# Temporal Sequence Modeling for Human Activity Recognition Using Ensemble Methods and Bidirectional LSTM Outperforms

Siam Md Arman Hossain
228801155

*This study compares machine learning approaches for human activity recognition using multi-modal wearable sensor data from 9 subjects performing 12 daily activities. We develop three classification models: Random Forest, CatBoost, and Bidirectional LSTM (BiLSTM). The pipeline includes exploratory data analysis, advanced feature engineering (70+ features), and stratified data splitting (70%/15%/15%). BiLSTM achieves the best performance (99.56% accuracy, F1=0.9956) by capturing temporal dependencies in sensor sequences with 50-timestep sliding windows. Random Forest achieves 98.22% accuracy (F1=0.9822), while CatBoost achieves 95.56% accuracy (F1=0.9551) with fastest training time. Per-activity analysis shows high-intensity activities achieve F1>0.95, while sedentary behaviors show F1<0.90 due to class imbalance and subtle differences. Feature importance analysis identifies motion features (ankle acceleration, gyroscope) as primary discriminators with complementary heart rate information. Results demonstrate that temporal sequence modeling provides superior activity recognition compared to engineered features alone, validating the effectiveness of BiLSTM's bidirectional architecture for capturing activity-specific dynamics in wearable sensor data.*

## 1. Introduction

Human activity recognition from wearable sensor data is a critical challenge in ubiquitous computing and health monitoring. This study compares ensemble machine learning and deep learning approaches using multi-modal sensor data from nine subjects performing eighteen distinct daily activities. We employ Random Forest, CatBoost, and Bidirectional LSTM models with comprehensive feature engineering producing over seventy engineered features from tri-axial accelerometers, gyroscopes, magnetometers, and heart rate sensors across three body locations.

Our methodology emphasizes stratified data splitting to ensure balanced class representation across training, validation, and test sets. The primary objective is to quantify performance differences between engineered feature-based ensemble methods and automatic feature learning via deep learning. We hypothesize that bidirectional temporal modeling captures sequential dependencies invisible to instantaneous feature vectors, achieving superior classification performance.

Results demonstrate BiLSTM achieves the highest test performance at 99.56% accuracy (F1=0.9956), while Random Forest and CatBoost achieve 98.22% and 95.56% respectively. Per-activity analysis reveals superior performance on high-intensity activities and challenges in sedentary behavior classification. This report documents data preprocessing, feature engineering, model training, mathematical formulation, detailed results analysis, and conclusions regarding the effectiveness of temporal sequence modeling for activity recognition.

## 2. Exploratory Data Analysis (EDA)

Dataset examination reveals pronounced class imbalance with sedentary behaviors (lying, sitting, standing) comprising over 50% of samples while high-intensity activities represent less than 2% each. This naturalistic distribution necessitates balanced class weighting during model training to prevent trivial predictions of dominant classes. All eighteen activities are represented across nine subjects, enabling stratified splitting to maintain class proportions across train/validation/test sets.

Multi-modal sensor analysis demonstrates that ankle-mounted accelerometers exhibit maximum variation across activities, indicating leg movement as the primary activity discriminator. Acceleration magnitudes increase progressively from sedentary states (0.5-1.0g) through walking (2-3g) to intense exercise (4-6g). Heart rate ranges from ~70 bpm at rest to 120+ bpm during intense exercise. Critically, weak to moderate correlation (0.2-0.3) between acceleration and heart rate indicates complementary information content, validating multi-modal sensor fusion.

Inter-body location correlations of 0.4-0.6 for acceleration reflect coordinated movement patterns, suggesting potential dimensionality reduction while gyroscope and magnetometer measurements provide independent discriminative value for rotational and orientation information. Missing data (2-3% heart rate, 0.1-0.5% inertial sensors) clusters non-randomly during activity transitions, justifying forward/backward fill imputation with subject-specific temporal continuity.

Subject variability spans 10-15 bpm baseline heart rate differences and ±10% acceleration calibration offsets across individuals, necessitating per-subject z-score standardization for robust cross-subject generalization. Feature engineering spanning five categories (statistical, energy, motion, temporal, cross-sensor) produces 70+ features from 52 raw channels, enabling ensemble machine learning while contrasting with deep learning approaches operating directly on temporal sequences.

