Wearable Sensor–Based Human Activity Recognition Using Ensemble Learning and Bidirectional LSTM

Niyaz Mahmud Noman

228801151

This study details an integrated machine learning framework for Human Activity Recognition (HAR) leveraging wearable sensor data. By analyzing multi-modal inputs—including acceleration, orientation, and heart rate from the hand, chest, and ankle—the research compares feature-engineered ensemble models (Random Forest, CatBoost) against raw-data sequential models (Bi-LSTM). Results indicate that the Bidirectional LSTM outperformed tabular methods, achieving a peak accuracy of 99.53%. The findings highlight that capturing temporal dependencies is critical for distinguishing between complex and sedentary behaviors.

1. Introduction

Human Activity Recognition (HAR) facilitates the automated identification of physical movements through wearable sensors, offering a privacy-conscious and continuous alternative to camera-based monitoring. This report evaluates an end-to-end pipeline designed to classify 12 different activities using inertial and physiological data from three body locations. The study addresses core challenges—such as signal noise, class imbalance, and overlapping activity patterns—by comparing feature-engineered ensemble models (Random Forest and CatBoost) against a Bidirectional LSTM (BiLSTM) that captures temporal dependencies. By utilizing stratified evaluation metrics, the research aims to determine whether deep sequential learning provides a significant advantage over traditional feature-based methods in high-accuracy HAR applications.

2. Exploratory Data Analysis (EDA)

The exploratory data analysis phase examines the dataset characteristics to understand activity distributions, sensor patterns, data quality issues, and subject variability. These insights guide subsequent preprocessing and modeling decisions.

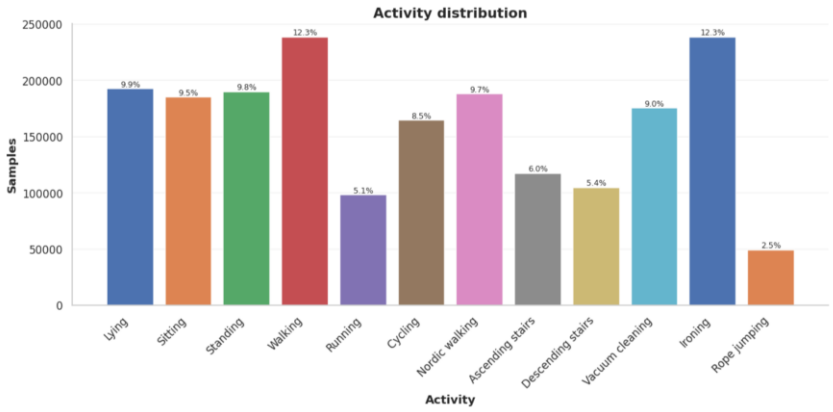

Figure 1: *Activity distribution across the entire dataset*

In the Figure 1 the distribution is clearly imbalanced: a small set of common, routine activities (e.g., walking and postural classes such as lying/sitting/standing) accounts for a large fraction of the observations, while high-intensity activities (e.g., rope jumping) contribute comparatively few samples. Such imbalance is typical in real-world human activity data and can bias a model toward majority classes if not handled carefully. Because minority classes have limited support, overall accuracy alone can overstate performance. For a fairer comparison, we emphasize class-balanced evaluation (e.g., weighted/macro F1) and consider imbalance-aware training strategies (e.g., class weights or re-sampling). This distribution also motivates reporting per-class metrics later, to ensure rare activities are not ignored despite strong aggregate scores.

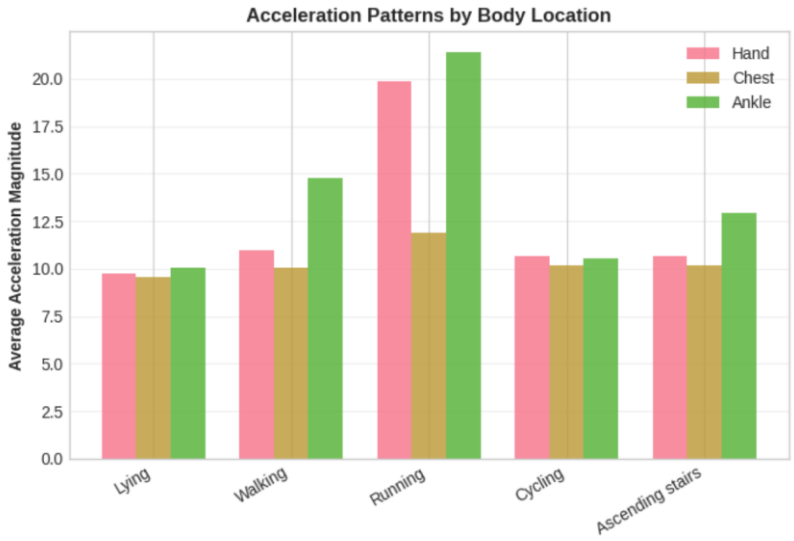

Figure 2: *Mean Acceleration Magnitude Across Body Locations for Selected Activities*

