International Journal of Scientific Research in Computer Science and Engineering
Vol.**12**, Issue.**6**, pp.**66-74**, December **2024**
E-ISSN: 2320-7639
Available online at: www.isroset.org

Research Article

A Deep Learning-based Hybrid CNN-LSTM Model for Human Activity Recognition

Damilola Akinola[1*], Adenike Oluyemisi Oyedemi[2], Micheal Olalekan Ajinaja[3]

[1]Dept. of Management Information System/College of Business, Bowie State University, Maryland, United State of America
[2]Dept. of Computer Science, University of Ilesa, Ilesa, Nigeria
[3]Dept. of Computer Science, Federal Polytechnic Ile Oluji, Ile Oluji, Nigeria

**Corresponding Author: akinoladamilola66@gmail.com*

Received: 27/Oct/2024; **Accepted**: 29/Nov/2024; **Published**: 31/Dec/2024

Abstract — Human Activity Recognition (HAR) involves identifying and classifying physical activities performed by individuals using data collected from sensors like accelerometers, gyroscopes, or cameras. HAR has broad applications in healthcare, fitness monitoring, and human-computer interaction, where accurate activity recognition can enhance user experiences and provide actionable insights. Despite progress in standalone deep learning models, limitations persist in capturing both spatial and temporal dependencies effectively. To address this, we developed a hybrid deep learning model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. CNNs were employed to extract spatial features from multivariate time-series sensor data, while LSTMs captured temporal dependencies crucial for accurate classification. The UCI HAR dataset, consisting of six activity labels, was used to benchmark the model. The implementation was carried out in Python, leveraging libraries like TensorFlow and Keras. The proposed hybrid CNN-LSTM model achieved an overall accuracy of 94%, with precision, recall, and F1-scores (macro and weighted averages) also reaching 94%. Individual activity labels recorded F1-scores ranging from 85% to 100%, demonstrating the model's robustness across diverse activities. These findings validate the effectiveness of the hybrid CNN-LSTM architecture in overcoming the limitations of standalone models. The ability to capture both spatial and temporal patterns in sensor data underscores the model's potential in advancing HAR applications. This study provides a foundation for future research in refining hybrid approaches, exploring additional datasets, and deploying such models in real-world applications. The results have significant implications for improving healthcare monitoring, fitness tracking, and human-computer interaction systems.

Keywords — human activity recognition, CNN-LSTM, hybrid deep learning, multivariate time-series, temporal dependencies, spatial feature extraction

1. Introduction

Human Activity Recognition has emerged as one of the key research areas, which allows the classification and analysis of physical activities using data captured by sensors such as accelerometers, gyroscopes, and cameras. This area offers a great potentiality in healthcare, fitness monitoring, human-computer interaction, and ambient-assisted living, where an accurate identification of activities may drive innovations in personalized applications and services [1]. HAR systems face challenges due to the complex and variable nature of human motion and the high dimensionality of sensor data. Effective HAR requires robust algorithms that accurately capture spatial and temporal features within activity data. Traditional machine learning methods, such as Support Vector Machines (SVM) and k-Nearest Neighbors (KNN), have been widely employed for HAR tasks. While these models achieved considerable success, their dependency on handcrafted feature extraction limits their scalability and generalization across diverse datasets [2]. In contrast, deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated superior performance by automating feature extraction and capitalizing on the hierarchical structure of data representations [3]. However, standalone CNNs and RNNs each have inherent limitations. While CNNs represent an outstanding performance in image-spatial feature extraction, their architecture lacks any effective way of modeling temporal dependencies. In contrast, RNNs and their advanced improvement by Long Short-Term Memory have outstanding performance on temporal pattern mining while having problems with complexities over spatial features.

In the recent years, to mitigate these limitations, several hybrid architectures have been developed that combine CNNs and LSTMs. These models combine the best of both - CNNs which is good at identifying patterns in spatial features, while

LSTMs excel at understanding how things change over time (temporal modeling). By using both, these models achieve better results in recognizing human activities [4]. Despite promising results, gaps remain regarding how to optimally adapt these hybrid models to multivariate time-series datasets, which are quite common in HAR applications. For example, sensor noise, activity overlap, and imbalanced datasets are some of the challenges that require further research and innovation. This paper aims to address the challenges in Human Activity Recognition (HAR) by introducing a new model that combines the strengths of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). The CNN part of the model analyzes the raw sensor data to identify important spatial patterns. The LSTM part then takes this information and learns how these patterns change over time, which is crucial for correctly classifying different activities. To test this new model, the researchers used the UCI HAR dataset, a well-known and widely used dataset in this field. This dataset is chosen because it includes a diverse range of real-world activities and data from various sensors, making it a good benchmark for evaluating the model's performance. The overarching objective of this study is to enhance classification accuracy while ensuring robustness across diverse activity types. By addressing the limitations of standalone models and optimizing the hybrid architecture, this research contributes to advancing HAR methodologies. These findings have a great impact on applications in healthcare, fitness tracking, and human-computer interaction, thus providing a basis for further research in this area.

