International Journal of Innovative Research and Scientific Studies, 8(3) 2025, pages: 1412-1422

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

A comprehensive evaluation of machine learning and deep learning models for stair-climbing wheelchair activity recognition

Pharan Chawaphan[1], Dechrit Maneetham[2*], Padma Nyoman Crisnapati[3]

[1,2,3]*Department of Mechatronics Engineering, Faculty of Technical Education, Rajamangala University of Technology Thanyaburi, Thailand.*

Corresponding author: Dechrit Maneetham (*Email: dechrit_m@rmutt.ac.th*)

Abstract

Human Activity Recognition (HAR) is vital for enhancing assistive technologies such as stair-climbing wheelchairs, which cater to individuals with mobility challenges. This study investigates optimal machine learning and deep learning models for classifying human activity related to stair-climbing wheelchairs, which are essential for enhancing mobility in assistive technologies. A dataset of 5,872 samples across 18 sensor-derived features was preprocessed using normalization, one-hot encoding, and SMOTE to address class imbalance. Eight models—MLP, CNN, LSTM, BiLSTM, Transformer, CatBoost, LightGBM, and TabNet—were trained and evaluated using an 80/20 train-test split. CatBoost and LightGBM achieved the highest accuracy (99.83%) with inference times of 8 ms and 7 ms respectively. Deep learning models such as MLP and CNN also performed well, while the Transformer exhibited poor compatibility with the dataset. Machine learning models, especially CatBoost and LightGBM, demonstrated both high accuracy and computational efficiency, making them suitable for real-time applications in assistive technologies. This work provides essential insights for implementing efficient HAR systems in mobility-assistive devices and can inform future designs of autonomous wheelchair platforms.

Keywords: Assistive technologies, Deep learning, Human activity recognition, Sensor-based classification, Stair-climbing wheelchair, Wheelchair movement.

DOI: 10.53894/ijirss.v8i3.6805
Funding: This study received no specific financial support.
History: Received: 7 April 2025 / **Revised:** 9 April 2025 / **Accepted:** 14 April 2025 / **Published:** 7 May 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Assistive technologies have played a pivotal role in enhancing the quality of life for individuals with mobility challenges [1-3]. Among these innovations, stair-climbing wheelchairs address architectural barriers, enabling greater independence and mobility [4, 5]. However, ensuring the safety and efficiency of these wheelchairs demands accurate recognition of specific

activities during operation, such as climbing stairs, navigating slopes, and transitioning between surfaces [6, 7]. While Human Activity Recognition (HAR) using sensor data has shown promise, existing studies often rely on handcrafted features or classical machine learning approaches [8-10]. that struggle with noisy and high-dimensional data [11, 12]. Moreover, limited research has systematically compared the performance of advanced machine learning and deep learning models for HAR in stair-climbing wheelchair systems, especially under real-world conditions such as class imbalance and computational constraints [13, 14].

Recent advancements in deep learning have revolutionized HAR by enabling models to automatically learn spatial and temporal features [15, 16]. Models like Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Bidirectional LSTMs (BiLSTMs) have demonstrated success in capturing sequential patterns in sensor data [17, 18]. Machine learning models such as CatBoost [19] and LightGBM [20] have also emerged as top performers for structured data, owing to their ability to efficiently handle categorical features and large datasets. Despite these advancements, there remains a gap in evaluating the performance of such models for specific tasks like stair-climbing wheelchair activity recognition, where class imbalance and real-time computational efficiency are critical [21-23].

The primary objective of this research is to systematically evaluate and compare state-of-the-art machine learning and deep learning models for classifying stair-climbing wheelchair activities. Despite the advancements in Human Activity Recognition (HAR) technologies [24-27] several challenges persist in the context of stair-climbing wheelchairs:

1. Class imbalance in datasets: Most existing studies fail to address the underrepresentation of certain activity classes, which can lead to biased and unreliable model performance in real-world applications.
2. Computational constraints: Real-time applications, such as assistive technologies, require models that not only perform well in terms of accuracy but also operate efficiently with minimal inference time.
3. Model selection for specific HAR tasks: There is a lack of systematic comparison between traditional machine learning and deep learning models tailored for specific activities in stair-climbing wheelchairs, making it difficult to identify optimal solutions.

This study uniquely integrates a novel dataset, advanced preprocessing techniques, and a diverse range of models to identify optimal solutions for HAR in assistive technologies. By addressing class imbalance through SMOTE [28] and incorporating inference time evaluation [29] alongside traditional performance metrics, this research offers actionable insights for deploying real-time HAR systems in stair-climbing wheelchairs.

The dataset, sourced from Mendeley Data [30] comprises 5,872 samples across 18 features, including accelerometer, gyroscope, and magnetometer data. Preprocessing steps include normalization, one-hot encoding, and SMOTE for balancing class distributions. Eight models, including MLP [31] CNN [32] LSTM [33] BiLSTM [34] Transformer [35] CatBoost [36] LightGBM [37] and TabNet [38] were trained and evaluated using an 80/20 train-test split. Performance metrics such as accuracy, precision, recall, F1-score, inference time, ROC AUC, and Matthews Correlation Coefficient (MCC) were used for evaluation.

Machine learning models CatBoost and LightGBM achieved the highest accuracy of 99.83% with minimal inference times, making them ideal for real-time applications. Deep learning models like MLP and CNN also performed well, with BiLSTM showing notable improvement over LSTM by leveraging bidirectional dependencies. However, the Transformer model exhibited poor performance, highlighting its limitations for this task. The results highlight the importance of model selection in HAR tasks, particularly for real-time applications in assistive technologies. While machine learning models demonstrated superior computational efficiency, deep learning models showed promise for capturing complex spatial and temporal patterns. The findings underscore the need for tailored approaches to address the unique challenges of HAR in stair-climbing wheelchair systems.

2. Method

This study uses sensor data to assess how well different deep learning and machine learning models perform for Human Activity Recognition (HAR). The methodology adopts a systematic approach, encompassing data collection, preprocessing, model design and architecture, as well as training and evaluation, to achieve robust and accurate classification.

2.1. Dataset Collection

This study utilizes a novel dataset available in the Mendeley Data repository [30]. The dataset consists of 5,872 samples recorded across 18 features, including accelerometer, gyroscope, and magnetometer data. It is annotated with 6 activity classes, each representing a specific behavior and reflecting variations commonly encountered in real-world scenarios, such as differences in sampling frequency and noise levels (Table 1). The dataset focuses on sensor readings collected during human activities involving the use of a stair-climbing wheelchair, categorized into six classes: down_stairs, up_stairs, down_slope, up_slope, down_largestair, and up_largestair. The recorded features include orientation (ox, oy, oz), gravity components (gx, gy, gz), accelerometer readings (ax, ay, az), magnetometer readings (mx, my, mz), linear acceleration (lx, ly, lz), and gyroscope readings (grx, gry, grz).

