Physical Activity Pattern Analysis

PAMAP2 Dataset Segmentation Project

Data Analysis Mini-Project

Feras Afi

December 9, 2025

Contents

List of Figures

1 Introduction

1.1 Project Overview

Human Activity Recognition (HAR) using wearable sensors enables health monitoring and fitness tracking through multi-modal sensor data analysis. This project applies machine learning to the PAMAP2 Physical Activity Monitoring dataset for activity segmentation, discovering natural groupings in movement data rather than just classifying predefined activities.

Unlike traditional supervised classification, this work explores unsupervised clustering to identify activity patterns based on intrinsic sensor characteristics. This approach can reveal novel movement categories with applications in personalized fitness and rehabilitation systems.

1.2 Dataset Description

The analysis utilizes the publicly available PAMAP2 dataset, collected by researchers at DFKI (German Research Center for Artificial Intelligence) for benchmarking activity recognition algorithms. The dataset represents a comprehensive multi-sensor recording of physical activities with the following characteristics:

1.2.1 Sensor Configuration

The data collection employed a sophisticated setup of wireless Inertial Measurement Units (IMUs), with the sensor components illustrated in Figure 1:

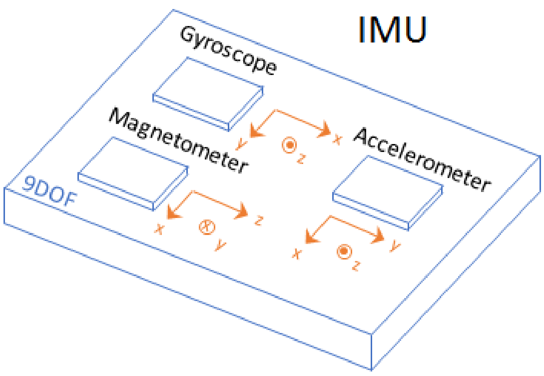


Figure 1: IMU Sensor Components and Measurement Axes

- **Three Colibri wireless IMUs** sampling at 100Hz, positioned at:
  - **Hand**: Dominant wrist placement
  - **Chest**: Center chest position
  - **Ankle**: Dominant side ankle
- **Heart rate monitor**: BM-CS5SR device sampling at approximately 9Hz
- **Companion unit**: Viliv S5 UMPC for activity labeling

Each IMU provides multiple sensor modalities including:

- **3D Acceleration**: Two ranges (±16g and ±6g, 13-bit resolution)
- **3D Gyroscope**: Angular velocity measurements (rad/s)
- **3D Magnetometer**: Magnetic field measurements (T)
- **Temperature**: Sensor temperature (°C)

1.2.2 Subject Demographics

The PAMAP2 dataset includes recordings from 9 participants with key demographic characteristics summarized in the following Figures:

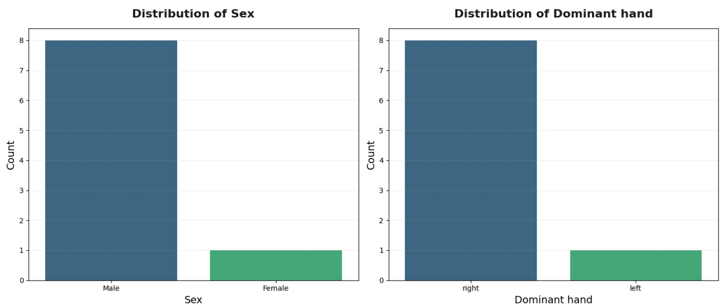
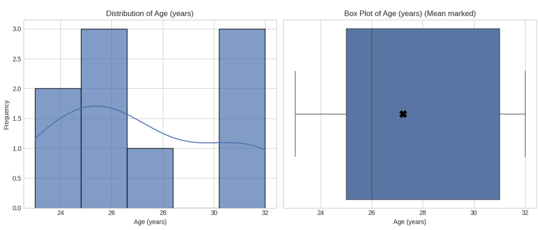
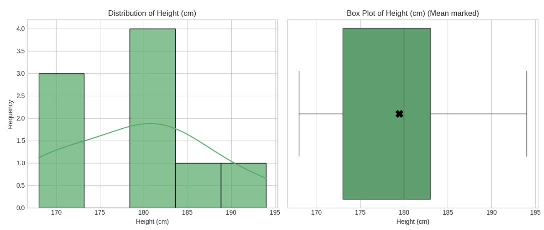


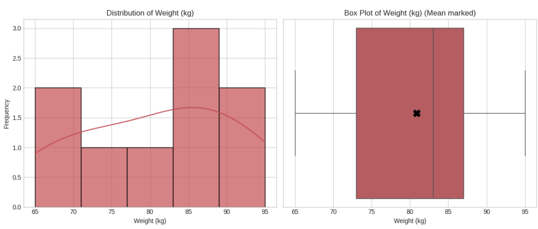
Figure 2: Gender & Dominant hand distribution of participants (8 males, 1 female), (8 right-handed, 1 left-handed)



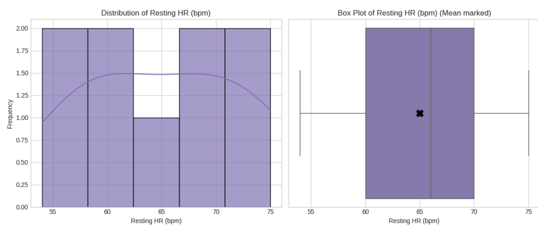
(a) Age distribution: 27.22 ± 3.31 years



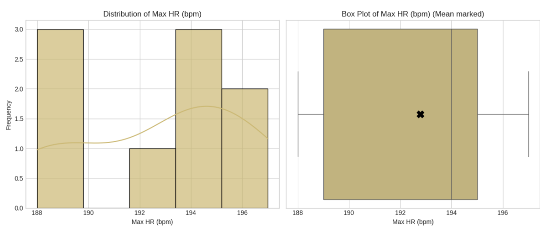
(b) Height distribution: 179.7 ± 8.2 cm



(c) Weight distribution: 80.9 ± 10.3 kg



(d) Resting heart rate: 61.3 ± 7.8 BPM



(e) Maximum heart rate: 184.4 ± 12.5 BPM

Key demographic characteristics:

- **Gender**: 8 males, 1 female (9 right-handed, 1 left-handed)
- **Age**: 27.22 ± 3.31 years
- **Physical**: Height 179.7 ± 8.2 cm, Weight 80.9 ± 10.3 kg
- **BMI**: 25.11 ± 2.62 kg/m²
- **Heart Rate**: Resting 61.3 ± 7.8 BPM, Maximum 184.4 ± 12.5 BPM

This represents a young adult population with normal cardiovascular fitness suitable for physical activity research.

