

Attention-based CNN-LSTM architecture for smartphone IMU-based human activity recognition in intelligent mobile monitoring systems

Mussa Turdalyuly^{1,†}, Nazerke Baisholan^{2,*}, Sergiy Gnatyuk^{3,†}, Ay Zholdassova^{2,†} and Aigerim Yerimbetova^{1,†}

¹META University, Dosmukhamedov Str., 45, Almaty, 050012, Kazakhstan

²International Engineering and Technological University, Al-Farabi Ave., 89/21, Almaty, 050060, Kazakhstan

³National University "Kyiv Aviation Institute", Liubomyra Huzara Ave., 1, Kyiv, 03058, Ukraine

Abstract

Smartphone-based Human Activity Recognition (HAR) is an important research direction for mobile health monitoring, elderly care, rehabilitation support, fitness tracking, and context-aware intelligent systems. Modern smartphones are equipped with embedded inertial sensors, making them a practical and low-cost platform for recognizing human physical activities without additional wearable devices. However, accurate HAR remains challenging due to inter-subject variability, differences in movement style, smartphone placement, sensor noise, and similarity between dynamic activities such as Walking, Upstairs, and Downstairs. This paper proposes an attention-based CNN-LSTM architecture for smartphone-based HAR using accelerometer signals. The model combines one-dimensional convolutional layers for local temporal feature extraction, bidirectional LSTM layers for sequential dependency modeling, and an attention mechanism for emphasizing the most informative time steps within each sensor window. The model is evaluated on the WISDM dataset, which contains tri-axial accelerometer recordings of six daily activities: Walking, Jogging, Upstairs, Downstairs, Sitting, and Standing. Experimental results show that the proposed CNN-LSTM-Attention model achieves 99.76% accuracy and an F1-score of 0.9976, outperforming the CNN baseline with 99.67% accuracy. Although the absolute improvement is 0.09%, misclassified test windows are reduced from 7 to 5, corresponding to a 28% reduction in classification errors. Per-class analysis demonstrates perfect recognition of static activities such as Sitting and Standing, while most remaining errors are associated with similar dynamic activities, especially Upstairs. The results confirm that convolutional feature extraction, recurrent temporal modeling, and attention-based weighting improve the reliability of accelerometer-based HAR while maintaining practical computational efficiency for real-time and mobile activity monitoring applications.

Keywords

human activity recognition, smartphone sensors, accelerometer data, deep learning, CNN-LSTM, attention mechanism, time-series classification, WISDM dataset

1. Introduction

Human Activity Recognition (HAR) has become an important research direction in artificial intelligence, mobile computing, healthcare analytics, rehabilitation monitoring, elderly care, sports analytics, and context-aware services. The main objective of HAR is to automatically identify human physical activities from sensor signals, such as accelerometer, gyroscope, magnetometer, or other inertial measurement unit (IMU) data. Compared with camera-based activity recognition, smartphone- and wearable-sensor-based HAR is less intrusive, more suitable for daily-life monitoring, and more practical for privacy-sensitive environments [1, 2]. Since modern smartphones are equipped with embedded inertial sensors, they provide a widely available platform for continuous physical activity monitoring without the need for

WCH'26: Fifth International Workshop on Cyber Hygiene, June 25–26, 2026, Krakow, Poland

*Corresponding author.

[†]These authors contributed equally.

✉ m.turdalyuly@gmail.com (M. Turdalyuly); nbaishol@purdue.edu (N. Baisholan); s.gnatyuk@kai.edu.ua (S. Gnatyuk); aa.zholdassova@metu.edu.kz (A. Zholdassova); aigerian8888@gmail.com (A. Yerimbetova)

🆔 0000-0002-1470-3706 (M. Turdalyuly); 0000-0002-8134-0466 (N. Baisholan); 0000-0002-9421-8566 (S. Gnatyuk); 0009-0003-4271-2823 (A. Zholdassova); 0000-0002-2013-1513 (A. Yerimbetova)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

dedicated wearable devices.

Early HAR systems were mainly based on hand-crafted statistical and frequency-domain features extracted from acceleration signals, followed by classical machine learning algorithms such as decision trees, support vector machines, k-nearest neighbors, and random forests [3, 4]. Although these methods demonstrated promising results on controlled datasets, their performance strongly depended on feature engineering, sensor placement, window size, and the quality of manually designed descriptors. Public datasets such as WISDM and UCI HAR accelerated the development of smartphone-based activity recognition by providing benchmark data for walking, jogging, stair climbing, sitting, standing, and other daily activities [3, 5, 6]. However, practical HAR systems still face several challenges, including inter-subject variability, differences in body movement patterns, changes in smartphone orientation, and confusion between similar dynamic activities.

Deep learning has significantly changed the design of HAR systems by enabling automatic representation learning directly from raw or minimally processed sensor signals. Convolutional Neural Networks (CNNs) are effective for extracting local temporal patterns and short-term motion signatures from sensor windows [7, 8]. Recurrent neural networks, especially Long Short-Term Memory (LSTM) networks, are able to model longer temporal dependencies and sequential structures in human movement data [9, 10]. Hybrid CNN-LSTM architectures combine the advantages of convolutional feature extraction and recurrent temporal modeling, making them suitable for recognizing both static activities and complex dynamic movements [9, 11]. Recent studies have also shown that cross-subject evaluation, window-length selection, and sensor modality configuration strongly affect the generalization ability of smartphone-based HAR models [12].

Despite the high accuracy of deep learning-based HAR models, several limitations remain. First, activities such as walking upstairs, walking downstairs, and normal walking may have overlapping acceleration patterns, especially when only accelerometer data are used. Second, models trained under subject-dependent protocols may overestimate real-world performance because the same participants can appear in both training and testing sets. Third, complex models may achieve high accuracy but can be difficult to deploy in real-time mobile or edge environments due to computational and memory constraints [13, 14, 15]. Therefore, HAR models should be evaluated not only by classification accuracy but also by generalization, computational efficiency, and suitability for resource-constrained deployment.

