling ability of a Transformer encoder. The overall workflow comprises data acquisition, preprocessing, feature learning, activity classification, and performance evaluation.

3.2 Dataset Description

The experiments conducted in this study utilized the Khulna University Human Activity Recognition (KU-HAR) dataset, a publicly available benchmark dataset specifically developed for wearable sensor-based Human Activity Recognition (HAR). The dataset was introduced by Sikder et al. (2021) to address the need for a comprehensive dataset containing both static and dynamic human activities collected under realistic conditions.

The KU-HAR dataset was collected from 90 healthy volunteers (75 males and 15 females) aged between 18 and 34 years using the built-in triaxial accelerometer and triaxial gyroscope sensors of a smartphone. The smartphone was attached to the participants' waist during data acquisition, enabling the collection of inertial signals representing body movements. A total of 1,945 raw activity recordings were obtained, from which 20,750 non-overlapping time-domain subsamples of 3 seconds each were extracted for machine learning and deep learning applications.

The dataset consists of 18 human activity classes, representing a mixture of daily living activities, transitional movements, locomotion, and exercise-related activities. These activities include:

1. Stand
2. Sit
3. Talk-Sit
4. Talk-Stand
5. Stand-to-Sit
6. Lay
7. Lay-to-Stand
8. Pick
9. Jump
10. Push-up
11. Sit-up

12. Walk
13. Walk Backward
14. Walk Circle
15. Run
16. Stair Up
17. Stair Down
18. Table Tennis

These activities were intentionally selected to include both simple and complex movement patterns, making the dataset suitable for evaluating robust Human Activity Recognition algorithms.

Each subsample contains synchronized inertial measurements from the smartphone's accelerometer and gyroscope. In this study, only the six raw sensor channels were utilized:

- Accelerometer X-axis
- Accelerometer Y-axis
- Accelerometer Z-axis
- Gyroscope X-axis
- Gyroscope Y-axis
- Gyroscope Z-axis

Each activity window contains 300-time samples per sensor axis, corresponding to approximately 3 seconds of recording at a sampling frequency of about 100 Hz. Consequently, every sample was represented as a matrix of 300 × 6, where 300 denotes the temporal sequence length and 6 represents the six inertial sensor channels. This input representation preserves the temporal dependencies required for deep learning models, particularly Transformer-based architectures.

For model development, the dataset was organized into an input tensor of shape (20,750, 300, 6), where each sample consists of a temporal sequence of 300 observations across six sensor channels. The activity labels were encoded into 18 categorical classes using one-hot encoding before training the proposed model. The dataset was then partitioned into training and testing subsets using an 80:20 split, resulting in approximately 16,600 training samples and 4,150 testing samples.

The KU-HAR dataset was selected because it contains a diverse range of static, dynamic, and transitional activities that require effective temporal feature extraction. Such characteristics make it highly suitable for evaluating the proposed Transformer-Guided Model for Temporal Feature Learning, whose attention mechanism is designed to capture long-range temporal dependencies that conventional CNN-based models may overlook.

Dataset Specification

Parameter	Description
Dataset	KU-HAR (Khulna University Human Activity Recognition)
Participants	90 (75 males, 15 females)
Sensors	Triaxial Accelerometer + Triaxial Gyroscope
Sensor Channels	6
Total Activities	18
Raw Recordings	1,945
Time-domain Subsamples	20,750
Window Length	3 seconds
Sampling Frequency	Approximately 100 Hz
Samples per Window	300
Input Shape	(300, 6)
Number of Classes	18

Parameter	Description
Train/Test Split	80% / 20%
Learning Task	Multi-class Human Activity Recognition

3.3 Data Preprocessing

The KU-HAR dataset was preprocessed to prepare it for training the proposed Transformer-guided Human Activity Recognition model. The dataset comprises 20,750 time-domain samples, each represented by 300-time steps and six sensor channels (three-axis accelerometer and three-axis gyroscope). The raw sensor signals were reshaped into an input tensor of 300 × 6, while the activity labels were converted into 18 one-hot encoded classes for multi-class classification. The dataset was then randomly divided into 80% training and 20% testing subsets. No handcrafted feature extraction was performed, as the proposed hybrid 1D CNN-Transformer architecture was designed to automatically learn both local spatial features and long-range temporal dependencies directly from the raw inertial sensor data.