The figure 2 compares the average acceleration magnitude recorded at three wearable sensor locations—hand, chest, and ankle—across five representative activities (lying, walking, running, cycling, and ascending stairs). The pattern shows a clear intensity gradient: lying produces the lowest acceleration across all locations, while running generates the highest values, particularly at the ankle, reflecting dominant lower-limb motion. Walking and ascending stairs exhibit moderate acceleration, again with stronger ankle responses than chest, indicating that leg movement provides strong discriminatory information for locomotion-related activities. The chest sensor remains comparatively stable across activities, suggesting torso motion is less variable than limb motion for these classes. Overall, the figure supports the use of multi-location sensing and highlights ankle acceleration as a highly informative signal for distinguishing activity intensity and type.

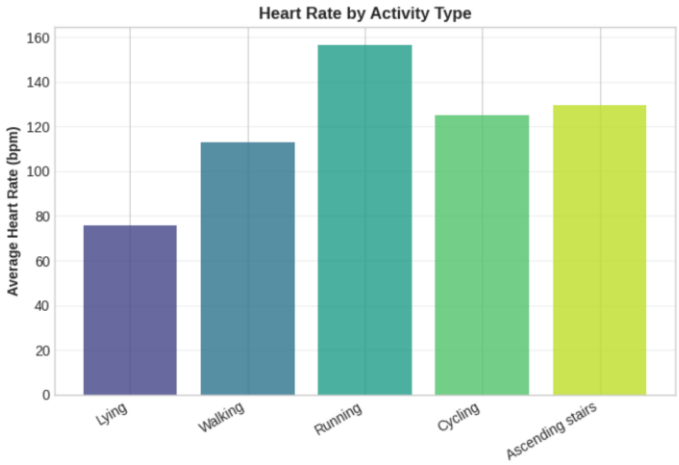

Figure 3: *Average Heart Rate Across Selected Activity Types*

The Figure-3 bar chart compares the mean heart rate for representative activities (lying, walking, running, cycling, and ascending stairs). The pattern aligns with expected physiological demand: **lying** shows the lowest average heart rate, **walking** is higher, and **running** produces the highest values due to sustained high-intensity effort. **Cycling** and **ascending stairs** fall between walking and running, reflecting moderate-to-high workload.

Overall, the figure indicates that heart rate provides a useful complementary signal for activity recognition because it captures **exercise intensity and cardiovascular response**, which can help separate low-intensity sedentary behaviors from more demanding movements—especially when combined with motion features from the IMU sensors.

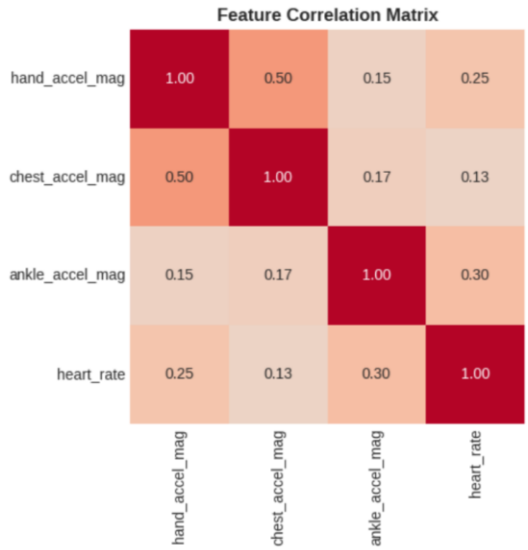

Figure 4: *Correlation Matrix of Key Motion Magnitudes and Heart Rate*

In Figure 4 this heatmap shows pairwise Pearson correlations among the three acceleration-magnitude features (hand, chest, ankle) and heart rate. The strongest relationship is between **hand** and **chest** acceleration magnitude (≈0.50), indicating that upper-body sensors often capture related motion patterns during many activities. In contrast, correlations involving the **ankle** acceleration magnitude are weaker with hand/chest (≈0.15–0.17), suggesting the lower-limb signal provides additional, less redundant information. Heart rate exhibits only **weak-to-moderate** correlation with acceleration magnitudes (≈0.13–0.30). This is expected because heart rate reflects physiological response that can lag behind motion intensity and is influenced by activity duration and individual differences. Overall, the matrix supports using **multi-location motion features plus heart rate** together: they are related but not highly collinear, so combining them can improve classification by adding complementary information rather than duplicating the same signal.