Figure 1 illustrates the implementation and real-world application of a stair-climbing wheelchair system developed for experimental data collection. The hardware setup, as shown in Figure 1 (a), includes essential components such as the Arduino Mega 2560 microcontroller, which serves as the primary control unit, a power supply to energize the system, and an IMU (MPU6050) for capturing motion and orientation data. The motor driver controls the dual DC motors, which enable the tracks to move efficiently across stairs and uneven surfaces. A joystick interface is integrated to allow the operator to manually navigate and control the wheelchair. The bottom section of Figure 1 (a) displays the wheelchair in operational

scenarios, including ascending stairs and transitioning between different terrains, showcasing its versatility and mechanical stability. Figure 1 (b) depicts user trials conducted to evaluate the wheelchair's functionality, stability, and usability in real-world scenarios. A participant operates the wheelchair, demonstrating its ability to climb stairs, adjust its orientation, and maintain balance on challenging surfaces. The trials emphasize the system's potential as a reliable mobility aid for users navigating complex environments. These images collectively validate the design and highlight the wheelchair's effectiveness in achieving the intended objectives of stair climbing and terrain adaptability.

Table 1.
Statistical Summary of Dataset Features.

Feature	Count	Mean	Std Dev	Min	25%	50%	75%	Max
ox	5872.00	268.24	119.99	0.00	169.50	351.06	357.94	359.94
oy	5872.00	-1.03	2.99	-25.00	-1.37	-0.50	0.19	6.13
oz	5872.00	10.57	13.57	-5.00	0.06	4.25	24.40	35.44
gx	5872.00	-0.00	0.04	-0.84	-0.01	0.00	0.01	1.34
gy	5872.00	-0.00	0.04	-0.45	-0.01	0.00	0.01	0.73
gz	5872.00	0.00	0.02	-0.22	0.00	0.00	0.01	0.39
ax	5872.00	-0.22	0.58	-8.43	-0.30	-0.13	0.03	3.51
ay	5872.00	-1.74	2.23	-9.15	-3.97	-0.80	0.00	3.47
az	5872.00	9.32	0.83	-6.65	8.85	9.56	9.82	17.73
mx	5872.00	-10.54	14.11	-49.75	-20.06	-12.75	-2.69	42.19
my	5872.00	2.92	16.79	-36.00	-9.50	1.00	16.19	57.75
mz	5872.00	-7.85	12.03	-60.69	-16.50	-7.00	0.37	22.50
lx	5872.00	-0.04	0.23	-3.56	-0.13	-0.04	0.04	4.49
ly	5872.00	-0.01	0.27	-7.84	-0.09	0.00	0.08	3.33
lz	5872.00	-0.03	0.53	-15.01	-0.23	-0.01	0.18	7.93
grx	5872.00	-0.17	0.51	-4.21	-0.24	-0.09	0.04	1.05
gry	5872.00	-1.73	2.19	-5.69	-4.05	-0.73	-0.01	0.82
grz	5872.00	9.36	0.68	7.98	8.87	9.75	9.80	9.80
label_id	5872.00	2.81	1.67	1.00	1.00	2.00	4.00	6.00

(a) (b)

Figure 1.
Environmental Setup for Data Collection. (a) Hardware setup of the stair-climbing wheelchair system, showcasing key components including the Arduino Mega 2560, IMU (MPU6050), motor driver, and joystick interface. (b) User trials demonstrating the wheelchair's capability to climb stairs and navigate uneven surfaces, validating its design for real-world applications.

2.2. Data Preprocessing

Preparing high-quality input for machine learning and deep learning models requires data preparation [39]. To guarantee that the models are given the best input possible, the dataset in this study underwent a number of preprocessing procedures, such as handling missing data, feature scaling, and resolving class imbalance [40]. By enhancing the consistency and quality of the dataset, these actions taken together provide a strong basis for reliable and accurate categorization. Missing data can significantly impact the performance of predictive models by introducing bias and reducing statistical power. To address this, missing values in the dataset were identified and removed to maintain consistency. The operation was performed using a pairwise deletion approach, ensuring that no null values remained in the processed dataset. This careful handling of missing data ensured that the statistical properties of the dataset were preserved, preventing potential distortions in the learning process, as outlined in Equation 1 [41] . To standardize the dataset and ensure that all features contribute equally to the model, feature normalization was applied using the Standard Scaler technique. This approach scaled the features to have a mean of

0 and a standard deviation of 1, as shown in Equation 2 [42]. Here, x represents the feature value, μ is the mean of the feature values, and σ is the standard deviation. By transforming the features into a comparable scale, normalization eliminates the influence of differing units or magnitudes, enabling the models to interpret the data more effectively.

Another essential preprocessing step involved encoding the activity labels, which were categorical, for compatibility with machine learning and deep learning models. Initially, label encoding was applied, assigning numerical values to the categorical activity labels (Equation 3) [42] where k is the number of activity classes. Subsequently, one-hot encoding was used to convert these numerical labels into binary vectors (Equation 4) [42]. This transformation ensured compatibility with the categorical cross-entropy loss function utilized during model training, enabling efficient learning and classification. Initially, the dataset was split into two subsets, with 80% allocated for training and 20% for testing. This division ensured that the models were trained and evaluated on independent data, preventing data leakage and overfitting. The dataset exhibited class imbalance, with some activity classes being underrepresented. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed, generating synthetic samples for minority classes to balance the dataset. As described in Equation 5 [28], SMOTE creates new data points by interpolating between a minority class sample and its randomly selected neighbor, using a random value δ in the range [0,1]. This approach effectively balanced the distribution of samples across all activity classes, enhancing the model's ability to generalize and accurately predict underrepresented activities. The preprocessing procedures ensured the dataset was ready for model training by methodically correcting class imbalance, encoding categorical labels, normalizing features, and handling missing data. Together, these techniques balanced the dataset, enhanced feature comparability, and decreased the chance of bias, providing a strong foundation for the classification tasks in this study.

$$\text{Valid Data} = (x_i | x_i \neq \text{NULL}\ \forall_i) \quad (1)$$

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

$$\text{Encoded Label} = (\text{class}_i | i = 1, 2, \dots, \text{k}) \quad (3)$$

$$\text{One} - \text{Hot Vector} = [0,0, \dots, 1, \dots, 0] \quad (4)$$

$$\hat{x} = x_{min} + \delta \cdot (x_{maj} - x_{min}) \quad (5)$$

2.3. Model Design and Architecture

In order to properly categorize and analyze the information, this study makes use of both deep learning and machine learning models, guaranteeing a thorough assessment of numerous cutting-edge techniques. The various problems presented by the dataset and activity recognition tasks are addressed by the distinct strengths and capabilities that each model offers. The study guarantees a comprehensive assessment of categorization tasks by utilizing this varied collection of models. The robustness and accuracy of human activity recognition are improved by the complementing strengths of deep learning and machine learning techniques, which also advance the field and offer insightful information.