The HAR dataset used in this study was created by recording 30 individuals performing everyday activities like walking, climbing stairs, and resting. These participants, aged 19-48, wore smartphones equipped with motion sensors (accelerometer and gyroscope) on their waists. The sensors captured detailed movement data, including acceleration and rotation, at a high rate of 50 times per second. To ensure accuracy, the recordings were accompanied by video footage, which was then manually reviewed to label each activity. This dataset, publicly available on [5], is diverse and well-suited for research due to its inclusion of various individuals and activities. This paper is organized as follows: Section 2 provides a review of relevant literature. Section 3 outlines the research methodology, including data sources, text preprocessing techniques, and the model architecture. Section 4 presents the results of the study and a discussion of their significance. Finally, Section 5 concludes the paper with a summary of the findings.

2. Related Works

HAR has been a subject of extensive research, with both traditional machine learning and more recent deep learning techniques being explored. While traditional methods played a crucial role in early HAR research, advancements in deep learning have led to significant improvements in terms of accuracy and reliability. This section presents a comprehensive review of existing research, examining their methodologies, underlying principles, and limitations. Additionally, it emphasizes how our proposed hybrid CNN-LSTM model overcomes some of these limitations.

[6] investigated the use of SVM and Decision Trees (DT) for activity recognition using smartphone sensors. The study relied on handcrafted feature extraction to process raw accelerometer and gyroscope data, which were then fed into the models for classification. While the approach provided interpretable results and achieved reasonable accuracy, its dependence on manual feature engineering limited its scalability and generalization to other datasets. Additionally, these models struggled with complex time-series patterns, such as overlapping activities. The hybrid CNN-LSTM model overcomes these issues by automating feature extraction and learning hierarchical representations, enabling better handling of high-dimensional, multivariate data.

HAR systems often rely on wearable sensors, which can be challenging for elderly individuals. To address this limitation, [7] proposed a vision-based approach using CCTV footage and camera images. Their system employed a HAR feature-based classifier to detect human poses in images and a Convolutional Neural Network (CNN) to recognize activities. Our research utilized a custom dataset of 5,648 images to train a similar HAR system. The results were promising, achieving a high detection accuracy of 99.86% and a recognition accuracy of 99.82%. Moreover, the system demonstrated real-time performance, processing approximately 22 frames per second after 20 training epochs.

In their work, [8] explored the recognition of complex human activities using a high-performance 1D - CNN model. HAR is a significant area of research with diverse applications in healthcare, social sciences, and human-computer interaction. Many human activities are complex and require monitoring to improve well-being, quality of life, and health. This study proposed a novel approach for recognizing complex HAR using a 1D - CNN model trained on data collected from a tri-axis accelerometer sensor embedded in a smartwatch. The dataset comprises three complex activities: studying, playing games, and mobile scrolling. 1D CNNs offer a compelling combination of high accuracy and reduced computational complexity for HAR tasks. The proposed model was trained and optimized using Python on a self-curated dataset, achieving an impressive accuracy of 98.28%. These promising preliminary results demonstrate the effectiveness of the 1D CNN model for recognizing these target activities and provide a strong foundation for further advancements in the field of HAR.

In their research, the authors in [9] proposed a novel approach for recognizing human activities using a one-dimensional Convolutional Neural Network (1D CNN). This method leverages triaxial accelerometer data collected from users' smartphones. The study focused on three distinct activities: walking, running, and staying still. To extract relevant features, the x, y, and z acceleration values were combined into a single vector magnitude, which served as the input for the 1D CNN model. The proposed 1D CNN model demonstrated superior performance, achieving an accuracy of

92.71% in recognizing these activities, surpassing the baseline random forest approach which attained an accuracy of 89.10%.

[10] pioneered the use of a hybrid CNN-RNN architecture for HAR, combining the strengths of CNNs for spatial feature extraction and RNNs for capturing temporal dependencies within multivariate sensor data. While this hybrid approach showed promising results with significant accuracy improvements over standalone models, it employed basic RNNs, limiting its ability to effectively handle long-term dependencies. Furthermore, the model's performance was impacted by imbalanced datasets, where certain activity classes were underrepresented. This study addresses these limitations by proposing a novel hybrid CNN-LSTM model. By incorporating Long Short-Term Memory (LSTM) networks instead of basic RNNs, the model enhances its ability to capture and learn from long-term dependencies within the sensor data. Additionally, the study employs data augmentation techniques to mitigate the impact of class imbalance, thereby improving the overall robustness and generalizability of the model.

Smartphones, with their increasing sophistication and widespread adoption, have become powerful platforms for activity recognition. This has led to the development of numerous applications aimed at monitoring daily routines and fostering healthier lifestyles. A key challenge in this domain lies in developing efficient methods for accurately recognizing various physical activities, such as walking, jogging, and sitting. In their work, [11] proposed a novel approach utilizing a Convolutional Neural Network (CNN) to identify human activities based on data collected from the three-axis accelerometer embedded in smartphones. The study focused on recognizing a range of activities, including walking, jogging, sitting, standing, and ascending/descending stairs. A unique aspect of their approach is the direct utilization of raw, three-dimensional accelerometer data as input for the CNN, eliminating the need for complex preprocessing steps. The CNN-based model demonstrated exceptional performance, achieving an impressive 91.97% accuracy in recognizing these activities. This significantly outperformed the traditional Support Vector Machine (SVM) approach, which achieved an accuracy of 82.27% when using six manually extracted features from the raw accelerometer data. These findings highlight the effectiveness of the proposed CNN model in delivering high recognition accuracy while minimizing computational overhead.

