*International Journal of Mechatronics and Applied Mechanics*
*2025, Issue 20, Vol. 1*

# ATTENTION-ENHANCED CNN-LSTM MODELS FOR SENSOR-BASED HUMAN ACTIVITY RECOGNITION: A COMPARATIVE STUDY

*Mohamed Khaled Farouk[1] [0009-0004-4420-3659], A. Abdellatif[2*] [0000-0003-1053-6100], Mohammed I. Awad[3] [0000-0002-0367-0187], and Mostafa R. A. Atia[2] [0000-0001-7485-4365]*
*[1]Mechatronics Engineering Department, Faculty of Engineering, Egyptian Russian University, Cairo 11829, Egypt.*
*[2]Mechanical Engineering Department, Arab Academy for Science Technology and Maritime Transport, Sheraton Branch, Cairo 11757, Egypt.*
*[3]Mechatronics Engineering Department, Faculty of Engineering, Ain Shams University, Cairo 11517, Egypt.*
**Corresponding Author, Email: a_abdellatif@aast.edu*

**Abstract -** Human Activity Recognition (HAR) using wearable sensors is an essential component in healthcare, rehabilitation, and assistive robotics. It enables intelligent systems to interpret and respond to human motion in real time. However, designing models that effectively capture both spatial and temporal dependencies in sensor data remains a major challenge, particularly when aiming for deployment in practical, real-world scenarios. This paper presents a comprehensive comparative study of various deep learning architectures for sensor-based HAR, including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and hybrid CNN-LSTM models with and without integrated attention mechanisms. The motivation behind this paper stems from the limitations of conventional machine learning approaches, which require manual feature engineering and struggle with modeling complex, sequential activity transitions. To overcome these challenges, we evaluate each model under consistent conditions using the ENABL3S dataset, a publicly available benchmark that captures bilateral neuromechanical signals from wearable sensors across multiple locomotor tasks. All models were trained using a standardized pipeline, including fixed window segmentation, identical hyperparameters, and equal epoch counts. The comparative results reveal that the TimeDistributed CNN-LSTM model enhanced with a Keras-based attention layer—applied after the LSTM block—achieved the highest test accuracy of 98.91%. The attention mechanism enabled the network to prioritize critical temporal segments in the sensor data, leading to improved classification accuracy. These findings highlight the importance of architectural enhancements in HAR systems and suggest that attention-augmented hybrid models are particularly well-suited for applications demanding high accuracy, such as intelligent prosthetics, real-time rehabilitation monitoring, and personalized mobility support technologies.

**Keywords**: Rehabilitation Robotics, Wearable sensors, Human Activity Recognition (HAR), Deep learning, Artificial neural network (ANN), Long-Term Short-Term Memory (LSTM), (Convolutional Neural Network) CNN.

## 1. Introduction

Limb loss significantly diminishes the quality of life, impacting individuals on physiological, psychological, and social levels due to the irreversible nature of amputation [1], [2]. Human mobility relies on bipedal movement. On average, individuals in urban areas take around 6,500 steps daily [3].

Due to the critical role of the limbs in daily functioning and the challenges posed by limb loss, prosthetics have evolved to actively assist in restoring natural movement [4]. Modern prosthetic devices are equipped with motors and advanced mechanisms designed to replicate human limb function, thereby facilitating more intuitive and efficient mobility for individuals with limb loss [5]. Another approach that addresses human movement is rehabilitation, which has advanced significantly through wearable robotics [6].

These technologies offer promising solutions for individuals affected by limb loss or muscle

 

*https://doi.org/10.17683/ijomam/issue20*

dysfunction, enabling improved mobility and facilitating recovery by enhancing motor function and supporting rehabilitation processes [7]. To effectively control devices such as exoskeletons or active prosthetics and restore natural human movement, it is essential to use algorithms that can interpret and respond to the user's movements.

These algorithms are crucial for determining how the device should adapt to different types of activities, allowing for seamless integration with the user's motions. This is where Human Activity Recognition (HAR) plays a pivotal role, enabling devices to recognize and adjust to the specific activity being performed by the individual.

Due to these promising applications, HAR has increasingly attracted attention across diverse fields such as sports analytics, smart homes, and industrial safety. In healthcare, HAR plays a critical role in elderly care, post-stroke rehabilitation, and monitoring chronic diseases by enabling real-time movement tracking and early detection of abnormalities [8].

HAR systems can be broadly classified into vision-based and sensor-based approaches. Vision-based HAR relies on cameras and computer vision techniques to track human motion, often utilizing pose estimation and action recognition algorithms [9]. While effective in controlled environments, these methods are prone to occlusions, require significant computational resources, and raise privacy concerns.

On the other hand, sensor-based HAR employs wearable or environmental sensors such as accelerometers, gyroscopes, electromyography (EMG), and inertial measurement units (IMUs) to track movements in real time [10]. These methods are less intrusive, portable, and robust to lighting variations, making them ideal for healthcare, rehabilitation, and assistive robotics. Traditional HAR systems relied on handcrafted features extracted from sensor signals, followed by machine learning classifiers like Support Vector Machines (SVM), Decision Trees, and Hidden Markov Models (HMM) [11]. However, these approaches require extensive feature engineering and often struggle with scalability and generalization across diverse datasets.

Recent advances in deep learning have significantly improved HAR performance by automating feature extraction and capturing complex spatial-temporal patterns in sensor data. CNNs have been successfully used in wearable based HAR to classify activities such as walking, running, and sitting, demonstrating higher accuracy than conventional ML models [12]. LSTM networks, designed for sequential data processing, have proven effective in recognizing complex movements and detecting activity transitions, for instance, Sun et al. [13] explored attention-based LSTMs for activity recognition, showing improved performance by allowing models to focus on key time steps in sensor sequences.

Similarly, Chen et al. [14] proposed spatio-temporal CNNs, demonstrating that multi-scale CNNs could effectively capture local and global motion patterns. HAR-based assistive devices have also benefited from deep learning advancements. Wang et al. [15] developed a CNN-LSTM model for lower-limb prostheses control, significantly improving real-time gait recognition. Similarly, Rv et al. [16] implemented deep reinforcement learning based HAR for exoskeleton control, optimizing movement assistance based on user intent.