Attention mechanisms provide a promising solution for improving the interpretability and discriminative capability of deep HAR models. Originally popularized in sequence modeling and transformer architectures [16], attention allows a model to assign different importance weights to different time steps, channels, or sensor modalities. In HAR, attention-based models can focus on the most informative fragments of motion signals and reduce the influence of noisy or less relevant segments [17, 18]. This is especially useful for smartphone accelerometer data, where discriminative patterns may occur only during specific phases of an activity, for example during foot impact, stair transition, or posture change.

In this paper, we propose an attention-based CNN-LSTM architecture for smartphone-based human activity recognition using accelerometer data. The proposed model first applies one-dimensional convolutional layers to extract local temporal features from segmented sensor windows. The extracted feature maps are then processed by bidirectional LSTM layers to capture sequential dependencies in both forward and backward directions. Finally, an attention mechanism is used to produce a weighted temporal representation by emphasizing the most informative time steps. The model is evaluated on the WISDM dataset and compared with a CNN baseline. The main contributions of this paper are as follows:

1. We propose a hybrid CNN-LSTM-Attention architecture for smartphone-based human activity recognition using accelerometer signals.
2. We provide a detailed methodological description of sensor windowing, deep feature extraction, temporal modeling, and attention-based weighting for activity classification.
3. We evaluate the proposed model on the WISDM dataset using a subject-independent experimental protocol to better assess generalization across users.

4. We compare the proposed architecture with a CNN baseline in terms of accuracy, F1-score, parameter count, and inference time.
5. We analyze class-wise performance and demonstrate that attention-based temporal modeling reduces misclassification between similar dynamic activities while maintaining practical computational efficiency.

The remainder of this paper is organized as follows. Section 2 reviews related work on classical machine learning, deep learning, CNN-LSTM architectures, and attention mechanisms for HAR. Section 3 describes the proposed methodology, including data preprocessing, window segmentation, model architecture, and training configuration. Section 4 presents the experimental setup and results. Section 5 discusses the findings, limitations, and practical implications. Section 6 concludes the paper and outlines directions for future research.

2. Related work

Human Activity Recognition (HAR) has been actively studied for more than two decades due to its importance for health monitoring, elderly care, rehabilitation support, sports analytics, mobile services, and intelligent environments. Existing HAR approaches can be broadly divided into three groups: classical machine learning methods based on hand-crafted features, deep learning methods for automatic representation learning, and attention-based neural architectures designed to improve the extraction of discriminative temporal and sensor-related patterns.

2.1. Classical machine learning approaches for HAR

Early sensor-based HAR systems relied mainly on manually designed features extracted from accelerometer or inertial measurement unit (IMU) signals. These features typically included time-domain characteristics such as mean, variance, standard deviation, signal magnitude area, zero-crossing rate, and correlation between axes, as well as frequency-domain descriptors obtained through Fourier transform or wavelet analysis [1, 2]. After feature extraction, classical machine learning algorithms such as decision trees, k-nearest neighbors, support vector machines, naïve Bayes, and random forests were commonly used for classification [3, 4].

One of the influential early smartphone-based HAR studies was conducted by Kwapisz et al. [3], who introduced the WISDM dataset and demonstrated that accelerometer data collected from smartphones could be used to classify daily activities such as walking, jogging, sitting, standing, walking upstairs, and walking downstairs. Similarly, Bao and Intille [4] showed that accelerometer-based wearable sensing can support the recognition of a wide range of human activities. Later, the UCI HAR dataset provided another important benchmark for smartphone-based activity recognition, contributing to the standardization of experimental evaluation in this area [5, 6].

Although classical approaches achieved acceptable performance in controlled scenarios, their effectiveness strongly depended on the quality of manually engineered features. Feature design usually required expert knowledge about signal processing and human motion patterns. In addition, such models often had limited generalization ability when applied to new subjects, different smartphone positions, or unseen movement styles. These limitations motivated the transition from hand-crafted feature engineering to deep learning-based representation learning.

2.2. Deep learning for sensor-based activity recognition

Deep learning has become a dominant approach in sensor-based HAR because it can automatically learn hierarchical representations from raw or minimally processed time-series signals. Convolutional Neural Networks (CNNs) have been widely used to extract local temporal patterns from accelerometer and gyroscope data. Zeng et al. [7] demonstrated that CNNs can effectively learn local dependencies from mobile sensor signals, while Yang et al. [8] showed that convolutional architectures can process

multichannel time-series data and achieve strong HAR performance. Ronao and Cho [19] further confirmed the effectiveness of deep neural networks for smartphone-based HAR using accelerometer and gyroscope signals.

CNN-based models are especially suitable for identifying short-term motion patterns, such as periodic acceleration changes during walking or jogging. However, human activities are not only local signal patterns but also temporal processes that unfold over time. Therefore, recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have been introduced to model long-term temporal dependencies in sensor data. Ordóñez and Roggen [9] proposed the DeepConvLSTM architecture, combining convolutional layers with LSTM layers for multimodal wearable activity recognition. Their work showed that CNN layers can extract local features, while LSTM layers can capture sequential dependencies over longer time intervals.

Other studies also confirmed that recurrent and hybrid deep learning models can improve HAR performance. Hammerla et al. [11] compared deep, convolutional, and recurrent models for wearable sensor-based HAR and showed that different architectures have different advantages depending on the dataset and activity structure. Chen and Xue [20] applied deep learning to single-accelerometer HAR, demonstrating that even limited sensor configurations can provide useful activity-related information. Murad and Pyun [10] investigated deep recurrent neural networks for HAR and showed that LSTM-based models are effective for modeling sensor time series. Guan and Plötz [21] further explored ensembles of LSTM learners to improve the robustness of activity recognition models.

Despite their strong performance, deep learning-based HAR models still face several practical challenges. First, high classification accuracy reported under subject-dependent evaluation protocols may not reflect real-world generalization, because the same subjects may appear in both training and testing sets. Second, sensor placement, device orientation, window size, and sampling rate can significantly affect model performance. Third, more complex models often require higher computational resources, which may limit their use in real-time smartphone or edge-device deployment [13, 22, 14].