3.4 Proposed Hybrid 1D-CNN–Transformer Architecture

The proposed Human Activity Recognition (HAR) framework integrates a lightweight one-dimensional Convolutional Neural Network (1D-CNN) with a Transformer encoder to jointly learn local motion characteristics and long-range temporal dependencies from wearable inertial sensor signals. As illustrated in Figure **1**, the implementation flow consists of five principal stages: **(i)** sensor input, **(ii)** CNN-based feature extraction, **(iii)** positional encoding, **(iv)** Transformer-based temporal learning, and **(v)** activity classification.

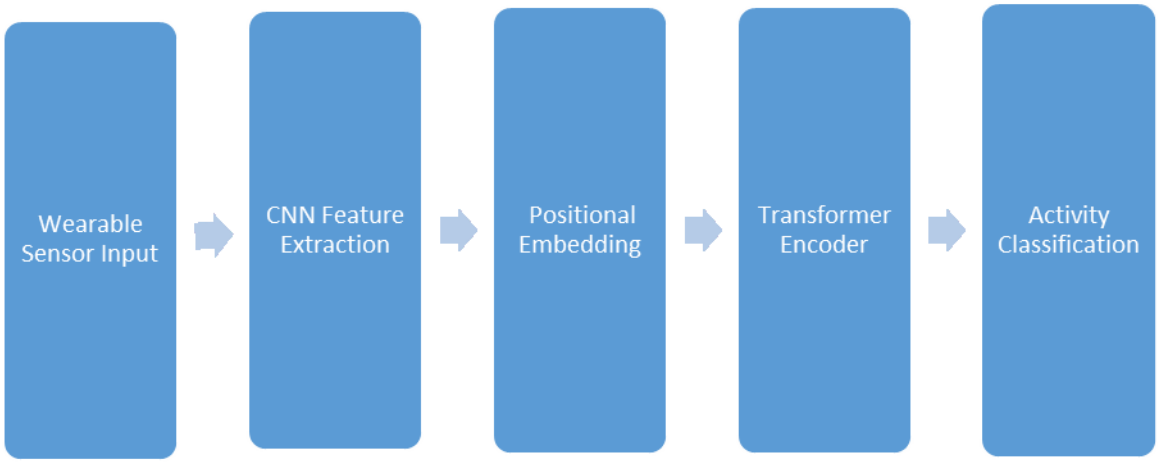

Figure 1: Implementation flow of the Proposed System

3.4.1 Wearable Sensor Input

Each activity sample is represented as a multivariate time-series sequence containing synchronized accelerometer and gyroscope measurements. The input to the network is as shown in equation 1 is expressed as

$$X \in R^{TxF} \quad (1)$$

where T denotes the number of temporal observations and F represents the number of sensor channels. For the KU-HAR dataset, T = 300 and F = 6, corresponding to the three-axis accelerometer and three-axis gyroscope signals.

3.4.2 CNN-Based Local Feature Extraction

The first stage of the proposed architecture employs a lightweight 1D-CNN to extract discriminative local motion features directly from the raw sensor signals. Two successive convolutional layers with Rectified Linear Unit (ReLU) activation are used to learn short-term temporal patterns associated with human movements. Max-pooling layers are incorporated after each convolution to reduce the temporal dimension, suppress redundant information, and improve computational efficiency.

The output of the CNN block as shown in equation 2 is represented as

$$H \in R^{T' xD} \quad (2)$$

where T′ denotes the reduced temporal dimension after pooling and D represents the learned feature dimension. Compared with the original sensor sequence, this representation contains higher-level motion descriptors that are more suitable for temporal modeling.

3.4.3 Positional Encoding

Since the Transformer architecture does not inherently preserve sequential order, positional encoding is added to the extracted CNN features to retain temporal information. The positional encoding as shown in equation 3 is computed as

$$PE_{(t,2i)} = \sin\left(\frac{t}{10000^{2i/d}}\right), p_{(t,2i+1)} = \cos\left(\frac{t}{10000^{2i/d}}\right) \quad (3)$$

The encoded feature representation in equation 4 is obtained by

$$z = H + PE \quad (4)$$

where PE denotes the positional encoding matrix. This operation enables the Transformer encoder to distinguish the temporal ordering of sensor observations, thereby improving recognition of activities with similar motion characteristics, particularly transitional movements.

3.4.4 Transformer-Based Temporal Learning

The encoded features are subsequently processed by a lightweight Transformer encoder consisting of a Multi-Head Self-Attention (MHSA) module, residual connections, layer normalization, and a position-wise feed-forward network.

The self-attention mechanism enables each time step to interact with every other time step in the sequence, allowing the network to capture long-range temporal dependencies that conventional CNNs may overlook. Residual connections and layer normalization improve optimization stability and gradient propagation, while the feed-forward network enhances the discriminative capability of the learned representations.