To evaluate machine learning techniques, open source HAR datasets are used before experimental testing. For example, UCI HAR dataset includes accelerometer and gyroscope data collected from smartphones worn by participants performing six different activities [17]. Researchers have extensively used this dataset to train CNN, LSTM, and hybrid models for HAR. Another research is proposed by Guan and Ploetz [18]. Deep learning CNNs models are used to classify activities from the UCI HAR dataset, achieving high accuracy while reducing the need for handcrafted features.

Another widely known dataset is the Opportunity dataset, which includes multi-modal sensor recordings of daily life activities, providing a rich set of movement patterns for HAR [19]. Researchers have used this dataset to train multi-sensor fusion models, improving HAR performance in real-world scenarios. In a study by Ordóñez and Roggen [20], deep recurrent neural networks (RNNs) were used to model complex motion sequences, demonstrating the effectiveness of LSTMs in learning long-term dependencies.

The Human Activity Recognition Trondheim dataset (HARTH) dataset, focusing on activity recognition from wearable IMUs and EMG sensors, has been used to develop models for medical applications such as rehabilitation monitoring [21]. Similarly, the ENABL3S dataset, used in this study, provides a comprehensive collection of bilateral neuromechanical signals from lower-limb movements, making it highly suitable for gait analysis, exoskeleton control, and prosthetic development [22]. These datasets collectively enable researchers to develop and benchmark robust HAR models for various applications.

While prior studies such as Sun et al. [13] and Wang et al. [15] have explored the use of attention-based LSTM or CNN-LSTM configurations for HAR, our work introduces a more systematic architectural investigation by evaluating multiple attention placements—after CNN layers and after LSTM layers—within a unified hybrid framework. Unlike existing research that applies attention at a single point or on limited datasets, this study offers a comparative analysis across several architectures under identical experimental conditions using the

ENABL3S dataset. We also evaluate both Keras-native and fine-tuned attention mechanisms, highlighting trade-offs between performance and computational cost.

In particular, the fine-tuned attention mechanism is implemented as a custom-designed layer that performs attention weighting over flattened feature maps. This layer includes learnable weights and biases to generate context vectors adapted to the HAR task. By integrating this layer after different stages of the CNN-LSTM pipeline, we demonstrate improved flexibility in emphasizing relevant temporal or spatial features. This dual-attention strategy—together with a standardized evaluation protocol—provides new insights into the architectural choices that most impact HAR model performance, especially in the context of wearable rehabilitation applications.

In this paper, a hybrid CNN-LSTM architecture is used for activity recognition on an open-source dataset (ENABL3s). This methodology leverages the strengths of both approaches, CNNs for spatial feature extraction and LSTMs for modeling temporal dependencies in sequential data. This approach is enhanced by implementing an attention mechanism to improve prediction accuracy.

## 1.1 Critical Comparison with Prior Literature

Several previous studies have employed deep learning architectures for sensor-based Human Activity Recognition (HAR). For instance, Sun et al. [13] utilized an attention-based LSTM, demonstrating improved model interpretability by focusing on crucial temporal segments. Similarly, Wang et al. [15] combined CNN-LSTM architectures to improve gait-phase recognition. However, these studies typically investigate a single model configuration without providing comprehensive comparisons across multiple variants of attention placement or evaluating different attention mechanisms under identical conditions. In contrast, our study systematically compares numerous architectures—including ANN, CNN, LSTM, hybrid CNN-LSTM models, and both standard and customized attention mechanisms—using the standardized ENABL3S dataset. Moreover, while prior studies such as Guan and Ploetz [18] and Ordóñez and Roggen [20] have shown benefits of deep learning approaches on other datasets, our analysis specifically evaluates bilateral neuromechanical data relevant for prosthetic and rehabilitation robotics, adding distinct practical implications. This critical comparison emphasizes our study's contribution in providing clearer insights into the optimal architectural and methodological choices for effective sensor-based HAR.

The work in this paper is outlined in six sections. Section 2 and 3 present the proposed methodology of work and system overview respectively. The utilized deep learning approaches are shown in section 4. The results of classification and its discussion are presented in section 5. Finally, the paper is concluded with conclusions and future work in section 6.

## 2. Methodology

The proposed methodology, depicted in Figure 1, outlines the overall approach. The dataset employed in this study ENABLE3s [22], published by Hu et al., is used for both training and testing the proposed models. In the pre-processing phase, the time series data was segmented into sequences. By restructuring the dataset into sequences, the time series data was transformed into a supervised learning problem. This pre-processing procedure was consistently applied across all approaches. After pre-processing, the data was used for training and testing the deep learning models, which are described in further details in the following sections.

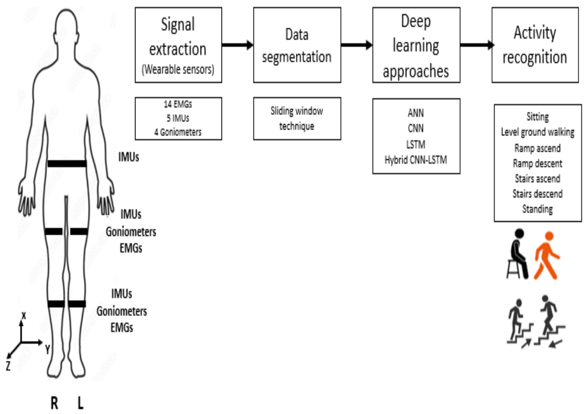


*Figure 1: Methodology of the proposed HAR deep learning models*

To provide a clear overview of the proposed methodology, Figure 2 presents a high-level workflow of the human activity recognition pipeline used in this study.

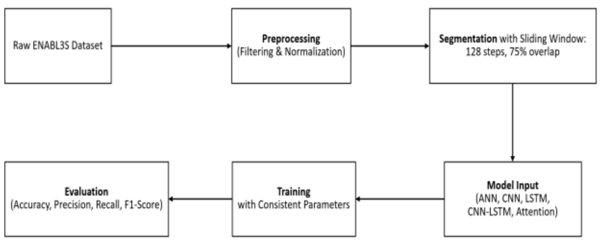


*Figure 2: A high-level workflow of the human activity recognition pipeline used in this study*

The process begins with collecting raw bilateral neuromechanical signals from the ENABL3S dataset.

