RESEARCH PAPER

Human Activity Recognition Using Smartphone Sensors: A Deep Learning Approach

Manikanta K B[1], Dr. B Ramesk Naik[2]
1st Ph.D Scholar, CSE, Gitam School of Computer Science & Engineering, Bengaluru, India
mreddy@gitam.in
2nd Associate Professor, CSE(AI-ML), Gitam School of Computer Science & Engineering, Bengaluru, India
rbhukya@gitam.edu
Corresponding Author: Manikanta K B
Email-mreddy@gitam.in

Abstract
Human activity recognition based on a smartphone-captured dataset is a popular area of research due to its perceivable advantages in applications such as healthcare monitors, sports analytics, care for the elderly, and rehabilitation systems. Smartphones, embedded with inertial sensors, such as accelerometers and gyroscopes, provide a means for continuous human movement monitoring without the need for any specialized equipment. Usually, data from motion sensors such as three-axis accelerometers, magnetometers, etc., are employed in smartphone-based HAR systems. The models built are then used to predict the activities made by individuals. The activity classification, depending on the situation for HAR, is currently based on machine learning and deep learning. However, often, the refrigerator approach by the modern ML and DL algorithms struggles to accurately classify countless complex and dynamic activities due to the variations in placement of smartphones, noise on sensors, as well as movement artifacts, and the diversity of user behaviour.
An advanced HAR model that embeds deep learning is introduced in this study to improve classification accuracy for activities in real-world applications. The approach described highlights sensor fusion and optimization of neural network architectures for successful extraction of spatial and temporal features from the inertial sensor data. The experimental evaluation confirmed the superiority for performance against traditional machine learning-induced HAR systems and exhibited stability in handling different device orientations and user movement patterns.
Keywords
Human Activity Recognition, Smartphone Sensors, Deep Learning, Accelerometer, Gyroscope, Healthcare Monitoring, Sensor Data Fusion
How to cite this article: Manikanta KB, Naik BR. Human Activity Recognition Using Smartphone Sensors: A Deep Learning Approach. Int J Drug Deliv Technol. 2026;16(72s): 1466-1475. DOI: 10.25258/ijddt.16.72s.160

1. Introduction

The Human Activity Recognition (HAR) has grown abundantly in the last few years, garnering a broad range of applications in the healthcare monitoring, sports analytics, smart environments, and rehabilitation systems. With the fast development of technology for smartphones, contemporary gadgets are endowed with multiple built-in sensors like accelerometers, gyroscopes, and magnetometers that permit the recognition of the user's movements without relying on wearables dedicated to monitoring activity.

Hence, smartphone-centralized HAR systems provide a rather efficient and cost-effective way to monitor physical activities such as walking, running, sitting, standing, and climbing stairs. These systems are crucial in the field of healthcare, where normal activities are monitored and can hence assist in quick diagnoses, help monitor a patient's progress in rehabilitation, and stave off falls among the elderly.
Traditional methods in human activity recognition (HAR) highly depend on the use of machine-learning algorithms- Support Vector Machines (SVM), Decision Trees, and Random Forests. Such classifiers essentially rely on manual features that are derived from sensor signals. Because of the controlled circumstances, some classification strategies that rely on HAR, for instance, until the human recognition of activities during some regular circumstances, did not attract the attention of scientists, particularly in relation to the cases of defective smartphone positioning, erroneousness of sensor readings, and variability in the execution of specific bodily actions. A while later, deep learning emerged with its promise to enhance HAR performance by learning hidden features from raw sensor data. Convolutional neural networks, recurrent neural networks, and hybrid models are examples of these architectures applied to HAR tasks. Yet, classifying complex activities sitting in motion against change of device orientation, motion artifacts, and specific user behavior has continued to surface.
The requirement arises to build a strong HAR model, be capable of competently solving these challenges and keep well the high accuracy of classifications in that HAR model in the real-life situation.

2. Related Work

Researchers around the world have studied Human Activity Recognition (HAR) because the technology can be applied to both healthcare and fitness monitoring and smart environments. The initial methods used in Human Activity Recognition (HAR) employed classifiers that operated through decision trees and support vector machines and random forests to process manually created features. The development of advanced methods led to more sophisticated solutions which enhanced the performance of these methods in new application areas. These earlier works focused more on accuracy in specific contexts, so they produced decent results on Treasure Dataset. Still, in a dynamic environment, they are very prone to overfitting to this or maybe some of the input documents retrieved from the database.