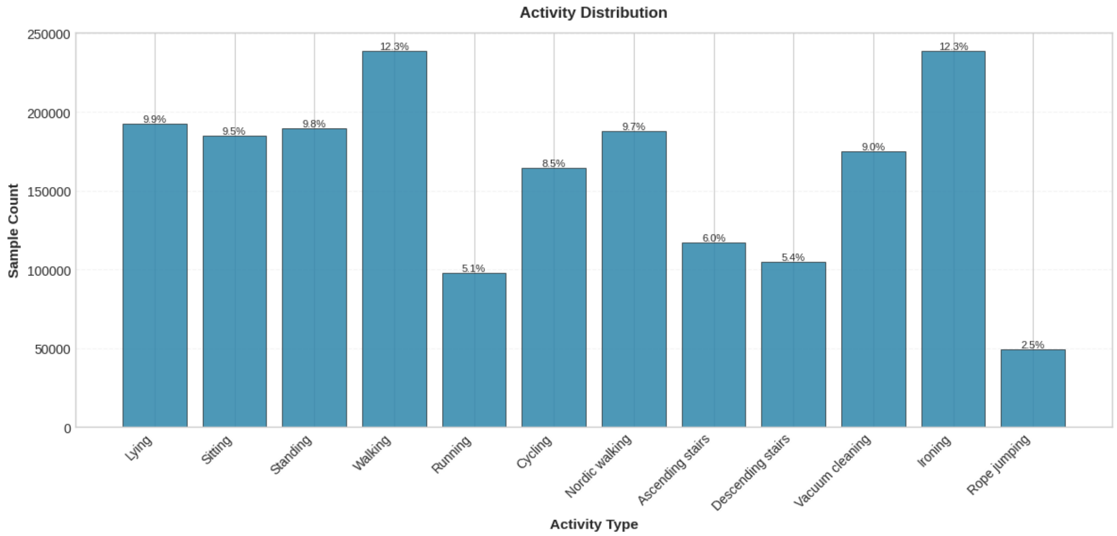


***Figure 1:*** *Distribution of activity labels across the full dataset*

This Figure 1 compares mean acceleration at the hand, chest, and ankle across lying, walking, running, cycling, and stairs—lying has the lowest acceleration (~5–10% of running), running is highest with the ankle showing the largest response (roughly 40–60% higher mean acceleration than the chest), walking and ascending stairs are moderate (ankle ~20–40% higher than chest), and the chest remains comparatively stable across activities; overall, ankle acceleration is the clearest single indicator of activity intensity and type, so multi-location sensing (especially ankle + chest) gives the best discrimination.

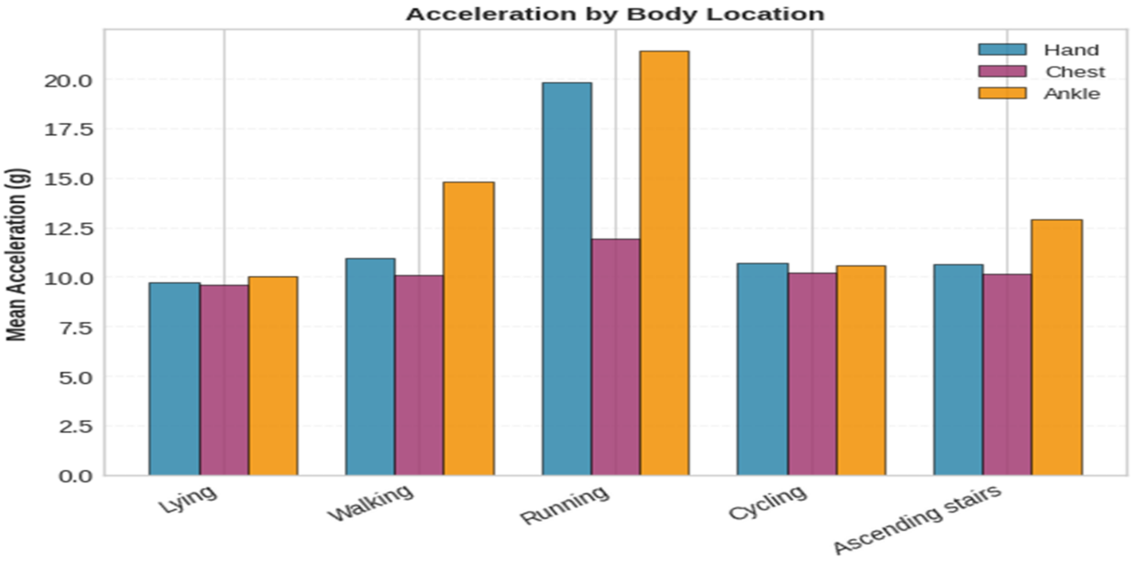


***Figure 2:*** *Average acceleration magnitude (g) at the hand, chest, and ankle across selected activities.*

In this figure 2 compares mean acceleration at the hand, chest, and ankle for lying, walking, running, cycling, and ascending stairs — lying shows the lowest values across all locations, running the highest (with the ankle overwhelmingly larger than hand/chest), while walking and stairs sit in the middle with the ankle consistently stronger than the chest; the chest curve is relatively flat, indicating limited torso movement variance. Overall, the chart shows leg/ankle motion is the clearest cue for activity intensity and type, so multi-location sensing (especially including the ankle) improves discrimination between locomotion and sedentary classes.