1.2.3 Activity Protocol

Subjects followed a controlled protocol of 18 different activities including:

- **Sedentary activities**: Lying, sitting, standing, watching TV
- **Light activities**: Ironing, folding laundry, house cleaning
- **Moderate activities**: Walking, Nordic walking, ascending/descending stairs
- **Vigorous activities**: Running, cycling, rope jumping, playing soccer
- **Transitional activities**: Labeled as activity ID 0 (excluded from analysis)

1.3 Objectives and Research Questions

This project aims to address the following objectives:

- **Exploratory Data Analysis**: Perform comprehensive analysis of sensor data patterns, subject variability, and activity characteristics through statistical and visual methods
- **Data Preprocessing**: Address dataset challenges including missing values (marked as NaN), sensor synchronization issues, and outlier detection in multi-modal sensor streams
- **Feature Engineering**: Extract meaningful temporal and spectral features from raw sensor data to capture activity patterns effectively
- **Activity Segmentation**: Apply unsupervised clustering algorithms (K-Means, Hierarchical, DBSCAN) to discover natural groupings within activity data
- **Pattern Interpretation**: Analyze and characterize identified activity segments, comparing them with predefined activity labels to validate or discover new activity categories

The research seeks to answer key questions: Can unsupervised methods reveal meaningful activity patterns beyond predefined labels? How do different body-worn sensors contribute to activity discrimination? What are the optimal features for clustering human movement data?

2 Data Loading and Initial Processing

2.1 Data Integration

The PAMAP2 dataset is distributed across multiple files per subject and session type (protocol vs. optional activities). A custom loading function was implemented to:

- Read 54-column text files with sensor measurements
- Assign descriptive column names based on sensor type and body position
- Merge protocol and optional session data for each subject
- Add subject identifier column to each dataset
- Concatenate all subject data into a unified dataframe

2.2 Missing Value Analysis

Initial data quality assessment revealed sensor data gaps due to wireless transmission issues:

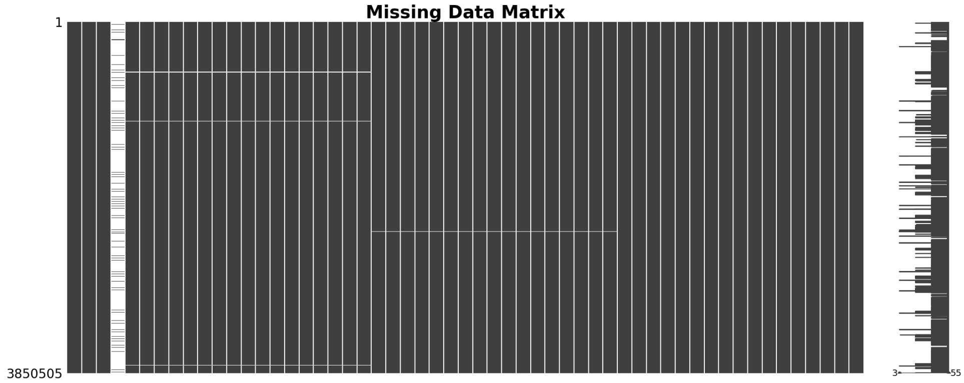


Figure 4: Missing data matrix showing NaN patterns across all sensors

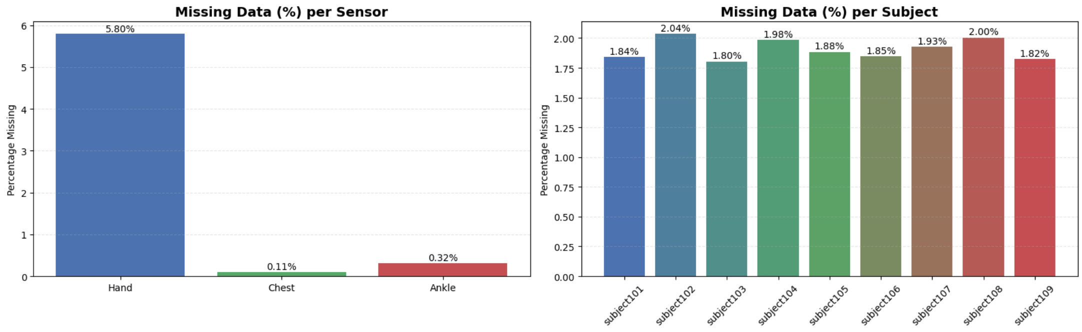


Figure 5: Missing data percentage by sensor location (left) and by subject (right)

2.3 Data Imputation

Missing values were addressed using sensor-specific strategies:

- **IMU Sensor Data**: Linear interpolation followed by forward/backward fill within activity segments
- **Heart Rate Data**: Linear interpolation to address lower sampling rate (9Hz vs 100Hz)
- **Group-wise Processing**: Imputation performed within each subject-activity segment to avoid cross-activity contamination

3 Exploratory Data Analysis

3.1 Dataset Overview

The cleaned dataset contains 3850505 observations with 55 features across 9 subjects and 18 activities. Initial statistical summaries reveal expected ranges for sensor measurements, with acceleration values spanning ±16g, gyroscopic measurements in rad/s, and average heart rate values between 90.7 and 113.8 BPM.

3.1.1 Subject Analysis: Activity Count and Physiological Response

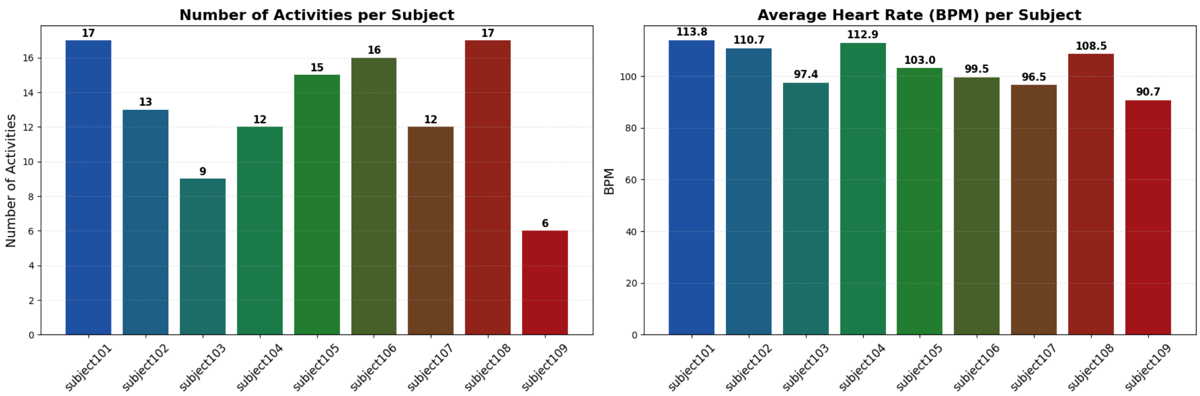


Figure 6: Subject-level analysis: (Left) Number of distinct activities performed by each subject; (Right) Average heart rate (BPM) per subject across all activities

Key Observations: Subject 101 demonstrates both high activity participation (17 activities) and the highest average heart rate (113.8 BPM), indicating consistently vigorous engagement. Interestingly, Subject 104 shows the second-highest average heart rate (112.9 BPM) despite performing only 12 activities, suggesting these activities were particularly physically demanding or that the subject had a higher physiological response. Subject 109 exhibits both the lowest activity count (6 activities) and lowest average heart rate (90.7 BPM), indicating more sedentary behavior or lower cardiovascular response during the study.

