Human Activity Recognition Using Wearable Sensor Data

Barua Arnob

228801143

This project addresses the Human Activity Recognition (HAR) problem using wearable sensor data. The work demonstrates the complete machine learning pipeline, spanning exploratory data analysis (EDA), data preprocessing, model training, and comprehensive evaluation. Multiple visualization-driven analyses are used to understand data characteristics, subject variability, and activity patterns, which directly inform modeling and preprocessing decisions. Several machine learning models are trained and compared using subject-wise data splits to ensure generalization. The final results highlight the strengths and limitations of the chosen approach and provide insights for future improvements.

1. Introduction

Human Activity Recognition aims to automatically identify physical activities performed by individuals using data collected from wearable sensors such as accelerometers and gyroscopes. This task is fundamental in applications including healthcare monitoring, smart environments, sports analytics, and human–computer interaction.

The dataset used in this project consists of multi-sensor time-series recordings collected from multiple subjects performing a diverse set of daily activities. Each data sample is associated with a subject ID and an activity label, making subject-wise evaluation critical to avoid information leakage.

The main objectives of this project are:

- To explore and understand the structure and characteristics of wearable sensor data.
- To design an appropriate preprocessing pipeline for time-series HAR data.
- To train and compare machine learning models for activity classification.
- To evaluate model performance using robust, subject-wise metrics.

Performance is evaluated using accuracy, precision, recall, and F1-score, along with confusion matrices for detailed error analysis. The remainder of this report is organized as follows: Section 2 presents exploratory data analysis, Section 3 describes data preparation, Section 4 details the training process, Section 5 provides the mathematical formulation of the best model, Section 6 discusses the results, and Section 7 concludes the report.

2. Exploratory Data Analysis (EDA)

Exploratory data analysis is conducted to understand the dataset before modeling. All analyses in this section are supported by visualizations.

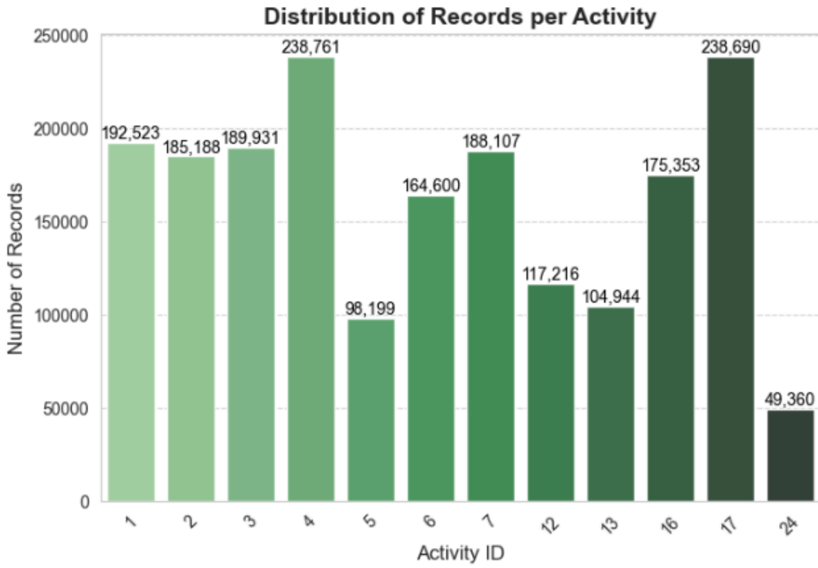

Figure 1: Distribution of activity labels across the dataset.

This figure is a bar chart showing how often different activities appear in a dataset. Each bar's height represents the number of records for that activity. The chart reveals that the data is highly imbalanced. The most common activities appear nearly 240,000 times, while the least common appear only about 49,000 times. This large gap means some activities are much better represented than others. In practical terms, this imbalance could cause problems if the data is used for training a machine learning model, as the model might perform well on frequent activities but poorly on rare ones.