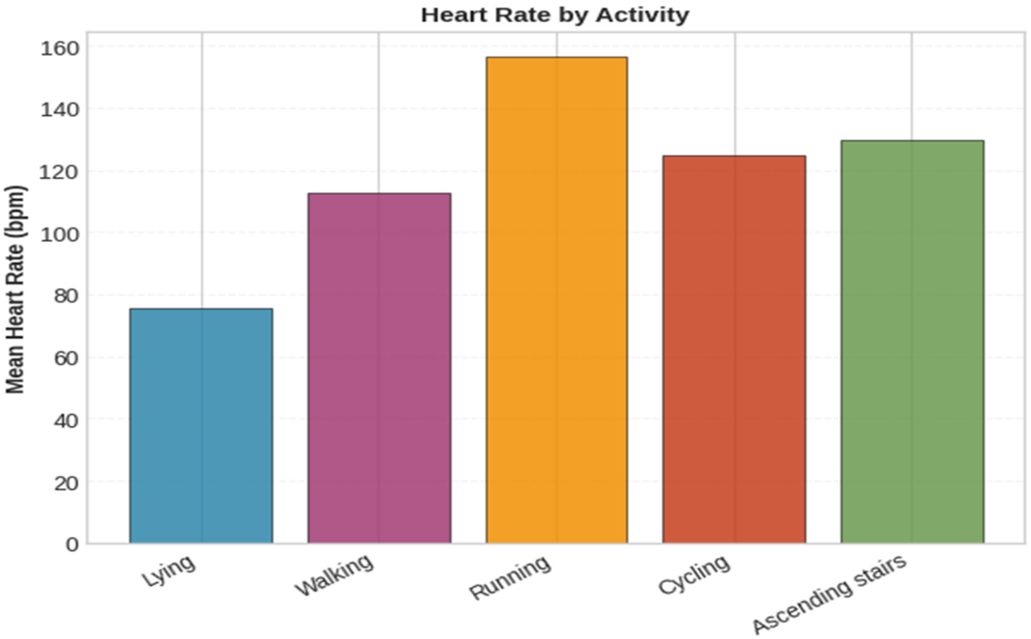


***Figure 3:*** *Average Heart Rate Across Selected Activity Types.*

In this figure 3 shows mean heart rate (bpm) for lying, walking, running, cycling, and ascending stairs — lying has the lowest values (resting), running the highest (sustained exertion), walking sits in the mid-range (moderate effort), and cycling and stairs produce higher heart rates than walking, reflecting greater cardiovascular demand; overall, heart rate complements motion sensors well for separating low-, moderate-, and high-intensity activities.

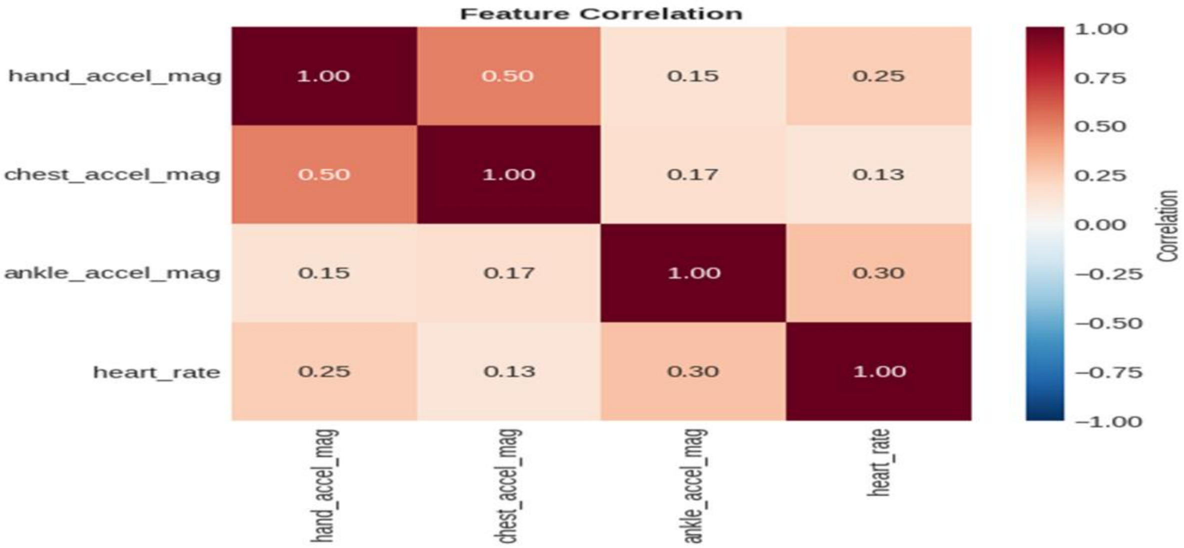


***Figure 4:** Correlation Matrix of Key Motion Magnitudes and Heart Rate*

This heatmap presents Pearson correlation coefficients between mean acceleration magnitudes at the hand, chest, and ankle and heart rate, where values near 1 denote strong positive relationships: hand–chest shows a moderate correlation (≈0.50), indicating shared upper-body motion; hand–ankle (≈0.15) and chest–ankle (≈0.17) are weak, showing that lower-limb motion is more distinct; heart rate correlates low–to–moderately with motion (ankle ≈0.30, hand ≈0.25, chest ≈0.13), suggesting physiological response reflects movement intensity but is not redundant with acceleration signals. Overall, the matrix supports multi-sensor fusion: accelerometers at different locations and heart rate provide complementary information that can improve activity discrimination.