In addition, recent surveys emphasize that HAR systems should be evaluated not only by accuracy, but also by robustness to sensor placement, subject variability, data quality, and computational cost [22, 14, 15]. Residual recurrent models and self-attention-based DeepConvLSTM architectures have also demonstrated that temporal modeling and attention can improve recognition of complex movement patterns [23, 24].

2.3. Cross-subject evaluation, window length, and sensor modality

A major issue in smartphone-based HAR is the difference between subject-dependent and subject-independent evaluation. In subject-dependent settings, data from the same participants may be included in both training and testing subsets. This can lead to overly optimistic results because the model may learn subject-specific movement patterns. In contrast, subject-independent evaluation separates participants between training and testing, making the task more realistic and more difficult. For practical HAR systems, cross-subject generalization is essential because the model must recognize activities of users who were not included during training.

Recent work by Turdalyuly et al. [12] directly investigated smartphone-based adult physical activity monitoring using IMU signals under a cross-subject deep learning protocol. Their study analyzed the influence of window length and sensor modality on recognition performance and showed that realistic evaluation settings are necessary for assessing the practical value of HAR models. This is particularly relevant for smartphone-based systems because users may carry devices differently, move at different speeds, and produce different acceleration patterns even when performing the same activity.

Sensor modality is another important factor. Accelerometer data are commonly used because they are available in almost all smartphones and directly reflect body movement. Gyroscope data can provide additional information about angular velocity and device rotation, while magnetometer data can support orientation estimation. However, using more sensors increases data dimensionality, energy consumption, and computational cost. Therefore, accelerometer-only HAR remains an important research direction, especially for lightweight real-time mobile applications. The present study focuses

on accelerometer-based recognition and aims to improve classification accuracy through architectural design rather than increasing the number of sensor modalities.

Window length also plays a significant role in HAR. Short windows can provide fast predictions and low latency but may not contain enough information to distinguish similar activities. Longer windows contain richer temporal context but increase computational cost and may delay recognition. In this work, fixed-length sliding windows are used to segment accelerometer signals into samples suitable for deep learning. The use of CNN, LSTM, and attention layers allows the model to extract both local and long-range movement information from each window.

2.4. Attention mechanisms in HAR

Attention mechanisms have recently become an important component of deep learning architectures for time-series classification. Originally popularized in sequence modeling and transformer architectures [16], attention allows a model to assign different importance weights to different parts of an input sequence. In HAR, this is useful because not all time steps contribute equally to activity recognition. For example, the most informative signal fragments for stair climbing may occur during the transition phase of foot movement, while other parts of the signal may be less discriminative.

Several studies have applied attention mechanisms to sensor-based HAR. Ma et al. [17] proposed AttnSense, a multi-level attention mechanism for multimodal HAR that combines convolutional and recurrent components to model spatial and temporal dependencies in sensor signals. Mahmud et al. [18] investigated self-attention for wearable sensor data and demonstrated that attention can help capture spatial and temporal relationships without relying only on recurrent structures. Almaslukh et al. [25] explored robust smartphone-based HAR under position-independent conditions, showing the importance of model robustness when smartphone placement varies. Recent recurrent and self-attention-based models also show that attention can improve temporal feature selection and classification robustness in sensor-based HAR [23, 24].

Attention-based CNN-LSTM models are particularly suitable for smartphone-based HAR because they combine three complementary advantages. CNN layers extract local temporal features from raw acceleration windows. LSTM layers model sequential dependencies and activity dynamics. Attention layers identify the most informative temporal segments and reduce the influence of less relevant signal parts. This combination is useful for distinguishing activities with similar movement patterns, such as walking, walking upstairs, and walking downstairs.

The proposed work follows this research direction by integrating a lightweight attention mechanism into a CNN-LSTM architecture for accelerometer-based HAR. Unlike approaches that rely on multiple sensors or highly complex architectures, the proposed model focuses on improving the discriminative capacity of accelerometer-only recognition while preserving practical computational efficiency. This makes the approach suitable for smartphone-based monitoring scenarios where real-time inference, low computational cost, and cross-user reliability are important.

3. Methodology

3.1. Problem formulation

The task of smartphone-based Human Activity Recognition (HAR) is formulated as a supervised multiclass time-series classification problem. Let

$$X = \{x_1, x_2, \dots, x_n\}$$

denote a continuous sequence of accelerometer readings collected from a smartphone, where each sample $x_t \in \mathbb{R}^3$ contains acceleration values along three axes: x -axis, y -axis, and z -axis. The objective is to learn a function f_θ that maps a segmented sensor window $X_w \in \mathbb{R}^{T \times 3}$ to one of K activity classes:

$$f_\theta : \mathbb{R}^{T \times 3} \rightarrow \{1, 2, \dots, K\}.$$

where T is the number of time steps in one window, $K = 6$ is the number of activity classes, and θ denotes the trainable parameters of the neural network. The six target activities in the WISDM dataset are Walking, Jogging, Upstairs, Downstairs, Sitting, and Standing [3].

The proposed model aims to classify each window into the correct activity class by learning discriminative temporal representations from accelerometer signals. Since smartphone-based accelerometer signals are noisy and may vary across subjects, the model should be able to extract robust features, capture temporal dependencies, and focus on the most informative parts of the signal.

3.2. Data preprocessing and window segmentation

Raw accelerometer signals are first cleaned and transformed into fixed-length windows suitable for deep learning. The WISDM dataset contains tri-axial accelerometer readings collected from smartphones carried by participants during daily physical activities [3]. Since neural networks require inputs of equal length, the continuous sensor stream is divided into overlapping sliding windows.

In this study, we use a window length of $T = 200$ samples with 50% overlap between consecutive windows. Considering the 20 Hz sampling rate of the WISDM dataset, each window corresponds to approximately 10 seconds of sensor data. This window length provides sufficient temporal context for distinguishing dynamic activities such as Walking, Jogging, Upstairs, and Downstairs, while still remaining suitable for real-time or near-real-time processing. The 50% overlap increases the number of training samples and helps preserve temporal continuity between adjacent windows.