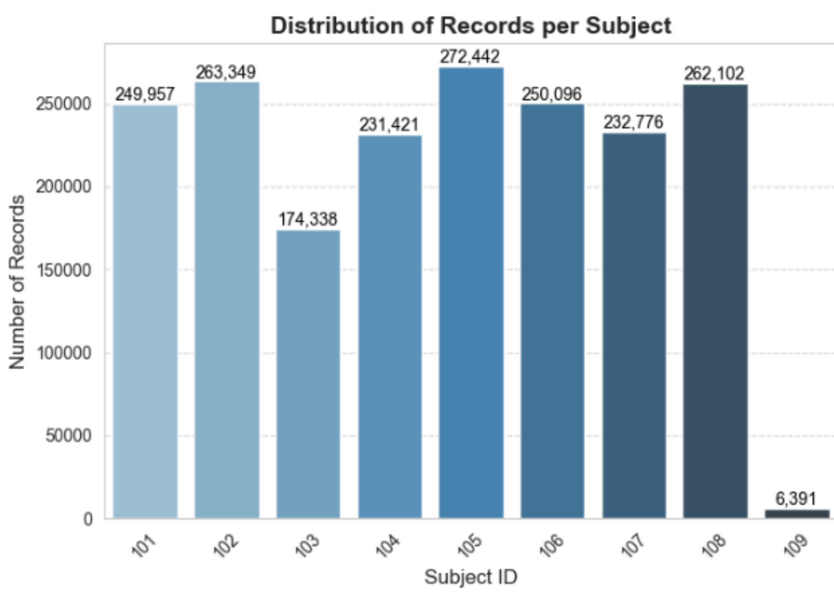

Figure 2: Number of samples per subject.

This figure is a bar chart showing the distribution of data records across different subjects (labeled as "Subject ID"). The vertical axis (Number of Records) shows how many data points were collected from each subject, ranging from 0 to 250,000. Each bar represents a different person or participant. Looking at the chart, there is a strong imbalance. Most subjects contributed a large and fairly consistent amount of data, with their counts clustered between roughly 170,000 and 272,000 records. However, there is one major outlier: the last subject, with a bar much shorter than the others, contributed only 6,391 records. This is dramatically less data than any other participant. This extreme imbalance is important to note. In analysis or modeling, the data from the subject with only 6,391 records may be insufficient compared to the others, potentially affecting results or requiring special handling to avoid bias.