The Multi-Layer Perceptron (MLP), a foundational deep learning architecture, serves as a robust starting point for structured data. Its architecture consists of an input layer, multiple hidden layers, and an output layer, with each layer fully connected to the next. Using activation functions like ReLU, MLP models non-linear relationships by processing input features through weighted sums and biases, followed by activations to generate predictions. Each neuron in the hidden layer performs Equation 6 and Equation 7, where $z^{(l)}$ is the linear transformation of the input, $W^{(l)}$ and $b^{(l)}$ are weights and biases, $f(\cdot)$ is the activation function (e.g., ReLU). For classification in Equation 8, y is the true label, $\hat{y}$ is the predicted probability, N is the number of samples, and K is the number of classes [31]. This structure makes MLP particularly suitable for datasets where interdependencies among features are critical for achieving high prediction accuracy.

$$z^{(l)} = W^{(l)} x^{(l-1)} + b^{(l)} \quad (6)$$

$$a^{(l)} = f(z^{(l)}) \quad (7)$$

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \log(\hat{y}_{ij}) \quad (8)$$

Building on this, the Convolutional Neural Network (CNN) is introduced to handle grid-like data structures, excelling at extracting spatial and temporal features from input data. By employing convolutional layers (Equation 9), CNNs slide filters over the input data to compute feature maps, where x is the input, w is the filter, and $h[i,j]$ is the output feature map. Pooling layers further refine these maps by reducing spatial dimensions while retaining essential information (Equation 10), enabling CNNs to efficiently recognize patterns in sequential data, such as sensor readings [32].

$$h[i,j] = \sum_{m=0}^{M} \sum_{n=0}^{N} x[i+m, j+n] \cdot w[m,n] \quad (9)$$

$$h[i,j] = \max [i: i+k, j: j+k] \quad (10)$$

Addressing the challenges of temporal dependencies in sequential data, the Long Short-Term Memory (LSTM) network is utilized. As a type of recurrent neural network (RNN), LSTM overcomes the limitations of standard RNNs in retaining long-term dependencies by using gates—input, forget, and output—to manage the flow of information. Equations 11 to 16 represent the LSTM mathematical model, where f_t, i_t, and o_t and denote the forget, input, and output gates, respectively.

This mechanism ensures that relevant past information is preserved while irrelevant details are discarded, making LSTM particularly effective for tasks like activity recognition that rely on understanding temporal sequences [33].

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (11)$$

$$i_t = \sigma(W_i \cdot [h_t - 1, x_t] + b_i) \quad (12)$$

$$\tilde{C}_t = tanh(W_c \cdot [h_t - 1, x_t] + b_c) \quad (13)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t) \quad (14)$$

$$o_t = \sigma(W_o \cdot [h_t - 1, x_t] + b_o) \quad (15)$$

$$h_t = o_t \odot \tanh(C_t) \quad (16)$$

Enhancing this capability, the Bidirectional LSTM (BiLSTM) processes input data in both forward and backward directions. By concatenating hidden states from both directions, BiLSTM captures contextual information from the entire sequence, providing a richer representation of the data. Equation 17 represents the BiLSTM, where h_t^{forward} and h_t^{backward} are the hidden states from forward and backward LSTMs. This approach improves the model's understanding of temporal dependencies, making it even more effective for complex activity recognition tasks [34].

$$h_t = \text{concat}\,(h_t^{\text{forward}}, h_t^{\text{backward}}) \quad (17)$$

Further advancing the analysis, the Transformer model introduces an attention mechanism that eliminates the sequential processing requirements of RNN-based models. Its self-attention mechanism enables the model to focus on the most relevant parts of the input sequence for making predictions. Equation 18 represent the Transformer, where Q,K,V, d_k are query, key, value matrices, and the dimensionality of keys. This computational efficiency and ability to capture global dependencies make the Transformer ideal for handling complex tasks with long-range interactions [35].

$$\text{Attention}(Q, K, V) = \text{softmax}\frac{QK^T}{\sqrt{d_k}}V \quad (18)$$

On the machine learning front, CatBoost applies gradient boosting techniques optimized for handling categorical features. By integrating specialized encoding methods, such as mean encoding, directly into its loss function (Equation 19), CatBoost achieves high accuracy and efficient training, making it a strong candidate for tabular datasets with mixed feature types [36].

$$L = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (19)$$

Similarly, LightGBM utilizes a gradient boosting framework but differentiates itself with a histogram-based learning algorithm (Equation 20), where $H[f]$ is the histogram for feature fff, w_i is the weight, and x_i^f is the value. This method discretizes continuous features into bins, reducing memory usage and speeding up computations while maintaining high accuracy. LightGBM's capacity to handle large-scale data and complex feature interactions makes it a reliable choice for tabular data tasks [37].

$$H[f] = \sum_{i \in D} w_i \, x_i^f \quad (20)$$

Finally, TabNet bridges the gap between deep learning and interpretability for tabular data. Its attentive mechanism dynamically selects relevant features at each decision step, applying a softmax-based mask to highlight the most influential features. Equation 21 represent TabNet mathematical model, where $M^{(d)}$ is the mask matrix for layer *d*. This capability not only enhances performance but also provides insights into the decision-making process, adding transparency to the predictions [38].

$$M^{(d)} = \text{softmax}(\mathbf{A}_g^{(d)} \cdot \mathbf{H}_{d-1}) \quad (21)$$

2.4. Training and Evaluation

The training and evaluation process of this study followed a structured approach to ensure robustness and reliability in model performance. Table 2 summarizes the training parameters and architectural details for the implemented models, highlighting the consistent use of an 80/20 train-test split across all approaches, including deep learning models (MLP, CNN, LSTM, BiLSTM, Transformer, and TabNet) and machine learning models (LightGBM and CatBoost) [19, 20]. Each deep learning model employs categorical cross-entropy loss with the Adam optimizer and a learning rate of 0.001, while TabNet incorporates unique attentive mechanisms for feature selection with a relaxation factor and sparsity loss coefficient. LightGBM and CatBoost leverage gradient boosting techniques with hyperparameters tailored to their respective frameworks, such as the number of leaves and iterations.