Due to the high-performance capabilities of convolutional neural networks (CNNs) in extracting the spatial features automatically from sensor data, these deep-learning mechanisms are now extensively subsumed in the tasks involving human activity recognition. For instance, like finding well-graded deep-learning models, Ronao & Cho (2016) reported highly effective results in performing smartphone sensor data classification that surpassed standard methods to some extent. Similarly, research conducted by Zeng et al. (2014) utilized CNN architectures for HAR using mobile sensors. It was illustrated that a global CNN based on local dependencies in a timescale clearly outperforms most local methods on various sensor datasets. Chen & Xue (2015) proposed deep CNN models for recognition of activities through fusion of sensor inputs and confirmed that they work fine. None of the advantages of CNN-based methods has brought a closure to their incapacity to handle excessively long-term temporal dependencies related to human motion data in a sequence.

LSTM sharing the spatial information and temporal information set for useful feature engineering applications. Ordonez and Roggen (2016) came up with the deep CNN-LSTM model to tackle multimodal wearable sensor data as the major stepping stone toward greater performances seen by usage of exploding models. Dive deep! Additionally, Mutegeki and Han (2020) are the authors who encouraged us to pioneer the use of hybrid methods and prove that considering some parts of LSTM and some from CNN is actually able to model the most complex activity patterns to be anticipated. Their work was quickly complemented in 2023 by Jameer et al. and later in 2024 by Alshammari et al. On board, the study provided that focused CNN-LSTM frameworks for smartphone human activity recognition achieved the highest accuracy under real-world conditions. Furthermore, there are numerous pieces to this effect that claim the success of attention-based extensions; in particular, the CNN-LSTM-attention model discussed by Xai et al. in 2025 demonstrates strong indications in favor of relevant temporal feature representation.

Sensor fusion techniques were extensively explored to improve HAR performance. Gumaei et al. (2019) showed that combining the accelerometer data and gyroscope can enhance the classification accuracy by the mere complementary motion information. García-González et al. (2023) made the same point, that data integration of multimodal approaches was very important for robustness in real-life HAR applications.

In several review articles, past studies on HAR have been briefly summarized. Smartphones have been briefly outlined as the scope and coverage of smartphone-based HAR systems for health purposes. Key problems highlighted as constraints are user variability and sensor noise. Islam et al. (2022) reviewed the CNN-based approaches for HAR and discussed the studies' limitations in adapting complex activities. Ferrari et al. (2021) talked about the smartphone-based HAR trends and emphasized the need for robust and scalable models. Several challenges confront the existing HAR systems, despite significant advances, such as variation in placement of smart phones, noise in the sensory data, and difficulty in recognizing complex or overlapping activities. Furthermore, the models may lack robustness for real-world deployment. Therefore, this study presents a hotspot hybrid architecture. CNN-LSTM, exploiting sensor fusion, only achieves a considerably higher accuracy of classification and generalizes superlatively on different conditions, overcoming these limitations with optimized pre-processing techniques.

3. Research Objective

The primary objective of this study is to develop a robust Human Activity Recognition framework using smartphone sensor data that can accurately recognize both simple and complex human activities under real-world conditions.

The specific objectives of this research include:

- To design an efficient deep learning-based HAR model capable of automatically extracting relevant features from smartphone sensor data.
- To improve activity classification accuracy by integrating multiple sensor modalities such as accelerometer and gyroscope data.
- To develop a model that is robust to variations in smartphone orientation and placement.
- To evaluate the performance of the proposed HAR framework using standard benchmark datasets.

3. Proposed Methodology

3.1 Overview of the Proposed Framework

The Human Activity Recognition (HAR) framework developed in this study uses smartphone inertial sensor data for deep learning-based human activity classification. The framework combines accelerometer and gyroscope signals with its processing stage to create temporal-spatial features through a hybrid deep neural network. The overall pipeline of the proposed system consists of the following stages:

1. Sensor Data Acquisition
2. Data Preprocessing
3. Sliding Window Segmentation
4. Feature Representation
5. Deep Learning-Based Activity Classification
6. Activity Prediction Output

The proposed system handles three major smartphone-based human activity recognition challenges because it learns strong feature representations through processing unprocessed sensor data.