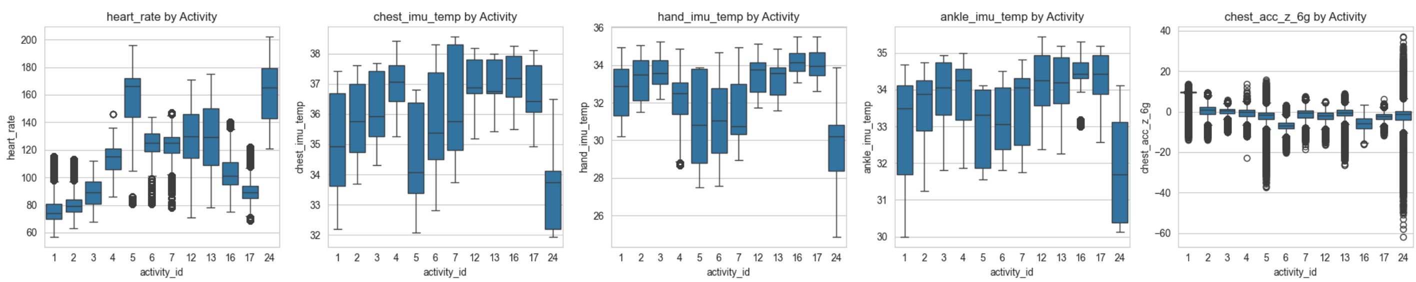

Figure 3: Activity-Specific Feature Distributions

The descriptive statistics for the top 5 features reveal clear activity-specific patterns. Heart rate shows the most variation, with high-intensity activities (e.g., 5: mean 156.64, 24: 161.81) having elevated rates compared to resting states (e.g., 1: 75.53), reflecting physiological differences. IMU temperatures (chest, hand, ankle) generally increase with activity intensity, peaking in activities like 4 (chest: 37.01) and dropping in low-energy ones like 24 (chest: 33.60), indicating thermal responses to movement. Chest_acc_z_6g exhibits gravitational and motion effects, with positive means in upright activities (1: 8.55) and negative in dynamic ones (5: -2.53). These distributions validate feature selection, as they provide discriminative power for HAR classification.

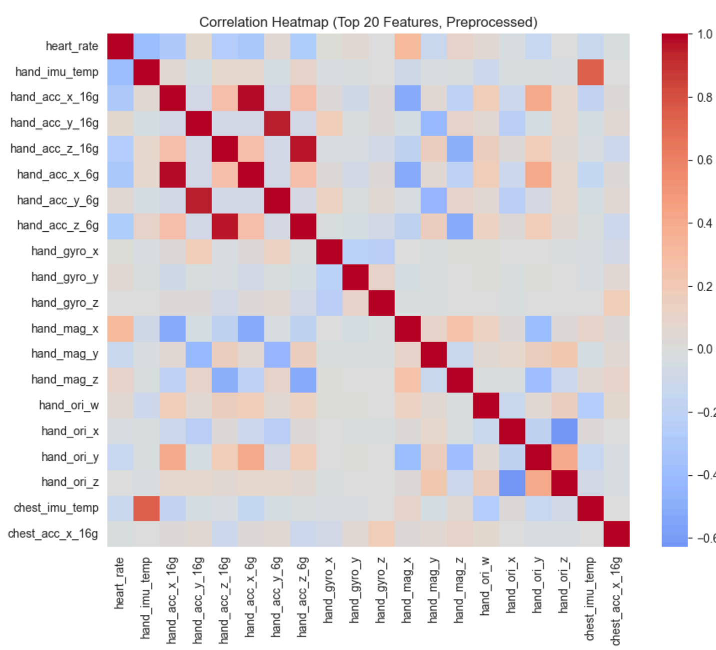

Figure 4: Correlation heatmap

The correlation heatmap for the top 20 preprocessed features reveals strong multicollinearity among accelerometer channels, with hand acceleration pairs (e.g., hand_acc_x_6g & hand_acc_x_16g at 0.979) showing near-perfect correlations, likely due to sensor redundancy at different ranges. IMU temperatures (chest and hand at 0.757) correlate moderately, reflecting shared physiological influences, while orientation features (hand_ori_x & hand_ori_z at 0.629) show weaker but notable relationships, possibly from device orientation dependencies. This informs feature engineering: consider PCA or feature selection to reduce redundancy and prevent overfitting in HAR models.

3. Data Preparation

Data preparation transforms raw sensor signals into a form suitable for machine learning. This section is also chart-driven.