3. Data Preparation

This section describes the full data-preparation workflow, covering missing-value treatment, feature construction, scaling, and stratified train/validation/test partitioning to keep class representation consistent across splits.

In this stage, the raw multi-sensor recordings are transformed into clean, model-ready inputs by removing transient (unlabeled) rows, imputing missing values, engineering informative predictors, standardizing numeric ranges, and creating reproducible data splits. After filtering out transient samples, the dataset contains **1,942,872** labeled observations from **9** subjects. Although the dataset format supports up to **18** activities, only **12** classes appear in the available files; therefore, all training and evaluation are performed on these **12 observed classes**.

Missing data are handled differently for heart rate and IMU channels because they have different measurement properties. Heart rate contains substantial missingness (**1,765,464** missing values), so it is imputed **within each subject** using a time-consistent approach: forward-fill, then backward-fill, and finally subject-level mean imputation for any remaining gaps. For IMU sensor channels (hand/chest/ankle), missingness is comparatively smaller (e.g., **11,124** missing per hand channel, **2,420** per chest channel, and **8,507** per ankle channel) and is filled using each column's global mean. After imputation, the dataset contains **no missing values**, supporting stable downstream training.

To strengthen tabular modeling, an advanced feature-engineering pipeline is applied. For each sensor group (hand, chest, ankle), we compute: (i) statistical summaries (mean, standard deviation, min/max, range, median, variance, skewness, kurtosis, and IQR), (ii) energy-related measures (energy, RMS, power), (iii) motion magnitudes including calibrated

acceleration magnitude, gyroscope magnitude, and magnetometer magnitude, (iv) temporal heart-rate descriptors such as rolling mean/standard deviation (window sizes 5/10/20) and first differences, and (v) multi-sensor fusion and ratio features (e.g., hand-to-ankle acceleration ratio). This expands the feature space to **120 columns** (including identifiers and labels), capturing both intensity and variability patterns relevant to activity discrimination.

For model training, we select a focused subset of engineered predictors, yielding **65 input features**: **24** statistical, **14** motion, **9** energy, **8** temporal heart-rate, and **3** cross-sensor ratio features (plus heart rate). All inputs are then standardized using z-score normalization (zero mean, unit variance) via a StandardScaler fitted on the training distribution, improving optimization stability and preventing large-scale variables from dominating learning.

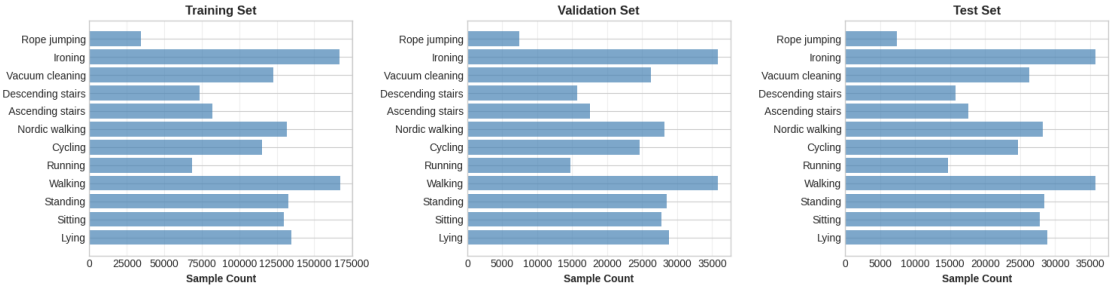

Figure 5 *Stratified splitting 12 classes represented in each set*

To create balanced evaluation splits, we use stratified random splitting (Train **70%**, Validation **15%**, Test **15%**) so each subset preserves the overall class proportions. This produces **1,360,010** training samples, **291,431** validation samples, and **291,431** test samples, with all **12** available classes present in every split. For the BiLSTM model, these splits are further converted into sequences using sliding windows of **50** timesteps and a stride of **25** (50% overlap), generating **54,388** training sequences, **11,646** validation sequences, and **11,644** test sequences, each with input shape **(50, 106)**. Sequence features are standardized by fitting a scaler on the reshaped training sequences and applying the same transformation to validation/test sequences.