This data undergoes preprocessing, including down-sampling, signal normalization, and segmentation using a sliding window approach. The segmented data is then structured and fed into various deep learning models—ranging from simple ANNs to complex hybrid CNN-LSTM architectures with attention mechanisms. Following training, the models are evaluated using standard metrics such as accuracy, precision, recall, and F1-score. This structured pipeline ensures consistency across experiments and facilitates direct comparison of model performance.

## 3. System Overview

The computational experiments conducted in this study were performed on a high-performance PC workstation equipped to handle the extensive demands of deep learning algorithms. The workstation features a Ryzen 9 5950x processor, which includes 16 cores and 32 threads, complemented by 32 GB of RAM, facilitating data processing and multitasking. All algorithms were executed within an Anaconda environment, leveraging Python 3.11 paired with TensorFlow 2.15, which facilitates data manipulation and complex numerical computations.

The dataset used in this study (ENABLE3s), stands for ENcyclopedia of Able-bodied Bilateral Lower Limb Locomotor Signals. It contains bilateral lower limb neuromechanical signals collected from wearable sensors under naturalistic conditions. The dataset was meticulously curated to record a wide range of locomotor activities using a sensor-based approach, incorporating IMUs, EMGs, and electro-goniometers. A total of 23 sensors were used in the data collection process, comprising 14 EMG sensors, 4 goniometers, and 5 IMUs.

The dataset includes bilateral signals and joint and limb kinematics collected from 10 healthy, able-bodied individuals (seven males, three females; mean age 25.5 ± 2 years; height 174 ± 12 cm; weight 70 ± 14 kg). These participants performed activities such as sitting, standing, and walking-related movements, including level-ground walking, stair ascent (SA)/descent (SD), and ramp ascent (RA)/descent (RD) as shown in Figure 3.

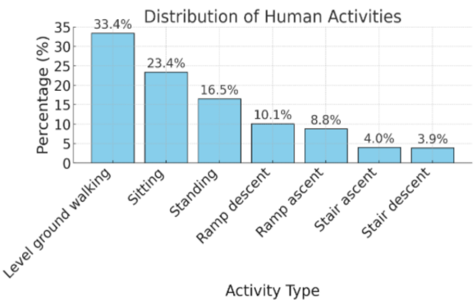


*Figure 3: Activity distribution percentage in ENABLE3s*

EMG signals were recorded using bipolar surface electrodes from seven muscles in each leg, while IMUs were positioned bilaterally on the thighs (below the rectus femoris) and shanks (near the tibialis anterior) and sampled at 500 Hz. Additionally, joint kinematic signals were captured using electro-goniometers placed on the knees and ankles, also sampled at 500 Hz.

The sliding window technique was applied to segment the continuous time-series data for Human Activity Recognition (HAR) [23]. By dividing sensor signals into fixed-length overlapping sequences, the model captures essential temporal dependencies within human motion. Each segment is treated as an independent training sample, standardizing input dimensions for deep learning architectures such as CNN-LSTM.

In this study, a window size of 128time steps—equivalent to 2.56 seconds at a 50 Hz down sampled rate—was selected, with a stride of 32 steps, resulting in a 75% overlap (96 samples) between windows. This configuration was chosen based on both the characteristics of the ENABL3S dataset and widely accepted practices in the HAR literature. The dataset contains high-resolution, multi-sensor neuromechanical recordings that span complete gait cycles across varied terrains. The selected window length ensures that each segment includes sufficient temporal context to capture full or nearly full locomotor phases (e.g., stance and swing). The 75% overlap increases the volume and continuity of training samples, which is particularly useful for learning smooth transitions and subtle distinctions between activity classes, while avoiding excessive redundancy. This balance has been shown to enhance model robustness and generalization.

Additionally, the placement of the attention mechanism was empirically tested at two locations: after the CNN layers to enhance spatial saliency, and after the LSTM layers to prioritize temporally relevant segments. The highest performance was achieved when applying the attention layer after the LSTM block, enabling the model to emphasize the most informative time steps.

This approach aligns with prior HAR studies such as [Sun et al., 2019] [13], where temporal attention improved classification accuracy and interpretability.

## 4. Deep Learning Approaches

This section presents an overview of the deep learning models employed in this study, each tailored to address specific challenges inherent to sensor-based Human Activity Recognition (HAR). The models are designed to effectively process and learn from multi-dimensional time-series data, capturing both spatial correlations within sensor channels and temporal dependencies across activity sequences.

The Artificial Neural Network (ANN) is used as a baseline model, providing a reference point for performance comparison with more complex architectures. The Convolutional Neural Network (CNN) focuses on extracting local spatial features, such as motion patterns and signal shapes, across sensor modalities. In contrast, the Long Short-Term Memory (LSTM) network is adept at modeling sequential dependencies and long-range temporal patterns, making it suitable for dynamic activity transitions.

To leverage the strengths of both architectures, hybrid CNN-LSTM models are implemented. These models utilize CNN layers for spatial feature extraction and LSTM layers for temporal sequence modeling, resulting in enhanced recognition performance.

## 4.1 ANN Models

The proposed ANN model consists of four fully connected layers with 265, 128, 64, and 32 units, each using the ReLU activation function. To match the ANN's required 2D input format (samples, features), the original 3D time-series data (samples, time steps, features) was reshaped by flattening the time and feature dimensions into a single vector (samples, time steps × features).

## 4.2 CNN Models

For CNN, both series and parallel CNN architectures are presented using Conv1D layers. All CNN models consist of three convolutional layers, followed by two dense layers with 100 and 64 units, respectively, using the ReLU activation function. A final Softmax layer with 7 output units is used for multi-class classification of human activities [24].

A series-based model was designed to extract spatial features from raw sensor signals. The input is structured as a 3D tensor with shape (samples, time steps, features), where filters are applied along the temporal dimension. The model uses three layers with 64, 128, and 256 filters, each with a kernel size of 3. Kernel weights were initialized using the Glorot Uniform initializer to ensure stable convergence during training. This configuration effectively captures local temporal patterns in sensor data.