[12] proposed a multimodal HAR system that combined data from multiple sensor types, such as accelerometers, gyroscopes, and magnetometers. The study demonstrated that multimodal approaches could improve classification accuracy by using corresponding information from diverse sensors. However, the integration of multiple sensor modalities required extensive preprocessing and synchronization, complicating the implementation. The proposed hybrid CNN-LSTM model simplifies the architecture by focusing on efficiently extracting and processing features from single-sensor multivariate data, providing a balance between simplicity and performance.

These studies reviewed, therefore, indicate a great deal of progress in the HAR area using both machine learning and deep learning techniques. Traditional approaches require handcrafted features, which can limit scalability and performance. In such a way, the proposed CNN-LSTM hybrid model ensures strong feature extraction on any given dataset. Another is Temporal Modeling. Standalone CNNs can't capture temporal dependencies that are present in sequential activities. By incorporating LSTMs, the proposed model effectively handles temporal dynamics. Other limitations include a limited ability for spatial feature extraction: Standalone LSTM models are not optimized for spatial feature extraction, which is quite critical for multivariate sensor data. The hybrid model uses CNNs to address this gap. Many of the existing models suffer in the case of rare activity classes. The hybrid model uses data augmentation and class weighting to mitigate this issue. Most of the attention-based and multimodal models are challenging to deploy in real time. The proposed hybrid architecture balances the trade-off between complexity and performance, making it appropriate for real-time applications. The hybrid CNN-LSTM model addresses these limitations and therefore provides a complete solution for a HAR task, paving the path toward more accurate, robust, and efficient activity recognition systems.

3. Methodology

In the next section, we will discuss four important methods used in our HAR system

3.1 Data Source and Splitting

The HAR dataset used in this study was collected from 30 volunteers aged 19-48 performing six daily activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) while carrying a waist-mounted Samsung Galaxy S II. The smartphone's embedded accelerometer and gyroscope captured 3-axis linear acceleration and angular velocity at a 50Hz sampling rate. Video recordings were used for manual data labeling. This publicly available dataset was randomly divided into training (70%) and testing (30%) sets. The raw sensor data (accelerometer and gyroscope) underwent pre-processing, including noise filtering. Subsequently, the data was segmented into 2.56-second windows with a 50% overlap, resulting in 128 readings per window. To isolate body acceleration from the acceleration signal, which includes both body motion and gravitational components, a Butterworth low-pass filter with a cutoff frequency of 0.3 Hz was applied, assuming gravitational forces primarily reside in low-frequency bands. Finally, a feature vector was extracted from each window, encompassing both time-domain and frequency-domain features.

3.2 Data Preprocessing

After loading the dataset, the next phase is data preprocessing. The results inf Figure 1 indicate that the dataset used for the study is clean and does not contain any problematic records,

which is an excellent starting point for training and testing your machine learning model. The duplicates in train = 0 and duplicates in test = 0 indicates that there are no duplicate rows in either the training or testing datasets. Duplicate rows could arise from repeated data entries, which can lead to bias in the model. Their absence ensures that the model is trained and evaluated on unique, diverse data points, enhancing the reliability of its performance. While the invalid values in Train = 0, and invalid values in test = 0

```
Duplicates in train = 0
Duplicates in test = 0
Invalid values in train = 0
Invalid values in test = 0
```

Figure 1. Checking for Duplicates and Nulls

"Invalid values" refer to missing or inappropriate entries in the dataset, such as NaN, null values, or data entries that do not make sense in the context of the problem (e.g., negative values for an attribute that should only be positive). The absence of invalid values confirms that the dataset is complete and well-prepared, with no gaps or errors that might adversely affect model training or testing. A clean dataset minimizes the risk of the model learning spurious patterns or being negatively impacted by anomalies, leading to more accurate and reliable predictions. It ensures that the model's performance during training and testing reflects its true capability, as there are no artifacts or errors to skew results and saves preprocessing time that would otherwise be spent cleaning the data.

Figure 2 shows visualization of user-provided data within the training set, grouped by activity labels. Each bar represents the distribution of activities such as Standing, Sitting, Laying, Walking, Walking Downstairs, and Walking Upstairs for different users. The activities are relatively well-distributed across the users, indicating that the dataset captures diverse activity patterns across different individuals. However, slight variations in bar heights suggest that some activities may have been performed more frequently than others by certain users.

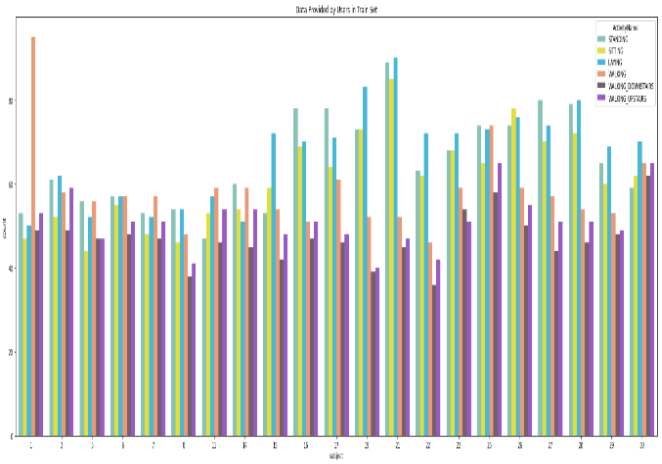

Figure 2. Data Provided by Users in Train Set

While most activities seem to have balanced contributions across users, a few peaks (taller bars) may reflect overrepresentation of specific activities for some individuals, potentially affecting model training. The variability in the height of bars for different users highlights inter-user differences in activity patterns. This is important for assessing the robustness of the HAR model when applied to new users. This balanced yet slightly variable distribution supports the generalizability of the hybrid CNN-LSTM model but underscores the need to mitigate biases caused by overrepresented activities or users.