Unlike recurrent architectures that process sequential data iteratively, the Transformer performs parallel sequence modeling, resulting in improved computational efficiency and better representation of complex activity patterns.

3.4.5. Activity Classification

The output of the Transformer encoder is converted into a fixed-length representation using Global Average Pooling (GAP), which significantly reduces the number of trainable parameters compared with conventional flattening operations. The pooled feature vector is then passed through a fully connected layer with ReLU activation and Dropout regularization to improve generalization.

Finally, activity probabilities as shown in equation 5 are computed using the SoftMax classifier,

$$\ddot{y} = \text{softmax}(Wz + b) \quad (5)$$

where **W** and **b** denote the trainable weight matrix and bias vector, respectively, and **z** represents the pooled feature vector. The predicted activity corresponds to the class with the highest probability among the 18 activity categories.

Overall, the proposed hybrid architecture effectively combines the strengths of convolutional feature extraction and Transformer-based temporal modeling. The CNN efficiently captures local motion characteristics, while the Transformer encoder models global contextual relationships through self-attention. This complementary design enables accurate recognition of static, dynamic, and transitional human activities while maintaining a lightweight architecture suitable for wearable and edge computing applications.

3.4.6 Architectural Configuration and Parameterization

Table 1. Layer-by-layer architectural configuration of the proposed lightweight 1D-CNN–Transformer Model.

Layer	Configuration	Otput Shape	Parameters
Input	300 time steps × 6 channels	300 × 6	0
Conv1D	64 filters, kernel = 5, padding = same	300 × 64	1,984
MaxPooling1D	Pool size = 2	150 × 64	0
Conv1D	128 filters, kernel = 3, padding = same	150 × 128	24,704
MaxPooling1D	Pool size = 2	75 × 128	0
Positional Embedding	Embedding dimension = 128 (learned)	75 × 128	26,112
Layer Normalization	Normalization over 128 features	75 × 128	256
Multi-Head Attention	4 heads	75 × 128	131,968
Residual add	Attention + normalized input	75 × 128	0
Layer Normalization	Normalization over 128 features	75 × 128	256
Dense	128 Units	75 × 128	16,512
Drop out	0.10	75 × 128	0
Dense	128 Units	75 × 128	16,512
Residual Add	Feed-forward + residual path	75 × 128	0
Global Average Pooling 1D	Average over 75 time steps	128	0
Dense	128 Units	75 × 128	16,512
Drop out	0.30	128	0

3.5 Performance Evaluation

The developed model was evaluated using commonly adopted classification metrics:

- Accuracy
- Precision
- Recall
- F1-score

- Confusion Matrix

Overall classification accuracy as shown in equation 6 is computed as

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Precision as shown in equation 7 is given by

$$\frac{TP}{TP+FP} \quad (7)$$

Recall as shown in equation 8 is computed as $\frac{TP}{TP+FN}$ (8)

while the F1-score as shown in equation 9 is computed as

$$F_1 = 2\,\frac{Precision * Recall}{Precision + Recall} \quad (9)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

These evaluation metrics provide a comprehensive assessment of the proposed model's ability to correctly recognize static, dynamic, and transitional human activities.

3.6 Experimental Environment

The proposed framework was implemented using the Python programming language with deep learning libraries such as TensorFlow/Keras. Model training and evaluation were conducted on a workstation equipped with high-speed processor to reduce computational time. Hyperparameters, including learning rate, batch size, number of epochs, kernel size, and dropout rate, were selected experimentally to achieve the best trade-off between recognition accuracy and computational efficiency.

4. RESULTS AND DISCUSSION

4.1 Experimental Results

The proposed hybrid 1D Convolutional Neural Network (1D-CNN) and Transformer model was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis. Model training was performed using the Adam optimizer with a learning rate of 0.001, a batch size of 64, and categorical cross-entropy loss for a maximum of 30 epochs. Early stopping and learning rate scheduling were employed to improve convergence and reduce overfitting.

The model achieved an overall classification accuracy of **96%**, with both macro-average and weighted-average F1-scores of **0.96**, demonstrating excellent recognition performance across the eighteen activity classes. Table 2 summarizes the overall performance of the proposed model.

Table 2. Overall Performance of the Proposed Model

Metric	Value
Accuracy	96%
Precision	96%
Recall	96%
Macro F1-score	96%
Best Performing Class	Class 4 (F1 = 1.00)
Lowest F1-score	Class 13 (F1 = 0.92)

The consistently high evaluation metrics indicate that the proposed architecture effectively learns discriminative representations from multivariate inertial sensor data while maintaining balanced classification performance across different activity categories.