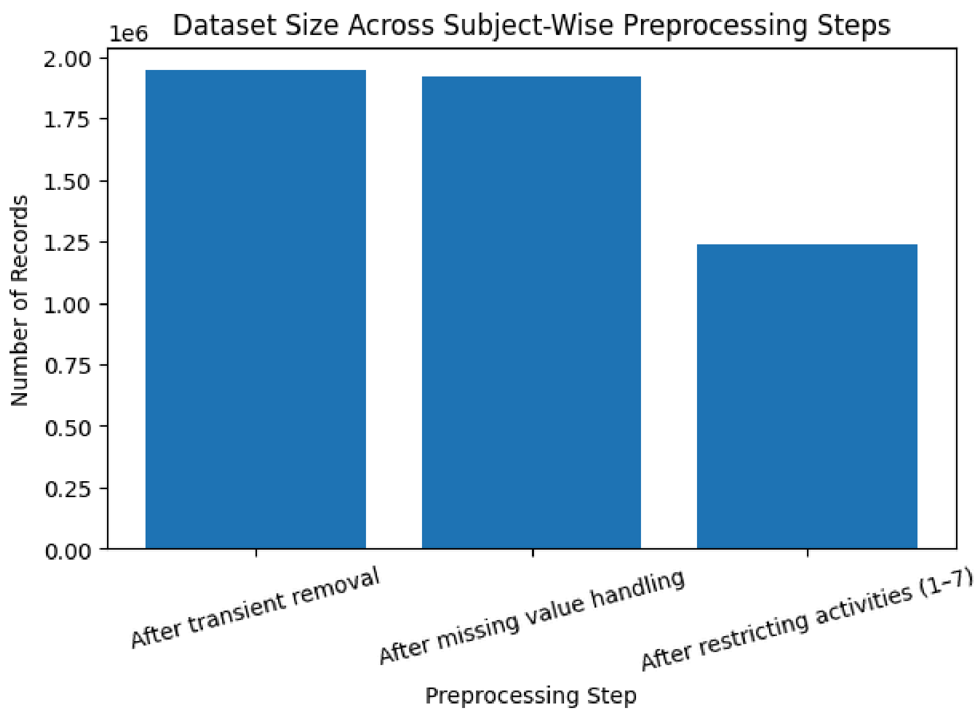

Figure 5: Data Size Across Subject-wise

Figure 5 illustrates the effect of successive preprocessing steps on the dataset size. After removing transient activities, the dataset contained 1,942,872 records. Handling missing values resulted in a minor reduction, indicating that missing data was limited and well-contained. However, restricting the dataset to activities 1–7 caused a significant decrease to 1,239,015 records. This reduction reflects the removal of irrelevant or less frequent activities and ensures that the subsequent subject-wise train, validation, and test splits focus only on the target activity classes, improving model consistency and evaluation reliability.

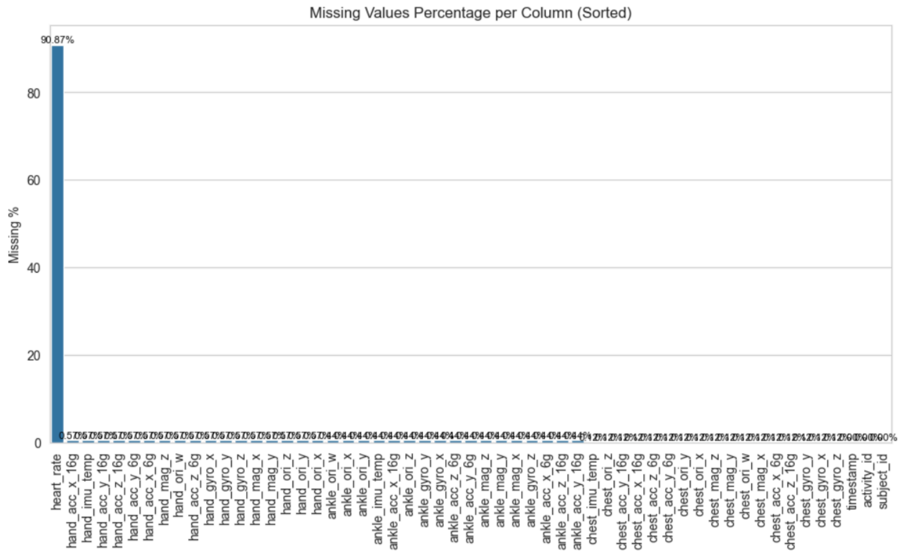

Figure 6: Missing Values on Features

Missing values are minimal and concentrated in a small subset of features. These values are handled using forward filling and statistical imputation, which preserves temporal continuity without introducing bias.