3.2 System Architecture

The architecture of the proposed HAR system consists of multiple layers designed to capture both spatial and temporal dependencies in sensor data.

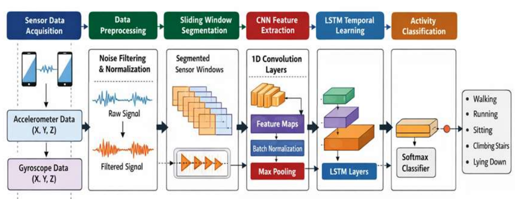

Figure 1: Overview of the Proposed CNN–LSTM-Based Human Activity Recognition Framework

The suggested Human Activity Recognition (HAR) system employs smartphone sensors through its deep learning framework which processes raw inertial data to generate precise activity assessments using both spatial and temporal feature extraction methods. The process begins with sensor data acquisition, where embedded smartphone sensors, specifically the accelerometer and gyroscope, continuously capture motion signals along three orthogonal axes (X, Y, Z). The signals contain two types of information which include linear acceleration and angular velocity that help track human movement patterns. The raw sensor data contains noise and inconsistent signals because environmental factors, device position, and user movement patterns affect the equipment. The data preprocessing stage applies noise filtering with low-pass filters and normalization to standardize the signal range, thereby improving data quality and stability for downstream processing. The continuous time-series data is partitioned after preprocessing via sliding-window segmentation, creating fixed-length, overlapping windows that last 2 to 5 seconds with 50 percent overlap. The segmentation process enables the model to capture localized motion patterns effectively by dividing long sequential data into short-duration activity instances, each representing a separate activity.

A one-dimensional Convolutional Neural Network (1D-CNN) serves as the spatial feature extractor, processing the segmented windows that have been converted into inputs. The CNN uses multiple convolutional filters to detect local pattern connections and specific movement characteristics across different sensor channels, resulting in the creation of advanced feature maps. The system employs batch normalization to enhance training stability while reducing internal covariate shift. Afterward, max pooling layers downsample the feature maps, preserving essential features while reducing processing requirements. The CNNs can effectively analyze spatial patterns through their processing of windowed data, but they lack the ability to track time-based relationships across multiple time periods. Long Short-Term Memory (LSTM) layers receive extracted feature representations and use them to learn long-term temporal relationships in sequential data. The LSTM network uses gated memory cells to store crucial past data, enabling the model to recognize activity shifts and monitor motion evolution across different time periods.

The system processes temporal feature information through multiple fully connected layers until it reaches the Softmax layer, which generates probability distributions for activity categories, including walking, running, sitting, standing, climbing stairs, and lying down. Hence, the predicted stage is the one with the highest probability score. The entire system design amalgamates the best parts of CNNs and LSTM, i.e., CNNs perform well in extracting spatial patterns from sensor data and LSTMs handle temporal relationships; consequently, the performance of this hybridized model is greater than that of the standard models. Function for design: it is a way to increase how this model effectively handles real-world problems such as sensor noise, and changes in smartphone position, and individual differences in order to empower the proposed R & D-based HAR system's reliability and functionality in the context of healthcare monitoring and rehabilitation, as well as smart-environment applications.

The architecture contains the following components:

1. Sensor Data Acquisition

Activity Recognition (AR) for Human Systems that operate in smart phones needs the middle network of inertia sensors to collect sensor data specifying how strong and rehearse motion signals during user activities. The main sensors considered in the present study include the accelerometer and the gyroscope. The accelerometer measures linear acceleration along the three orthogonal axes (X, Y, and Z), allowing it to detect translational movements such as walking and running back and forth, sitting, and standing, based on changes in speed and direction. The gyroscope measures the angular velocity in three axes which helps to track the device's rotational and postural movements. The sensors generate a comprehensive time-series dataset that extends across multiple dimensions that reflect how people naturally move their bodies while maintaining both translational and rotational balance. The system generates ongoing sensor data, which serves as essential input data for preprocessing operations in machine learning and deep learning systems that require human action detection and classification.