Another CNN architecture employed is the multi-headed design. This approach employs three parallel branches, each consisting of three convolutional layers with 64 filters, but using different kernel sizes (3, 5, and 11). This design enables the model to capture temporal features across multiple resolutions, enhancing its ability to distinguish between subtle activity variations. The outputs from each convolutional branch are concatenated along the feature axis before being passed to the dense layers.

## 4.3 LSTM Models

The Long Short-Term Memory (LSTM) network is a specialized form of Recurrent Neural Network (RNN) designed to overcome the vanishing gradient problem commonly encountered in standard RNNs [25]. Unlike traditional RNNs, LSTMs are capable of learning long-range temporal dependencies, making them highly effective for modeling sequential patterns in time-series data. This capability is enabled by a gated memory structure that regulates the flow of information through three key components:

- The forget gate, which determines what information to discard from the previous cell state.
- The input gate, which decides what new information to store.
- The output gate, which controls the flow of the current cell state to the output.

These gates work together to maintain and update the cell state, allowing the network to retain important historical context across time steps as can be seen in Figure 4.

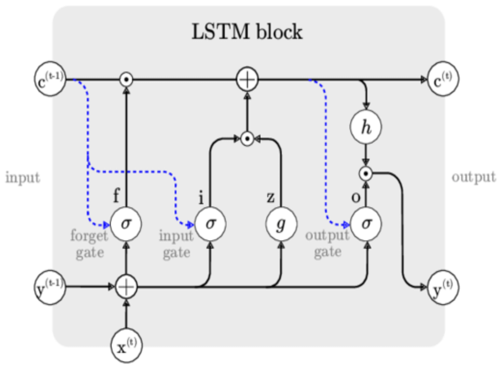


*Figure 4: The implementation of LSTM in HAR system, LSTM main blocks*

The proposed model consists of three LSTM layers with 512, 256, and 128 units, each using a ReLU activation function. Following the LSTM layers, two fully connected dense layers with 100 and 64 units, respectively, are applied.

The model concludes with a SoftMax layer consisting of seven output neurons to classify the activities. To fit the LSTM's input requirements, the data was reshaped into a 3D format of (samples, time_steps, num_features).

## 4.4 Hybrid CNN-LSTM: Sequential Configuration

The proposed integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) layers aims to leverage their complementary strengths as CNNs are effective at capturing spatial features from sensor signals, while LSTMs excel at modeling temporal dependencies across time-series data [26]. This hybrid approach is evaluated in two configurations: Sequential (Series) Models, where the LSTM layers are stacked directly after the CNN layers, and Parallel (Multi-Headed) Models, where CNN and LSTM layers process the same data simultaneously. Both configurations share the same underlying architecture, as can be seen in Figure 5, consisting of (CNN Layers, LSTM Layers and Dense Layers):

- CNN Layers: Two convolutional layers with 64 and 128 filters, a kernel size of 3, and weights initialized with glorot_uniform.
- LSTM Layers: Two stacked LSTM layers with 256 and 128 units, respectively.
- Dense Layers: Two fully connected dense layers with 100 and 64 units, followed by a SoftMax layer for classification.

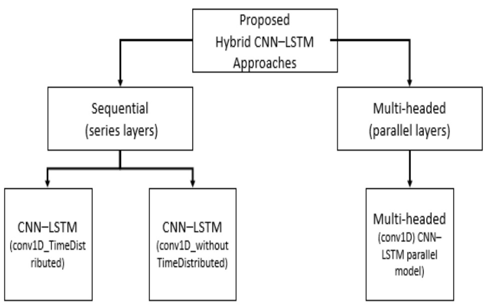


*Figure 5: The Proposed Hybrid CNN-LSTM Architectures*

In the sequential configuration, LSTM layers are placed directly after CNN layers to form a pipeline that extracts spatial features followed by temporal modeling. The CNN layers process raw time-series windows to capture local spatial dependencies, while the subsequent LSTM layers analyze the resulting feature sequences to learn dynamic activity transitions. However, this architecture introduces a structural challenge due to the incompatibility between the output shape of CNN layers and the input expectations of LSTM layers. To resolve this, two strategies were explored: the TimeDistributed wrapper and manual reshaping.

TimeDistributed Wrapper is a high-level technique that applies the same convolutional operations independently across each time step of the input sequence. This preserves the temporal structure and generates a sequence of feature maps, which are inherently compatible with the LSTM's required input shape (samples, time steps, features).

Additionally, in Manual reshaping, the output of CNN layers can be manually reshaped to match the LSTM input format. This involves flattening the feature maps and restructuring them into a 3D tensor of shape (samples, time steps, features). While this method allows for more control over intermediate features, it may risk disrupting temporal continuity if not properly implemented. Nonetheless, it is effective in scenarios where customized representation learning or advanced transformations are desired before sequence modeling.

## 4.5 Hybrid CNN-LSTM: Multi-headed Configuration

In the parallel configuration, the CNN and LSTM layers process the same input data simultaneously, with each branch—or "head"—independently learning complementary features. The CNN branch focuses on extracting spatial features from sensor data, often at multiple scales, while the LSTM branch captures temporal dependencies across time steps. The outputs from both branches are then concatenated and passed through fully connected layers for final classification. This dual-path architecture enables the model to integrate spatial and temporal information into a unified feature representation, improving its ability to recognize complex human activities. By learning in parallel, the network can capture multi-scale patterns and dynamic transitions that may not be effectively modelled by a single-stream design, making it particularly well-suited for datasets with rich spatiotemporal structure.

### 4.6 Attention Mechanism

To further improve the feature extraction and temporal sequence modeling capabilities of the CNN-LSTM architecture, attention mechanisms were integrated at two critical stages of the model: after the CNN layers and after the LSTM layers. These mechanisms enable the model to dynamically focus on the most significant and task-relevant features, thus enhancing its ability to recognize complex human activities in sensor-based data [27].

The attention layers operate by assigning learnable weights to different parts of the input features, allowing the network to prioritize more informative components. This adaptive weighting mechanism refines the flow of spatial and temporal information through the model, ensuring that the network learns to emphasize critical regions and time steps during classification [28]. Two distinct attention strategies were implemented and compared in this study: the Keras-based attention and Fine-tuned attention.