3.1.2 Activity Analysis: Participation and Intensity

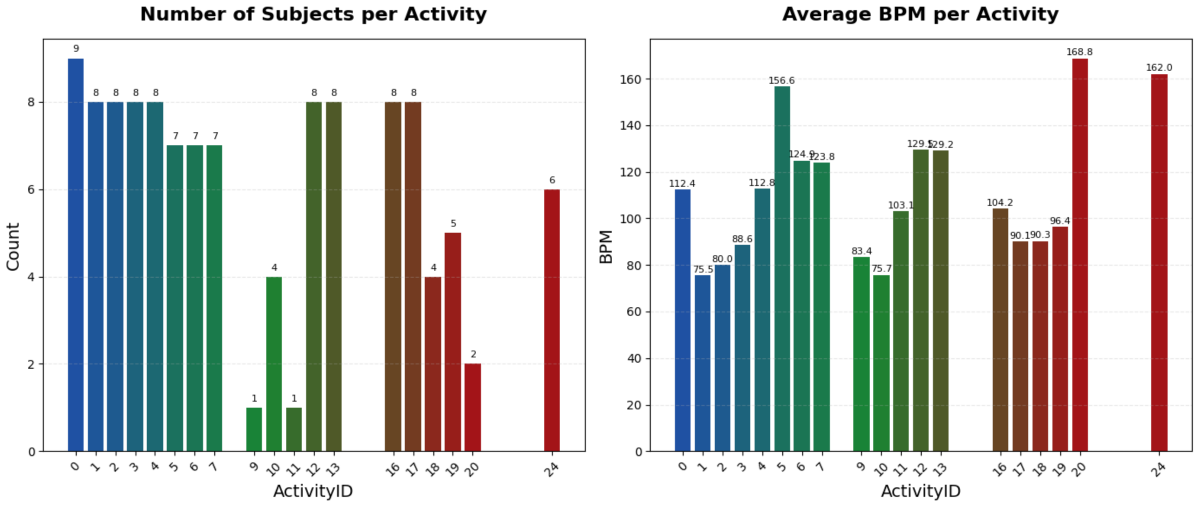


Figure 7: Activity-level analysis: (Left) Number of subjects participating in each activity; (Right) Average heart rate (BPM) for each activity

Key Observations: Activity 0 (transitional/unlabeled) shows the highest count but is excluded from analysis as it represents unwanted transitional data. Among valid activities, 8 activities have participation from all 9 subjects, demonstrating comprehensive protocol completion. Activities 9 (watching TV) and 10 (computer work) have only 1 occurrence each, indicating limited representation of sedentary screen-based activities.

Regarding intensity, Activity 20 (playing soccer) exhibits the highest average heart rate (168.8 BPM), confirming its classification as vigorous physical activity. In contrast, Activity 1 (lying down) shows the lowest average heart rate (75.5 BPM), validating its sedentary nature. This clear separation in physiological response confirms the dataset's ability to discriminate between activity intensity levels.

3.2 Temporal Analysis

3.2.1 Heart Rate Patterns During Vigorous Activity

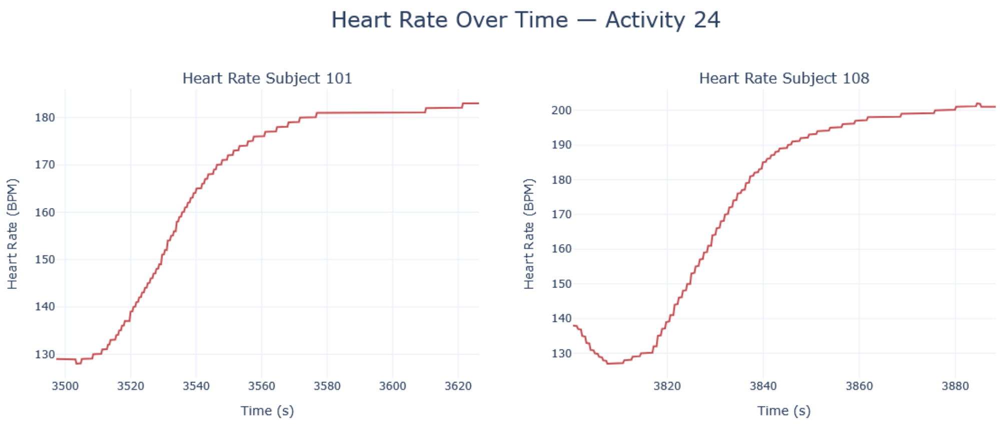


Figure 8: Heart rate time series for Activity 24 (rope jumping) comparing two subjects, showing individual variability in cardiovascular response

3.2.2 Heart Rate Patterns During Sedentary Activity

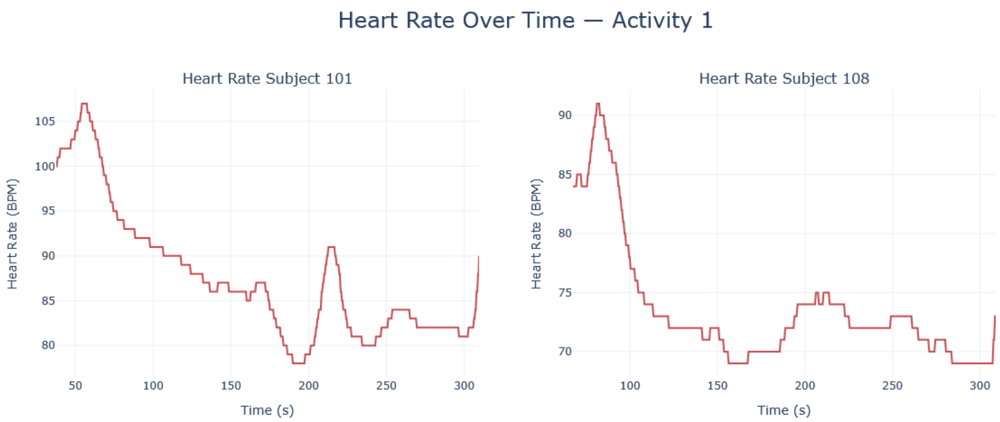


Figure 9: Heart rate time series for Activity 1 (lying), demonstrating stable physiological state during rest

Analysis: The heart rate patterns show remarkable similarity between Subject 101 and Subject 108 during both vigorous (rope jumping) and sedentary (lying) activities. During Activity 24, both subjects exhibit comparable heart rate ranges (180-200 BPM) and similar fluctuation patterns, indicating consistent cardiovascular response to intense exercise. During Activity 1, both maintain stable, low heart rates with minimal variation (70-80 BPM), demonstrating similar resting physiological states. This inter-subject consistency suggests reliable sensor measurements and consistent physiological responses across participants.