4.2 Class-wise Performance Analysis

Table 3 presents the class-wise precision, recall, and F1-score obtained by the proposed model.

Class label	Precision	Recall	F1-Score	Support
0	0.91	0.94	0.92	327
1	0.93	0.91	0.92	399
2	0.94	0.95	0.94	381
3	0.96	0.99	0.98	340
4	1.00	1.00	1.00	443
5	0.97	0.94	0.95	359
6	0.99	0.97	0.98	326
7	0.97	0.98	0.98	260
8	0.98	0.99	0.98	134
9	0.98	0.98	0.98	102
10	0.99	0.99	0.99	218
11	0.95	0.96	0.96	163
12	0.94	0.93	0.93	67
13	0.92	0.92	0.92	62
14	0.99	0.98	0.99	140
15	0.98	0.92	0.95	157
16	0.96	0.96	0.96	175
17	0.96	0.97	0.96	97

Class 4 achieved perfect classification performance, recording precision, recall, and F1-score values of **1.00**, indicating that no samples belonging to this activity were misclassified. This suggests that the activity possesses highly distinctive motion characteristics that were effectively captured by the proposed hybrid architecture.

Similarly, Classes 3, 6, 7, 8, 9, 10, and 14 achieved F1-scores between **0.98 and 0.99**, demonstrating the ability of the model to accurately distinguish both static and dynamic activities. These results indicate that the integration of convolutional feature extraction with Transformer-based temporal modeling enables the network to learn highly discriminative representations.

Relatively lower F1-scores were observed for Classes 0, 1, 12, and 13, with values ranging between **0.92 and 0.94**. These activities are likely characterized by similar movement patterns or transitional motions, increasing the likelihood of confusion between related classes. Nevertheless, the achieved performance remains high, indicating strong generalization even for more challenging activity categories. The summary table of the class-wise performance analysis is shown in Table 4

Table 4: Summary Table: Class-Wise Performance Analysis.

Class- Performance Range	Activity Classes
F1-score = 1.00	Class 4
F1-score ≥ 0.98	Classes 3, 6, 7, 8, 9, 10 and 14
F1-score = 0.95–0.97	Classes 5, 11, 15, 16 and 17
F1-score = 0.92–0.94	Classes 0, 1, 2, 12 and 13

4.3 Precision–Recall Analysis

The proposed model exhibits a close agreement between precision and recall across all activity classes, with both metrics averaging **96%**. This balance indicates that the classifier simultaneously minimizes false-positive and false-negative predictions.

High precision demonstrates that predicted activities are generally correct, while high recall confirms that the majority of true activity instances are successfully identified. Such balanced performance is particularly important in Human Activity Recognition applications involving healthcare monitoring, assisted living, and rehabilitation, where missed or incorrect activity recognition may adversely affect decision-making processes.

4.4 Training Performance

The learning curves demonstrate stable optimization throughout the training process. Training accuracy increased rapidly during the initial epochs before gradually converging toward **100%**, while validation accuracy stabilized at approximately **96%**. The relatively small gap between the training and validation curves indicates good generalization capability and minimal overfitting.

Similarly, both training and validation losses decreased consistently with increasing epochs before reaching stable convergence. The smooth decline in validation loss suggests that the proposed model successfully learned meaningful temporal representations without exhibiting unstable learning behaviour.

4.5 Confusion Matrix Analysis

The confusion matrix in Figure 2 further confirms the effectiveness of the proposed model. Most samples are concentrated along the principal diagonal, indicating correct classification across the eighteen activity classes. Only a small number of misclassifications occur between activities exhibiting similar movement characteristics, particularly among transitional activities.

The limited off-diagonal entries demonstrate that the hybrid architecture effectively separates activity classes despite similarities in motion patterns. These findings further validate the robustness of the proposed Transformer-guided temporal feature learning framework.