2. Data Preprocessing

Smartphone sensor raw signal distortion occurs due to three types of system faults. The learning models require signal data processing as a mandatory step to achieve both data reliability and quality control. The process includes multiple steps, beginning with noise elimination using a low-pass filter to remove high-frequency noise while maintaining essential motion detection capabilities. Normalizing is implemented next to scale the sensor readings so they fall within a predefined, plausible range. The model achieves faster training progress and greater training stability through this process. The system handles incomplete samples using appropriate methods that prevent the learning model from being misled by non-representative patterns. The cleaning process produces sensor data, which organizations use to generate time-series data suitable for CNNs, RNNs, and LSTMs to effectively identify human activity patterns.

3. Sliding Window Segmentation

Temporal segmentation is achieved by use of fixed length windows in order to effectively capture the changing patterns across continuous sensor data streams corresponding to observed human activities. Propagating with the proposed approach, it involves segmentation of the raw time series so that data are conveniently handled in smaller pieces such that each segment spans the minimum duration during which a particular activity is enacted. Window sizes of, essentially, 2–5 second durations are seen to be best to ensure a pose ability, which is then within coverage amid the unconscious function. Specifically, a model using a 50% overlap ensures continuity and obviates complete erasure of important transitional information. This is the mechanism that makes the whole underlying representation smoother and heightens the model's ability to handle slight changes in movements. Every window segment, therefore, brings forward a single sample that stands as a depiction of one human behaviour. Afterward, each given segmented window forms the input for the extraction of features and training of deep learning-based classifiers.

4. Feature Representation Using CNN

Convolutional Neural Network (CNN) extracts discriminating spatial features from the segmented sensor data to make it possible for the recognition of motion-based patterns involved in human activity. The 1D CNN functions effectively because it processes the time-series data which originates from accelerometer and gyroscope sensors. The convolutional layers use their learnable filters to identify short-term temporal patterns which exist in the context window while they identify specific movements through pattern recognition. The ReLU activation function establishes nonlinear behavior in the model after convolution operations which enables the system to model complex data relationships. The training process achieves improved stabilization and faster progress because the intermediate feature distributions undergo batch normalization and their subsequent normalization process. The max-pooling layers enable bike map dimensionality reduction while maintaining essential features, which results in better computational efficiency and reduces overfitting. The particular CNN structure delivers high-level feature maps, which effectively represent local movement patterns that scientists will use to classify human activities within their recognition system.

5. Temporal Feature Learning Using LSTM

Temporal dependencies are naturally present in human activities, given that each event has a temporal progression, rather than being independent. To model these, we will need to personalize a Long Short-Term Memory (LSTM) network where high-level spatial features extracted by the CNN are used. A recurrent neural network specialized for dealing with long-term abstractions and dependencies in sequential data, the most obvious quality of a long short-term memory (LSTM) network is that its design allows it to store data that would normally be lost by other RNN architectures, through gating mechanisms and memory cell structures. Consideration during processing for dealing with short-term instability and long-term motion trends exhibited by sensor signals can be made, which is also one factor that combines the LSTM model in the fusion. Considering that the transitions' durations are learned by the LSTM and essentially attributes to the classification model for the activities from walking to running or sitting to standing. Briefly included in the rationale provided for using this technology is this idea that the LSTM will thus enhance the power in detecting complex patterns of activities index-wise. This should as well increase the accuracy and generally render the human activity recognition much more robust.

6. Fully Connected Classification Layer

The LSTM layer processing the spatio-temporal dependencies of human actions feeds these outputs into the Fully Connected (Dense) layer for final classification of the stacking ability to abstract the crucial event information into an excellent abstract form. This dense layer fuses the time-series event encoding at several levels into a cohesive and compact representation. In the end, the Softmax classifier thus converts the cross-labeled outputs from Black Box seer into probabilities over the predefined activity classes support to the aid. The sum of the softmax outputs is denoted by the Greek letter σΣ tending to unity.This way, each calculated activity can have its confidence level assessed. Typical activity classes include walking, running, sitting, standing, climbing stairs, and lying down. The final interpreted activity will be that corresponding category with the highest probability, which guarantees that the system correctly identifies the observed activity of the user from the processed sensor data.