Figure 3 illustrates the distribution of activity data contributed by users in the test set, categorized into activities or events such as walking upstairs, sitting, and standing. Compared to the training set, the test set shows a more noticeable variation in activity representation among users. Some activities, such as Sitting and Laying, appear to be overrepresented for certain users, while others like Walking Downstairs seem less frequent overall.

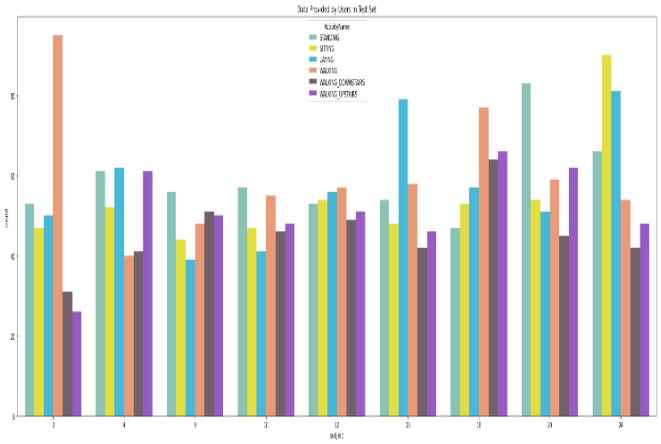

Figure 3. Data Provided by Users in Test Set

The observed variation in the number of instances for different activities across the test set indicates potential class imbalances. This uneven distribution raises concerns about the model's generalization performance, as underrepresented activities may be less accurately predicted. While the diversity of activity distribution across users is crucial for assessing the robustness of the hybrid CNN-LSTM model, it necessitates careful consideration of performance metrics to ensure fair evaluation of all activity classes.

Figure 4 displays the total count of activities available in the training set, where each activity is represented by its total occurrence. The activity distribution in the training set is relatively balanced, with similar counts across most activities. However, slight differences suggest that certain activities have been performed more frequently, which could influence the training process by biasing the model toward these activities. This overall balance supports the creation of a robust model capable of generalizing well to different activity classes, though minor adjustments might still be needed for underrepresented activities to ensure uniform performance.

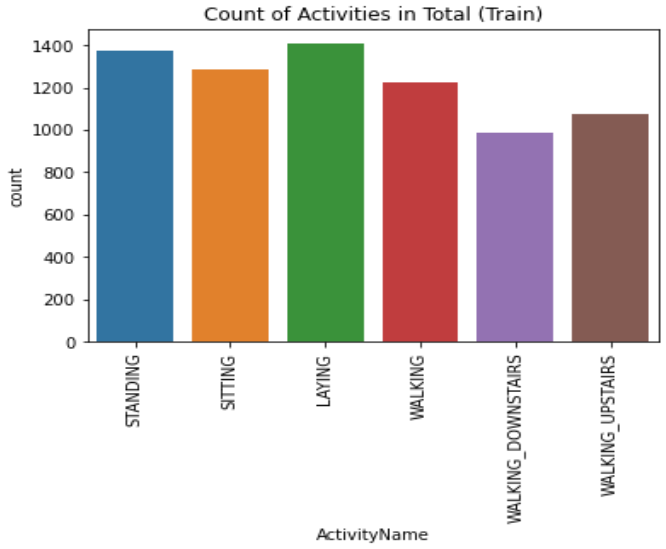

Figure 4. Count of Activities in Total (Train)

The bar chart in Figure 5 illustrates the distribution of activities performed in the test dataset. The activity "STANDING" has the highest count, while "WALKING_DOWNSTAIRS" shows the lowest. The activities are generally well-represented, indicating a balanced dataset for most activities, though slight variations in count may affect model performance, particularly for underrepresented activities. This balance suggests sufficient data diversity for robust activity recognition, though optimization might focus on improving accuracy for less frequent activities.

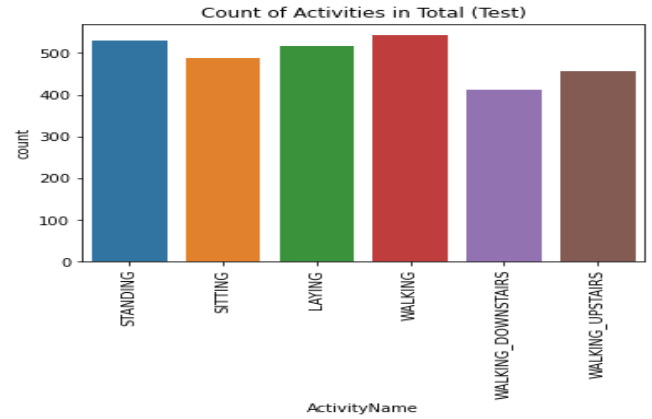

Figure 5. Count of Activities in Total (Test)