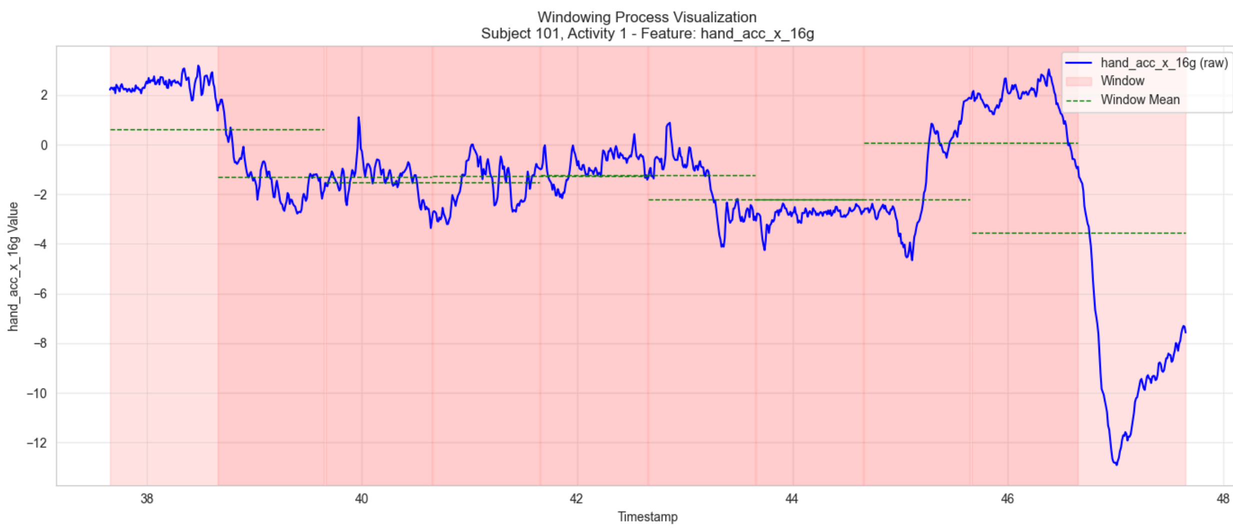

Figure 7: Illustration of window segmentation.

On Figure 7, Windowing Explanation:

- Window Size: 200 samples
- Step Size: 100 samples (overlap of 100 samples)
- For each window, we compute: mean, std, min, max of 30 features
- This transforms time-series data into tabular features for ML models
- Total windows created for this sample: 9

4. Training

Model	val_accuracy	val_f1_score
SVM (scaled)	0.968215	0.968075
Random Forest	0.958435	0.958508