Keras-Based attention enables the network to assign weights during training automatically and dynamically. This method proves to be computationally lightweight and effective for enhancing performance without the need for additional customization.

As for Fine-Tuned attention, it allows for fine control over the attention weights and biases, enabling the model to adapt to the unique characteristics of the dataset. The custom layer computes attention scores, applies them to the feature maps or sequence outputs, and generates a context vector representing the weighted combination of inputs. Unlike the Keras-based implementation, this custom layer exposes trainable parameters that are optimized during training, providing more flexibility and control over how the model prioritizes input features.

# 5. Results and Discussion

## 5.1 Performance Results

All proposed techniques were tested to obtain prediction results. The prediction was done on 30000 samples where the training and testing percentages were 70% to 30% respectively.

The achieved prediction accuracies for training and testing data are demonstrated in Figure 6. These results show that integrating attention mechanisms after the CNN and LSTM layers led to a noticeable improvement in model performance. These mechanisms enabled the network to dynamically focus on the most informative spatial and temporal features, enhancing its classification capabilities. The CNN-LSTM Keras attention model demonstrated the highest accuracy in training and testing data, reaching 98.91% and 97% respectively. This model outperformed the other models by learning task-specific attention weights tailored to the dataset's characteristics.

Another aspect of comparison is the training/testing times. The comparison times of all algorithms are depicted in Figure 7. The training time analysis reveals a clear trade-off between model complexity and computational efficiency. Simpler models like the ANN and basic CNN trained rapidly (7.92 min and 10.53 min, respectively), making them well-suited for real-time applications. More advanced architectures, such as multi-headed CNNs and LSTM-based models, required longer training durations due to their deeper structure and temporal processing capabilities. The highest training times were observed in attention-enhanced CNN-LSTM models, particularly the Keras attention CNN-LSTM, with a duration of 65.7 minutes, highlighting the added computational cost required to improve model precision.

Lightweight models like ANN and CNN achieved fast testing times, achieving predictions in just 2–4 seconds. In contrast, attention-based architectures exhibited the longest testing times, peaking at 21 seconds, due to their layered complexity and attention calculations.

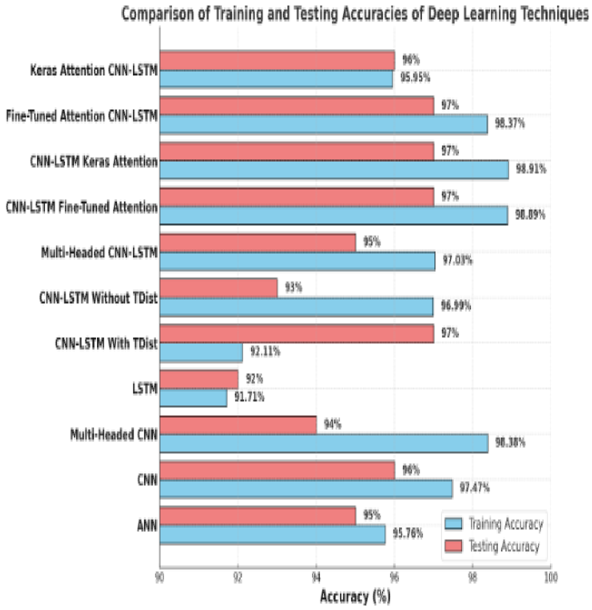


*Figure 6: The resultant prediction accuracies of the utilized deep learning techniques.*

Figure 6 presents the training and testing accuracy for all evaluated models. It is evident that models incorporating the attention mechanism—particularly the TimeDistributed CNN-LSTM with post-LSTM attention—achieve the highest accuracy (98.91%), outperforming simpler architectures like ANN (95.76%) and standalone LSTM (91.71%).

This indicates that combining spatial and temporal features, along with a mechanism for focusing on the most informative time steps, significantly enhances classification performance. Notably, the gap between training and testing accuracy remains small for the top models, suggesting good generalization and minimal overfitting.

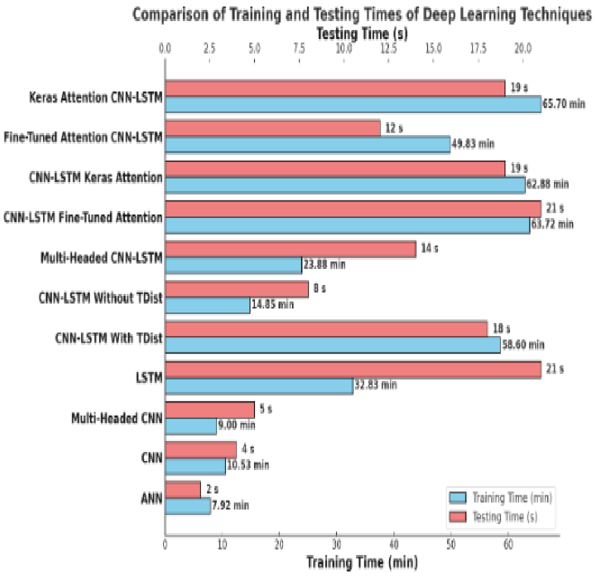


*Figure 7: The resultant prediction times of the utilized deep learning techniques*

Figure 7 compares the training time required for each model across 30 epochs. As expected, complex architectures like the CNN-LSTM with attention demand significantly more time—up to 65.7 minutes—due to their multi-layered structure and additional attention computations. In contrast, simpler models such as ANN and Conv1D-CNN train in under 11 minutes.

## 5.2 Performance Analysis

The results achieved in section 5.1 were analyzed using confusion matrices which can be seen in Figure 8. Attention mechanism enhanced algorithms are analyzed to show the validity of classification. Hence, reliable performance indicators like Precision (P), Recall (R) and F1-Score [29] are calculated. The detailed results achieved by the utilized algorithms are shown in Figure 9.