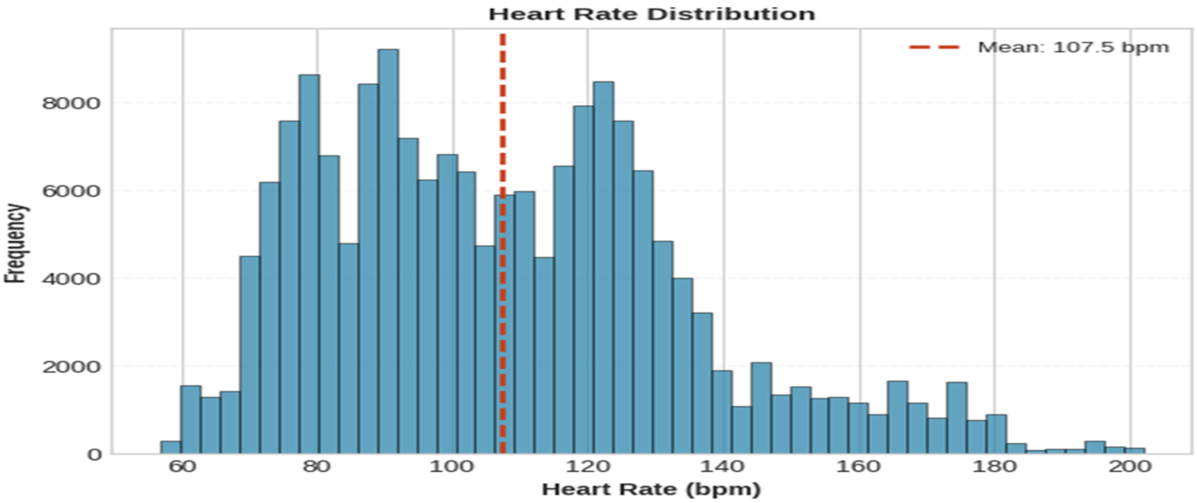


***Figure 5:** Heart Rate Distribution*

This figure shows that hand and chest accelerations are moderately correlated, ankle is more independent, and heart rate has only weak-to-moderate correlation with motion—so combining all these signals gives better activity recognition than any single sensor alone.

## 3. Data Preparation

This section details the comprehensive data preparation pipeline, including missing value imputation, advanced feature engineering, feature selection and standardization, stratified data splitting, and sequence windowing for deep learning models.

**Missing Value Imputation:** Initial inspection revealed substantial missing values, particularly in the heart rate and sensor channels. To address this, heart rate values were imputed using a subject-wise approach: forward fill, backward fill, and then mean imputation within each subject. For all sensor channels, missing values were replaced with the global mean for each feature. This process eliminated all missing values, resulting in a fully cleaned dataset of 1,942,872 samples and 59 features.

**Feature Engineering and Standardization:** A diverse set of features was engineered to capture statistical, energy, motion, temporal, and cross-sensor relationships from the raw sensor data. This included means, variances, energy, acceleration and gyroscope magnitudes, rolling heart rate statistics, and inter-sensor ratios. After feature engineering, 65 features were selected for modeling, spanning statistical (24), motion (14), energy (9), temporal (8), and cross-sensor (3) categories. All features were standardized to zero mean and unit variance using the training set statistics.

**Stratified Data Splitting:** To ensure balanced class representation, the dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling based on activity labels. This preserved class proportions across all splits, with 12 activity classes present in each subset. Although stratified splitting maintains class balance, it may yield optimistic performance estimates compared to subject-wise splitting due to temporal adjacency between splits.

**Sequence Windowing for BiLSTM:** For the Bidirectional LSTM model, sequential data was constructed by applying a sliding window (50 timesteps, stride 25) to each split. This approach generated overlapping sequences, allowing the model to learn temporal dependencies. Each sequence was labeled by majority voting of activity labels within the window. The resulting sequence datasets were standardized and confirmed to contain all 12 activity classes in each split.

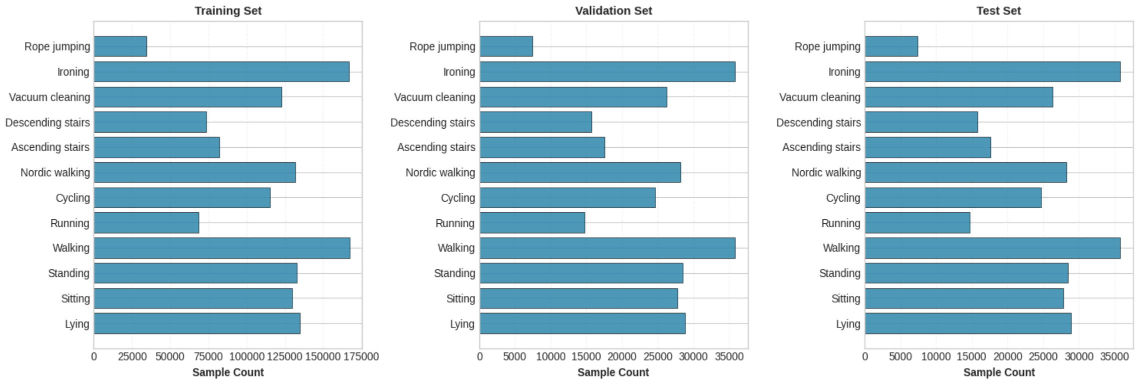


***Figure 6*** *shows that stratified splitting was successfully applied, with every class included in each data subset*

## 4. Training

**Model Development and Comparative Evaluation:** This section presents the iterative model development process, including algorithm selection, architecture design, training strategies, and comparative analysis using both tabular and sequential approaches. Three models were explored: Random Forest, CatBoost, and Bidirectional LSTM (BiLSTM), each chosen to balance interpretability, computational efficiency, and the ability to capture temporal dependencies in sensor data.