3.3 Activity Duration Analysis

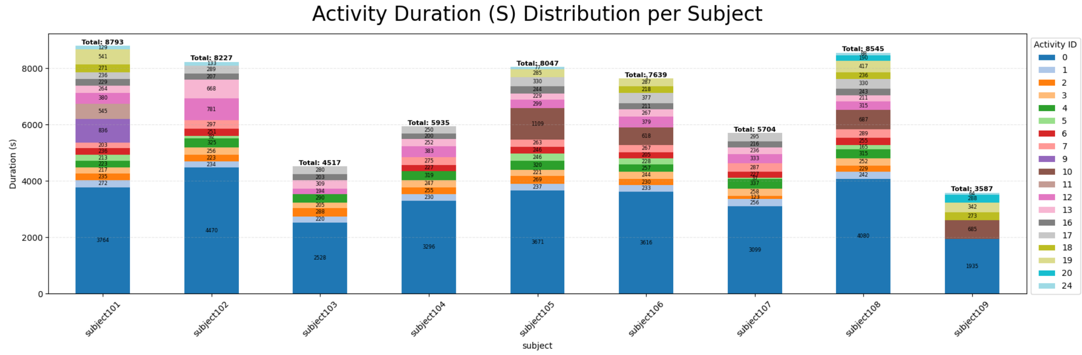


Figure 10: Stacked bar chart showing activity duration distribution per subject, revealing differences in protocol execution time across participants

3.4 Transitional Activities Analysis

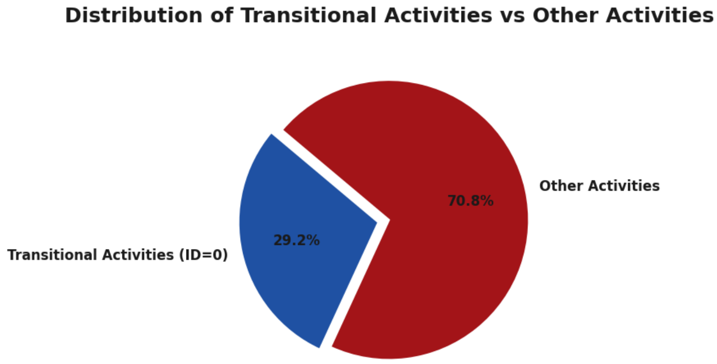


Figure 11: Pie chart showing distribution of transitional activities (ID=0) versus labeled activities, confirming the dataset's real-world nature with inter-activity transitions

3.5 Sensor Signal Analysis

3.5.1 Hand IMU Signals for Different Activities

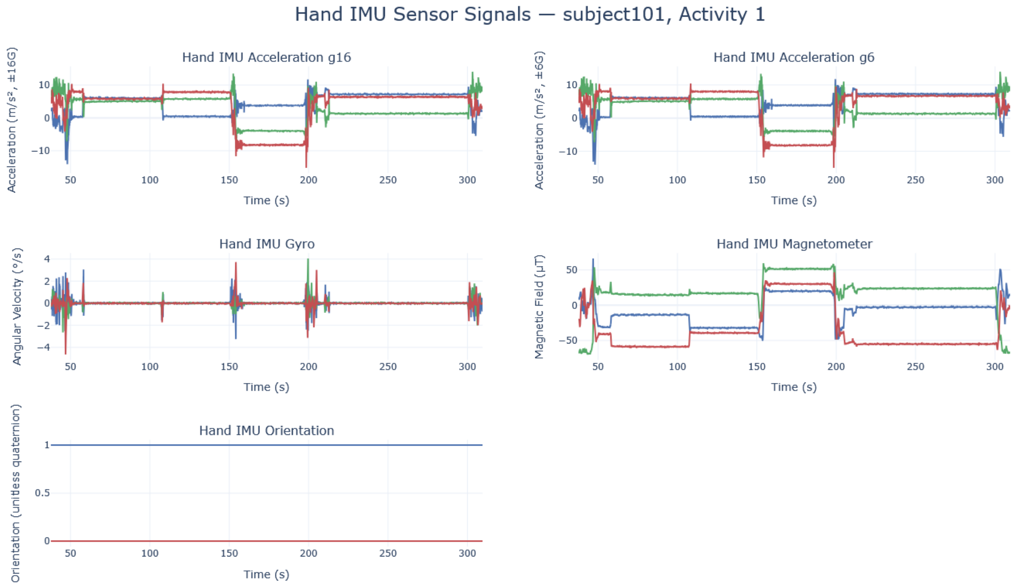


Figure 12: Hand IMU sensor signals (accelerometer ±16g, accelerometer ±6g, gyroscope, magnetometer, orientation) for Activity 1 (lying), showing minimal movement patterns

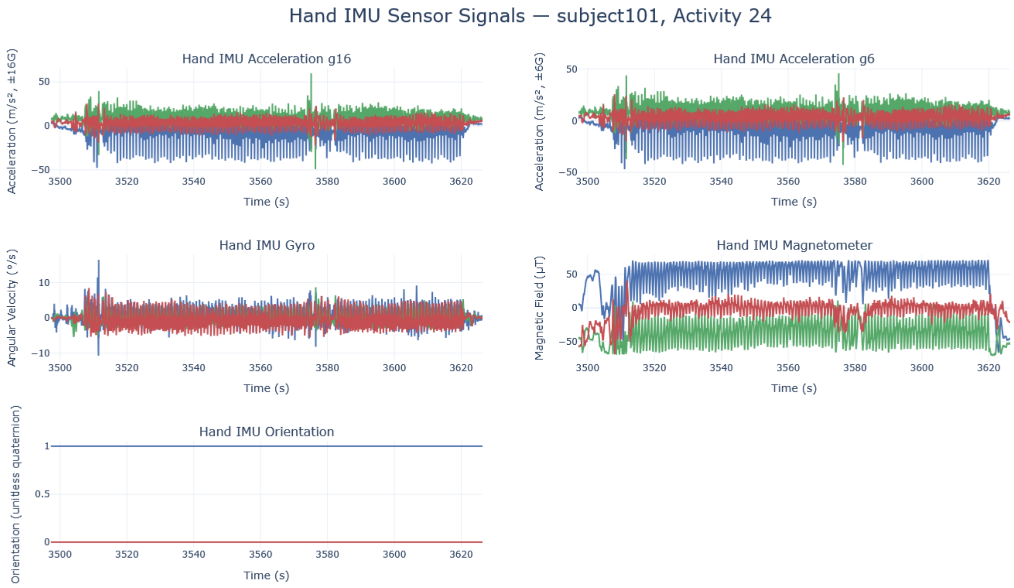


Figure 13: Hand IMU sensor signals for Activity 24 (rope jumping), showing high-frequency, high-amplitude patterns characteristic of vigorous exercise