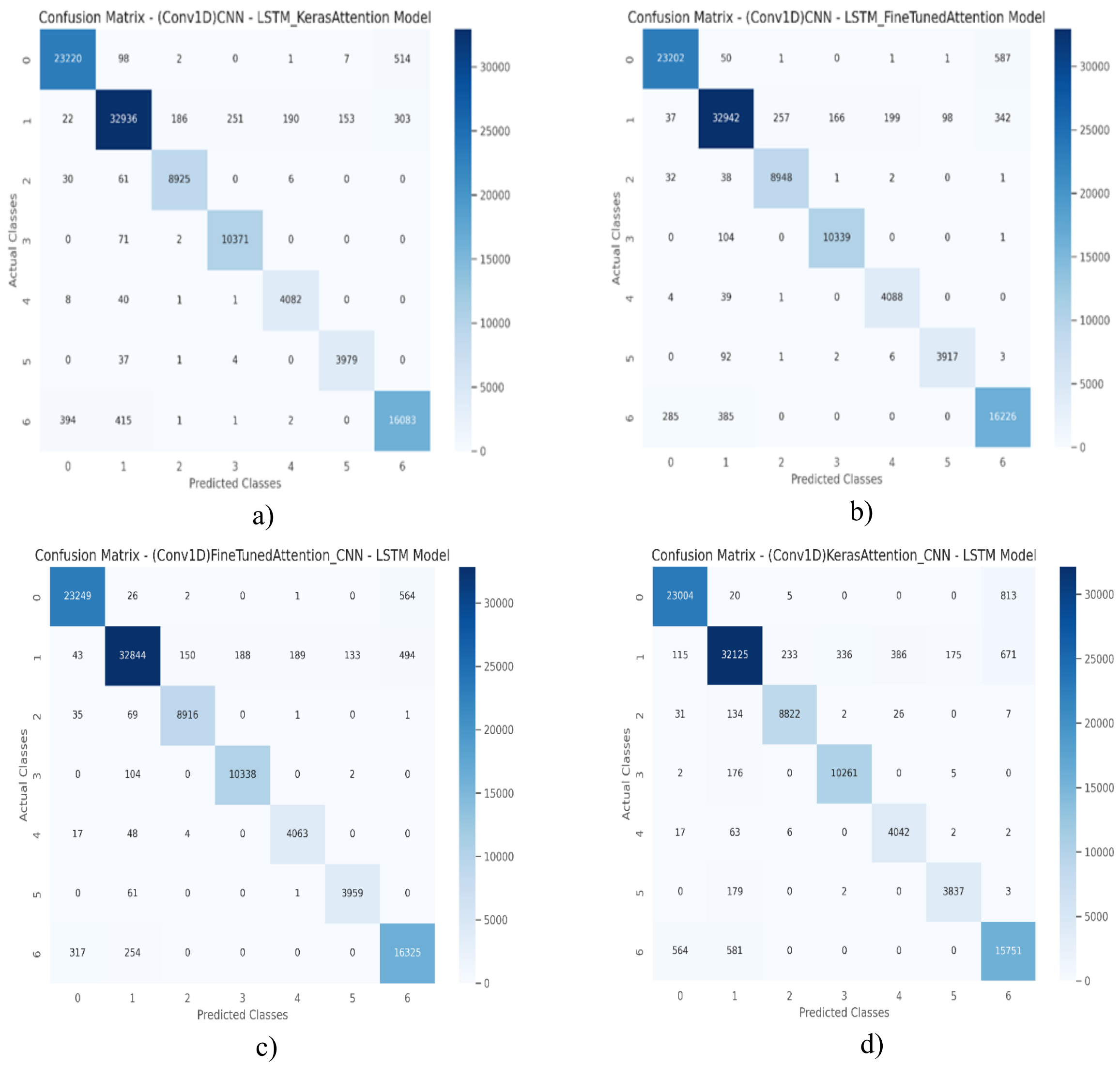


*Figure 8. Confusion matrices for all Attention-enhanced CNN-LSTM models a) CNN-LSTM Fine-Tuned Attention, b) CNN-LSTM Keras Attention, c) Fine-Tuned Attention CNN-LSTM and d) Keras Attention CNN-LSTM.*

The confusion matrices in Figure 8 reveal that most misclassifications occur between activities with similar movement patterns, such as walking vs. ramp descent or stair ascent vs. descent. These overlaps are reduced in models using attention, particularly post-LSTM attention, which helps the network focus on fine-grained temporal cues that distinguish subtle transitions. This improvement is crucial in rehabilitation, where precise recognition of functional movement phases (e.g., controlled descent vs. uncontrolled fall) can affect device response and therapy outcomes.

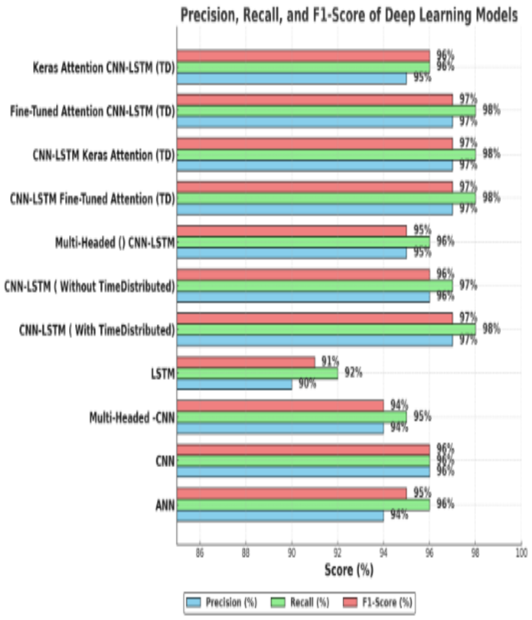


*Figure 9. Comparison of Precision, Recall, and F1-Score for Deep Learning Models.*

The evaluation of models' performance across precision, recall, and F1-score reveals that advanced architecture consistently outperforms simpler baselines. Models such as the Multi-Headed CNN-LSTM and those enhanced with Attention mechanisms achieved the highest precision values, approaching 97%, indicating strong confidence in their positive predictions.

In terms of recall, LSTM-based models, especially those integrated with attention layers, demonstrated superior performance, with top models reaching up to 98%.