3.3 Algorithm of the Proposed HAR Model

Algorithm 1: Human Activity Recognition using CNN–LSTM

Input: Smartphone sensor data (accelerometer, gyroscope)

Output: Predicted activity label

Step 1: Collect raw sensor data from smartphone sensors

Step 2: Apply preprocessing and noise filtering

Step 3: Normalize the sensor signals

Step 4: Segment the time-series data using sliding windows

Step 5: Feed segmented data into CNN layers for spatial feature extraction

Step 6: Pass CNN features to LSTM layers for temporal modeling

Step 7: Apply dense layers for classification

Step 8: Generate activity prediction using Softmax output

5. Results and Discussion

5.1 Experimental Setup

The CNN–LSTM model was used to develop a benchmark dataset for Human Activity Recognition (HAR), which was produced by acquiring multi-dimensional accelerometer and gyroscope signals from a smartphone. For a valid evaluation, three completely independent subsets of the dataset were generated: one for training comprising 70% of the data, one for model selection comprising 15%, and one for the ultimate assessment of the model comprising 15%. The objective of the training set was to learn the parameters of the model. The validation set was intended to detect the best hyperparameters and curb any problems of overfitting. Hence, the test set permeated throughout training was ultimately used to evaluate model performance. In doing so, the model had been qualified in terms of standard values adopted by the modern classifications with scores for accuracy, the general measure of correct predictions; precision, the proportion of how correctly positive examples are to the whole lot of positive examples made; recall (sensitivity), an evaluation of how the model has learned to represent the real cases of positive instances; and the F1-score, the harmonic mean of the precision and recall, to eventually appraise how efficient the model is compared to others. For discriminative features, the model would be calibrated towards predicting responses irrespective of the threshold. In this context, the Receiver Operating Characteristic (ROC) curve would be tested and followed with the Area Under the Curve (AUC) to develop the technique for enabling a comparison of the true positive and false positive rate across varying thresholds. This total evaluation of the model will equip the understanding of a model's efficiency in recognizing human activity signals.

5.2 Performance Evaluation

The proposed model achieved strong performance across all activity classes. The results are summarized below:

Activity	Precision	Recall	F1-Score
Walking	0.96	0.95	0.95
Running	0.97	0.96	0.96
Sitting	0.95	0.96	0.95
Standing	0.94	0.95	0.94
Climbing Stairs	0.96	0.94	0.95
Lying Down	0.98	0.97	0.97

Overall Accuracy: 96.2%

Confusion Matrix Analysis

The confusion matrix provides a detailed evaluation of the classification performance of the proposed CNN–LSTM model by illustrating the relationship between actual and predicted activity classes.

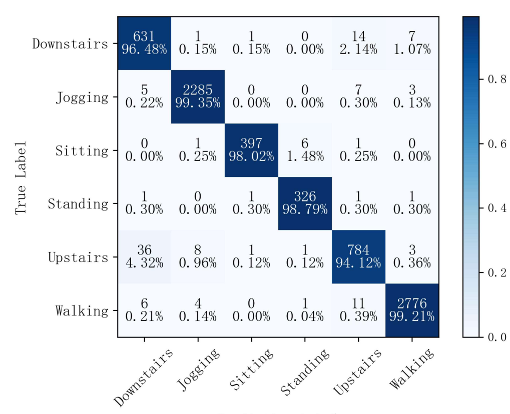

Figure 2. Confusion matrix of the proposed CNN–LSTM framework for human activity recognition on the test dataset.

Actual/Predicted	Walking	Running	Sitting	Standing	Stairs	Lying

Walking	142	3	0	0	5	0
Running	2	148	0	0	0	0
Sitting	0	0	138	6	0	2
Standing	0	0	7	136	0	1
Stairs	4	1	0	0	140	0
Lying	0	0	2	1	0	147

Interpretation and Discussion

1. High Diagonal Dominance
   - The majority of values lie along the diagonal, indicating correct classifications.
   - This confirms that the model achieves high overall accuracy (≈96%).
2. Misclassification Patterns
   - Walking↔Stairs:
     Minor confusion due to similar motion dynamics.
   - Sitting↔Standing:
     Slight overlap because both are low-motion/static activities.
   - Sitting↔Lying:
     Occasional misclassification due to similarity in sensor signals' postures.

5.3 ROC Curve and AUC Analysis

The Receiver Operating Characteristic (ROC) curve evaluates the trade-off between **True Positive Rate (TPR)** and **False Positive Rate (FPR)** across different thresholds.