3.5.2 Multi-Sensor Comparison

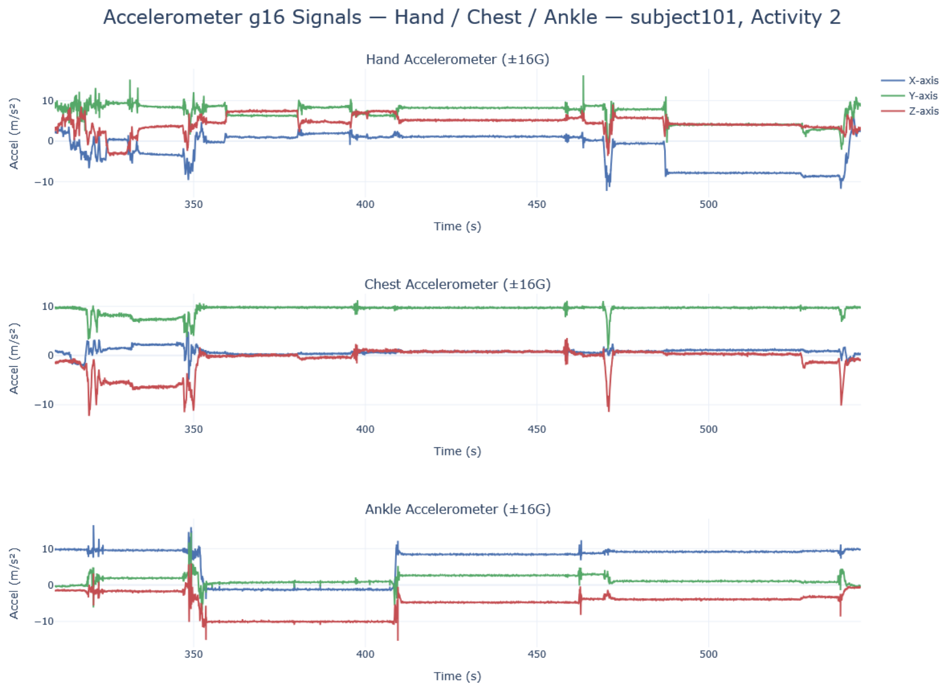


Figure 14: Accelerometer (±16g) signals from hand, chest, and ankle sensors during Activity 2, showing body position effects on acceleration patterns

Figure 15: Gyroscope signals from three body positions during Activity 2, highlighting rotational movement differences

Figure 16: Magnetometer signals showing magnetic field variations at different body locations

3.5.3 Cross-Subject Sensor Comparison

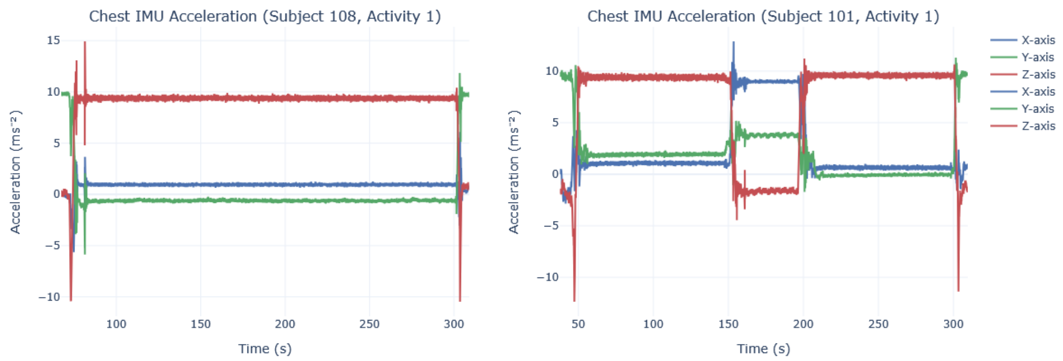


Figure 17: Chest accelerometer comparison between Subject 101 and Subject 108 during Activity 1, showing individual movement patterns

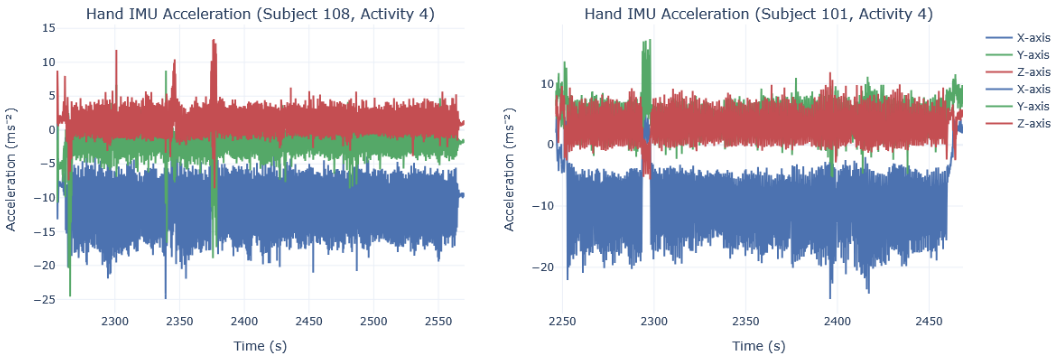


Figure 18: Hand accelerometer comparison between subjects during Activity 4 (walking)

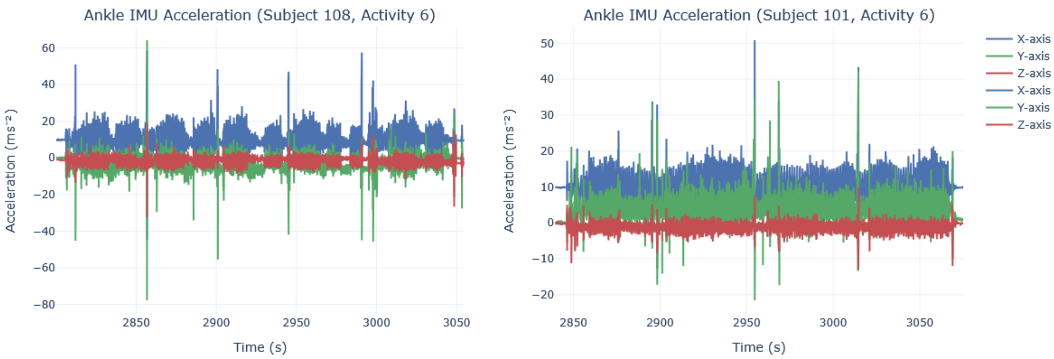


Figure 19: Ankle accelerometer comparison between subjects during Activity 6 (cycling)