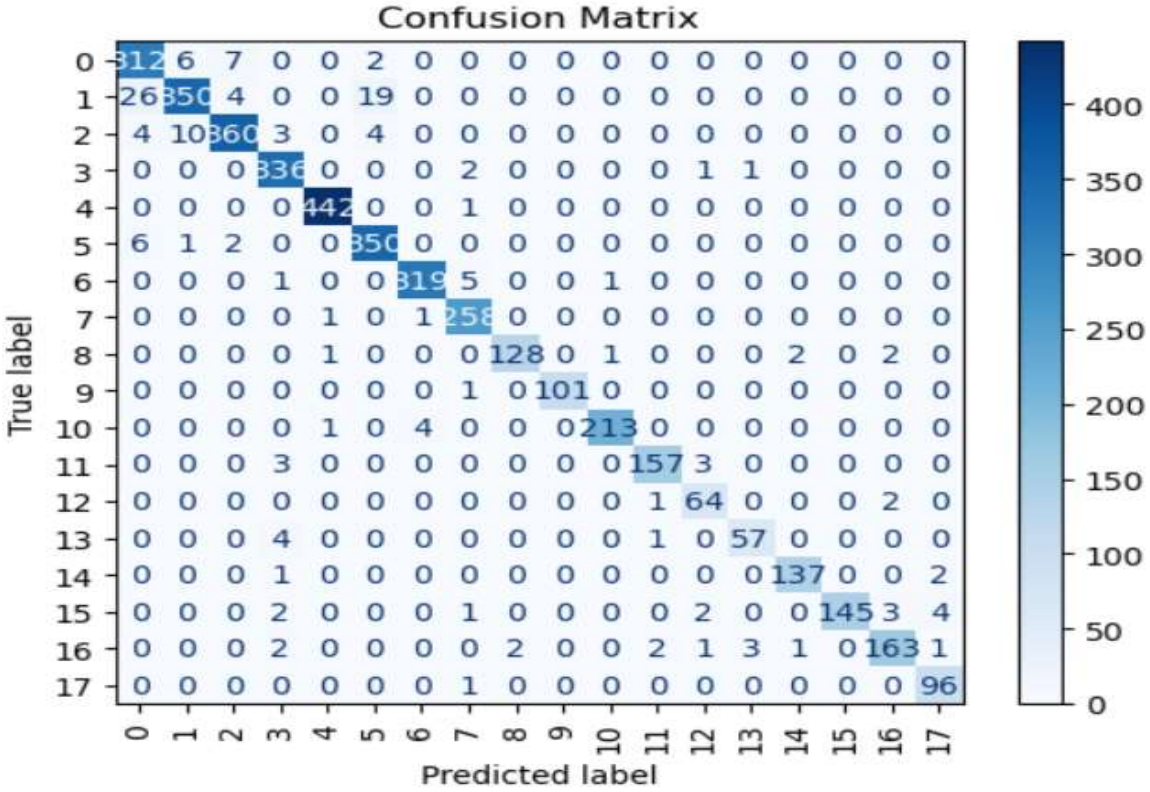

Figure 2: Confusion Matrix

4.6 Discussion

The experimental results demonstrate that combining a lightweight 1D-CNN with a Transformer encoder significantly improves Human Activity Recognition performance by jointly learning local motion features and long-range temporal dependencies. The convolutional layers effectively extract low-level spatial representations from raw accelerometer and gyroscope signals, while the Transformer encoder models contextual relationships within the temporal sequence through self-attention.

The achieved **96% classification accuracy** and **0.96 macro F1-score** compare favorably with several existing deep learning approaches reported in the HAR literature, confirming the competitiveness of the proposed framework. Unlike conventional CNN models that primarily capture local dependencies or recurrent models that process sequences sequentially, the proposed architecture benefits from the Transformer's ability to model global temporal relationships while maintaining computational efficiency.

Furthermore, the consistently high F1-scores across all eighteen activity classes indicate strong robustness to variations in movement patterns and class imbalance. The model successfully recognizes static, dynamic, and transitional activities with high reliability, making it suitable for practical wearable sensor applications, including healthcare monitoring, elderly assistance, rehabilitation, fitness tracking, and smart environments.

Overall, the experimental findings validate the effectiveness of the proposed Transformer-guided temporal feature learning model as an accurate, lightweight, and computationally efficient solution for wearable Human Activity Recognition.

4.7 Performance Comparison of the Proposed Model with existing methods on the KU-HAR Dataset

Table 5. Comparison of the proposed 1D-CNN–Transformer hybrid with existing work evaluated on the KU-HAR dataset.