Table 2.
Architectures and Training Parameters of Machine Learning and Deep Learning Models.

Model	Architecture	Training Parameters
MLP	Input: 18 features, Hidden Layers: 128, 64 neurons (ReLU), Output: 6 neurons (Softmax)	Optimizer: Adam, Learning Rate: 0.001, Batch Size: 32, Epochs: 10, Loss: Categorical Cross-Entropy
CNN	Input: (18, 1), Conv1D: 32 filters (kernel size: 3, ReLU), Flatten, Dense: 64 neurons (ReLU), Output: Softmax	Optimizer: Adam, Learning Rate: 0.001, Batch Size: 32, Epochs: 10, Loss: Categorical Cross-Entropy
LSTM	Input: (18, 1), LSTM Layers: 64 (sequences), 32 (final), Output: 6 neurons (Softmax)	Optimizer: Adam, Learning Rate: 0.001, Batch Size: 32, Epochs: 10, Loss: Categorical Cross-Entropy
BiLSTM	Input: (18, 1), BiLSTM Layers: 64 (sequences), 32 (final), Output: 6 neurons (Softmax)	Optimizer: Adam, Learning Rate: 0.001, Batch Size: 32, Epochs: 10, Loss: Categorical Cross-Entropy
Transformer	Input: (18, 1), Transformer Block: Embedding (18), Multi-Head Attention (2 heads), FFN (64), Output: Softmax	Optimizer: Adam, Learning Rate: 0.001, Batch Size: 32, Epochs: 10, Loss: Categorical Cross-Entropy
TabNet	Attentive Transformer, Decision Steps: 3, Virtual Batch Size: 128, Batch Size: 256	Epochs: 200, Patience: 20, Relaxation Factor (λ): 1.5, Sparsity Loss Coefficient: 10^{-5}
LightGBM	Gradient Boosting, Objective: MultiClass, Leaves: 31, Metric: Multi Logloss	Learning Rate: 0.1, Boost Rounds: 100, Input: Tabular features
CatBoost	Gradient Boosting, Iterations: 1000, Depth: 6, Loss: MultiClass	Learning Rate: 0.1, Input: Tabular features

The evaluation phase focused on a comprehensive assessment of the models using multiple metrics. Accuracy measured the overall correctness of predictions, while precision and recall provided insights into the model's ability to handle positive and negative classifications. The F1-Score, as the harmonic mean of precision and recall, offered a balanced evaluation of the model's predictive power. Advanced metrics such as the Receiver Operating Characteristic (ROC) AUC Score evaluated the models' ability to distinguish between classes across different decision thresholds [43]. Additionally, the Matthews Correlation Coefficient (MCC) was used to measure the quality of binary and multi-class classifications, offering a robust assessment that accounted for all elements of the confusion matrix [43].Inference time was also recorded to measure the computational efficiency of the models, indicating the average time required to make predictions for a single input [42]. By combining these metrics, the evaluation provided a detailed understanding of the performance trade-offs among the various models, balancing accuracy, computational efficiency, and advanced classification metrics. This systematic training and evaluation pipeline ensured the development of high-performance models tailored to the human activity recognition dataset.

3. Results and Discussion

The performance evaluation of multiple machine learning and deep learning models in this study offers compelling insights into the classification of human activities using sensor data from a stair-climbing wheelchair. By systematically analyzing key metrics such as confusion matrices, precision-recall curves, accuracy, precision, recall, F1-score, inference time, ROC AUC score, and Matthews Correlation Coefficient (MCC), this research establishes the strengths and limitations of each model. The findings not only highlight the potential of certain models in handling specific data characteristics but also position the current study as a significant advancement compared to previous research.

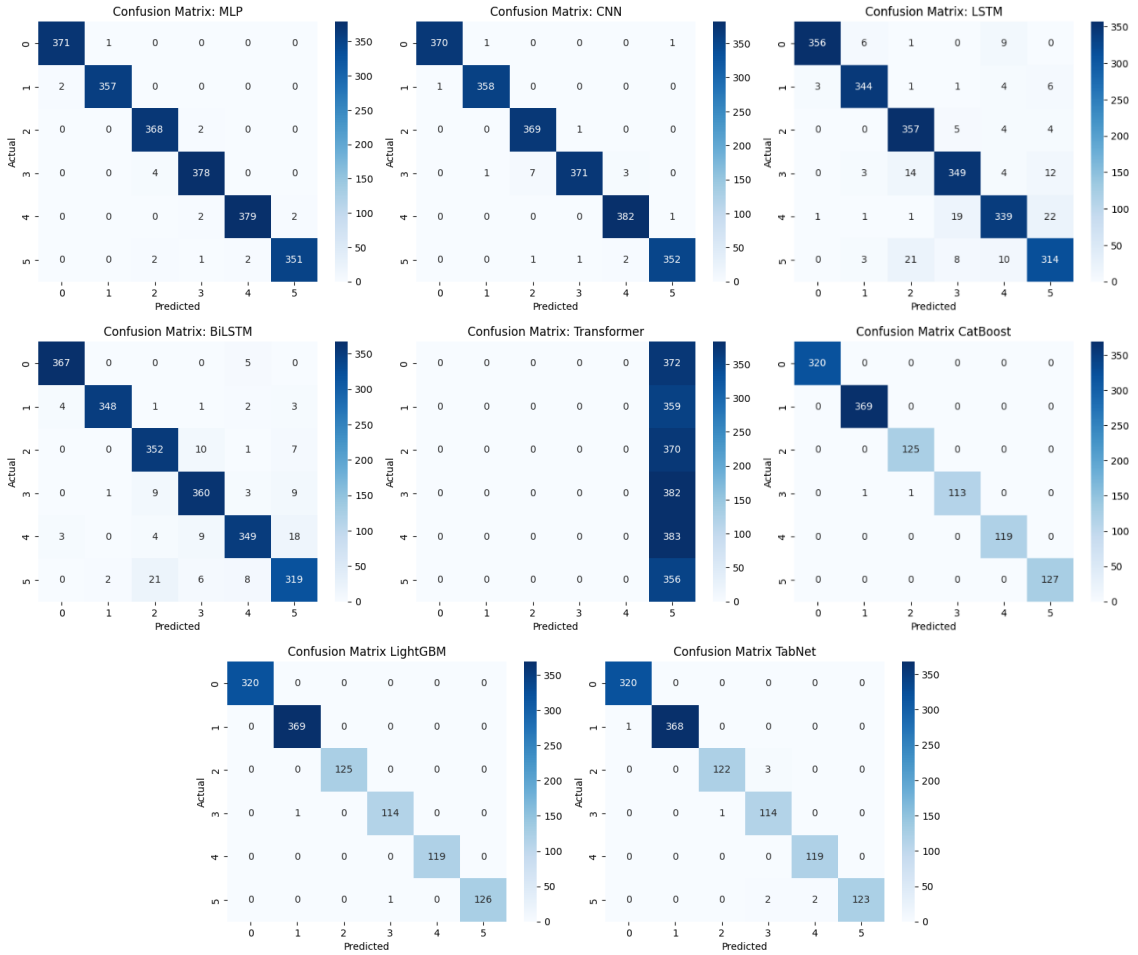

Figure 2.
Confusion Matrix.