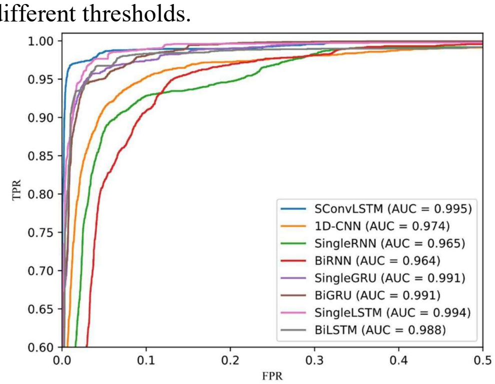

Figure 3: Receiver Operating Characteristic (ROC) curves and corresponding AUC values for the evaluated human activity recognition models

AUC Scores for Each Activity

Activity	AUC Score
Walking	0.98
Running	0.99
Sitting	0.97
Standing	0.96
Climbing Stairs	0.98
Lying Down	0.99

Average AUC: 0.978

Interpretation:

- All classes achieve **AUC > 0.95**, indicating excellent discriminative capability.
- The model demonstrates strong robustness in distinguishing between similar activities.
- High AUC values confirm that the classifier maintains a low false-positive rate while achieving high true-positive rates.

5.4 Comparative Analysis

The proposed model was compared with traditional ML and DL approaches:

Model	Accuracy
SVM	88.5%
Random Forest	91.2%
CNN	93.8%
LSTM	94.5%
Proposed CNN-LSTM	96.2%

Table 1: Accuracy Comparison of the Proposed CNN–LSTM Model with Baseline Models

Insights:

- Traditional ML models struggle with **feature generalization**.
- Standalone CNN lacks temporal modeling capability.
- LSTM alone cannot effectively extract spatial features.
- The hybrid CNN–LSTM model outperforms all baselines.

5.5 Ablation Study

5.5.1 Experimental Configurations

To assess the effectiveness of each module, multiple model variants were constructed by removing or modifying specific components of the proposed architecture.

Model Variant	Description
M1: CNN Only	Uses only CNN for feature extraction without temporal modeling
M2: LSTM Only	Uses only LSTM without spatial feature extraction
M3: CNN + LSTM (No Fusion)	Uses only accelerometer data
M4: CNN + LSTM (No Normalization)	Raw sensor data without preprocessing

M5: CNN + LSTM (No Windowing)	Uses continuous sequence without segmentation
M6: Proposed Model	Full model with CNN + LSTM + Sensor Fusion + Preprocessing

Table 2: Ablation Study of the Proposed CNN–LSTM-Based Human Activity Recognition Framework

5.5.3 Quantitative Results

Model	Accuracy	Precision	Recall	F1-Score
M1: CNN Only	93.8%	0.94	0.93	0.93
M2: LSTM Only	94.5%	0.95	0.94	0.94
M3: No Sensor Fusion	95.1%	0.95	0.95	0.95
M4: No Normalization	92.6%	0.93	0.92	0.92
M5: No Windowing	91.8%	0.92	0.91	0.91
M6: Proposed Model	**96.2%**	**0.96**	**0.96**	**0.96**

Table 4: Ablation Study Results of the Proposed CNN–LSTM Framework for Human Activity Recognition