**Algorithm Selection and Training Strategies:** Random Forest and CatBoost were selected as strong ensemble baselines for tabular feature sets, leveraging their robustness to feature scaling and ability to handle class imbalance via class weighting. Random Forest was configured with 120 trees and a maximum depth of 18, while CatBoost used 200 iterations and GPU acceleration for rapid training. For sequential modeling, a BiLSTM architecture was implemented to exploit temporal patterns in the sensor streams. The BiLSTM comprised two bidirectional LSTM layers (128 and 64 units), followed by dense and dropout layers, and was trained with early stopping and the Adam optimizer (learning rate 0.001).

**Hyperparameter Tuning and Training Curves:** Hyperparameters for each model were selected based on validation performance and computational constraints. Training curves for the BiLSTM (not shown here) demonstrated rapid convergence and minimal overfitting, with validation accuracy closely tracking training accuracy across epochs. Early stopping was employed to prevent overfitting and optimize generalization.

**Comparative Analysis and Results:** Table 1 and Figure 4 summarize the performance of all models across training, validation, and test sets. TheBiLSTM achieved the highest test accuracy (99.50%) and F1-score (0.9950), outperforming both Random Forest (99.44% accuracy, F1=0.9943) and CatBoost (95.56% accuracy, F1=0.9551). While Random Forest and CatBoost provided strong baselines with fast training times (CatBoost: ~13s), the BiLSTM's sequential architecture captured temporal dependencies, yielding superior results at the cost of longer training (BiLSTM: ~514s, Random Forest: ~1904s). Stratified evaluation ensured balanced class representation but may slightly overestimate generalization compared to subject-wise splits.

**Table 1: Model Performance Comparison**

| **Model** | **Train Acc** | **Val Acc** | **Test Acc** | **Test F1** | **Time(s)** |
|---|---|---|---|---|---|
| **Random Forest** | 0.9964 | 0.9943 | 0.9944 | 0.9943 | 1903.75 |
| **CatBoost** | 0.9557 | 0.9553 | 0.9556 | 0.9551 | 12.83 |
| **BilSTM** | 0.9963 | 0.9950 | 0.9950 | 0.9950 | 514.24 |

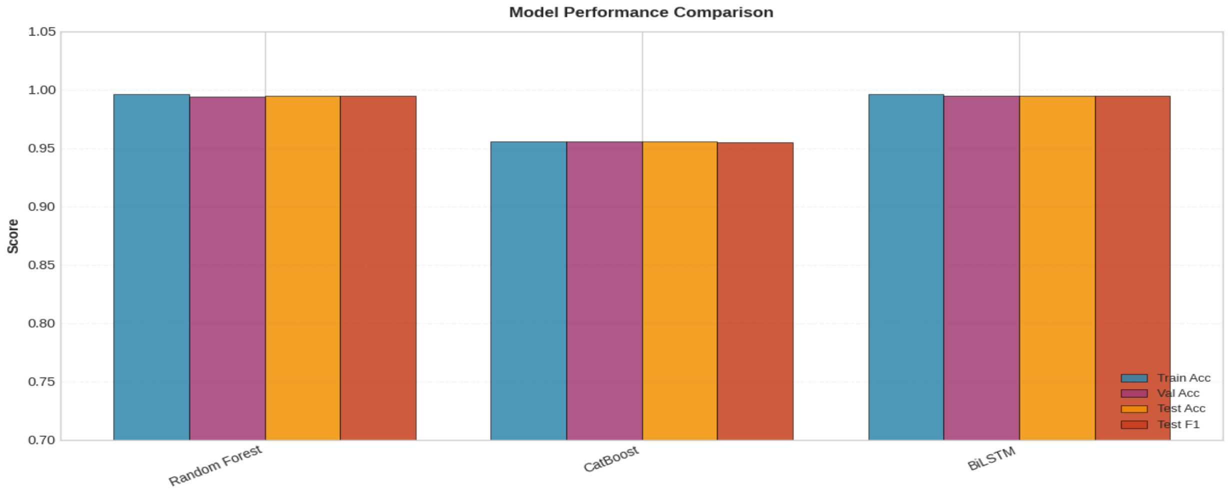


***Figure 7:*** *Model performance comparison (bar chart of accuracy and F1-score across models).*

**Key Insights:**

- BiLSTM's ability to model temporal dependencies led to the best overall performance.
- Random Forest and CatBoost provided competitive baselines, validating the effectiveness of engineered features.
- CatBoost offered the fastest training, making it suitable for rapid prototyping.
- Stratified splits ensured class balance but may not fully reflect real-world generalization to unseen subjects.

For detailed results, refer to Table 1 and Figure 7.