The F1-score confirms the effectiveness of these hybrid models, particularly the Fine-Tuned attention CNN-LSTM, which achieved F1-scores as high as 97%, reflecting their overall robustness and reliability for real-world human activity recognition applications.

### 5.3 Comparative Analysis with Prior Studies

To contextualize the effectiveness of the proposed models, Table 1 presents a performance comparison with several recent studies that employed CNN-LSTM architectures with attention mechanisms for sensor-based human activity recognition. Yin et al. [30] reported an accuracy of 96.71% on the UCI dataset using a parallel CNN-BiLSTM model enhanced with attention. Mekruksavanich and Jitpattanakul [31] achieved 98.80% accuracy on the WISDM dataset using a hybrid convolutional neural network with a channel attention mechanism. Deng et al. [32] proposed a self-attention-based deep convolutional LSTM framework, achieving 97.83% accuracy on a custom badminton activity dataset.

In comparison, our best-performing model—the TimeDistributed CNN-LSTM with Fine-Tuned attention—achieved 98.91% accuracy on the ENABL3S dataset. While these studies used different datasets, our model's performance is competitive, particularly considering the complexity and variability of bilateral neuromechanical signals captured in ENABL3S, which are more representative of prosthetic control and rehabilitation scenarios. This comparative evaluation demonstrates that the proposed architecture not only achieves high recognition accuracy but also maintains robustness across complex, real-world sensor inputs, validating its suitability for healthcare, rehabilitation, and assistive robotic systems.

*Table 1 – Comparison of prediction results achieved by different research studies that used open-source datasets*

| Study | Dataset | Model Architecture | Accuracy (%) |
|---|---|---|---|
| **Yin et al. (2022) [30]** | UCI | CNN-BiLSTM with attention | 96.71 |
| **Mekruksavanich & Jitpattanakul (2023) [31]** | WISDM | CNN-LSTM with Self-Attention | 98.80 |
| **Deng et al. (2023) [32]** | Custom Dataset | Dilated CNN + Bi-LSTM + attention | 97.83 |
| **Our Study** | ENABL3S | Time Distributed CNN-LSTM with Fine-Tuned attention | 98.91 |

### 5.4 Practical Applicability and Limitations

While the proposed models achieved high accuracy on the ENABL3S dataset, it is important to recognize the limitations and challenges that affect their practical deployment. First, the dataset comprises only healthy individuals under controlled conditions, which may not fully represent real-world or clinical scenarios where sensor noise, subject variability, or pathological gait patterns are present. Second, the current implementation was evaluated in an offline setting, whereas real-time deployment—particularly on embedded or wearable hardware—imposes strict constraints on inference speed, energy efficiency, and robustness. These factors must be considered when transitioning from experimental results to functional systems for rehabilitation robotics or assistive devices. Addressing these limitations through clinical testing, embedded deployment, and validation under noisy or unconstrained conditions is a necessary step for realizing the model's practical impact.

## 6. Conclusions and Future Work

In this paper, a deep learning-based methodology is used to predict activities of daily living (ADLs) which is obtained from the open-source dataset (ENABLE3s). A comprehensive evaluation of various deep learning architectures for sensor-based human activity recognition (HAR) is presented. This evaluation includes implementation of CNN and LSTM based models, hybrid CNN-LSTM configurations, and attention-enhanced variants. The results demonstrate that models integrating both spatial and temporal feature extraction as CNN-LSTM Keras Attention model demonstrated the highest accuracy in training and testing data, reaching 98.91% and 97% respectively. On the other hand, the longest training times were observed in the same Attention-enhanced CNN-LSTM models, particularly the Keras attention CNN-LSTM, with a duration of 65.7 minutes, verifying the added computational cost of improving model precision.

Future work will explore the integration of more advanced attention mechanisms, such as transformer-based encoders and self-supervised pretraining, to further enhance recognition performance with less labelled data. Additionally, a custom-made HAR dataset has been published by the authors in [33]. Hence, this HAR framework can be implemented experimentally and extended to support online learning and continuous activity adaptation, enabling real-time responsiveness in dynamic environments. Furthermore, future work will include clinical validation of the proposed models and their deployment on embedded hardware to assess performance in real-world rehabilitation and assistive scenarios.

## References


[1] Serpush, F., Menhaj, M. B., Masoumi, B., & Karasfi, B. (2022). Wearable sensor-based human activity recognition in the smart healthcare system. *Computational Intelligence and Neuroscience, 2022*,1391906.https://doi.org/10.1155/2022/1391906

[2] Halim, A., Abdellatif, A., Awad, M. I., & Atia, M. R. A. (2021). Prediction of human gait activities using wearable sensors. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 235*(6), 676–687.https://doi.org/10.1177/09544119211001238

[3] Alluhydan, K., Siddiqui, M. I. H., & Elkanani, H. (2023). Functionality and comfort design of lower-limb prosthetics: A review. *Journal of Digital Rehabilitation, 2*(3), 10–23. https://doi.org/10.57197/JDR-2023-0031

[4] Chopra, S., & Emran, T. B. (2024). Advances in AI-based prosthetics development: Editorial. *International Journal of Surgery, 110*(8), 4538–4542.https://doi.org/10.1097/JS9.0000000000001573

[5] Halim, A., Ibrahim, A. A., Awad, M. I., & Atia, M. R. A. (2020). Optimization of sensor number for lower limb prosthetics using genetic algorithm. *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*, 210–215. https://doi.org/10.1109/ITCE48509.2020.9047780

[6] Schrader, L., et al. (2020). Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people. *Journal of Population Ageing, 13*(2), 139–165. https://doi.org/10.1007/s12062-020-09260-z.

[7] Abdelbar, M., Mohamed, I., Abdellatif, A., & Hegaze, M. M. (2022). Towards the mechatronic development of a new upper-limb exoskeleton (SAMA). *Designs, 6*(5), 80. https://doi.org/10.3390/designs6050080

[8] Osmani, V., Balasubramaniam, S., & Botvich, D. (2008). Human activity recognition in pervasive healthcare: Supporting efficient remote collaboration. *Journal of Network and Computer Applications, 31*(4), 628–655. https://doi.org/10.1016/j.jnca.2007.11.002.