For each window, the corresponding activity label is assigned according to the dominant activity in that segment. Since WISDM contains pre-labeled activity intervals, windows are generated only from continuous segments with consistent activity labels. This prevents the creation of ambiguous windows that contain transitions between different activities.

Before training, the accelerometer values are normalized using z -score standardization:

$$x' = \frac{x - \mu}{\sigma},$$

where μ and σ are the mean and standard deviation computed from the training set. The same normalization statistics are then applied to the validation and test sets. This prevents information leakage from the test data and ensures that the model is evaluated under a realistic experimental protocol. Normalization is performed independently for each accelerometer axis.

3.3. Proposed CNN-LSTM-attention architecture

The proposed architecture consists of three main components: convolutional feature extraction, bidirectional recurrent temporal modeling, and attention-based temporal weighting. This design combines the strengths of CNNs, LSTMs, and attention mechanisms for sensor-based time-series classification.

The overall processing pipeline is as follows:

Input accelerometer window $\rightarrow$ CNN feature extractor $\rightarrow$ Bidirectional LSTM $\rightarrow$ Attention layer $\rightarrow$
 $\rightarrow$ Fully connected classifier $\rightarrow$ Activity label

3.3.1. CNN-based local feature extraction

The first component of the model is a one-dimensional convolutional neural network. CNN layers are used to extract local temporal patterns from accelerometer signals. Such patterns may correspond to repeated motion cycles, short acceleration peaks, posture changes, or impact-related signal fragments. CNNs are widely used in HAR because they can automatically learn local features from raw sensor windows without requiring manual feature engineering [7, 8, 19].

The input window $X_w \in \mathbb{R}^{200 \times 3}$ is passed through three one-dimensional convolutional layers. The convolutional layers use 64, 128, and 256 filters, respectively, with a kernel size of 5. Each convolutional layer is followed by batch normalization and a rectified linear unit (ReLU) activation function. The first two convolutional blocks include max-pooling with a pooling factor of 2 to reduce the temporal resolution and computational cost while preserving the most important local features.

The convolution operation can be represented as

$$z_t = \text{ReLU}(W_c * x_{t:t+k-1} + b_c),$$

where W_c denotes the convolutional kernel, k is the kernel size, b_c is the bias term, and $*$ denotes the convolution operation. The output of the CNN block is a sequence of learned feature vectors that encode local motion characteristics.

3.3.2. Bidirectional LSTM for temporal modeling

Although CNN layers are effective in extracting local temporal features, they do not explicitly model long-range dependencies across the entire window. Human activities are sequential processes, and their correct recognition often requires analyzing how motion patterns evolve over time. For example, walking upstairs and walking downstairs may contain similar local acceleration peaks, but their temporal dynamics and phase structure can differ.

To capture such dependencies, the CNN feature sequence is passed to a two-layer Bidirectional Long Short-Term Memory network. LSTM networks are designed to model sequential data and overcome the vanishing gradient problem through gating mechanisms. They are widely used in HAR tasks where temporal dependencies play an important role [9, 11, 10].

A bidirectional LSTM processes the input sequence in both forward and backward directions. The forward LSTM captures information from past to future time steps, while the backward LSTM captures information from future to past time steps. The hidden representation at each time step is obtained by concatenating the forward and backward hidden states:

$$h_t = [\vec{h}_t; \overleftarrow{h}_t].$$

3.3.3. Attention-based temporal weighting

Not all time steps within a sensor window contribute equally to activity recognition. Some fragments may contain highly discriminative information, while others may include noise, redundant motion, or less informative signal parts. To address this issue, an attention mechanism is added after the bidirectional LSTM layer.

The attention layer computes an importance score for each LSTM hidden state. The attention score for each time step is computed as:

$$e_t = \mathbf{u}^\top \tanh(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}_a),$$

where $\mathbf{W}_a$ and $\mathbf{b}_a$ are trainable parameters, and $\mathbf{u}$ is a learnable context vector. The attention weights are obtained using the softmax function:

$$\alpha_t = \frac{\exp(e_t)}{\sum_j \exp(e_j)}.$$

Finally, the context vector $\mathbf{c}$ is computed as a weighted sum of the hidden states:

$$\mathbf{c} = \sum_t \alpha_t \mathbf{h}_t.$$

The context vector $\mathbf{c}$ represents the entire sensor window, with greater emphasis on the most informative time steps. This mechanism helps the model focus on discriminative temporal fragments,

which is especially important for separating similar dynamic activities such as Walking, Upstairs, and Downstairs. Attention mechanisms have shown strong potential in sensor-based human activity recognition by improving both discriminative performance and interpretability [16, 17, 18].

3.3.4. Classification layer

The attention-generated context vector is passed to a fully connected classification head. The final layer uses the softmax activation function to produce class probabilities:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_o \mathbf{c} + \mathbf{b}_o),$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are the trainable parameters of the output layer. The predicted activity class is selected as the class with the highest probability.

3.4. Training configuration

The proposed model is trained using categorical cross-entropy loss:

The model is trained by minimizing the categorical cross-entropy loss:

$$\mathcal{L} = - \sum_i \sum_k y_{i,k} \log(\hat{y}_{i,k}),$$

where $y_{i,k}$ is the true label and $\hat{y}_{i,k}$ is the predicted probability for class k . The Adam optimizer is used with an initial learning rate of 0.001. The learning rate is reduced by a factor of 0.5 when the validation loss does not improve. Early stopping is applied to prevent overfitting, with training stopped if the validation performance does not improve for 10 consecutive epochs.

To improve generalization, dropout with a probability of 0.5 and L_2 regularization with a coefficient of 0.0001 are used. The model is trained with a batch size of 512. Training is performed using GPU acceleration. In the experimental setup, DataParallel training across three NVIDIA L40 GPUs is used to accelerate the training process.