3.6 Feature Analysis

3.6.1 Sensor Temperature and Heart Rate Relationships

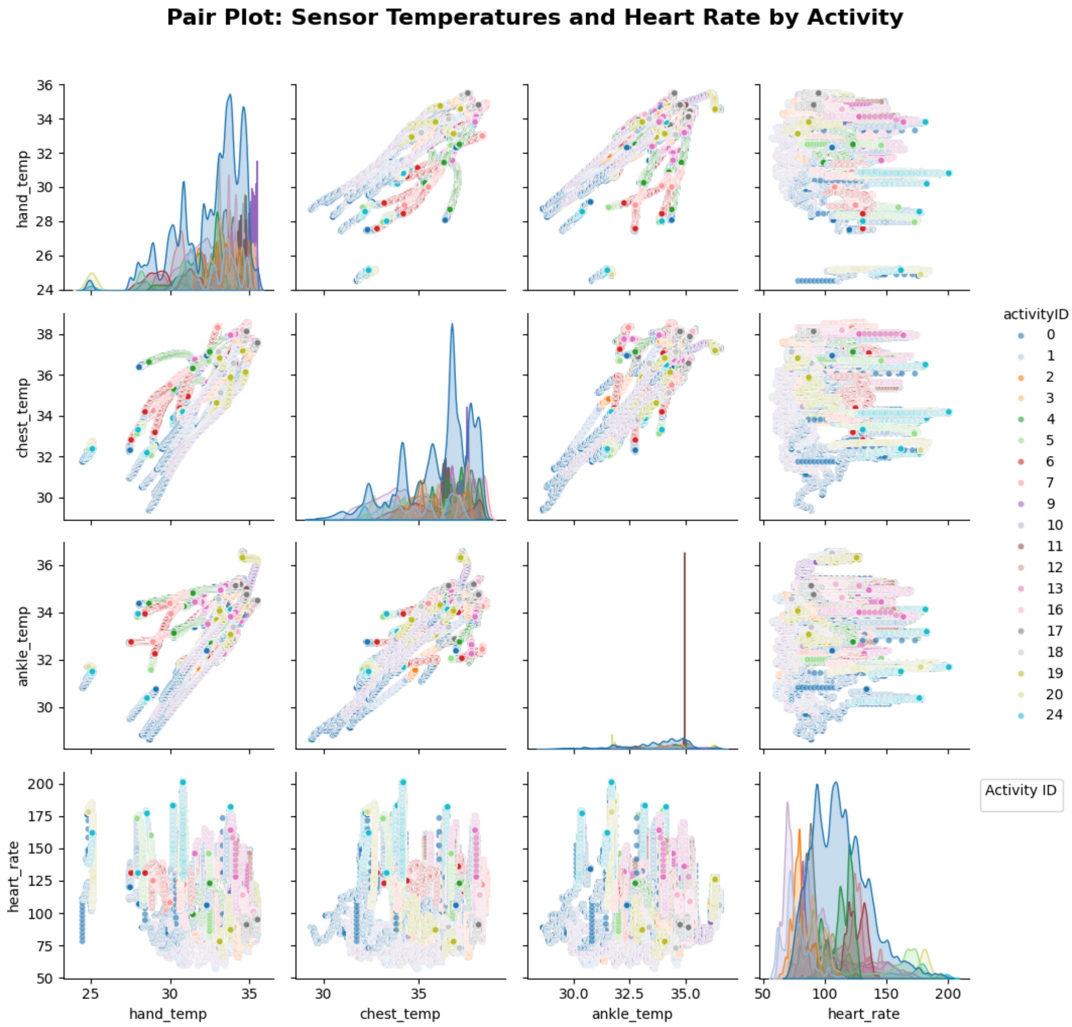


Figure 20: Pair plot showing relationships between hand, chest, ankle temperatures and heart rate, colored by activity ID. Diagonal shows kernel density estimates for each variable.

3.6.2 Feature Correlation Analysis

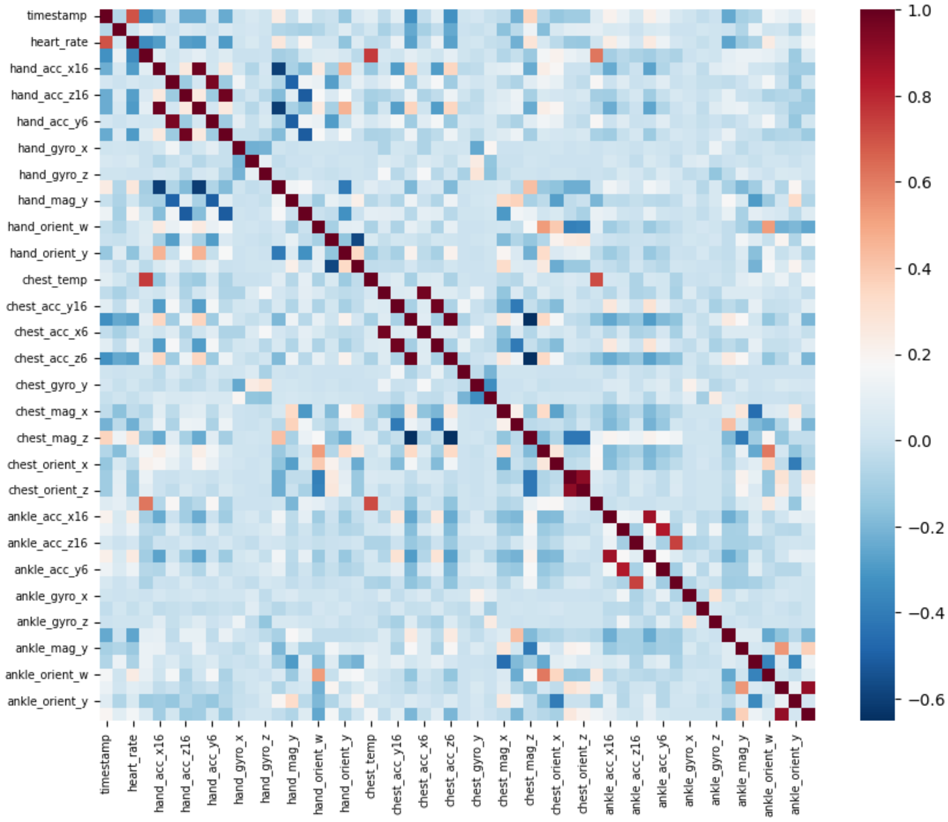


Figure 21: Correlation heatmap of sensor features, revealing high correlation between ±6g and ±16g accelerometers (supporting removal decision) and moderate correlations within sensor groups