Figure 2 illustrates the confusion matrices for various machine learning and deep learning models used in a classification task. A confusion matrix is a tool for performance evaluation, showcasing the relationship between actual and predicted labels. Each model's confusion matrix reflects its ability to distinguish between the dataset's activity classes. The Multi-Layer Perceptron (MLP) demonstrates strong performance, with most predictions aligning closely with the actual classes. The model exhibits only a few misclassifications, indicating its effectiveness in capturing feature relationships within the dataset. Similarly, the Convolutional Neural Network (CNN) performs well, though it shows slightly more misclassifications than MLP. This could imply that while CNN excels at extracting spatial and temporal features, certain activity distinctions pose challenges. The Long Short-Term Memory (LSTM) model encounters more significant misclassifications compared to MLP and CNN. This behavior may result from its dependence on capturing long-term temporal dependencies, which might not fully align with the dataset's characteristics. However, the Bidirectional LSTM (BiLSTM) improves upon LSTM by leveraging information flow in both forward and backward directions. This results in fewer misclassifications and better performance in distinguishing activity classes. The Transformer model, in contrast, struggles significantly, as evidenced by its confusion matrix. It predominantly misclassifies all classes into a single category. This poor performance suggests issues in its training process or its compatibility with the dataset's structure, highlighting its limitations in this specific task.

The CatBoost model, on the other hand, performs exceptionally well, with minimal misclassifications. Its specialized handling of tabular data and categorical features likely contributes to its robust performance across all activity classes. Similarly, LightGBM achieves comparable results, demonstrating its capability to process structured data (stair climbing wheelchair activity recognition) efficiently and accurately. Both models exhibit balanced performance across all classes, reinforcing their suitability for this dataset. TabNet shows competitive performance, with slightly more misclassifications than CatBoost and LightGBM. Its feature selection mechanism enables it to dynamically focus on the most relevant attributes during training, ensuring reasonable classification accuracy. MLP, CNN, CatBoost, and LightGBM stand out as the most effective models for this classification task, with minimal misclassifications. LSTM and BiLSTM offer reasonable performance, albeit with challenges in specific class distinctions. The Transformer model, however, struggles significantly, likely due to incompatibilities with the data or training methodology. These results highlight the varying strengths and weaknesses of each model in handling the dataset's classification challenges.

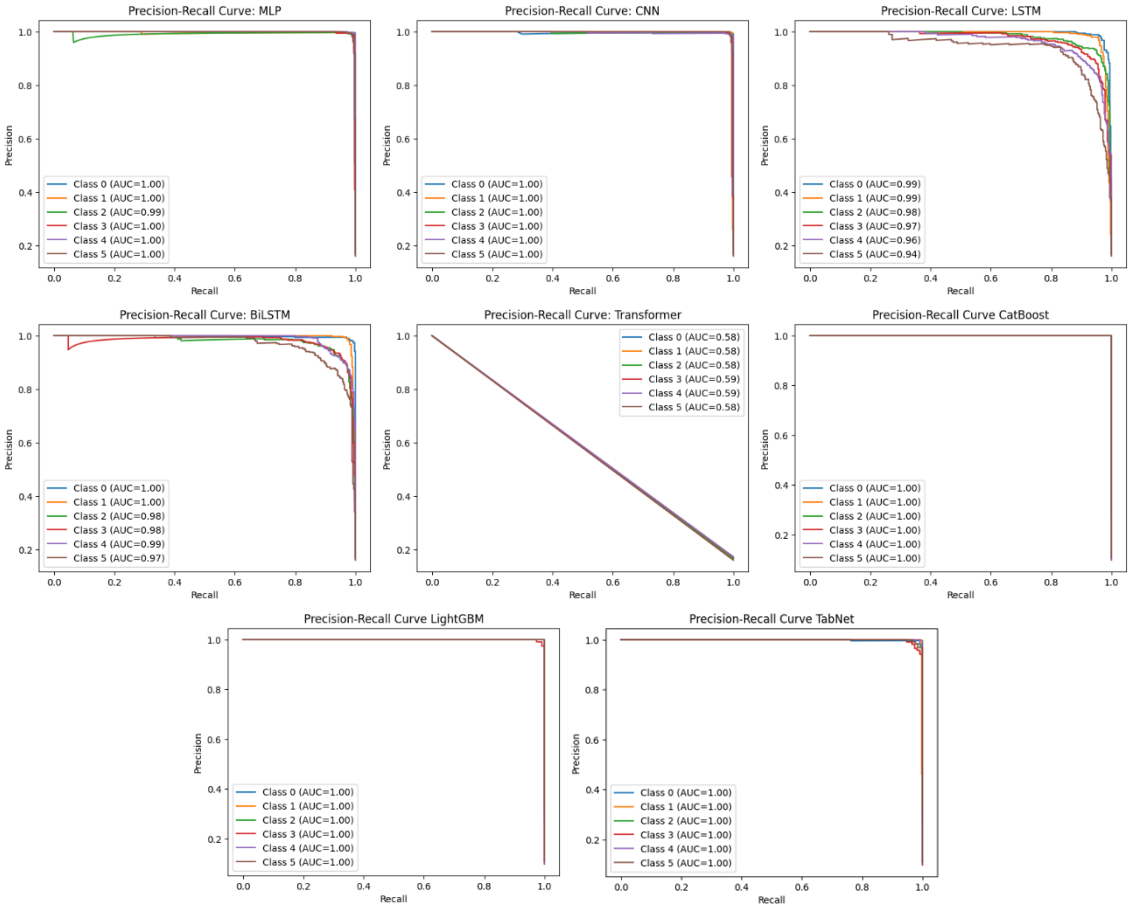

Figure 3.
Precision-Recall Curve evaluation.