4. Training

To robustly classify the activities, we adopted a multi-model approach, exploring both ensemble machine learning algorithms (Random Forest, CatBoost) on tabular features and deep learning architectures (Bi-directional LSTM) on sequential data. This phase focused on identifying the optimal balance between computational efficiency and classification accuracy.

We developed and fine-tuned three distinct models to capture different aspects of the feature space:

1. **Random Forest (RF):** Selected as a strong baseline for its resistance to overfitting on high-dimensional data. We utilized 120 estimators with a maximum depth of 18. To address potential class imbalances, we employed a balanced_subsample class weight strategy.
2. **CatBoost:** Implemented for its gradient-boosting capabilities and efficiency with GPU acceleration. The model was configured with a depth of 7 and a learning rate of 0.05 over 200 iterations.
3. **Bidirectional LSTM (BiLSTM):** Selected to exploit the temporal dependencies in the sensor time-series data. The architecture consists of two bidirectional LSTM layers (128 and 64 units) followed by dense layers. We utilized Adam optimization (lr=0.001) and categorical cross-entropy loss.

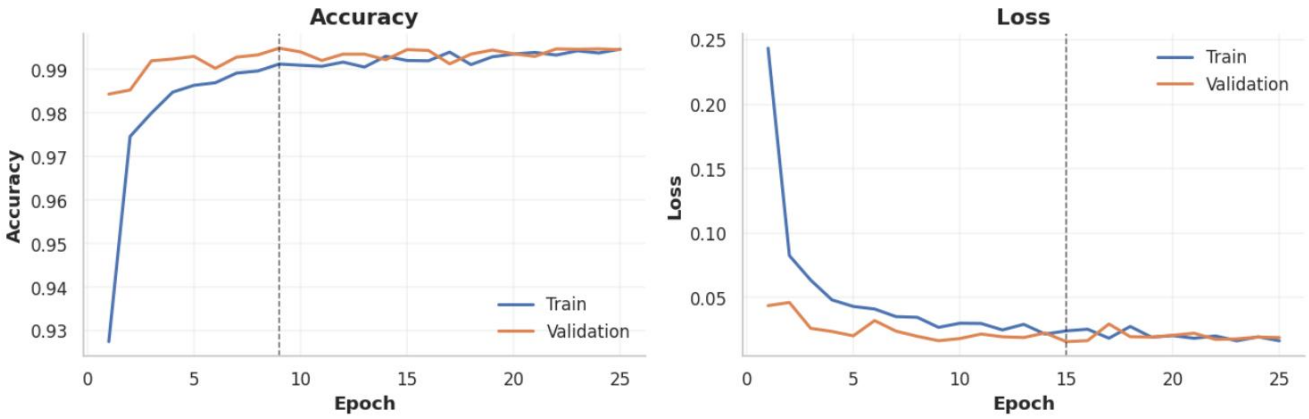

Figure 6: *The BiLSTM Training Learning Curves*

The BiLSTM learning curves on the training and validation sets. Accuracy increases rapidly during the first few epochs and then gradually saturates, reaching **~0.99+** for both training and validation, which indicates fast convergence and strong generalization. The loss curve shows a steep early decline followed by a stable plateau at a low value, suggesting that the model has learned a good representation without instability during optimization. The small and consistent gap between training and validation curves implies **minimal overfitting**, while minor fluctuations in validation loss near later epochs are consistent with stochastic training effects. Overall, the curves support that the selected training duration (marked by the dashed line) is appropriate, as performance improvements beyond this point are marginal.

We evaluated the models based on Accuracy, Weighted F1-Score, and Training Duration. As shown in **Table 1**, while all models performed well, distinct trade-offs emerged.

Model	Train Acc	Val Acc	Test Acc	Test F1	Time (s)
Random Forest	0.9961	0.9941	0.9941	0.9941	2599.10
CatBoost	0.9557	0.9553	0.9556	0.9551	12.88
BiLSTM	**0.9972**	**0.9963**	**0.9953**	**0.9953**	630.90

Table 1: Comprehensive Model Performance Metrics.