3.3 Method 1 - Convolutional Neural Networks

CNNs are a class of deep learning models specifically designed to process grid-like data, such as images or time-series data. In the context of HAR, CNNs excel at extracting spatial features from multi-sensor data, including accelerometer and gyroscope readings, by applying convolutional operations to the input data. Given an input sensor data matrix $X \in R^{n \times m}$ (where n is the time dimension and m can be represented as the number of sensor channels), a kernel $K \in R^{k \times m}$ (where k is the kernel size) slides over the input, performing element-wise multiplications followed by summation. The output feature map Y is mathematically represented as:

$$Y(i, j) = \sum_{p=0}^{k-1} \sum_{q=0}^{m-1} X(i+p, q).K(p, q) \quad (1)$$

This operation extracts local patterns from the sensor data, such as transitions between activities. After convolution, an activation function f, typically $ReLU$, is applied to introduce non-linearity:

$$Y'(i, j) = f(Y(i, j)) = \max(0, Y(i, j)) \quad (2)$$

Pooling layers (e.g., max-pooling) decrease the spatial dimensions of feature maps, while retaining the most prominent features while reducing computational complexity:

$$P(i, j) = max_{p=0}^{s-1}\ Y'(i+p, j:j+s) \quad (3)$$

Where s is the pooling size. Flattened feature maps are passed to fully connected layers for classification, producing activity probabilities using a *softmax* function:

$$\ddot{Y}_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \quad (4)$$

where z is represented as the output of the last dense layer, and C is known as the number of activity classes. CNNs excel in extracting spatial correlations between sensor channels, identifying unique patterns for different activities (e.g., walking vs. running).

3.4 Method 2 – Long Short-Term Memory

LSTM networks are a specialized type of RNN specifically designed to effectively capture long-range dependencies within sequential data. In the context of HAR, LSTMs excel at analyzing temporal patterns in sensor data streams to accurately classify activities that unfold over multiple time steps. The core of an LSTM unit comprises of three key gates: the input gate, the forget gate, and the output gate, which meticulously control the flow of information within the network. Let x_t be the input vector at time t, h_{t-1} the hidden state from the previous time step, and C_{t-1} the cell state. The forget gate determines which information to remove from the previous cell state:

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (5)$$

where W_f and b_f are the weight matrix and bias, respectively, and σ is the sigmoid activation function. The input gate decides which new information to store:

$$i_t = \sigma(W_f . [h_{t-1}, x_t] + b_i) \quad (6)$$

Updates the cell state with candidate values:

$$\hat{C}_t = \tanh(W_C . [h_{t-1}, x_t] + b_C) \quad (7)$$

Cell state update combines the previous cell state and new candidate values:

$$C_t = f_t \Theta C_{t-1} + i_t \Theta C_t \quad (8)$$

Output gate computes the new hidden state based on the updated cell state:

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \Theta \tanh(C_t) \quad (10)$$

LSTMs are particularly effective in modeling temporal dependencies, such as recognizing sequences of movements

(e.g., standing to walking). However, they may struggle to extract spatial correlations within multivariate sensor data, limiting their standalone effectiveness for HAR tasks.

3.5 Method 3 – Proposed Hybrid Model

The hybrid CNN-LSTM architecture uses the strengths of both CNN and LSTM. CNNs are basically used to extract spatial features from multivariate sensor data, and these features are fed as input into LSTM layers to capture temporal dependencies for activity classification. For the input representation,

Sensor data $X \in R^{n \times m}$ (time steps n, sensor channels m) is fed into CNN layers for spatial feature extraction. To do CNN feature extraction, the output from the CNN layers is a high-dimensional feature map $F \in R^{n' \times d}$, where n′ is represented as the reduced time dimension after pooling, and d is the number of features. The next phase is the LSTM Temporal Modeling where the feature map F is passed to LSTM layers to model temporal dependencies:

$$h_t = \text{LSTM}(F_t, h_{t-1}, C_{t-1}) \quad (11)$$

For the classification layer, the final hidden state h_T (at the last time step T) is passed to fully connected layers for classification, producing activity probabilities using softmax:

$$\ddot{y} = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \quad (12)$$

Hybrid CNN-LSTM architectures exhibit a significant advantage in terms of feature representation. By leveraging the spatial feature extraction capabilities of CNNs and integrating them with the temporal modeling capabilities of LSTMs, these hybrid models effectively address the limitations inherent in using CNNs or LSTMs independently. This synergistic approach results in a more comprehensive and informative representation of the underlying activity patterns. Also, the architecture efficiently handles the high-dimensionality of sensor data while combining spatial and temporal modeling ensures higher accuracy, particularly for complex activities involving transitions. Mathematically, given input X, CNN output: F = CNN(X), LSTM output: h_T = LSTM(F), the hybrid activity probabilities can be represented as given input X:

$$\ddot{y}_i = \frac{\exp(W_h . h_t + b_h)}{\sum_{j=1}^{C} \exp(W_h . h_t + b_h)} \quad (13)$$

This hybrid approach when implemented should demonstrate superior performance on benchmark datasets, achieving high accuracy and robustness in classifying diverse activities. The architecture of the Hybrid CNN-LSTM Model for HAR shown in Figure 4 visualizes the flow of data:

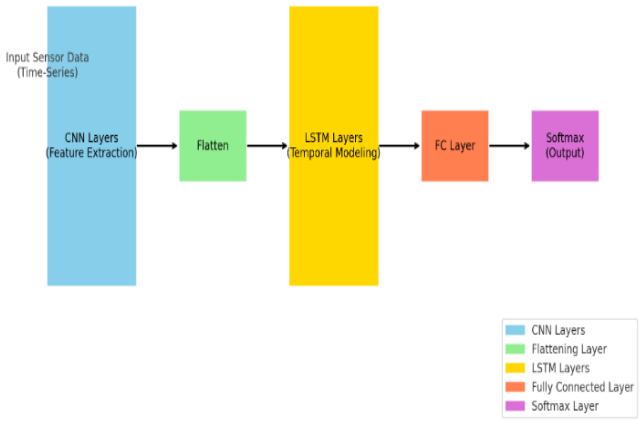

Figure 5. Architecture of the system

The input sensor data takes the raw multivariate time-series data collected from sensors and the CNN layers then extract spatial features from the input data by identifying patterns across sensor channels. The next phase is the flattening Layer that converts the CNN feature maps into a 1-dimensional vector for temporal processing. The LSTM Layers captures temporal dependencies in the sequence of features for activity classification, then the fully connected (FC) layer processes the temporal features and outputs activity class scores. Lastly, softmax Layer converts the class scores into probabilities for final classification. This architecture combines the strong point of CNNs for spatial analysis and LSTMs for temporal modeling, offering robust performance for HAR tasks.

4. Discussion of Findings

The classification report offers a comprehensive evaluation of the hybrid CNN-LSTM model's performance. It provides detailed metrics, including precision, recall, F1-score, and accuracy, for each activity class and overall. These metrics are essential for a thorough understanding of the model's performance on both individual activities and across the entire dataset.

4.1 Evaluation Metrics

Table 1 summarizes the evaluation metrics employed in this research, along with their corresponding values. Precision, one of the key metrics, measures the proportion of true positive predictions among all instances predicted as positive. Essentially, it quantifies the model's ability to minimize false positive classifications.

$$\text{Precision} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \quad (14)$$

True Positives (TP) represent instances where the model correctly identifies a class, while False Positives (FP) occur when the model incorrectly classifies an instance as belonging to a particular class. A high precision value (approximately 1.00) for class 0 indicates a very low rate of false positives, suggesting that the model rarely misclassifies instances as belonging to class 0. Conversely, the slightly lower precision of 0.87 for class 4 implies a higher rate of false positives, indicating that more instances were mistakenly classified as belonging to class 4, as evident in Table 1.

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances that are correctly identified by the model. It essentially reflects the model's ability to detect all relevant cases within the dataset.

$$\text{Recall} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Negatives\ (FN)} \quad (15)$$

False Negatives (FN) occur when the model incorrectly fails to predict an instance as belonging to its true class. Class 2 demonstrates perfect recall (1.00), signifying that all actual instances of class 2 were correctly identified by the model. In contrast, Class 3 exhibits a lower recall (0.84), indicating that the model missed some instances belonging to this class, resulting in a higher number of false negatives. These findings are further illustrated in Table 1.

The F1-score represents the harmonic mean of precision and recall, providing a single, balanced measure of a model's overall performance. A high F1-score signifies that the model exhibits both high precision and high recall.

$$F1 = 2 \cdot \frac{Precision\ .\ Recall}{Precision+Recall} \quad (16)$$

Class 5 achieves an F1-score of 1.00, demonstrating near-perfect balance between precision and recall for this class. In contrast, Class 3 exhibits a lower F1-score (0.85), reflecting the combined impact of its lower precision and recall values, as evident in Table 1.

Support refers to the number of instances belonging to each class within the dataset, representing the total occurrences of that class used for evaluation. Class 5 exhibits the highest support with 537 instances, providing the model with more opportunities to learn and generalize from this class. Conversely, Class 2 has lower support with 420 instances. Despite this, Class 2 still achieves a high F1-score, demonstrating the model's robustness and ability to generalize effectively even with limited data for this class, as shown in Table 1.

Accuracy measures the overall proportion of correctly predicted instances across all classes within the dataset. It provides a general assessment of the model's overall performance.

$$\text{Accuracy} = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Instances} \quad (17)$$

The overall accuracy is 0.94, indicating that 94% of all predictions were correct as shown in table 1.

Macro average calculates the average of precision, recall, and F1-score across all classes, giving equal weight to each class regardless of their class distribution or support.

$$\text{Macro Avg} = \frac{\sum_{i=1}^{C} Metric_i}{C} \quad (18)$$

where C is represented as the number of classes. Macro averages for recall, precision, and F1-score are all 0.94, showing consistent performance across classes.

Weighted average computes the average of precision, recall, and F1-score across all classes, giving greater weight to classes with higher support (more instances). It accounts for class imbalance.

$$\text{Weighted Avg} = \frac{\sum_{i=1}^{C} Metric_i\ .\ Support_i}{Total\ Instances} \quad (19)$$

The weighted average for recall, precision, and F1-score is 0.94, indicating strong overall performance, particularly considering the class distribution within the dataset, as shown in Table 1.