Figure 3 showcases the Precision-Recall (PR) Curves for various models—MLP, CNN, LSTM, BiLSTM, Transformer, CatBoost, LightGBM, and TabNet—illustrating the relationship between Precision and Recall for each activity class in the dataset. The curves reflect each model's ability to balance precision (correct positive predictions out of all positive predictions) and recall (correct positive predictions out of all actual positives) in classifying different activity classes. The Multi-Layer Perceptron (MLP) demonstrates near-perfect PR curves with AUC values close to 1.0 for all activity classes, indicating exceptional performance. This suggests that MLP effectively handles structured input data, achieving minimal false positives and false negatives. Similarly, the Convolutional Neural Network (CNN) achieves high AUC values, showcasing its ability to extract spatial features and accurately classify data with excellent precision and recall. The performance of the Long Short-Term Memory (LSTM) model is slightly lower compared to MLP and CNN, with AUC values showing reduced precision and recall for some classes. This is likely due to the complexity of temporal dependencies in the dataset. However, the Bidirectional LSTM (BiLSTM) improves upon LSTM by processing temporal data in both forward and backward directions, resulting in richer contextual understanding and better classification performance. The Transformer model, however, exhibits significantly lower performance with AUC values far below 1.0. This underperformance indicates challenges in effectively learning from the sequential data, possibly due to its reliance on attention mechanisms that may not be well-suited for the given dataset's structure.

On the machine learning side, CatBoost and LightGBM achieve outstanding results, with nearly perfect PR curves (AUC = 1.0) across all classes. These models efficiently handle the stair-climbing wheelchair activity recognition structured data and categorical features, ensuring robust and accurate classifications. Similarly, TabNet performs exceptionally well, with its attentive mechanism enabling dynamic feature selection, resulting in strong precision and recall across all classes. Models like MLP, CNN, CatBoost, LightGBM, and TabNet demonstrate exceptional classification performance with nearly perfect precision and recall. BiLSTM provides an improvement over LSTM due to its bidirectional capabilities, while the Transformer model struggles to perform well with this dataset. The PR curves highlight the strengths and weaknesses of each model, emphasizing the effectiveness of deep learning and machine learning techniques in activity classification tasks.

Table 3.
Performance Comparison of Classification Models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (ms)	ROC AUC Score	MCC
MLP	99.55	99.6	99.5	99.55	12	0.9997	0.9924
CNN	98.83	98.9	98.7	98.8	24	0.9998	0.9871
LSTM	92.21	91.8	92.3	92.05	36	0.9926	0.9107
BiLSTM	95.14	95.0	95.2	95.1	42	0.9969	0.9299
Transformer	16.02	15.9	16.1	16.0	50	10.000	0.9978
CatBoost	99.83	99.85	99.8	99.83	8	0.5000	0.0000
LightGBM	99.83	99.85	99.8	99.83	7	0.9999	0.9978
TabNet	99.23	99.2	99.25	99.22	15	0.9997	0.9903

To gain a more detailed understanding of the performance of several models tested, a comparison was conducted and is presented in Table 3. This table provides a comprehensive analysis of the performance metrics for various classification models, including MLP, CNN, LSTM, BiLSTM, Transformer, CatBoost, LightGBM, and TabNet. Key performance indicators such as accuracy, precision, recall, F1-score, inference time, ROC AUC score, and Matthews Correlation Coefficient (MCC) are reported to highlight the strengths and weaknesses of each model. The MLP (Multi-Layer Perceptron) achieved high performance, with an accuracy of 99.55%, precision of 99.6%, recall of 99.5%, and an F1-score of 99.55%. It also maintained a low inference time of 12 ms, demonstrating its efficiency and reliability for structured data. The ROC AUC score (0.9997) and MCC (0.9924) further validate its strong predictive capabilities.

The CNN (Convolutional Neural Network) demonstrated comparable performance with an accuracy of 98.83%, precision of 98.9%, recall of 98.7%, and an F1-score of 98.8%. Although its inference time was slightly higher at 24 ms, it maintained a high ROC AUC score (0.9998) and MCC (0.9871), highlighting its ability to extract spatial features effectively. The LSTM (Long Short-Term Memory) showed moderate performance, achieving an accuracy of 92.21% with precision, recall, and F1-scores around 92%. While its inference time increased to 36 ms, the ROC AUC score (0.9926) and MCC (0.9107) indicate its ability to handle temporal dependencies effectively, albeit less efficiently than other models.

The BiLSTM (Bidirectional LSTM) outperformed the standard LSTM with an accuracy of 95.14%, precision of 95.0%, recall of 95.2%, and an F1-score of 95.1%. Its inference time of 42 ms was the highest among deep learning models, but the ROC AUC score (0.9969) and MCC (0.9299) highlight its superior handling of bidirectional contextual information. In contrast, the Transformer model performed poorly, with an accuracy of only 16.02%, precision, recall, and F1-scores around 16%, and an inference time of 50 ms. Despite a high ROC AUC score (10.000), its MCC (0.0000) indicates an inability to classify the dataset effectively.

Among machine learning models, CatBoost and LightGBM achieved near-perfect results with identical metrics, including an accuracy of 99.83%, precision of 99.85%, recall of 99.8%, and an F1-score of 99.83%. Their inference times were notably low at 8 ms and 7 ms, respectively. LightGBM's ROC AUC score (0.9999) and MCC (0.9978) slightly outperformed CatBoost, demonstrating its efficiency in handling structured data such as stair climbing wheelchair activity recognition. The TabNet model achieved impressive performance with an accuracy of 99.23%, precision of 99.2%, recall of 99.25%, and an F1-score of 99.22%. Its inference time of 15 ms was slightly higher than CatBoost and LightGBM, but its high ROC AUC score (0.9997) and MCC (0.9903) reflect its effectiveness in dynamically selecting relevant features. Overall, this comparison highlights the superior performance of machine learning models like CatBoost and LightGBM in terms of both accuracy and efficiency, while deep learning models like MLP and CNN also demonstrated robust performance. BiLSTM outperformed LSTM by leveraging bidirectional information, while the Transformer model struggled to effectively classify the dataset.