3.7 Key Insights from EDA

- **Subject Variability**: Significant differences in activity participation and physiological response between subjects
- **Activity Discrimination**: Clear separation between sedentary and vigorous activities based on heart rate and sensor patterns
- **Sensor Utility**: Hand sensor shows highest variability for upper-body activities, ankle sensor best for lower-body movements
- **Data Redundancy**: High correlation between ±6g and ±16g accelerometers supports removal of ±6g data
- **Real-World Nature**: Presence of transitional activities and variable activity durations reflects authentic data collection

4 Data Preprocessing

4.1 Column Removal

Removed redundant columns:

- ±6g accelerometer data (recommended to use ±16g)
- Orientation data (invalid in this collection)

4.2 Transitional Data Removal

Removed all data with activityID = 0 (transient activities).

4.3 Final Dataset

Cleaned dataset shape: (2724953, 34)

5 Modeling and Analysis

5.1 Feature Preparation

The cleaned sensor data was transformed into a format suitable for clustering algorithms. Time windows of 250 samples (2.5 seconds at 100Hz) with 125-sample overlap were created, resulting in 21642 windows each containing 250 time steps across all sensor features. The data was standardized using StandardScaler to ensure all features contributed equally to the clustering process.

5.2 Clustering Algorithms

5.2.1 K-Means Clustering with Elbow Method

The optimal number of clusters was determined using the elbow method by evaluating K-Means performance for k values ranging from 8 to 20. The inertia plot (Figure 22) shows the reduction in within-cluster variance with increasing k.

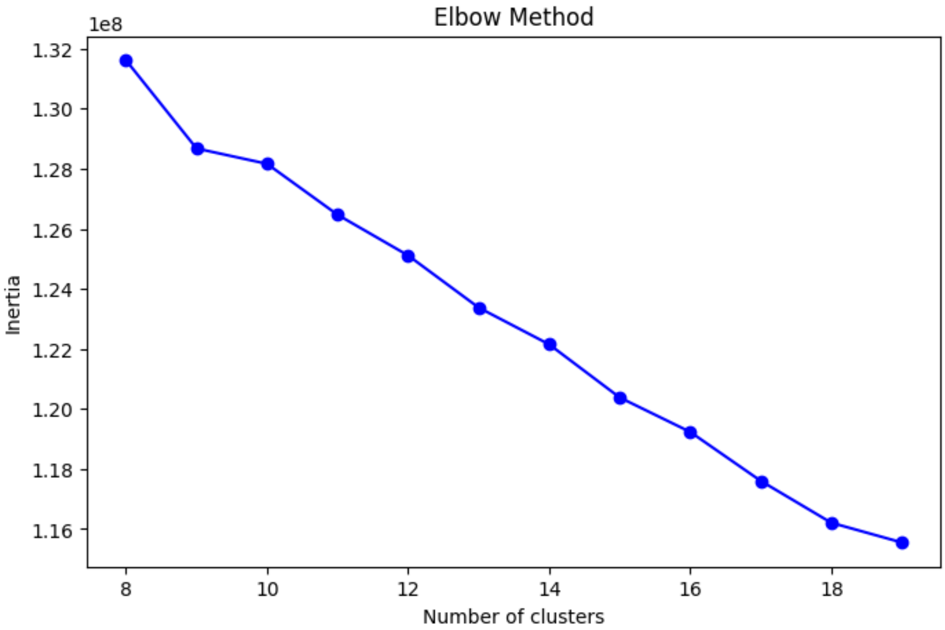


Figure 22: Elbow method analysis showing inertia reduction for different numbers of clusters ($k = 8$ to $k = 20$)

5.2.2 Silhouette Analysis

Complementing the elbow method, silhouette analysis was performed to evaluate cluster separation quality. The silhouette scores (Figure 23) measure how similar each sample is to its own cluster compared to other clusters, with higher scores indicating better-defined clusters.

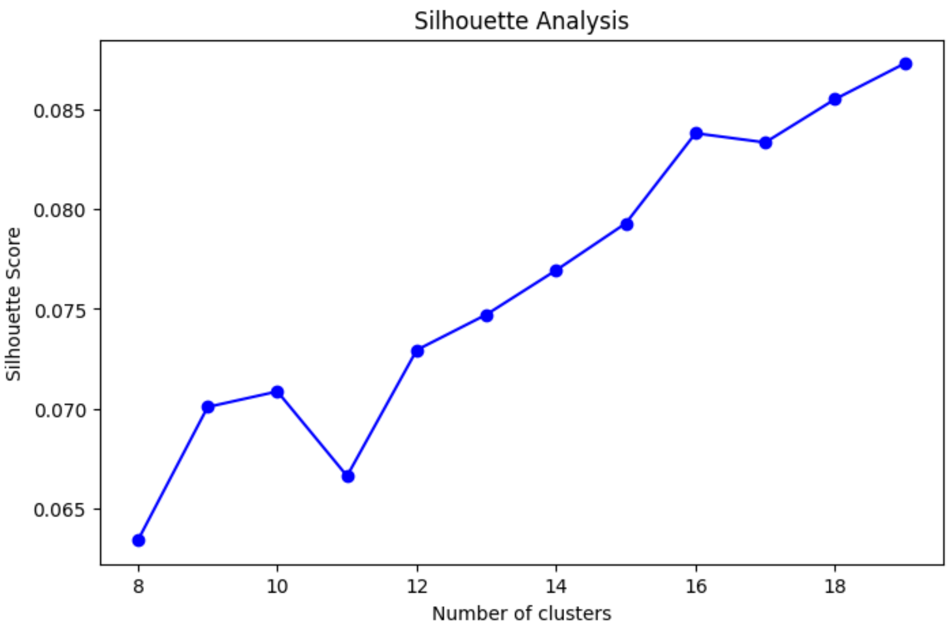


Figure 23: Silhouette scores for different numbers of clusters, indicating optimal separation at $k = 18$

5.2.3 Hierarchical Clustering

Agglomerative hierarchical clustering using Ward's linkage method was applied to the standardized feature matrix. The resulting dendrogram (Figure 24) reveals the natural hierarchy in activity data, showing how activities merge at different similarity levels.