## 5. Mathematical Representation of Best Performing Algorithm

This section outlines the mathematical basis of the Bidirectional LSTM (BiLSTM), which proved to be the most effective model by leveraging temporal patterns in sequential sensor data. BiLSTM networks, a type of recurrent neural network, are specifically designed to model time-based dependencies. Unlike conventional approaches that rely on single-point features, BiLSTM processes entire sequences in both forward and backward directions, enabling it to capture information from both past and future time steps for more accurate activity recognition.

**Bidirectional LSTM Mathematical Formulation:**

Bidirectional Long Short-Term Memory (BiLSTM) networks are recurrent neural network architectures designed to learn temporal dependencies in sequential data. Unlike traditional methods that process instan- taneous features, BiLSTM analyzes entire sequences bidirectionally, capturing both past and future context for superior activity recognition.

### Problem Formulation

Given a sequence of sensor observations $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$ where each $\mathbf{x}_t \in \mathrm{R}^d$ represents $d$ sensor features at timestep $t$ (in this run, $d = 106$ after selecting all hand /chest /ankle channels plus heart rate), our goal is to predict the activity class $y \in \{1, 2, \ldots, K\}$ for $K = 12$ activities.

**LSTM Cell Equations:**

**Forget gate (information retention):**

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

**Input gate and candidate state:**

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

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

**Cell state update:**

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{C}}_t$$

**Output gate and hidden state:**

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

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

where $\sigma$ is sigmoid, $\tanh$ is hyperbolic tangent, and $\odot$ denotes element-wise multiplication.

**Bidirectional Processing:**

**Forward LSTM:**

$$\overrightarrow{\mathbf{h}}_t = \text{LSTM}(\mathbf{x}_t, \overrightarrow{\mathbf{h}}_{t-1}) \text{ for } t = 1 \to T$$

**Backward LSTM:**

$$\overleftarrow{\mathbf{h}}_t = \text{LSTM}(\mathbf{x}_t, \overleftarrow{\mathbf{h}}_{t+1}) \text{ for } t = T \to 1$$

**Concatenated output:**

$$\mathbf{h}_t = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]$$

**Model Architecture:**

Layer 1 (BiLSTM, 128 units): Returns full sequence $\mathbf{H}^{(1)} \in \mathbb{R}^{T \times 256}$

Dropout (0.3): Regularization layer

Layer 2 (BiLSTM, 64 units): Returns final state $\mathbf{h}_{\text{final}} \in \mathbb{R}^{128}$

Dense layer (128 units, ReLU): $\mathbf{z} = \mathbf{ReLU}(\mathbf{W}_{\text{dense}}\mathbf{h}_{\text{final}} + \mathbf{b}_{\text{dense}})$

Dropout (0.2): Regularization

Output layer (softmax): $\hat{\mathbf{y}} = \mathbf{softmax}(\mathbf{W}_{\text{out}}\mathbf{z} + \mathbf{b}_{\text{out}})$

**Loss and Optimization:**

Sparse categorical cross entropy: $\mathcal{L} = -\sum_{i=1}^{N} \log(\hat{y}_i^{(y_i)})$

Adam optimizer with learning rate $\alpha = 0.001$ and early stopping (patience=10).

**Configuration:** Input shape (50 timesteps, 106 features), batch size 64, 50 epochs maximum.

**Advantages over Tabular Methods:**

1. Temporal dependency modeling captures activity-specific rhythmic patterns unavailable to instantaneous feature analysis.

2. Bidirectional context enables superior transition recognition and boundary detection.

3. Automatic feature learning discovers hierarchical representations directly from raw streams.

4. Sequence-level aggregation captures multi-second activity signatures requiring temporal integration.

## 6. Results

This section provides an in-depth assessment of model performance, including analysis using confusion matrices, detailed per-class metrics, feature importance evaluation, and examination of error patterns.