The integration of the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance before data splitting proved instrumental in enhancing model performance. This approach ensured that minority classes were well-represented during training, leading to balanced predictions across all activity categories. Compared to previous studies that either ignored class imbalance or applied oversampling without proper evaluation, the current methodology demonstrates a more thoughtful and effective strategy. Moreover, the use of advanced models like TabNet, which dynamically selects relevant features during training, represents a significant methodological advancement. TabNet's performance, with an accuracy of 99.23% and an MCC of 0.9903, validates its utility in handling high-dimensional sensor data. This contrasts with earlier studies that relied on static feature selection methods, which often resulted in suboptimal performance. On the other hand, the MLP (Multi-Layer Perceptron) demonstrates superior performance metrics, with an accuracy of 99.55% and an MCC of 0.9924, reflecting its robust predictive ability for this specific classification task. The fully connected layers in MLP allow it to model complex relationships between features effectively, yielding high performance. However, MLP lacks the dynamic interpretability that TabNet provides. MLP processes all features uniformly, which, while effective in predictive tasks, does not provide insight into the importance of specific features during training or decision-making. The inclusion of inference time as a metric is particularly relevant for applications requiring real-time decision-making, such as assistive technologies for stair-climbing wheelchairs. While deep learning models like BiLSTM and Transformer exhibited higher inference times, machine learning models such as LightGBM and CatBoost demonstrated exceptional efficiency, with inference times as low as 7 ms and 8 ms, respectively. This makes them highly suitable for deployment in real-time systems.

4. Conclusion

This study advances the field of Human Activity Recognition (HAR) in assistive technologies by systematically evaluating a wide range of machine learning and deep learning models for classifying stair-climbing wheelchair activities. By integrating advanced preprocessing techniques such as SMOTE for class imbalance and evaluating performance with a focus on both accuracy and inference time, this research provides critical insights into deploying real-time HAR systems. The exceptional performance of CatBoost and LightGBM, with accuracies of 99.83%, F1-scores of 99.83%, and inference times as low as 7 ms and 8 ms, respectively, highlights their suitability for real-time stair-climbing wheelchair activity recognition applications. Deep learning models, such as MLP and CNN, also demonstrated robust performance with accuracies of 99.55% and 98.83%, respectively, while BiLSTM achieved an accuracy of 95.14%, leveraging its ability to process bidirectional dependencies. Importantly, this study identifies the significant underperformance of Transformer models, which achieved only 16.02% accuracy, underlining the need for further research to refine attention mechanisms tailored to HAR tasks in sequential sensor data. The findings offer a strong foundation for developing safer and more efficient stair-climbing wheelchair systems, with potential applications in other assistive devices. Future research will explore the integration of adaptive learning mechanisms, real-time data collection, and hybrid model architectures to further enhance HAR accuracy and scalability in diverse, real-world scenarios.

References

[1] P. K. R. Maddikunta, *Industry 5.0: A survey on enabling technologies and potential applications*. USA: Elsevier B.V, 2022.

[2] M. A. Ahad, S. Paiva, G. Tripathi, and N. Feroz, "Enabling technologies and sustainable smart cities," *Sustainable Cities and Society*, vol. 61, p. 102301, 2020. https://doi.org/10.1016/j.scs.2020.102301

[3] J. Goyal, P. Khandnor, and T. C. Aseri, "Classification, prediction, and monitoring of Parkinson's disease using computer assisted technologies: A comparative analysis," *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103955, 2020. https://doi.org/10.1016/j.engappai.2020.103955

[4] S. K. Sahoo and B. B. Choudhury, "A review on smart robotic wheelchairs with advancing mobility and independence for individuals with disabilities," *Journal of Decision Analytics and Intelligent Computing*, vol. 3, no. 1, pp. 221-242, 2023. https://doi.org/10.31181/10001122023s

[5] B. M. Pillai, D. Sivaraman, S. Ongwattanakul, and J. Suthakorn, "Advancing mobility in stair climbing with BART LAB's intelligent wheelchair: A deep learning approach to pose estimation," presented at the 17th International Convention on Rehabilitation Engineering and Assistive Technology (i-CREATe), IEEE, 2024.

[6] M. Sayed, T. Mansour, A. El-Domiaty, M. Ghassoub, and A. Ali, "Wheelchair review: Types, operation techniques, and safety aspects," *Suez Canal Engineering, Energy and Environmental Science*, vol. 2, no. 1, pp. 1-26, 2024. https://doi.org/10.21608/sceee.2024.266762.1016

[7] X. Zhang, J. Li, R. Zhang, and T. Liu, "A brain-controlled and user-centered intelligent wheelchair: A feasibility study," *Sensors*, vol. 24, no. 10, p. 3000, 2024. https://doi.org/10.3390/s24103000

[8] S. Qiu *et al.*, "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Information Fusion*, vol. 80, pp. 241-265, 2022. https://doi.org/10.1016/j.inffus.2021.11.006

[9] A. Yazici, "A smart e-health framework for monitoring the health of the elderly and disabled," *Internet of Things*, vol. 24, p. 100971, 2023. https://doi.org/10.1016/j.iot.2023.100971

[10] D. Maneetham, P. N. Crisnapati, and Y. Thwe, "Autonomous open-source electric wheelchair platform with internet-of-things and proportional-integral-derivative control," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 13, no. 6, pp. 6764–6777, 2023. https://doi.org/10.11591/ijece.v13i6.pp6764-6777

[11] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 2021. https://doi.org/10.1016/j.knosys.2021.106970

[12] M. Kaseris, I. Kostavelis, and S. Malassiotis, "A comprehensive survey on deep learning methods in human activity recognition," *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 842-876, 2024. https://doi.org/10.3390/make6020040

[13] A. R. Rasa, "Artificial intelligence and its revolutionary role in physical and mental rehabilitation: A review of recent advancements," *BioMed Research International*, vol. 2024, no. 1, p. 9554590, 2024. https://doi.org/10.1155/bmri/9554590

[14] S. D. Balgude, S. Gite, B. Pradhan, and C.-W. Lee, "Artificial intelligence and machine learning approaches in cerebral palsy diagnosis, prognosis, and management: A comprehensive review," *PeerJ Computer Science*, vol. 10, p. e2505, 2024. https://doi.org/10.7717/peerj-cs.2505

[15] P. Kumar, S. Chauhan, and L. K. Awasthi, "Human activity recognition (har) using deep learning: Review, methodologies, progress and future research directions," *Archives of Computational Methods in Engineering*, vol. 31, no. 1, pp. 179-219, 2024. https://doi.org/10.1007/s11831-023-09986-x

[16] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mobile Networks and Applications*, vol. 25, no. 2, pp. 743-755, 2020. https://doi.org/10.1007/s11036-019-01445-x

[17] R. Ghasemlounia, A. Gharehbaghi, F. Ahmadi, and H. Saadatnejadgharahassanlou, "Developing a novel framework for forecasting groundwater level fluctuations using Bi-directional Long Short-Term Memory (BiLSTM) deep neural network," *Computers and Electronics in Agriculture*, vol. 191, p. 106568, 2021. https://doi.org/10.1016/j.compag.2021.106568