Table 1 Classification Report

	precision	recall	f1-score	support
0	1.00	0.99	0.99	496
1	0.98	0.95	0.97	471
2	0.93	1.00	0.96	420
3	0.87	0.84	0.85	491
4	0.87	0.88	0.87	532
5	0.99	1.00	1.00	537
accuracy			0.94	2947
macro avg	0.94	0.94	0.94	2947
weighted avg	0.94	0.94	0.94	2947

The model demonstrates high overall performance (~0.94 for all metrics) with some variation across classes. High precision and recall for certain classes (e.g., 0, 5) indicate strong performance, while slightly lower scores for others (e.g., 3, 4) highlight areas for improvement.

4.2 Confusion Matrix

Figure 6 presents a confusion matrix that provides a detailed breakdown of the human activity recognition model's performance across six activity categories. Each row corresponds to the true activity label, while each column represents the predicted label. The diagonal entries indicate correct classifications, while off-diagonal entries signify misclassifications. Activities like "WALKING," "LAYING," and "STANDING" show strong performance with minimal misclassifications. For instance, "LAYING" achieved 537 correct predictions with only three misclassifications as "SITTING," reflecting the model's robustness in distinguishing this activity.

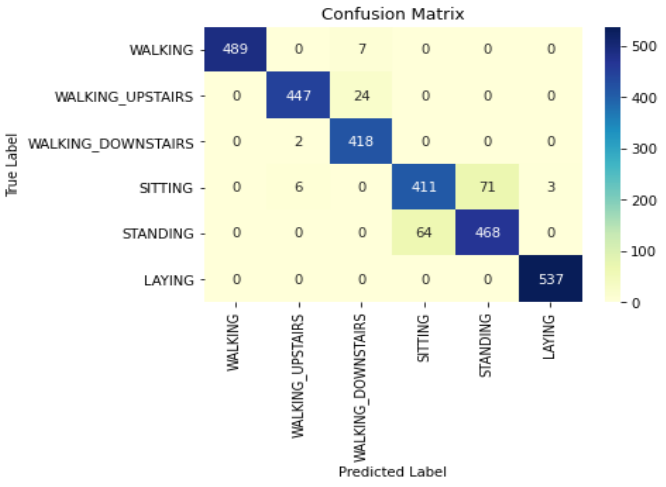

Figure 6. Confusion Matric

"SITTING" and "STANDING" show higher confusion, with several "SITTING" instances misclassified as "STANDING" (71 cases). This observation suggests that the model may have difficulty distinguishing between these two activities, potentially due to similarities in their motion or postural profiles as captured by the sensors. For "WALKING_UPSTAIRS" and "WALKING_DOWNSTAIRS," the confusion is more noticeable. "WALKING_UPSTAIRS" had 24 instances misclassified as "WALKING_DOWNSTAIRS," indicating that the model finds it challenging to distinguish between these closely related activities. This might be due to overlapping patterns in the sensor data for these movements. The majority of predictions are on the diagonal, highlighting the model's ability to generalize across most activity types. This is a positive indicator of its effectiveness in recognizing diverse activities. The misclassifications, particularly between "SITTING" and "STANDING" as well as "WALKING_UPSTAIRS" and "WALKING_DOWNSTAIRS," point to areas where the model could be refined. Techniques such as incorporating additional features, improving preprocessing steps, or adjusting model hyperparameters might help mitigate these issues. In summary, the confusion matrix demonstrates that the model performs well overall, with high accuracy for most activities.

5. Conclusion

This study aimed to develop a robust hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model for HAR. HAR, a critical task in domains such as healthcare, fitness, and human-computer interaction, involves identifying physical activities from sensor data. While significant progress has been made in HAR methodologies, existing models often struggle to effectively capture both spatial and temporal features, leading to suboptimal performance. To address this limitation, we propose a novel hybrid architecture that leverages the strengths of both CNNs and LSTMs. CNN layers were employed to extract intricate spatial features from raw sensor data, while LSTM layers were incorporated to model the temporal dependencies crucial for distinguishing between sequential activities. The model was evaluated on the widely-used UCI HAR dataset, achieving a high classification accuracy of 94% and strong F1-scores across all activity classes. These results demonstrate the superior performance of the proposed hybrid model compared to standalone CNN or LSTM models, particularly in handling complex activity sequences. The hybrid approach effectively mitigates the limitations observed in prior work, such as difficulty in learning long-term dependencies and insufficient generalizability across diverse activity types. Furthermore, the combination of CNN and LSTM layers enhances the model's robustness in handling datasets with imbalanced class distributions, as evidenced by its consistent precision and recall metrics across all activity classes.

While this study demonstrates the effectiveness of the proposed hybrid CNN-LSTM model, several avenues for future research remain. Firstly, incorporating attention mechanisms into the architecture could significantly enhance performance. By enabling the model to focus on the most critical temporal or spatial features within the sensor data, attention mechanisms have shown promise in improving activity classification accuracy in recent studies [12]. Secondly, using transfer learning techniques could address the challenge of data scarcity. By fine-tuning pre-trained models trained on large datasets, the model's performance can be improved even with limited data available for the specific HAR task. Transfer learning has shown promising results in previous activity recognition research. Thirdly, exploring multi-modal sensor fusion by integrating data from multiple sources, such as accelerometers, gyroscopes, and cameras, can significantly enhance the model's ability to classify diverse activities accurately. This approach has the potential to significantly benefit HAR systems in complex real-world scenarios, including healthcare monitoring and industrial environments. Finally, research efforts should be directed towards optimizing the deployment of hybrid CNN-LSTM models on edge devices, such as smartphones. This involves exploring techniques such as model pruning and quantization to minimize latency, reduce energy consumption, and ensure real-time performance without compromising classification accuracy. By addressing these research directions, we can further advance the field of Human Activity Recognition by developing more accurate, efficient, and robust HAR systems that are applicable to a wider range of real-world scenarios.