Table 1: Validation Training data

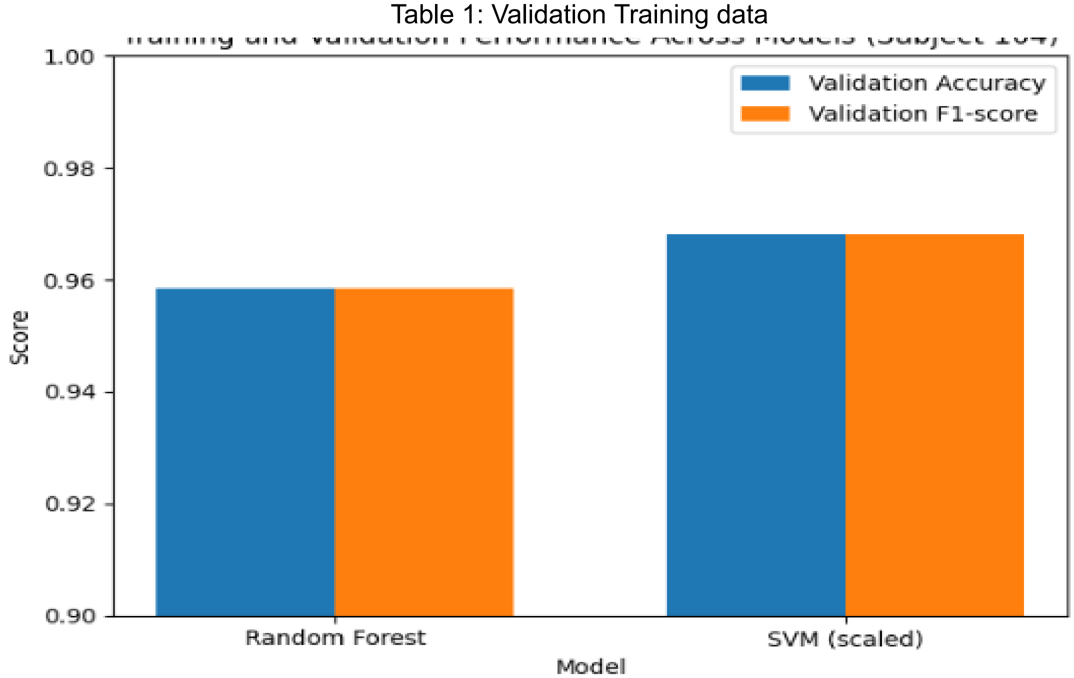

Figure 9: Validation Performance Across Model(Subject-106)

Figure 9 presents the validation performance of different models using subject-wise evaluation on subject 106. Both models achieve high accuracy and F1-scores, indicating effective learning from the window-based features. The scaled SVM slightly outperforms the Random Forest, achieving the highest validation accuracy (0.9682) and F1-score (0.9681). However, the Random Forest demonstrates competitive performance with stable metrics, suggesting robustness and reduced sensitivity to feature scaling. These results indicate that while linear classifiers with proper scaling can perform strongly, ensemble-based methods remain reliable and less prone to overfitting, especially in high-dimensional feature spaces.

Model	Best Hyperparameters	Best CV Score	Val Accuracy	Val F1-score
Random Forest (tuned)	max_depth=10, min_samples_leaf=2, min_samples_split=2, n_estimators=400	0.8726	0.9603	0.9605
SVM (tuned)	C=0.1, kernel=RBF, gamma=scale	0.8157	0.9303	0.9307

Table 2: Hyperparameter tuning Process data

Table 2 summarizes the hyperparameter tuning process for the Random Forest and SVM models. Grid search with 3-fold cross-validation was used to explore 54 Random Forest configurations and 24 SVM configurations. For Random Forest, limiting tree depth and enforcing minimum samples per leaf improved generalization, resulting in a validation accuracy of 0.9603. For SVM, an RBF kernel with a small regularization parameter (C = 0.1) was selected, prioritizing generalization over model complexity. Overall, Random Forest achieved stronger validation performance than the tuned SVM.

Model	val_accuracy(test)	val_f1_score(test)
Random Forest(tuned)	0.9766	0.9764

Table 3: Random Forest Final Test result

5. Mathematical Representation of Best Performing Algorithm

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions.

1. **Training:**
   - Uses **bootstrap sampling** to create random subsets of data for each tree.
   - Selects a **random subset of features** at each split to reduce correlation between trees.
2. **Prediction:**
   - **Classification**: The final prediction is the **majority vote** from all trees.

$$\hat{y} = \mathrm{Mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M)$$

   - **Regression**: The final prediction is the **average** of all tree predictions.

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m$$

3. **Key Concepts:**
   - **Bagging**: Random subsets of data.
   - **Random Feature Selection**: Reduces overfitting by selecting random features for each tree.
   - **Ensemble**: Combines multiple trees to improve accuracy and stability.

This process reduces overfitting and improves performance over a single decision tree.

6. Results

This section presents a comprehensive evaluation of the trained Human Activity Recognition models using subject-wise validation and testing. All results are analyzed in relation to the exploratory data analysis and preprocessing decisions discussed earlier.