3.5. Evaluation protocol

The model is evaluated using a subject-independent protocol. In this setting, subjects used for testing are not included in the training set. This evaluation strategy is more realistic than subject-dependent splitting because it measures how well the model generalizes to unseen users. Cross-subject evaluation is especially important in smartphone-based HAR because movement style, walking speed, phone placement, and body characteristics may differ significantly across individuals [12].

The dataset is divided into training, validation, and test subsets. The training set is used to optimize the model parameters, the validation set is used for hyperparameter selection and early stopping, and the test set is used only for final performance evaluation.

The following metrics are used:

The evaluation metrics are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Accuracy provides an overall measure of classification correctness, whereas precision, recall, and F1-score provide more detailed information about class-wise performance. The confusion matrix is also analyzed to identify which activities are most frequently confused by the model.

4. Experimental results

4.1. Dataset description

The proposed CNN–LSTM–Attention model was evaluated on the WISDM Activity Recognition dataset, which is one of the most widely used benchmark datasets for smartphone-based human activity recognition [3]. The dataset contains tri-axial accelerometer data collected from smartphones carried by 36 participants while performing six daily activities: Walking, Jogging, Upstairs, Downstairs, Sitting, and Standing. These activities represent both dynamic motion patterns and static postures, making the dataset suitable for evaluating the ability of HAR models to distinguish between different types of human movement.

The WISDM dataset contains 1,098,207 accelerometer samples collected at approximately 20 Hz. Each sample consists of acceleration values along the x -, y -, and z -axes together with the corresponding activity label. In this study, the continuous accelerometer stream was segmented into fixed-length windows of 200 samples with 50% overlap. Since the sampling frequency is approximately 20 Hz, each window corresponds to about 10 seconds of sensor data. This window size provides sufficient temporal context for recognizing periodic dynamic activities while preserving practical computational requirements.

The six activity classes have different signal characteristics. Sitting and Standing are static activities with low-variance acceleration patterns. Walking and Jogging contain periodic motion cycles but differ in intensity and acceleration amplitude. Upstairs and Downstairs are more challenging because their signals may overlap with Walking and with each other during certain movement phases. Therefore, per-class analysis and confusion matrix inspection are important for understanding model behavior.

4.2. Experimental setup

The dataset was divided into training, validation, and testing subsets. The training set was used for model optimization, the validation set was used for early stopping and learning-rate scheduling, and the test set was used only for final evaluation. The resulting split contained 10,590 training windows, 1,513 validation windows, and 2,119 test windows.

To avoid overly optimistic evaluation, the experiment was organized using a subject-independent protocol, where the test data were separated from the training data at the subject level. This evaluation setting is more realistic than random window-level splitting because it measures the ability of the model to generalize to unseen users. Cross-subject evaluation is especially important for smartphone-based human activity recognition, since acceleration signals may vary depending on body characteristics, walking style, phone placement, and individual movement habits [12].

All accelerometer channels were normalized using z -score standardization based on the training-set statistics. The same mean and standard deviation values were then applied to the validation and test sets. This procedure prevents data leakage and ensures that the reported performance reflects generalization rather than information transfer from the test set.

4.3. Baseline model

To evaluate the effectiveness of the proposed temporal modeling and attention components, we compared the proposed CNN–LSTM–Attention architecture with a CNN baseline. The baseline model used the same convolutional feature extraction block as the proposed model but replaced the bidirectional LSTM and attention mechanism with a fully connected classification head. This comparison allows us to measure the contribution of recurrent temporal modeling and attention-based temporal weighting.

The CNN baseline is computationally lighter and contains fewer trainable parameters. However, it mainly captures local temporal features and does not explicitly model long-range dependencies across the entire sensor window. In contrast, the proposed model uses bidirectional LSTM layers to analyze temporal dependencies and an attention mechanism to focus on the most informative time steps. Therefore, the comparison between these two models is useful for evaluating whether the additional temporal components improve classification reliability.

4.4. Evaluation metrics

The models were evaluated using accuracy, precision, recall, F_1 -score, parameter count, inference time, and confusion matrix analysis. Accuracy measures the overall proportion of correctly classified windows. Precision and recall provide class-specific information about false positives and false negatives. The F_1 -score is used as a balanced measure between precision and recall, which is especially useful when analyzing class-wise performance.

In addition to the classification metrics, computational characteristics were also considered. The parameter count was used to estimate model size, while inference time was measured to assess the practical feasibility of the model. Since inference time was measured on an NVIDIA L40 GPU, it should be interpreted as a laboratory performance indicator rather than as a direct smartphone deployment result. Nevertheless, the obtained latency provides useful evidence that the proposed architecture has potential for real-time or near-real-time human activity recognition applications after optimization for edge or mobile platforms.

4.5. Overall performance comparison

Table 1 presents the performance comparison between the CNN baseline and the proposed CNN-LSTM-Attention model.

Table 1

Performance Comparison Between the Baseline and Proposed Model

Method	Accuracy (%)	F_1 -score	Parameters	Inference time
CNN baseline	99.67	0.9967	421K	~ 5
Proposed CNN-LSTM-Attention	99.76	0.9976	1.24M	~ 12

The proposed CNN-LSTM-Attention model achieved an accuracy of 99.76% and an F_1 -score of 0.9976 on the WISDM test set. The CNN baseline achieved 99.67% accuracy and an F_1 -score of 0.9967. Although the absolute improvement in accuracy is 0.09%, the number of misclassified test windows was reduced from 7 to 5 out of 2,119 test samples. This corresponds to an approximately 28% reduction in classification errors.

The improvement demonstrates that recurrent temporal modeling and attention-based weighting provide additional discriminative power, even when the CNN baseline already achieves very high performance. In human activity recognition tasks, small absolute improvements may still be meaningful when the baseline accuracy is close to saturation. In this case, the proposed architecture reduces the remaining classification errors and improves the reliability of the model.

The proposed model contains approximately 1.24 million trainable parameters, which is approximately three times more than the CNN baseline. The inference time increased from approximately 5 ms to 12 ms in the GPU-based experimental environment. This increase is expected because the proposed model includes bidirectional LSTM and attention layers. Nevertheless, the latency remains sufficiently low to indicate practical potential for real-time human activity recognition, particularly after model compression, quantization, or pruning techniques are applied before deployment on mobile or edge devices.