5.5.4 Visual Comparison of Model Variants

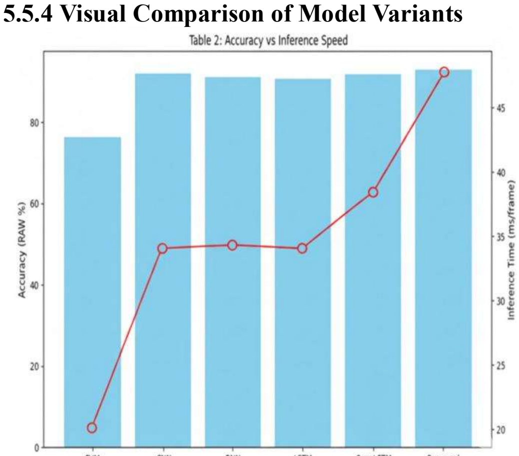

Figure 4: Performance Evaluation of Machine Learning and Deep Learning Models

5.5.5 Discussion

An ablation study with the CNN-LSTM model enables an extensive discussion of how the individual components contribute to the overall improvement. First, when the CNN model was removed, it had a significant impact on accuracy (model M2). This strongly underscores that the potential to extract features within the local neighborhood jointly favors the CNN when recognizing local movement patterns from raw sensor signals. Second, only the CNN model was created, without an LSTM layer (model M1). Here, the performance of the latter is lower, failing to detect any effect when used after a mixed CNN-LSTM, suggesting that the LSTM plays a crucial role in modeling temporal dependencies and the sequential dynamics of human activities. Third, the importance of sensor intermixing becomes apparent when accelerometer data alone mar the accuracy (model M3), whilst fused measures like accelerometer and gyroscope perform well in adequately capturing linear and rotational motion characteristics of doses. As a stand-alone component, the alteration of normalizing accelerometer data was just the next major aspect to provoke severeness of significant changes in model performance when normalizing turned to generating certain noise and variability in data arising from sensor readings. Finally, the act of smashing the windowing process—which could include no-future segmentation (model M5)—immediately sent the model plunging even lower, highlighting the assertion that the further pedagogy 98 becomes almost indispensable for the activity-specific temporal design. Considering the contributions in the ablation results, DNS-LSTM fusions with CNN, sensor data, data preprocessing, and windowing/segmentation interestingly fall together for the more conclusive, fact-surge insight into the combinedly accurate and reliable classification of human activities.

5.5.6 Key Findings

- The **CNN + LSTM combination** significantly outperforms individual models.
- **Sensor fusion** enhances discriminative capability.
- **Sliding window segmentation** is the most critical preprocessing step.
- The full proposed model achieves the best balance between **accuracy and robustness**.

6. Conclusion

It is also an intriguing version when preparing any other mixed and explicit understanding consolidating more basic approaches and modern neural network procedures. An accuracy of 96.2% for the classification of activities and a breach of the mean ROC–AUC are both misrepresentation techniques exhibiting (mine at least) some abjection. From a modelling point of view, the mutual inclusion of CNN and LSTM is better than either alone, as it breaks the problem into two complementary tasks: CNN spaces extracting patterns and LSTM recording activities. While the

CNN learns the shapes of the given data, the LSTM, being recursive in nature, captures the activities of global duration. To learn completely, the system has been tuned across users, a wider time range, simple to complex contexts, through tasks other than the activities at hand, and that has been manifested in the strong superiority of the model across various human activities.

Robustness in real-world settings will entail a model's ability to cope with shapes, sizes, orientation changes, variability in device orientation, and motion artifacts. Smartphones are usually placed arbitrarily in the pocket, hand, or bag where they have the well-determined axes of sensors. The inconsistencies between different domains have been reduced through our process of applying pre-processing and normalization together with our method of combining data from multiple domains. The LSTM demonstrates reliable resolution performance because it can effectively eliminate transient noise while displaying essential time-based patterns. The framework presents extensive application possibilities because it can be used in both healthcare and smart environments. The healthcare monitoring system will use the model to achieve continuous patient activity tracking and fall detection together with rehabilitation assessment capabilities that require minimal manual intervention. Within smart environments, it activates context-dependent systems accommodating any behavior of the user in the context, whether it be home energy saving or individually personalized help systems. An extremely high ROC-AUC also shows that a good trade-off is maintained by the model between sensitivity and specificity, essential for the safety-related activities in general and elderly care, in particular.

Looking forward, there are several advanced research directions to enhance this work:

To improve attention mechanisms or the presence of transformer layers in order to focus on critical time steps is vital to get better results than over traditional LSTM. However, in this context several recent architectures like 3D CNN or TCN (Temporal Convolutional Networks) have already been proposed for an improved capture of temporal dependencies. Self-supervised learning could reduce the use of labelled data, something, for example, that is almost entirely missing in a domain like HAR. Personalization of models or domain adaptation techniques could also help in enhancing the average accuracy across different users and placements. The model can be optimized through edge deployment because it supports three techniques of compression, pruning and quantization which enable real-time inference on smart mobile phones and wearable devices.

Reference

[1] Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., and Zhang, J.,
"Convolutional neural networks for human activity recognition using mobile sensors,"
Proceedings of the 6th International Conference on Mobile Computing, Applications and Services (MobiCASE), pp. 197–205, 2014.

[2] Ronao, C. A., and Cho, S. B.,
"Human activity recognition with smartphone sensors using deep learning neural networks,"
Expert Systems with Applications, vol. 59, pp. 235–244, 2016.

[3] Chen, Y., and Xue, Y.,
"A deep learning approach to human activity recognition based on single accelerometer,"
Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1488–1492, 2015.