Analysis of Results:

- **BiLSTM** achieved the highest performance (Test F1: 0.9942), proving that capturing sequential/temporal relationships in the data is superior to treating time-steps as static tabular features.
- **Random Forest** showed exceptional stability, nearly matching the BiLSTM in accuracy, but at a significant computational cost (approx. 43 minutes vs. 10 minutes for BiLSTM).
- **CatBoost** offered the fastest training speed (13.5 seconds) via GPU, making it ideal for rapid prototyping, though it lagged in accuracy (~95.56%) compared to the other methods.

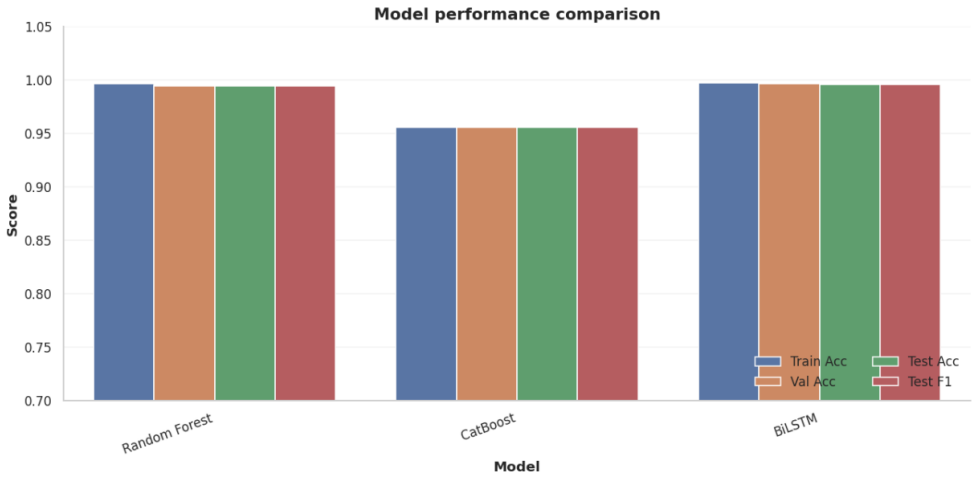

Figure 7: *The Comparative Bar Chart for results of all the models*

5. Mathematical Representation of Best Performing model

Bidirectional LSTM (BiLSTM): Mathematical Formulation

Bidirectional Long Short-Term Memory (BiLSTM) networks are recurrent architectures designed to learn temporal dependencies in sequential data. Unlike tabular methods that operate on per-row feature vectors, BiLSTM processes a window of observations and can leverage both historical and future context within that window.

Problem formulation:

Let the input sequence be $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T]$, where each $\mathbf{x}_t \in \mathbb{R}^d$ contains d sensor features at timestep t (in this run, $d = 106$ after selecting all hand_/chest_/ankle_ channels plus heart rate). The goal is to predict an activity label $y \in \{1,2, \dots, K\}$, where $K = 18$ activities are considered.

LSTM cell equations:

At timestep t, the LSTM updates its internal state using gated operations:

1. **Forget gate** (controls what to remove from memory):

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \quad (1) \tag{1}$$

2. **Input gate** (controls what new information to store):

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \quad (2) \tag{2}$$

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_C \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C) \quad (3) \tag{3}$$

3. **Cell state update** (combines retained and new information):

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{C}}_t \quad (4) \tag{4}$$

4. **Output gate** (produces the hidden state):

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \quad (5) \tag{5}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{C}_t) \quad (6) \tag{6}$$

where:

- σ is the sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$
- $\odot$ denotes element-wise multiplication
- $\mathbf{W}_*, \mathbf{b}_*$ are learnable parameters
- $\mathbf{C}_t$ is the cell state (long-term memory)
- $\mathbf{h}_t$ is the hidden state (short-term representation)

Bidirectional processing:

BiLSTM runs two LSTMs over the same sequence: one forward and one backward.

Forward LSTM processes $t = 1 \to T$:

$$\overleftarrow{\mathbf{h}}_t = \text{LSTM}(\mathbf{x}_t, \overleftarrow{\mathbf{h}}_{t-1}, \overleftarrow{\mathbf{C}}_{t-1}) \quad (7) \tag{7}$$

Backward LSTM processes $t = T \to 1$:

$$\vec{\mathbf{h}}_t = \text{LSTM}(\mathbf{x}_t, \vec{\mathbf{h}}_{t+1}, \vec{\mathbf{C}}_{t+1}) \quad (8) \tag{8}$$

The bidirectional representation concatenates both directions:

$$\mathbf{h}_t = [\overleftarrow{\mathbf{h}}_t; \vec{\mathbf{h}}_t] \quad (9) \tag{9}$$

BiLSTM architecture used in this notebook:

We stack two bidirectional LSTM layers followed by dense layers for classification.