4.6. Per-class classification performance

Table 2 presents the class-wise precision, recall, and F_1 -score of the proposed CNN-LSTM-Attention model.

Table 2

Per-Class Performance of the Proposed CNN-LSTM-Attention Model

Activity	Precision	Recall	F_1 -score
Downstairs	0.9948	0.9948	0.9948
Jogging	0.9985	0.9985	0.9985
Sitting	1.0000	1.0000	1.0000
Standing	1.0000	1.0000	1.0000
Upstairs	0.9957	0.9873	0.9915
Walking	0.9976	1.0000	0.9988

The proposed model achieved perfect classification for the static activities Sitting and Standing. This result is expected because these activities have stable and low-variance acceleration patterns that are clearly different from dynamic activities. The model also achieved very high performance for Walking and Jogging, with F_1 -scores of 0.9988 and 0.9985, respectively. These results indicate that the CNN layers effectively captured periodic acceleration patterns associated with locomotion.

The lowest F_1 -score was observed for Upstairs, with a value of 0.9915. This activity is more difficult to classify because its acceleration pattern can be similar to Walking and Downstairs during certain phases of movement. The lower recall for Upstairs suggests that some Upstairs windows were classified as other dynamic activities. This confirms that stair-related activities remain among the most challenging classes in smartphone-based human activity recognition.

4.7. Confusion matrix analysis

The confusion matrix provides a more detailed view of classification errors. Figure 1 shows the row-normalized confusion matrix of the proposed CNN-LSTM-Attention model on the WISDM test set. The model demonstrates strong discriminative capability, with most confusion occurring between Upstairs and similar dynamic activities.

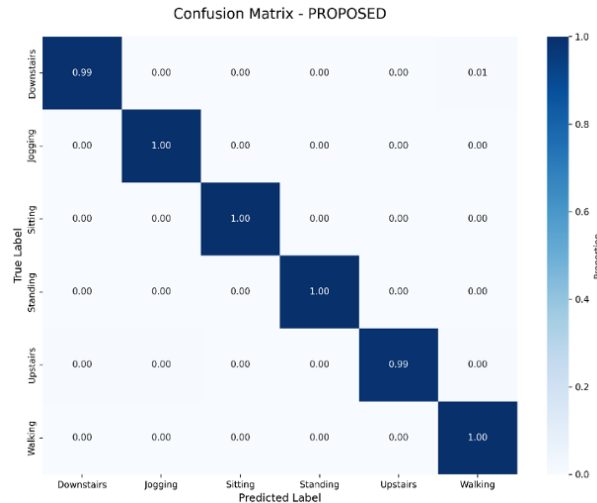


Figure 1: Confusion matrix for the proposed CNN-LSTM-Attention model on the WISDM test set.

The confusion matrix confirms that most activity classes were recognized with near-perfect accuracy. Static activities were classified without errors, while the few remaining misclassifications were mainly associated with dynamic activities. In particular, most errors involved the Upstairs class, which is

consistent with the per-class performance results. This indicates that the most difficult cases occur when the motion pattern of stair climbing resembles walking or descending stairs.

The attention mechanism helps reduce such confusion by assigning higher importance to discriminative temporal fragments within each window. For example, in stair-related activities, the most informative parts of the signal may correspond to changes in vertical acceleration and specific phases of foot movement. By emphasizing these fragments, the model can better distinguish between activities that have similar local acceleration patterns.

4.8. Computational efficiency analysis

Computational efficiency is an important requirement for smartphone-based HAR because practical systems must operate under limited memory, processing power, and battery capacity. The proposed model is more complex than the CNN baseline, but the increase in computational cost is moderate relative to the improvement in classification reliability.

The CNN baseline required approximately 421K parameters and achieved an inference time of about 5 ms in the experimental GPU environment. The proposed CNN-LSTM-Attention model required approximately 1.24M parameters and achieved an inference time of about 12 ms. Although the proposed model is larger, it remains compact compared with many deep neural architectures used for time-series classification. The attention mechanism introduces only a limited number of additional parameters while improving temporal focus and classification robustness.

For practical mobile deployment, further optimization may be required. Techniques such as weight pruning, 8-bit quantization, knowledge distillation, and architecture simplification can reduce memory usage and inference latency. Therefore, the proposed model can be considered a strong candidate for real-time HAR; however, direct benchmarking on smartphone hardware should be conducted in future work.

4.9. Summary of results

The experimental results show that the proposed CNN-LSTM-Attention architecture improves smartphone-based human activity recognition performance compared with a CNN baseline. The model achieved 99.76% accuracy and reduced the number of classification errors by 28%. The best performance was observed for static activities, whereas the most challenging class was Upstairs because of its similarity to other dynamic movements. The combination of CNN-based feature extraction, bidirectional LSTM temporal modeling, and attention-based weighting provides a balanced solution for accurate and computationally feasible accelerometer-based activity recognition.

5. Discussion and limitations

5.1. Discussion

The experimental results demonstrate that the proposed CNN-LSTM-Attention architecture is effective for smartphone-based Human Activity Recognition (HAR) using accelerometer data. The model achieved 99.76% accuracy and an F1-score of 0.9976 on the WISDM test set, outperforming the CNN baseline. Although the absolute improvement over the baseline was 0.09%, the number of classification errors was reduced from 7 to 5 cases, corresponding to a 28% reduction in misclassification errors.

This result is important because the CNN baseline already achieved very high accuracy. When a baseline model reaches near-saturated performance, even small absolute improvements may indicate that the proposed architecture is more effective at handling difficult borderline cases. In this study, the additional improvement is mainly associated with the ability of the bidirectional LSTM and attention mechanism to model temporal dependencies more effectively and identify discriminative signal fragments.

The obtained results also confirm that different activity types have different levels of recognition difficulty. Static activities such as Sitting and Standing were classified perfectly. This can be explained by their stable, low-variance acceleration patterns, which are clearly different from dynamic movements. In contrast, activities such as Upstairs and Downstairs are more difficult to recognize because they share similar motion components with Walking. This confirms previous findings that stair-related activities are among the most challenging classes in smartphone-based HAR [3, 12].