Data Availability
The datasets are available in open sources UCI

Conflict of Interest
The authors declare no conflicts of interest.

Funding Source
There was no outside funding for the study

Authors' Contributions
The research's conception and design were influenced by the work of all contributors. Data gathering, material preparation, and analysis were completed by Ajinaja M.O. The first draft of the manuscript was written by Akinola A. and Oyedemi A.A. All the authors read and approved the final manuscript and also agreed to all the content of the article including the author list and contributions.

Acknowledgements
We would like to express our sincere gratitude to God. We would like to thank the International Journal of Scientific Research in Computer Science and Engineering for their valuable suggestions, which significantly improved the quality of our research paper.

References

[1] A. Ignatov. "Real-time human activity recognition from accelerometer data using Convolutional Neural Networks". *Applied Soft Computing*, Issue *62*, **pp.915-922, 2017**. https://doi.org/10.1016/j.asoc.2017.09.027

[2] A. Sharma, P. Singh, C. Madhu, N. Garg and G. Joshi, "Human Activity Recognition with smartphone sensors data using CNN-LSTM," *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Gwalior, India, 2022, **pp.1-6, 2022**. doi: 10.1109/IATMSI56455.2022.10119372.

[3] L. Chen, X. Liu, L. Peng, and L. Wu. "Deep learning based multimodal complex human activity recognition using wearable devices." *Applied Intelligence* Issue **51**, **pp.4029–4042, 2021**. https://doi.org/10.1007/s10489-020-02005-7

[4] F.J. Ordóñez, and D. Roggen. "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition". *Sensors*, Vol.*16, Issue* 1 **pp.1–25, 2016.** https://doi.org/10.3390/s16010115

[5] C.A. Ronao, and S. Cho. "Human activity recognition with smartphone sensors using deep learning neural networks". *Expert Systems with Applications*, Vol.*59*, **pp.235-244, 2016.** https://doi.org/10.1016/j.eswa.2016.04.032

[6] M. Atikuzzaman, T. R. Rahman, E. Wazed, M. P. Hossain and M. Z. Islam, "Human Activity Recognition System from Different Poses with CNN," *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, Dhaka, Bangladesh, **pp.1-5, 2020** doi: 10.1109/STI50764.2020.9350508.

[7] R. Maurya, T. H. Teo, S. H. Chua, H. -C. Chow and I. -C. Wey, "Complex Human Activities Recognition Based on High Performance 1D CNN Model," *2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, Penang, Malaysia, **pp.330-336, 2022.** doi: 10.1109/MCSoC57363.2022.00059.

[8] Lee, S., Yoon, S.M., & Cho, H. "Human activity recognition from accelerometer data using Convolutional Neural Network". *IEEE International Conference on Big Data and Smart Computing (BigComp),* **pp.131-134, 2017.**

[9] A. Gumaei, M. M. Hassan, A. Alelaiwi and H. Alsalman, "A Hybrid Deep Learning Model for Human Activity Recognition Using Multimodal Body Sensing Data," in *IEEE Access*, vol. 7, **pp.99152-99160, 2019**, doi: 10.1109/ACCESS.2019.2927134.

[10] W. Xu, Y. Pang, Y. Yang and Y. Liu, "Human Activity Recognition Based on Convolutional Neural Network," *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, **pp.165-170, 2018.** doi: 10.1109/ICPR.2018.8545435.

[11] Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. "Deep learning for sensor-based activity recognition: A survey." *Pattern Recognition Letters*, Vol.*119*, **pp.3-11, 2019**. https://doi.org/10.1016/j.patrec.2018.02.010

AUTHORS PROFILE

Damilola Akinola holds an M.Sc. in Management Information Systems and brings over a decade of experience in cybersecurity and data analytics. He has made significant contributions to the field through impactful work in areas such as vulnerability management, threat intelligence, and compliance with international standards, including IEC 62443 and ISO 21434. With expertise spanning critical industries like power generation and oil and gas, Damilola has successfully published actionable strategies and integrated advanced security tools, including SIEMs and cloud security platforms, to safeguard sensitive systems. He also holds multiple certifications and remains dedicated to advancing national and global cybersecurity resilience.

Adenike Oluyemisi Oyedemi earned her B.Tech in Computer Engineering from Ladoke Akintola University of Technology, Ogbomoso in 2003. She earned her M.Sc and PhD in Computer Science from Obafemi Awolowo University, Ile-Ife in 2017 and 2022 respectively. She has published more than 10 research papers in reputable journals. She is a lecturer at the University of Ilesa, Ilesa, Osun State in the Department of Computer Science.

Micheal Olalekan Ajinaja earned his B.Sc. and M. Tech. in Computer Science from Obafemi Awolowo University and Federal University of Technology, Akure in 2017, and 2024, respectively. He is currently working a lecturer at Federal Polytechnic Ile Oluji in the department of Computer Science. He has published more than 20 research papers in reputed international journals.