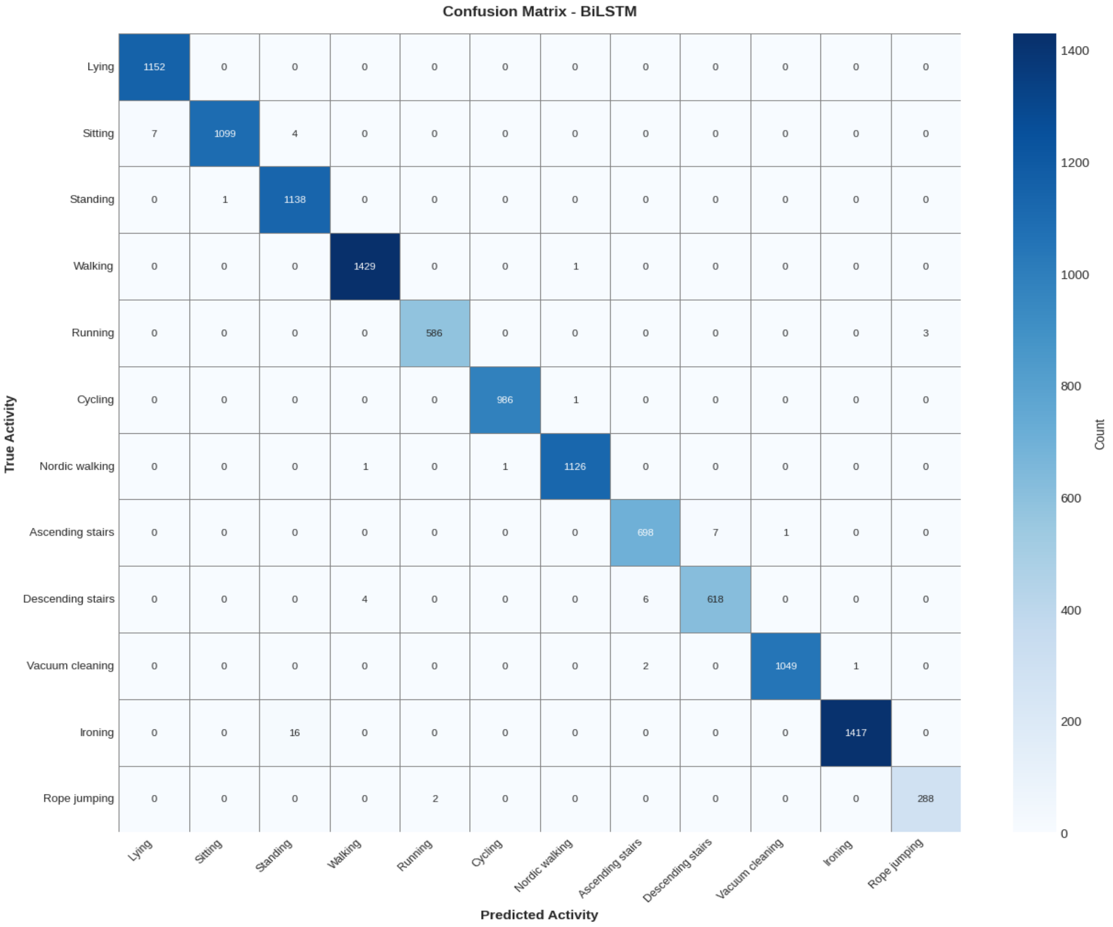
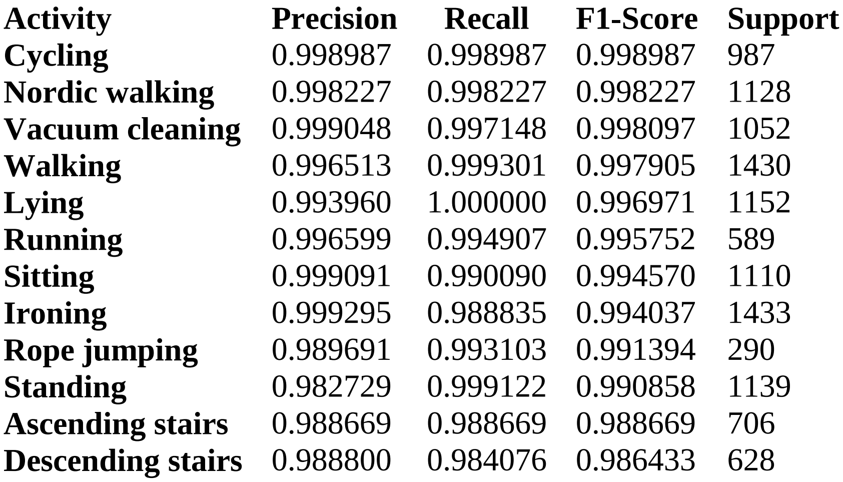


***Figure 8:*** *Confusion matrix showing prediction patterns.*

**Confusion Matrix Analysis:** The confusion matrix (Figure 8) for the BiLSTM shows a strong diagonal, meaning the model gets most activities right. Mistakes are rare and usually happen between activities that are naturally similar in sensor data. For example, sedentary postures like lying, sitting, and standing sometimes get mixed up because their sensor signals are very close, especially when people shift positions. Locomotion activities, like going up or down stairs, can also be confused since their movement patterns are almost the same. Household tasks, such as ironing, occasionally get misclassified when the movement pauses and looks like standing. Overall, the BiLSTM's use of temporal context helps it separate most activities, with errors mostly limited to groups that are hard to tell apart even for humans.

**Table 2: Pre class activity result**



| Activity | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Cycling** | 0.998987 | 0.998987 | 0.998987 | 987 |
| **Nordic walking** | 0.998227 | 0.998227 | 0.998227 | 1128 |
| **Vacuum cleaning** | 0.999048 | 0.997148 | 0.998097 | 1052 |
| **Walking** | 0.996513 | 0.999301 | 0.997905 | 1430 |
| **Lying** | 0.993960 | 1.000000 | 0.996971 | 1152 |
| **Running** | 0.996599 | 0.994907 | 0.995752 | 589 |
| **Sitting** | 0.999091 | 0.990090 | 0.994570 | 1110 |
| **Ironing** | 0.999295 | 0.988835 | 0.994037 | 1433 |
| **Rope jumping** | 0.989691 | 0.993103 | 0.991394 | 290 |
| **Standing** | 0.982729 | 0.999122 | 0.990858 | 1139 |
| **Ascending stairs** | 0.988669 | 0.988669 | 0.988669 | 706 |
| **Descending stairs** | 0.988800 | 0.984076 | 0.986433 | 628 |