The attention mechanism plays an important role in improving the model's ability to distinguish between similar activities. In accelerometer-based HAR, not all parts of a sensor window are equally informative. For example, during stair climbing, the most discriminative patterns may appear during specific phases of vertical movement, whereas other parts of the window may resemble normal walking. By assigning higher weights to informative time steps, the attention layer helps the model focus on the most relevant temporal fragments. This supports the use of attention mechanisms in HAR tasks, as also demonstrated in attention-based and self-attention-based sensor activity recognition studies [17, 18, 23, 24].

Another important aspect of the proposed model is its balance between accuracy and computational complexity. The proposed architecture contains approximately 1.24 million parameters, which is larger than the CNN baseline containing approximately 421,000 parameters. However, the inference time remains practical in the experimental GPU environment. This indicates that the model has potential for real-time HAR applications after additional optimization. For smartphone deployment, techniques such as model compression, pruning, quantization, and knowledge distillation may be used to reduce computational requirements [13, 14, 15].

The results are also relevant in the context of cross-subject HAR. In real-world applications, HAR models should recognize the activities of new users who were not included in the training data. Subject-independent evaluation is therefore more realistic than random window-level splitting. Recent research has shown that window length, sensor modality, and cross-subject evaluation protocols significantly influence the reliability of smartphone-based physical activity monitoring [12]. The present study follows this direction by focusing on accelerometer-based recognition and evaluating the model using a generalization-oriented experimental protocol.

5.2. Practical implications

The proposed CNN-LSTM-Attention model can be useful for several practical applications. In mobile health monitoring, it can support automatic tracking of daily physical activity and sedentary behavior. In elderly care, it may help monitor movement patterns and detect changes in activity routines. In rehabilitation, activity recognition can be used to assess patient compliance with prescribed physical exercises. In fitness and wellness applications, the model can support automatic logging of walking, jogging, and other daily movements.

The use of smartphone accelerometer data makes the proposed approach practical because it does not require external wearable devices. This is especially important for large-scale deployment, where users may not be willing to wear additional sensors. Smartphone-based HAR can therefore provide a low-cost and accessible solution for continuous activity monitoring.

At the same time, practical implementation requires careful consideration of battery consumption, data privacy, sensor noise, and device placement. Real-world smartphone data may be more variable than benchmark dataset recordings because users can carry phones in pockets, bags, hands, or different body positions. Therefore, further validation under uncontrolled daily-life conditions is necessary before deployment in commercial or clinical systems.

5.3. Limitations

Despite the strong classification results, this study has several limitations.

First, the experiments were conducted on the WISDM dataset, which is a widely used benchmark but still represents a limited set of activities and participants. The dataset includes six basic activities:

Walking, Jogging, Upstairs, Downstairs, Sitting, and Standing. Real-world activity recognition may require a larger number of activities, including transitions, complex movements, household tasks, transportation modes, and abnormal events such as falls.

Second, this study used only accelerometer signals. Accelerometer-only recognition is useful for lightweight smartphone-based HAR, but additional sensors such as gyroscopes and magnetometers may provide complementary information. Gyroscope data can capture angular velocity and device rotation, which may help distinguish activities with similar acceleration patterns. However, using additional sensors may increase energy consumption and computational cost.

Third, the inference time was measured in a GPU-based experimental environment. Although the obtained latency indicates the potential for real-time recognition, direct testing on smartphone hardware was not performed. Mobile deployment may produce different latency and energy consumption results depending on the device processor, operating system, background processes, and implementation framework.

Fourth, the study did not evaluate robustness to different smartphone placements. In practical scenarios, users may carry smartphones in front pockets, back pockets, hands, bags, or jackets. Changes in sensor orientation and placement can significantly affect accelerometer patterns. Therefore, future work should include placement-independent evaluation or data augmentation strategies to improve robustness.

Fifth, the proposed model was evaluated on segmented windows rather than continuous real-time streams containing activity transitions. In real applications, activities may change gradually, and windows may include transition periods. Future research should investigate continuous HAR without strict segmentation and incorporate transition-aware modeling.

Finally, although the attention mechanism improves temporal focus, the interpretability of the model remains limited. Attention weights can provide an approximate indication of important time steps, but they do not fully explain the model's decision-making process. Future studies may combine attention visualization with explainable artificial intelligence methods to improve transparency and trust in HAR systems.

6. Conclusions

This paper presented an attention-based CNN-LSTM architecture for smartphone-based Human Activity Recognition using accelerometer signals. The proposed model combines convolutional feature extraction, bidirectional recurrent temporal modeling, and attention-based temporal weighting to improve the classification of daily human activities from time-series sensor data.

The experimental evaluation was conducted on the WISDM dataset, which includes six activity classes: Walking, Jogging, Upstairs, Downstairs, Sitting, and Standing. The proposed CNN-LSTM-Attention model achieved 99.76% accuracy and an F1-score of 0.9976, outperforming the CNN baseline, which achieved 99.67% accuracy. Although the absolute improvement was 0.09%, the number of misclassification cases decreased from 7 to 5 out of 2,119 test windows. This corresponds to a 28% reduction in classification errors and confirms that the proposed architecture improves recognition reliability in difficult cases.

The per-class analysis showed that static activities such as Sitting and Standing were classified with perfect precision, recall, and F1-score. Dynamic activities were also recognized with very high accuracy, although Upstairs remained the most challenging class because of its similarity to Walking and Downstairs. The confusion matrix analysis confirmed that most remaining errors occurred between similar dynamic activities. These results demonstrate that the attention mechanism helps the model focus on discriminative temporal fragments of accelerometer signals.

The proposed model also provides a reasonable balance between classification accuracy and computational complexity. Although the CNN-LSTM-Attention architecture contains more trainable parameters than the CNN baseline, its inference time remained practical in the GPU-based experimental environment. This indicates that the model has potential for real-time or near-real-time HAR applications after

additional optimization for mobile and edge platforms.