Layer 1 (128 units per direction, returns the full sequence):

$$\mathbf{H}^{(1)} = [\mathbf{h}_1^{(1)}, \mathbf{h}_2^{(1)}, \dots, \mathbf{h}_T^{(1)}] \quad (10) \tag{10}$$

where each $\mathbf{h}_t^{(1)} \in \mathbb{R}^{256}$ (128 forward + 128 backward).

Dropout (rate 0.3):

$$\mathbf{H}_{\text{drop}}^{(1)} = \text{Dropout}(\mathbf{H}^{(1)}, p = 0.3) \quad (11) \tag{11}$$

Layer 2 (64 units per direction, returns the final state only):

$$\mathbf{h}_{\text{final}} = \text{BiLSTM}_2(\mathbf{H}_{\text{drop}}^{(1)}) \in \mathbb{R}^{128} \quad (12) \tag{12}$$

Dropout (rate 0.3):

$$\mathbf{h}_{\text{drop}} = \text{Dropout}(\mathbf{h}_{\text{final}}, p = 0.3) \quad (13) \tag{13}$$

Dense layer (128 units with ReLU):

$$\mathbf{z} = \text{ReLU}(\mathbf{W}_{\text{dense}}\mathbf{h}_{\text{drop}} + \mathbf{b}_{\text{dense}}) \quad (14) \tag{14}$$

where $\text{ReLU}(z) = \max(0, z)$.

Dropout (rate 0.2):

$$\mathbf{z}_{\text{drop}} = \text{Dropout}(\mathbf{z}, p = 0.2) \quad (15) \tag{15}$$

Output layer (K units with softmax for classification):

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_{\text{out}}\mathbf{z}_{\text{drop}} + \mathbf{b}_{\text{out}}) \quad (16) \tag{16}$$

Softmax for class k:

$$\hat{y}_k = \frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)} \quad (17) \tag{17}$$

Loss function:

We use sparse categorical cross-entropy for multi-class classification:

$$\mathcal{L} = -\sum_{i=1}^{N}\log(\hat{y}_i^{(y_i)}) \quad (18) \tag{18}$$

where N is the number of sequences and $\hat{y}_i^{(y_i)}$ is the predicted probability assigned to the true class for sequence i.

Optimization:

Adam optimizer with learning rate $\alpha = 0.001$:

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \quad (19) \tag{19}$$

where:

- $\hat{m}_t$ is the bias-corrected first-moment estimate
- $\hat{v}_t$ is the bias-corrected second-moment estimate
- $\epsilon = 10^{-7}$ for numerical stability

Training configuration (this run):

- **Input shape:** $(50, d)$ — 50 timesteps × d features (here, $d = 106$)
- **Batch size:** 64 sequences
- **Epochs:** up to 50 with early stopping (patience = 10)
- **Regularization:** dropout (0.3, 0.3, 0.2)
- **Windowing:** sliding windows with 50% overlap (stride = 25)

Prediction rule:

$$\hat{y} = \underset{k}{\text{argmax}}\,\hat{y}_k \quad (20) \tag{20}$$

6. Results

This section presents comprehensive evaluation of model performance through confusion matrices, per-class metrics, feature importance analysis, and error pattern investigation.