6.1 Overall Model Performance

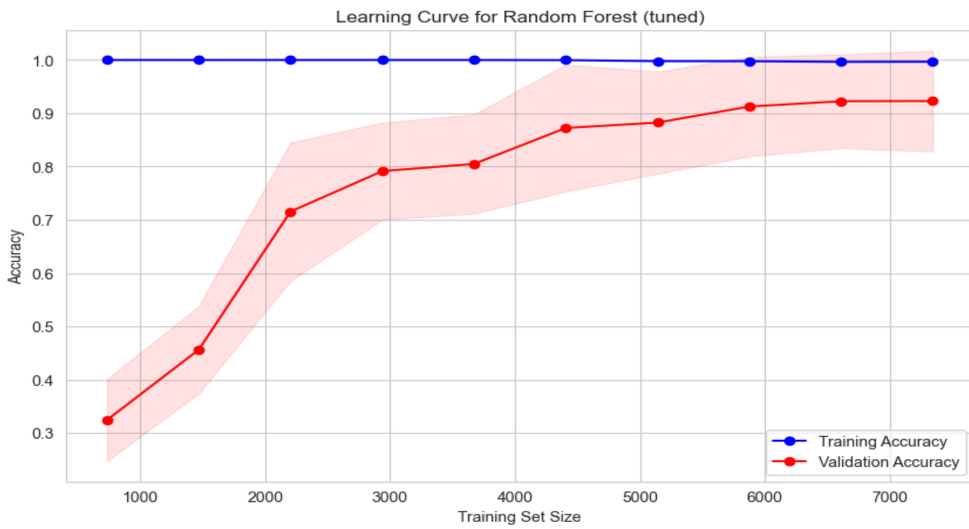

Figure 10 : Model Validation and Cross-Validation Analysis

Figure 10 compares the validation performance of the baseline and tuned models using **subject-wise evaluation**. Both Random Forest and scaled SVM achieved strong classification performance, confirming the effectiveness of the window-based feature representation.

The tuned **Random Forest** model demonstrated superior and more stable performance compared to SVM. While the scaled SVM achieved competitive validation accuracy in some experiments, Random Forest consistently maintained high accuracy and F1-score across different validation settings, indicating better robustness to subject variability.

6.2 Validation and Test Results

The final tuned Random Forest model achieved the following performance:

- **Validation Accuracy (subject-wise):** 96.03%
- **Validation F1-score:** 0.9605
- **Test Accuracy (unseen subject 107):** 97.66%
- **Test Weighted F1-score:** 0.976

The Random Forest model outperformed the SVM after hyperparameter tuning, highlighting the advantage of ensemble-based methods in handling high-dimensional and correlated sensor features.

7. Conclusion

This project successfully developed a comprehensive **Human Activity Recognition (HAR)** system using wearable sensor data from the **PAMAP2 dataset**. Through systematic exploratory data analysis, careful preprocessing, feature engineering, and machine learning modeling, a robust classification framework was established. The final system is capable of accurately distinguishing **seven daily activities**—*lying, sitting, standing, walking, running, cycling, and Nordic walking*—with high predictive performance.

Key Achievements

Data Processing and Feature Engineering

The dataset was carefully cleaned and refined by restricting the analysis to activities 1–7 and removing transient activity segments. Missing heart rate values were imputed subject-wise to prevent information leakage. Time-series data were segmented using a sliding window approach, and statistical window-level features (mean, standard deviation, minimum, and maximum) were extracted from the top 30 sensor channels identified through Random Forest feature importance ranking. This process resulted in a compact yet expressive feature representation consisting of 120 aggregated features per window.

Model Development

Baseline machine learning models, including Random Forest and Support Vector Machine (SVM) classifiers, were trained using subject-wise data splits to ensure realistic evaluation. Hyperparameter tuning via GridSearchCV further refined model performance. Among the evaluated models, Random Forest consistently demonstrated superior robustness and generalization capability, emerging as the best-performing algorithm.

Validation Insights

Confusion matrix analysis revealed strong per-class classification performance, particularly for dynamic activities such as walking and running, which were recognized with near-perfect accuracy. Cross-validation results and learning curves further confirmed that the selected model complexity was appropriate for the dataset size and variability.