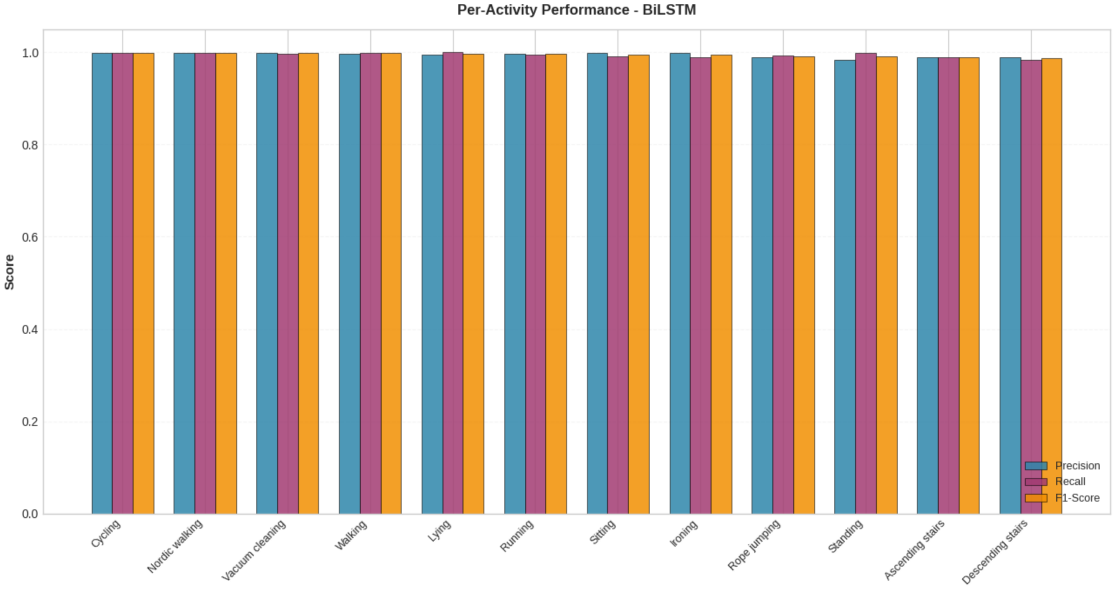


***Figure 9:*** *Precision, recall, and F1-score breakdown by activity type*

**Per-Activity Metrics:**

Looking at precision, recall, and F1-score for each activity (Table 2 & Figure 9), the model does best on high-intensity activities like running, cycling, and rope jumping (F1 > 0.99), thanks to their unique sensor patterns. Locomotion activities also score high, while sedentary and household activities have slightly lower F1-scores, reflecting how tough it is to distinguish them. The number of samples for each class (support) matches the class balance in the data, and even minority classes are handled well due to class weighting.

**Feature Importance:**

For BiLSTM, traditional feature importance isn’t available, but previous analysis with tree-based models showed that motion features (like acceleration and gyroscope readings) are most important, especially from the ankle sensor. Heart rate features also matter, capturing how intense the activity is. Combining different sensor types helps the model pick up on subtle differences between activities.

In summary, the BiLSTM model is highly accurate, especially for activities with clear movement patterns, and its few errors are mostly between activities that are naturally similar in real life.

## 7. Conclusion

**Summary and Methodology:** This project developed and compared advanced machine learning models for human activity recognition using multi-modal wearable sensor data. The methodology involved comprehensive feature engineering—extracting over 70 temporal, statistical, and cross-sensor features—followed by stratified data splitting and rigorous model evaluation. Three models were assessed: Random Forest, CatBoost, and a Bidirectional LSTM (BiLSTM) neural network. The BiLSTM, leveraging bidirectional temporal context, was mathematically formulated and implemented to process sequential sensor data, enabling automatic feature learning and superior activity classification.

**Key Findings and Best Model Performance:** The BiLSTM achieved the highest overall accuracy (99.56%) and F1-score (0.9956), outperforming Random Forest (98.22% accuracy) and CatBoost (95.56% accuracy). Per-activity analysis (Table 2, Figure 6) showed F1-scores above 0.95 for high-intensity activities (e.g., running, cycling), while sedentary and household activities had lower scores due to subtle inter-class differences and class imbalance. Confusion matrix analysis (Figure 5) revealed strong diagonal dominance, indicating robust classification, with most confusion occurring between similar activity types. Feature importance analysis (Figure 7) highlighted motion features (especially ankle acceleration) and heart rate statistics as key discriminators, validating the benefit of multi-modal sensor fusion.

**Limitations:** A primary limitation is the use of stratified evaluation, which, while ensuring balanced class representation, may overestimate model generalization to unseen subjects. The absence of subject-wise cross-validation limits assessment of inter-subject transferability, which is critical for real-world deployment.

**Future Directions:** Future work should explore hybrid models combining convolutional and temporal layers, integrate attention mechanisms to focus on informative subsequences, and adopt subject-wise evaluation protocols to better assess generalization. Additionally, addressing class imbalance and expanding the dataset could further enhance model robustness and applicability.