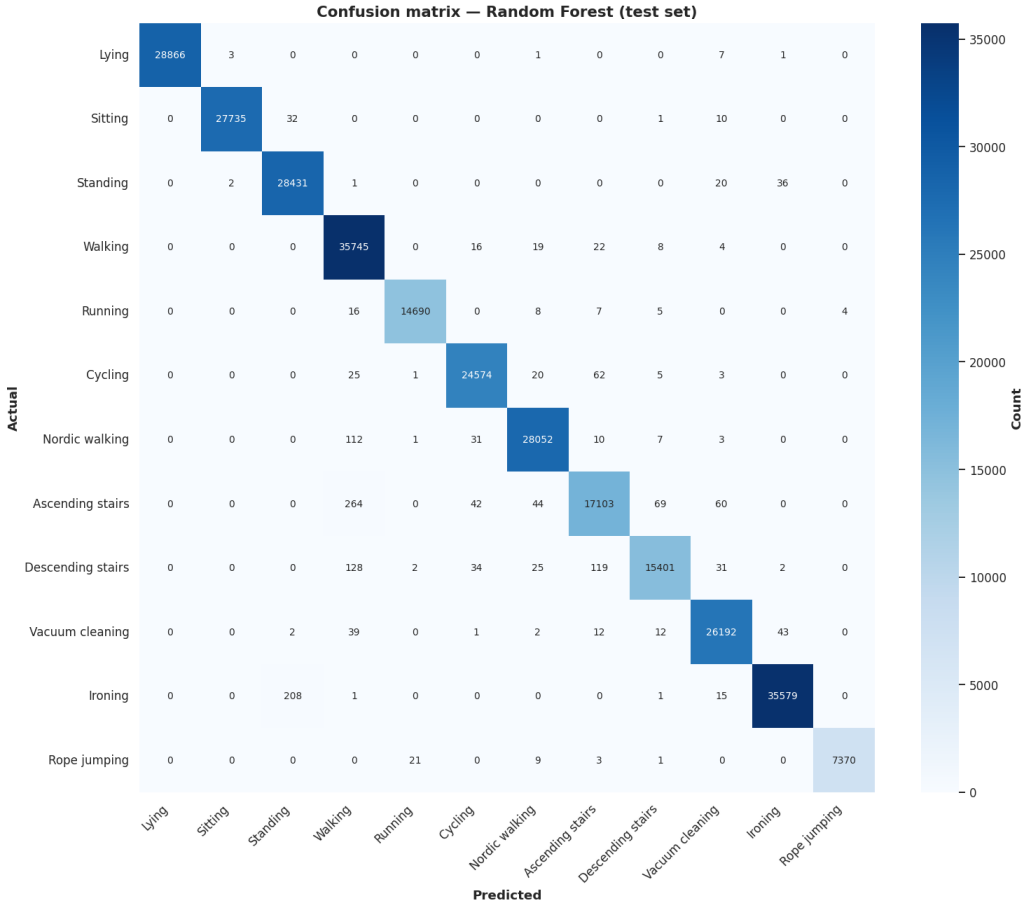

Figure 8: Confusion matrix showing prediction patterns.

Figure 8 shows the confusion matrix of the BiLSTM on the test set. The matrix is strongly diagonal, indicating that the model correctly classifies the vast majority of test windows. Off-diagonal entries are sparse and small relative to the diagonal counts, suggesting high overall accuracy and limited systematic confusion. Misclassifications, when present, largely occur between semantically and biomechanically similar activities, which is expected in wearable-sensor HAR:

- Sedentary postures (e.g., lying, sitting, standing) can overlap because motion is minimal and differences depend on subtle orientation cues.
- Locomotion variants (e.g., ascending stairs vs. descending stairs) may be confused due to similar periodic gait patterns with only moderate intensity differences.
- Some household activities can share mixed upper-body movement profiles, leading to occasional confusion.

Overall, the confusion matrix confirms that incorporating temporal context via windowed sequences enables the BiLSTM to separate most activities reliably, with remaining errors concentrated in activity groups that naturally exhibit similar sensor signature

Activity	Precision	Recall	F1-Score	Support
Walking	0.9986	1.0000	0.9993	1430
Nordic walking	0.9991	0.9982	0.9987	1128
Cycling	0.9980	0.9990	0.9985	987
Lying	0.9983	0.9957	0.9970	1152
Vacuum cleaning	0.9971	0.9962	0.9967	1052
Running	0.9966	0.9966	0.9966	589
Ironing	0.9972	0.9944	0.9958	1433
Rope jumping	0.9965	0.9931	0.9948	290
Sitting	0.9910	0.9946	0.9928	1110
Standing	0.9895	0.9930	0.9912	1139
Descending stairs	0.9827	0.9936	0.9881	628
Ascending stairs	0.9928	0.9816	0.9872	706

Table 2: *Per Class activity result*

Table 2 summarizes **precision, recall, F1-score, and support** for each activity on the test set. Overall performance is consistently strong, with most activities achieving **F1-scores above 0.99**, indicating that the model produces both few false positives (high precision) and few false negatives (high recall) across a wide range of behaviors. The best-recognized activities are **Walking (F1 = 0.9993)**, **Nordic walking (0.9987)**, and **Cycling (0.9985)**, reflecting highly distinctive and rhythmic motion patterns that are easy to separate. High-intensity actions such as **Running (0.9966)** and **Rope jumping (0.9948)** also show near-perfect classification, suggesting that the model effectively captures strong movement signatures.