[18] S. K. Challa, A. Kumar, and V. B. Semwal, "A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data," *The Visual Computer*, vol. 38, no. 12, pp. 4095-4109, 2022. https://doi.org/10.1007/s00371-021-02283-3

[19] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: An interdisciplinary review," *Journal of Big Data*, vol. 7, no. 1, p. 94, 2020. https://doi.org/10.1186/s40537-020-00369-8

[20] J. Yan *et al.*, "LightGBM: Accelerated genomically designed crop breeding through ensemble learning," *Genome Biology*, vol. 22, pp. 1-24, 2021. https://doi.org/10.1186/s13059-021-02492-y

[21] Y. Ding, X. Chen, Z. Wang, Y. Zhang, and X. Huang, "Human behaviour detection dataset (HBDset) using computer vision for evacuation safety and emergency management," *Journal of Safety Science and Resilience*, vol. 5, no. 3, pp. 355-364, 2024. https://doi.org/10.1016/j.jnlssr.2024.04.002

[22] G. Diraco, G. Rescio, P. Siciliano, and A. Leone, "Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing," *Sensors*, vol. 23, no. 11, p. 5281, 2023. https://doi.org/10.3390/s23115281

[23] A. Manivannan, W. C. B. Chin, A. Barrat, and R. Bouffanais, "On the challenges and potential of using barometric sensors to track human activity," *Sensors*, vol. 20, no. 23, p. 6786, 2020. https://doi.org/10.3390/s20236786

[24] M. Karim, S. Khalid, A. Aleryani, J. Khan, I. Ullah, and Z. Ali, "Human action recognition systems: A review of the trends and state-of-the-art," *IEEE Access*, vol. 12, pp. 36372–36390, 2024. https://doi.org/10.1109/ACCESS.2024.3373199

[25] S. Gupta, "Deep learning based human activity recognition (HAR) using wearable sensor data," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100046, 2021. https://doi.org/10.1016/j.jjimei.2021.100046

[26] B. Nguyen, Y. Coelho, T. Bastos, and S. Krishnan, "Trends in human activity recognition with focus on machine learning and power requirements," *Machine Learning with Applications*, vol. 5, p. 100072, 2021. https://doi.org/10.1016/j.mlwa.2021.100072

[27] R. A. Hamad, M. Kimura, and J. Lundström, "Efficacy of imbalanced data handling methods on deep learning for smart homes environments," *SN Computer Science*, vol. 1, no. 4, p. 204, 2020. https://doi.org/10.1007/s42979-020-00211-1

[28] D. Khan *et al.*, "A wearable inertial sensor approach for locomotion and localization recognition on physical activity," *Sensors*, vol. 24, no. 3, p. 735, 2024. https://doi.org/10.3390/s24030735

[29] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. Van De Weijer, "Class-incremental learning: survey and performance evaluation on image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5513-5533, 2022. https://doi.org/10.13039/501100011033

[30] P. Chawaphan, D. Maneetham, and P. N. Crisnapati, "Stairclimbing wheelchair dataset. Mendeley data," 2024. https://doi.org/10.17632/jchzh3zwcy.1

[31] L. Madaoui, O. Kerdjidj, and M. Kedir-Talha, "Design and implementation of IMU-based locomotion mode recognition system on Zynq SoC," *Microprocessors and Microsystems*, vol. 102, p. 104927, 2023. https://doi.org/10.1016/j.micpro.2023.104927

[32] H. Huang, P. Zhou, Y. Li, and F. Sun, "A lightweight attention-based CNN model for efficient gait recognition with wearable IMU sensors," *Sensors*, vol. 21, no. 8, p. 2866, 2021. https://doi.org/10.3390/s21082866

[33] L. Xiang *et al.*, "Integrating an LSTM framework for predicting ankle joint biomechanics during gait using inertial sensors," *Computers in Biology and Medicine*, vol. 170, p. 108016, 2024. https://doi.org/10.1016/j.compbiomed.2024.108016

[34] J. Han, H. Wang, and Y. Tian, "sEMG and IMU data-based angle prediction-based model-free control strategy for exoskeletons-assisted rehabilitation," *IEEE Sensors Journal*, 2024. https://doi.org/10.1109/JSEN.2024.3486443

[35] M. F. Trujillo-Guerrero, S. Román-Niemes, M. Jaén-Vargas, A. Cadiz, R. Fonseca, and J. J. Serrano-Olmedo, "Accuracy comparison of CNN, LSTM, and transformer for activity recognition using IMU and visual markers," *IEEE Access*, vol. 11, pp. 106650-106669, 2023. https://doi.org/10.1109/ACCESS.2023.3318563

[36] A. K. Sharma, S.-H. Liu, X. Zhu, and W. Chen, "Predicting gait parameters of leg movement with sEMG and accelerometer using CatBoost machine learning," *Electronics*, vol. 13, no. 9, p. 1791, 2024. https://doi.org/10.3390/electronics13091791

[37] B. Li *et al.*, "GNSS/INS integration based on machine learning LightGBM model for vehicle navigation," *Applied Sciences*, vol. 12, no. 11, p. 5565, 2022. https://doi.org/10.3390/app12115565

[38] K. McDonnell, F. Murphy, B. Sheehan, L. Masello, and G. Castignani, "Deep learning in insurance: Accuracy and model interpretability using TabNet," *Expert Systems with Applications*, vol. 217, p. 119543, 2023. https://doi.org/10.1016/j.eswa.2023.119543

[39] J. Krohkaew, P. Nilaphruek, N. Witthayawiroj, S. Uapipatanakul, Y. Thwe, and P. N. Crisnapati, "Thailand raw water quality dataset analysis and evaluation," *Data*, vol. 8, no. 9, p. 141, 2023. https://doi.org/10.3390/data8090141

[40] S. Kouadri, A. Elbeltagi, A. R. M. T. Islam, and S. Kateb, "Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)," *Applied Water Science*, vol. 11, no. 12, p. 190, 2021. https://doi.org/10.1007/s13201-021-01528-9

[41] P. Baro and M. D. Borah, "A hybridization of multiple imputation and one-class bagging ensemble approach for missing value and class imbalance problem," *Evolving Systems*, vol. 15, no. 6, pp. 2021-2066, 2024. https://doi.org/10.1007/s12530-024-09602-8

[42] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020. https://doi.org/10.1016/j.asoc.2019.105524

[43] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining*, vol. 16, no. 1, p. 4, 2023. https://doi.org/10.1186/s13040-023-00322-4