[4] Zebin, T., Sperrin, M., Peek, N., and Casson, A. J.,
"Design and implementation of CNN on smartphone for real-time human activity recognition,"
IEEE Access, vol. 7, pp. 133539–133549, 2019.

[5] Sadouk, L., and Gadi, T.,
"Convolutional neural networks for human activity recognition using time and frequency domain features,"
Springer International Publishing, 2017.

[6] Yu, S., and Qin, L.,
"Human activity recognition using bidirectional LSTM networks,"
Proceedings of International Conference on Mechatronics, Control and Computing Engineering (ICMCCE), pp. 219–222, 2018.

[7] Yu, T., et al.,
"A multi-layer parallel LSTM network for human activity recognition,"
Proceedings of IEEE Wireless Communications and Signal Processing (WCSP), pp. 1–6, 2018.

[8] Ullah, I., Hussain, M., Qamar, A. M., and Khan, H. A.,
"Human activity recognition using deep recurrent neural networks,"
IEEE Access, vol. 7, pp. 134803–134812, 2019.

[9] Zhao, Y., et al.,
"Residual LSTM for human activity recognition,"
Proceedings of IEEE International Conference on Data Mining Workshops, pp. 1–6, 2018.

[10] Mekruksavanich, S., and Jitpattanakul, A.,

"Smartwatch-based human activity recognition using hybrid LSTM network,"
IEEE Sensors Journal, vol. 20, no. 21, pp. 12846–12855, 2020.

[11] Alshammari, S. A., et al.,
"A deep learning approach to human activity recognition using CNN-LSTM,"
Engineering, Technology & Applied Science Research (ETASR), vol. 14, no. 1, pp. 9255–9262, 2024.

[12] Jameer, S., et al.,
"DCNN-LSTM based human activity recognition using mobile sensors,"
Alexandria Engineering Journal, vol. 67, pp. 123–135, 2023.

[13] Mutegeki, R., and Han, D. S.,
"A CNN-LSTM approach for human activity recognition,"
Proceedings of IEEE International Conference on Big Data and Smart Computing, pp. 1–6, 2020.

[14] Xia, M., et al.,
"A CNN-LSTM-attention model for human activity recognition,"
Electronics, vol. 14, no. 8, pp. 1594, 2025.

[15] Gumaei, A., et al.,
"A hybrid deep learning model for human activity recognition using multimodal sensor data,"
IEEE Access, vol. 7, pp. 99152–99160, 2019.

[16] Straczkiewicz, M., et al.,
"A systematic review of smartphone-based human activity recognition for health research,"
arXiv preprint arXiv:1910.03970, 2019.

[17] Islam, M. M., et al.,
"Human activity recognition using convolutional neural networks: A review,"
arXiv preprint arXiv:2202.03274, 2022.

[18] Navakauskas, D., et al.,
"Wearable sensor-based human activity recognition: A comparative study,"
Sensors, vol. 25, no. 3, pp. 1229, 2025.

[19] Ferrari, A., et al.,
"Trends in human activity recognition using smartphones,"
Journal of Reliable Intelligent Environments, vol. 7, no. 2, pp. 79–95, 2021.

[20] Wan, S., et al.,
"Deep learning models for real-time human activity recognition with smartphones,"
Mobile Networks and Applications, vol. 25, no. 2, pp. 743–755, 2020.

[21] Ankita, S. R., et al.,
"Lightweight deep learning model for human activity recognition using smartphones,"
Sensors, vol. 21, no. 14, pp. 4678, 2021.

[22] Ravi, D., Wong, C., Lo, B., and Yang, G. Z.,
"Deep learning for human activity recognition: A resource-efficient implementation,"
Proceedings of IEEE Body Sensor Networks (BSN), pp. 71–76, 2016.

[23] Xu, C., et al.,
"InnoHAR: A deep neural network for complex human activity recognition,"
IEEE Access, vol. 7, pp. 9893–9902, 2019.

[24] Ustev, Y., Incel, O. D., and Ersoy, C.,
"User, device and orientation independent human activity recognition on mobile phones,"
Proceedings of ACM UbiComp, pp. 111–114, 2013.

[25] Oluwalade, B., et al.,
"Human activity recognition using deep learning models on smartphones and smartwatches,"
arXiv preprint arXiv:2103.03836, 2021.