[9] Beddiar, D. R., Nini, B., Sabokrou, M., et al. (2020). Vision-based human activity recognition: A survey. *Multimedia Tools and Applications, 79*, 30509–30555.https://doi.org/10.1007/s11042-020-09004-3.

[10] Alarfaj, M., Al Madini, A., Alsafran, A., Farag, M., Chtourou, S., Afifi, A., Ahmad, A., Al Rubayyi, O., Al Harbi, A., & Al Thunaian, M. (2024). Wearable sensors based on artificial intelligence models for human activity recognition. *Frontiers in Artificial Intelligence, 7*. https://doi.org/10.3389/frai.2024.1424190

[11] Shah, S. H. H., Karlsen, A. S. T., Solberg, M., & Hameed, I. A. (2024). An efficient and lightweight

multiperson activity recognition framework for robot-assisted healthcare applications. *Expert Systems with Applications, 241*, 122482.https://doi.org/10.1016/j.eswa.2023.122482.

[12] Gupta, S. (2021). Deep learning-based human activity recognition (HAR) using wearable sensor data. *International Journal of Information Management Data Insights, 1*(2), 100046. https://doi.org/10.1016/j.jjimei.2021.100046.

[13] Sun, B., Liu, M., Zheng, R., & Zhang, S. (2019). Attention-based LSTM network for wearable human activity recognition. *2019 Chinese Control Conference(CCC)*,86778682. https://doi.org/10.23919/ChiCC.2019.8865360

[14] Chen, X., et al. (2021). Deep analysis of CNN-based spatio-temporal representations for action recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6161–6171. https://doi.org/10.1109/CVPR46437.2021.00610

[15] Wang, H., Wang, M., Li, D., Deng, F., Pan, Z., & Song, Y. (2025). Gait phase recognition of hip exoskeleton system based on CNN and HHO-SVM model.*Electronics,14*(1),107.https://doi.org/10.3390/electronics14010107.

[16] Rv, M., & Rakshit, S. (2024). Deep reinforcement learning-based control of lower limb exoskeleton. *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–6. https://doi.org/10.1109/IJCNN60899.2024.10651057.

[17] Reyes-Ortiz, J., et al. (2013). Human activity recognition using smartphones. *UCI Machine Learning Repository*. https://doi.org/10.24432/C54S4K.

[18] Guan, Y., & Ploetz, T. (2017). Ensembles of deep LSTM learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(2), 1–16. https://doi.org/10.1145/3090076.

[19] Ciliberto, M., Fortes, R. V., Calatroni, A., Lukowicz, P., & Roggen, D. (2021). Opportunity++: A multimodal dataset for video- and wearable, object and ambient sensors-based human activity recognition. *Frontiers in Computer* Science,3.https://doi.org/10.3389/fcomp.2021.792065

[20] Ordóñez, P., & Roggen, D. (2016). Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition.*Sensors,16*(1),115.https://doi.org/10.3390/s16010115.

[21] Logacjov, A., Bach, K., Kongsvold, A., Bårdstu, H. B., & Mork, P. J. (2021). HARTH: A human activity recognition dataset for machine learning. *Sensors,* 21(23),7853.https://doi.org/10.3390/s21237853

[22] Hu, B., Rouse, E., & Hargrove, L. (2018). Benchmark datasets for bilateral lower-limb neuromechanical signals from wearable sensors during unassisted locomotion in able-bodied individuals. *Frontiers in Robotics and AI, 5*. https://doi.org/10.3389/frobt.2018.00014.

[23] Baraka, A. M. A., & Mohd Noor, M. H. (2023). Similarity segmentation approach for sensor-based activity recognition. *IEEE Sensors Journal, 23*(17),19704–19716. https://doi.org/10.1109/JSEN.2023.3295778

[24] Sudhamathy, G., & Valliammal, N. (2023). The Bayesian CNN-LSTM classification model to predict and evaluate learner's performance. *International Journal of Applied Science and Engineering,20*,2023282.https://doi.org/10.6703/IJASE.202312_20(4).007

[25] Shruti, P., Venkatesh, M., & Pooja, K. (2022). LSTM-based ensemble network to enhance the learning of long-term dependencies in chatbot. *International Journal of Automation and Smart Technology,12*(1). https://doi.org/10.5875/ausmt.v12i1.2286

[26] Mutegeki, R., & Han, D. S. (2020). A CNN-LSTM approach to human activity recognition. *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 362–366.https://doi.org/10.1109/ICAIIC48513.2020.9065078

[27] Liu, L., Feng, J., Li, J., Chen, W., Mao, Z., & Tan, X. (2024). Multi-layer CNN-LSTM network with self-attention mechanism for robust estimation of nonlinear uncertain systems. *Frontiers in Neuroscience,18*.https://doi.org/10.3389/fnins.2024.1379495.

[28] Wei, X., & Wang, Z. (2024). TCN-Attention-HAR: Human activity recognition based on attention mechanism time convolutional network. *Scientific Reports,14*,7414.https://doi.org/10.1038/s41598-024-57912-3

[29] Elwaly, A., Abdellatif, A., & El-Shaer, Y. (2024). New eldercare robot with path-planning and fall-detection capabilities. *Applied Sciences, 14*(6), 2374.https://doi.org/10.3390/app14062374.

[30] Yin, X., Liu, Z., Liu, D., & Ren, X. (2022). A Novel CNN-based Bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data. Scientific Reports, 12(1),7878.https://doi.org/10.1038/s41598-022-11880-8

[31] Mekruksavanich, S., & Jitpattanakul, A. (2023). Hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition. Scientific Reports, 13(1), 12067.https://doi.org/10.1038/s41598-023-39080-y

[32] Deng, J., Zhang, S., & Ma, J. (2023). Self-attention-based deep convolution lstm framework for sensor-based badminton activity recognition. *Sensors*, *23*(20),8373.https://doi.org/10.3390/s23208373

[33] Abdellatif Hamed IBRAHIM, A., Atef El-Khoreby, M., Kamal Moawad, A., Hussien Issa, H., Ismail Fawaz, S., Ibrahim Awad, M., & Khaled Farouk, M. (2025). DATASET_SDALLE (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14841611.