Model/Authors	Accuracy	Precision	Recall	F1-Score	Params
Random Forest (Sikder et al., 2021) – original KU-HAR baseline	89.67%	-	-	-	NR
Aquila Optimizer + RF (Al-(Qaness et al., 2022)	88.53%	-	-	-	NR
GJO-XGB (metaheuristic feature selection) (Sarker et al.,2025)	93.55%	-	-	-	NR
WARSO-XGB (Sarker et al.,2025)	94.04%	93.47%	92.40%	92.88%	NR
Trnsformer-Only(Guo et al., 2024)	83.2 %	-	-	87.4%	NR
Transformer +vector-based Relative Position Embedding + Convolutional Feature Extractor Block (vRPE +CFEB) (Guo et al., 2024)	96.80%	97.50%	97.51%	97.50%	NR
1D-CNN–Transformer hybrid (Proposed)	**96.00%**	**96.00%**	**96.00%**	**96.00%**	**237,138**
ResLSTM (AlMuhaideb et al., 2024)	97.05%	-	-	-	386,038
CWT + DenseNet121 (Pavliuk et al. 2023)	97.48%	-	-	97.52%	NR
BiLSTM (AlMuhaideb et al., 2024)	98.63%	-	-	-	569,462

4.7.1 Discussion

Discussion

Table 5 compares the performance of the proposed lightweight 1D-CNN–Transformer hybrid with existing methods evaluated on the KU-HAR dataset. The proposed model achieved an overall accuracy, precision, recall, and F1-score of 96.00%, demonstrating that the combination of convolutional feature extraction and Transformer-based temporal modelling can effectively recognize human activities from wearable inertial sensor data.

The proposed model achieved substantially higher accuracy than the original Random Forest KU-HAR baseline of 89.67% and the Aquila Optimizer–Random Forest approach of 88.53%. It also outperformed the GJO-XGB and WARSO-XGB approaches, which reported accuracies of 93.55% and 94.04%, respectively. These results indicate that the proposed deep learning architecture can learn discriminative temporal representations directly from the multichannel sensor sequences without relying primarily on conventional machine-learning classifiers.

The results also demonstrate the benefit of incorporating both convolutional and Transformer-based components. The Transformer-only model reported by Guo et al. (2024) achieved 83.20% accuracy, whereas the Transformer architecture incorporating a vector-based Relative Position Embedding and Convolutional Feature Extractor Block (vRPE + CFEB) achieved 96.80% accuracy. The improvement suggests that convolutional feature extraction can provide useful local representations before or alongside Transformer-based temporal modelling. The proposed model follows a similar general principle by using a 1D-CNN frontend to extract local temporal features before applying a Transformer encoder to model temporal dependencies.

Although the proposed model does not obtain the highest accuracy in Table 5, its main advantage is its model compactness. The proposed architecture contains only 237,138 trainable parameters, compared with 386,038 parameters for ResLSTM and 569,462 parameters for BiLSTM. Thus, the proposed model uses approximately 38.6% fewer parameters than ResLSTM and 58.4% fewer parameters than BiLSTM, while still achieving 96.00% accuracy. This reduction in parameter count is important for the intended lightweight HAR application, particularly where computational resources, memory, and energy consumption may be constrained.

The CWT + DenseNet121 and BiLSTM approaches reported higher accuracy values of 97.48% and 98.63%, respectively. Similarly, the vRPE + CFEB model achieved 96.80% accuracy, slightly above the proposed model. However, these higher-performing approaches should not be interpreted as universally superior because the reported results originate from different experimental implementations and their parameter counts are not consistently available. Consequently, direct comparison based solely on accuracy may not fully reflect differences in computational complexity and deployment suitability.

Overall, the results demonstrate that the proposed 1D-CNN–Transformer hybrid provides a competitive balance between recognition performance and model complexity. Its 96.00% accuracy, combined with only 237,138 parameters, supports the objective of developing a lightweight deep learning architecture for wearable-sensor-based HAR. Therefore, the contribution of the proposed model is not necessarily achieving the absolute highest accuracy, but rather achieving strong and competitive recognition performance while maintaining a relatively small parameter footprint suitable for resource-constrained applications.

Note: NR = not reported in the original study; "–" = metric not reported. Reported results are reproduced from the respective studies and are presented for comparative purposes. Differences in experimental protocols, data partitioning, preprocessing, and evaluation procedures should be considered when interpreting the comparison.

5. CONCLUSION

This study presented a lightweight Transformer-guided temporal feature learning framework for Human Activity Recognition (HAR) using wearable inertial sensor data. The proposed architecture integrates a one-dimensional Convolutional Neural Network (1D-CNN) for extracting local motion features with a

Transformer encoder for modeling long-range temporal dependencies through self-attention. This hybrid design effectively exploits the complementary strengths of convolutional and attention-based learning, resulting in improved recognition performance while maintaining computational efficiency.