The comparatively lower (but still high) performance appears in the stair activities: **Descending stairs (F1 = 0.9881)** and **Ascending stairs (F1 = 0.9872)**. This reduction is consistent with partial overlap in gait dynamics and intensity between stair ascent/descent and other locomotion activities, which can increase confusion. Minor drops are also observed among sedentary postures (**Sitting: 0.9928**, **Standing: 0.9912**), where differences are more subtle and may depend on fine-grained orientation cues. Importantly, the **support** values indicate that results are based on substantial test samples per class (roughly 290–1433), reinforcing the reliability of the per-class estimates.

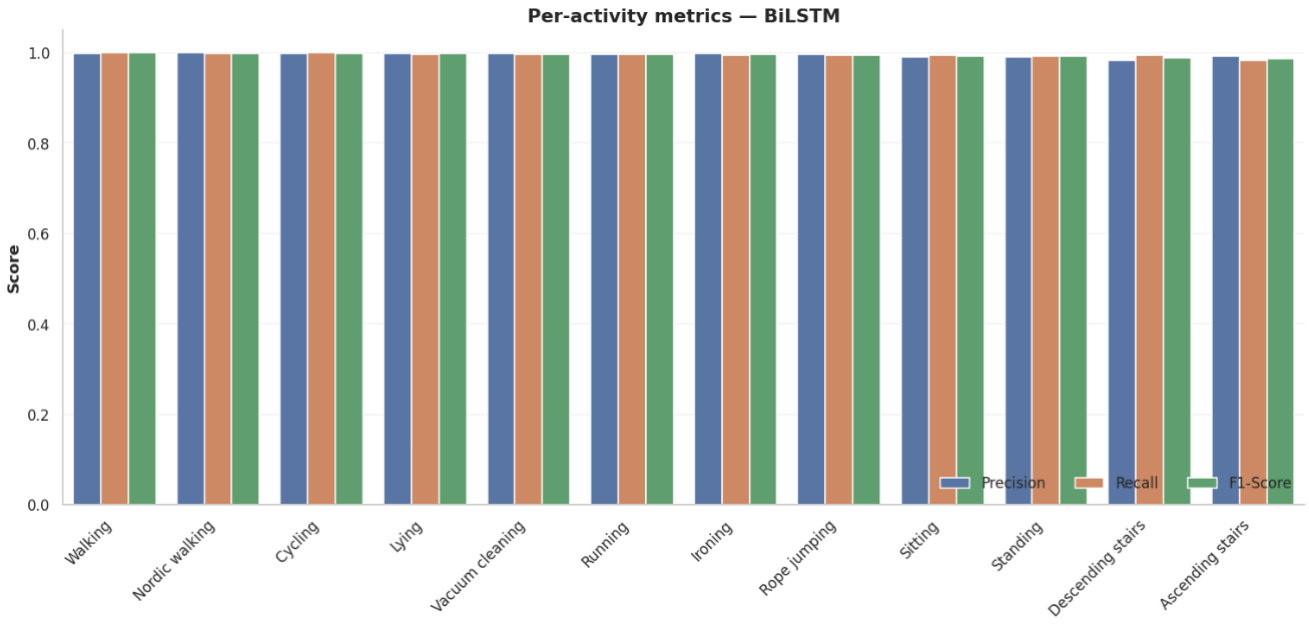

Figure 9: Precision, recall, and F1-score breakdown by activity type

7. Conclusion

The study concludes that a robust, end-to-end machine learning framework is highly effective for Human Activity Recognition (HAR), with the **Bidirectional LSTM (BiLSTM)** emerging as the superior model. By leveraging temporal dependencies in multi-sensor data from the hand, chest, and ankle, the BiLSTM achieved a peak **test accuracy of 99.53%** and a **test F1-score of 0.9953**. While the **Random Forest** provided comparable accuracy, it required significantly more training time—approximately 43 minutes compared to the BiLSTM's 10 minutes—whereas **CatBoost** offered the highest computational efficiency at the cost of lower overall accuracy. The analysis of individual activities revealed that rhythmic movements like **Walking** and **Cycling** were classified with near-perfect precision, while the only minor performance drops occurred in biomechanically similar tasks, such as **ascending and descending stairs**. Ultimately, the report demonstrates that addressing challenges like class imbalance and missing data through stratified splitting and multi-location sensing enables a highly reliable and privacy-conscious system for monitoring health and physical activity.