The main limitation of this study is that the model was evaluated on one benchmark dataset and only accelerometer signals were used. Future work will focus on evaluating the model on additional public datasets, incorporating gyroscope and magnetometer data, testing robustness to different smartphone placements, and deploying optimized versions of the model on real smartphone hardware. Further research will also investigate continuous activity recognition with transition-aware modeling and explainable AI techniques for improved interpretation of attention weights.

Overall, the results confirm that the integration of CNN, LSTM, and attention mechanisms is an effective approach for smartphone-based HAR. The proposed architecture improves the reliability of accelerometer-based activity recognition and can serve as a basis for practical mobile health monitoring, elderly care, rehabilitation support, and context-aware intelligent systems.

Data availability

The experiments in this study were conducted using the publicly available WISDM Activity Recognition dataset [3].

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

- [1] O. D. Lara, M. A. Labrador, A survey on human activity recognition using wearable sensors, *IEEE Communications Surveys & Tutorials* 15 (2013) 1192–1209. doi:10.1109/SURV.2012.110112.00192.
- [2] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, P. J. M. Havinga, A survey of online activity recognition using mobile phones, *Sensors* 15 (2015) 2059–2085. doi:10.3390/s150102059.
- [3] J. R. Kwapisz, G. M. Weiss, S. A. Moore, Activity recognition using cell phone accelerometers, *ACM SIGKDD Explorations Newsletter* 12 (2011) 74–82. doi:10.1145/1964897.1964918.
- [4] L. Bao, S. S. Intille, Activity recognition from user-annotated acceleration data, in: A. Ferscha, F. Mattern (Eds.), *Pervasive Computing*, volume 3001 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2004, pp. 1–17. doi:10.1007/978-3-540-24646-6_1.
- [5] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, A public domain dataset for human activity recognition using smartphones, in: *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Ciaco, Bruges, Belgium, 2013, pp. 437–442.
- [6] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, A. Ghio, X. Parra, D. Anguita, Transition-aware human activity recognition using smartphones, *Neurocomputing* 171 (2016) 754–767. doi:10.1016/j.neucom.2015.07.085.
- [7] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, J. Zhang, Convolutional neural networks for human activity recognition using mobile sensors, in: *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services*, Austin, TX, USA, 2014, pp. 197–205. doi:10.4108/icst.mobibase.2014.257786.
- [8] J. Yang, M. N. Nguyen, P. P. San, X. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition, in: *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, AAAI Press, Buenos Aires, Argentina, 2015, pp. 3995–4001.
- [9] F. J. Ordóñez, D. Roggen, Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition, *Sensors* 16 (2016) 1–25. doi:10.3390/s16010115.

- [10] A. Murad, J.-Y. Pyun, Deep recurrent neural networks for human activity recognition, *Sensors* 17 (2017) 1–18. doi:10.3390/s17112556.
- [11] N. Y. Hammerla, S. Halloran, T. Plötz, Deep, convolutional, and recurrent models for human activity recognition using wearables, in: *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, AAAI Press, New York, NY, USA, 2016, pp. 1533–1540.
- [12] M. Turdalyuly, A. Zholdassova, T. Turdalykyzy, A. Doshybekov, Wearable sensor-free adult physical activity monitoring using smartphone IMU signals: Cross-subject deep learning with window-length and sensor modality studies, *Information* 17 (2026) 1–24. doi:10.3390/info17040368.
- [13] A. Ignatov, Real-time human activity recognition from accelerometer data using convolutional neural networks, *Applied Soft Computing* 62 (2018) 915–922. doi:10.1016/j.asoc.2017.10.022.
- [14] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, Y. Liu, Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities, *ACM Computing Surveys* 54 (2021) 1–36. doi:10.1145/3457607.
- [15] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, U. R. Alo, Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges, *Expert Systems with Applications* 105 (2018) 233–261. doi:10.1016/j.eswa.2018.03.056.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems*, volume 30, 2017, pp. 5998–6008.
- [17] H. Ma, W. Li, X. Zhang, S. Gao, S. Lu, Attnsense: Multi-level attention mechanism for multimodal human activity recognition, in: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macau, China, 2019, pp. 3109–3115.
- [18] S. Mahmud, M. T. H. T., K. K. Bhaumik, A. K. M. M. Rahman, M. Amin, M. Shoyaib, M. A. H. Khan, A. Ali, Human activity recognition from wearable sensor data using self-attention, in: *Proceedings of the 24th European Conference on Artificial Intelligence*, 2020.
- [19] C. A. Ronao, S.-B. Cho, Human activity recognition with smartphone sensors using deep learning neural networks, *Expert Systems with Applications* 59 (2016) 235–244. doi:10.1016/j.eswa.2016.04.032.
- [20] Y. Chen, Y. Xue, A deep learning approach to human activity recognition based on single accelerometer, in: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Hong Kong, China, 2015, pp. 1488–1492. doi:10.1109/SMC.2015.208.
- [21] Y. Guan, T. Plötz, Ensembles of deep LSTM learners for activity recognition using wearables, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (2017) 1–28. doi:10.1145/3131672.
- [22] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: A survey, *Pattern Recognition Letters* 119 (2019) 3–11. doi:10.1016/j.patrec.2018.10.013.
- [23] Y. Zhao, R. Chen, G. Li, Y. Chen, Deep residual bidir-LSTM for human activity recognition using wearable sensors, *Mathematical Problems in Engineering* (2018) 1–12. doi:10.1155/2018/7316954.
- [24] S. P. Singh, M. K. Sharma, A. Lay-Ekuakille, D. Gangwar, S. Gupta, Deep ConvLSTM with self-attention for human activity decoding using wearable sensors, *IEEE Sensors Journal* 21 (2021) 8575–8582. doi:10.1109/JSEN.2020.3045135.
- [25] B. Almaslukh, A. M. Artoli, J. Al-Muhtadi, A robust deep learning approach for position-independent smartphone-based human activity recognition, *Sensors* 18 (2018) 1–19. doi:10.3390/s18113726.