Experimental evaluation demonstrated that the proposed model achieved an overall classification accuracy of 96%, with corresponding precision, recall, macro F1-score, and weighted F1-score of 96%. The class-wise analysis further revealed excellent recognition capability across all eighteen activity classes, with several activities attaining F1-scores between 0.98 and 1.00, while the lowest recorded F1-score remained high at 0.92. In particular, Class 4 achieved perfect classification performance, indicating that the proposed architecture effectively learned highly discriminative temporal and spatial representations from wearable sensor signals.

The training and validation curves demonstrated stable convergence with minimal overfitting, confirming the effectiveness of the adopted optimization strategy, regularization techniques, and model architecture. Furthermore, the confusion matrix showed that the majority of samples were correctly classified, with only a few misclassifications occurring between activities exhibiting similar movement characteristics. These findings demonstrate the robustness and generalization capability of the proposed framework.

Overall, the results confirm that the proposed Transformer-guided temporal feature learning model provides an accurate, reliable, and computationally efficient solution for wearable Human Activity Recognition. The model is therefore suitable for real-time applications in healthcare monitoring, elderly care, rehabilitation, fitness tracking, and other context-aware intelligent systems where accurate recognition of human activities is essential.

5. Future Work

Although the proposed Transformer-guided framework demonstrated excellent performance, several opportunities remain for further improvement.

Future research should investigate lightweight Transformer variants specifically designed for resource-constrained wearable devices to improve computational efficiency without compromising recognition accuracy. Additionally, incorporating self-supervised and transfer learning techniques may reduce the dependence on large labeled datasets while improving model generalization across different users and sensor configurations.

Future studies should also focus on improving the recognition of transitional activities, which remain among the most challenging categories because of their short duration and similarity to adjacent activities. Attention mechanisms specifically designed for transition modeling may further enhance recognition performance in such cases.

Another promising direction involves developing personalized HAR models capable of adapting to individual movement characteristics through online or continual learning. Such adaptive systems would improve robustness to inter-subject variability while maintaining long-term performance.

Finally, integrating the proposed model into Internet of Things (IoT) and edge-based healthcare environments would facilitate continuous, low-latency activity monitoring for smart homes, assisted living, rehabilitation, and other pervasive healthcare applications.

REFERENCES

Aggarwal, J.K., & Ryoo, M.S. (2011). Human activity analysis. *ACM Computing Surveys (CSUR), 43*, 1 - 43.

AlMuhaideb, S., AlAbdulkarim, L., AlShahrani, D. M., AlDhubaib, H., & AlSadoun, D. E. (2024). Achieving More with Less: A Lightweight Deep Learning Solution for Advanced Human Activity Recognition (HAR). *Sensors*, *24*(16), 5436. https://doi.org/10.3390/s24165436

Al-Qaness, M. A. A., Helmi, A. M., Dahou, A., & Elaziz, M. A. (2022). The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis. *Biosensors, 12*(10), 821. https://doi.org/10.3390/bios12100821

Bao, L., & Intille, S. S. (2004). Activity recognition from user-annotated acceleration data. In International Conference on Pervasive Computing (pp. 1–17). Springer.

Bian, S., Liu, M., Zhou, B., & Lukowicz, P. (2022). The State-of-the-Art Sensing Techniques in Human Activity Recognition: A Survey. *Sensors*, *22*(12), 4596. https://doi.org/10.3390/s22124596

Ciortuz, G., Hozhabr Pour, H., Irshad, M. T., Nisar, M. A., Huang, X., & Fudickar, S. (2025). Machine learning models for wearable-based human activity recognition: A comparative study. *Neurocomputing*, *650*, Article 130911. https://doi.org/10.1016/j.neucom.2025.130911.

Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, *117*(6), 633–659. https://doi.org/10.1016/j.cviu.2013.01.013

Chibuike, O., & Yang, X. (2024). Convolutional Neural Network-Vision Transformer Architecture with Gated Control Mechanism and Multi-Scale Fusion for Enhanced Pulmonary Disease Classification. *Diagnostics (Basel, Switzerland)*, *14*(24), 2790. https://doi.org/10.3390/diagnostics14242790

Contoli, C., Freschi, V., & Lattanzi, E. (2024). Energy-aware human activity recognition for wearable devices: A comprehensive review. *Pervasive and Mobile Computing*, *104*, Article 101976. https://doi.org/10.1016/j.pmcj.2024.101976

Da Silva, T. B. C., da Cunha Santiago, H., & Fernandes, B. J. T. (2026). From controlled benchmarks to unconstrained environments: A pipeline-oriented survey of models, domains, and emerging trends on human activity recognition. *Artificial Intelligence Review*, *59*(5), Article 170. https://doi.org/10.1007/s10462-026-11619-1

Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition, 108*, 107561. https://doi.org/10.1016/j.patcog.2020.107561

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

Guo, X., Kim, Y., Ning, X., & Min, S. D. (2025). Enhancing the Transformer Model with a Convolutional Feature Extractor Block and Vector-Based Relative Position Embedding for Human Activity Recognition. *Sensors*, *25*(2), 301. https://doi.org/10.3390/s25020301

Hammerla, N. Y., Halloran, S., & Ploetz, T. (2016). Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv*. https://arxiv.org/abs/1604.08880

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. doi.org

Ige, A. O., & Sibiya, M. (2024). State-of-the-art in 1D convolutional neural networks: A survey. *IEEE Access, 10*, 109447–109465. https://doi.org/10.1109/ACCESS.2022.3214512

Incel, D. O., & Bursa, S. Ö. (2023). On-device deep learning for mobile and wearable sensing applications: A review. *IEEE Sensors Journal*, *23*(6), 5501–5512. doi.org

Kaur, H., Rani, V., & Kumar, M. (2024). Human activity recognition: A comprehensive review. *Expert Systems*, 41(11), e13680. https://doi.org/10.1111/exsy.13680

Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, *12*(2), 74–82. https://doi.org/10.1145/1964897.1964918

Lamaakal, I., Yahyati, C., Maleh, Y., El Makkaoui, K., Ouahbi, I., & Abd El-Latif, A. A. (2025). A tiny inertial transformer for human activity recognition via multimodal knowledge distillation and explainable AI. *Scientific Reports*, *15*(1), Article 42335. https://doi.org/10.1038/s41598-025-26297-2

Ławryńczuk, M., & Zarzycki, K. (2025). LSTM and GRU type recurrent neural networks in model predictive control: A review. *Neurocomputing, 632*, 129712. https://doi.org/10.1016/j.neucom.2025.129712

Luptáková, I. D., Kubovčík, M., & Pospíchal, J. (2022). Wearable sensor-based human activity recognition with Transformer model. *Sensors, 22*(5), 1911. https://doi.org/10.3390/s22051911

Li, J., Tian, L., Wang, H., An, Y., Wang, K., & Yu, L. (2019). Segmentation and recognition of basic and transitional activities for continuous physical human activity. *IEEE Access, 7*, 42565–42576. https://doi.org/10.1109/ACCESS.2019.290557

Lohia, H. K., Dari, S. K., Kalwal, S., Singh, R., & Donald, A. D. G. (2022). Human activity recognition using machine learning techniques. *Journal of Emerging Technologies and Innovative Research, 9*(6), f619–f624. https://www.jetir.org/papers/JETIR2206579.pdf

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)* (pp. 1310–1318). PMLR. https://proceedings.mlr.press/v28/pascanu13.html

Pavliuk, O., Mishchuk, M., & Strauss, C. (2023). Transfer learning approach for human activity recognition based on continuous wavelet transform. *Algorithms, 16*(2), 77. https://doi.org/10.3390/a16020077

Rabbi, J., Fuad, M. T. H., & Awal, M. A. (2021). Human activity analysis and recognition from smartphones using machine learning techniques. *arXiv*. https://doi.org/10.48550/arXiv.2103.16490.

Sarker, P., Tiang, J.-J., & Nahid, A.-A. (2025). Metaheuristic-driven feature selection for human activity recognition on KU-HAR dataset using XGBoost classifier. *Sensors, 25*(17), 5303. https://doi.org/10.3390/s25175303

Shaikh, M. R., Shahryar, N., Mahboob, K., Naiyar, Q., & Talha, M. (2025). Transformers as the foundation of large language models: A comprehensive review. *International Journal of Information System and Technology*, *7*(4), 2705–2717. https://journal.50sea.com/index.php/IJIST/article/view/1632

Shavit, Y., & Klein, I. (2021). Boosting inertial-based human activity recognition with transformers. *IEEE Access*, *9*, 53540–53547. doi.org. For more information, visit IEEE Access. [1]

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

Trabelsi, I., Françoise, J., & Bellik, Y. (2022). Sensor-based activity recognition using deep learning: A comparative study. In *Proceedings of the 8th International Conference on Movement and Computing (MOCO '22)* (Article 25, pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3537972.3537996

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Wang, H., Zhao, J., Li, J., Tian, L., Tu, P., Cao, T., An, Y., Wang, K., & Li, S. (2020). Wearable Sensor-Based Human Activity Recognition Using Hybrid Deep Learning Techniques. *Security and Communication Networks, 2020,* 1–12. https://doi.org/10.1155/2020/2132138