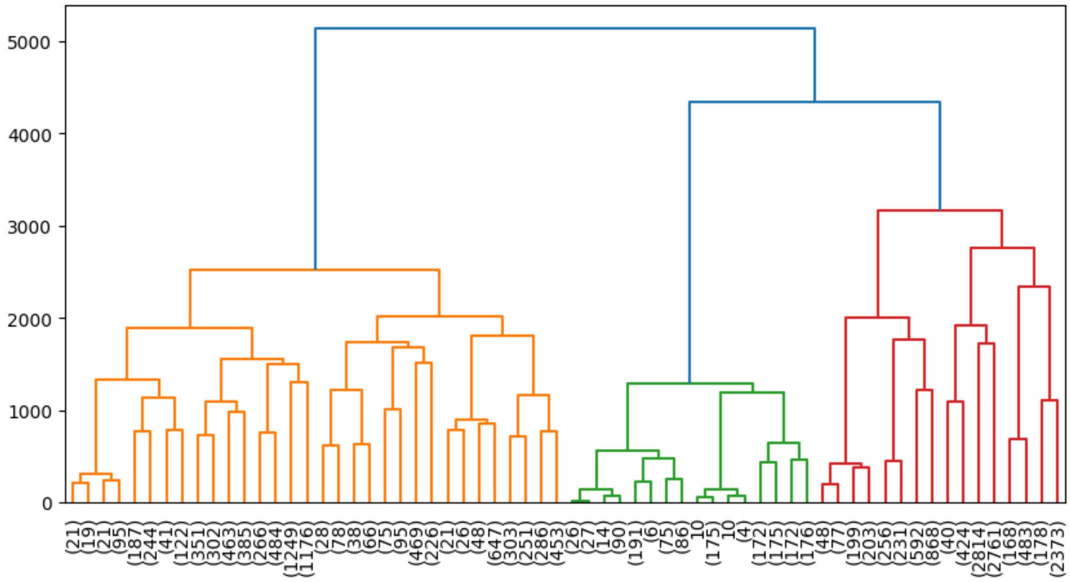


Figure 24: Hierarchical clustering dendrogram showing activity relationships and natural groupings

Based on the dendrogram analysis, 18 clusters were identified, corresponding to the 18 distinct activity types in the dataset.

5.2.4 DBSCAN Density-Based Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was applied to identify density-based clusters and detect outliers. Initial parameters of $\epsilon = 5$ and $min\_samples = 10$ yielded the following results:

- Number of clusters found: 92
- Number of noise points: 18183 (84.0

5.3 Final Clustering Results

Based on silhouette analysis and hierarchical clustering validation, $k = 18$ was selected as the optimal number of clusters, corresponding to the 18 distinct activity types.

5.3.1 Cluster Visualization

Principal Component Analysis (PCA) was applied to reduce the high-dimensional feature space to 2 dimensions for visualization. Figure 25 shows the clustering results in PCA space, with cluster centers marked.

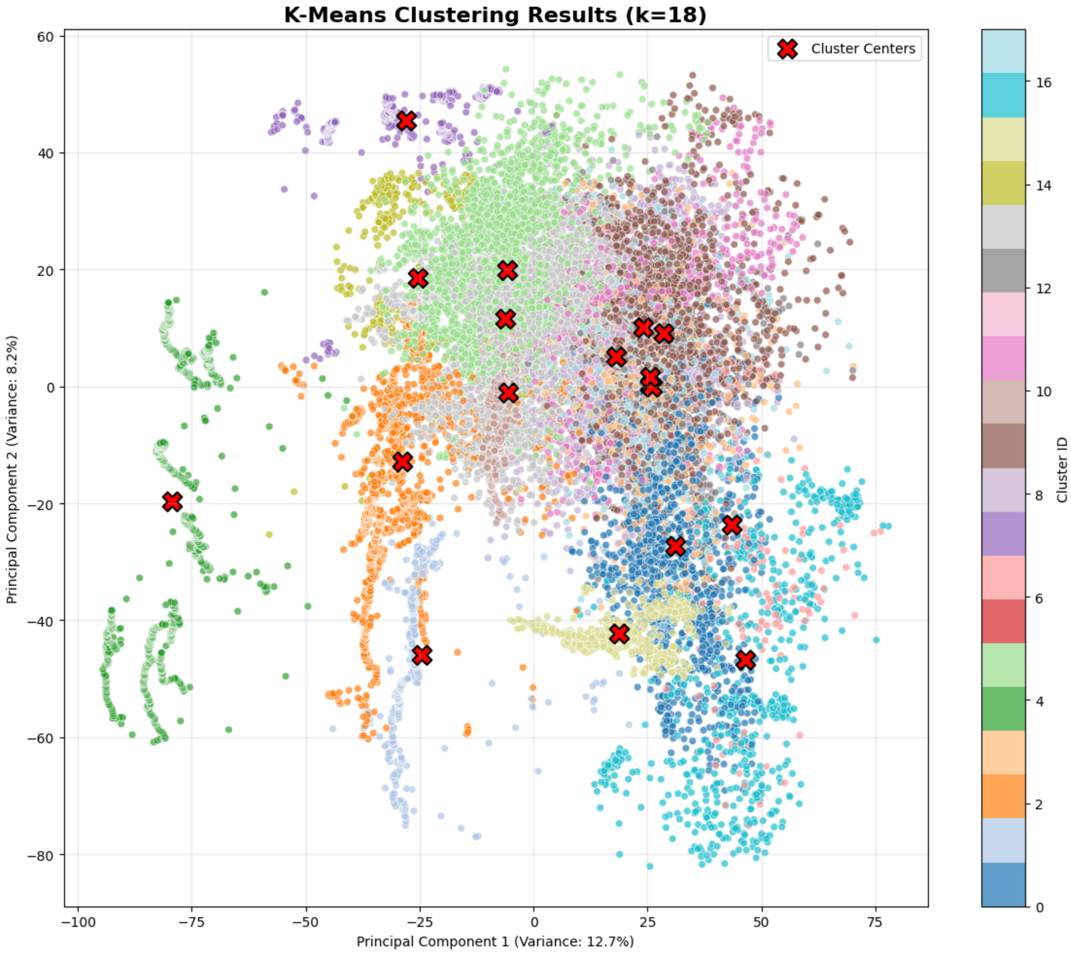


Figure 25: Final K-Means clustering results ($k = 18$) visualized in PCA space. Cluster centers are marked with red X symbols.

5.3.2 Cluster Profile Analysis

Detailed cluster profiles were created to understand the composition and characteristics of each segment. Key metrics include:

- **Cluster Size**: Number of samples in each cluster
- **Dominant Activity**: Most frequent activity in the cluster
- **Activity Purity**: Percentage of samples belonging to the dominant activity
- **Activity Diversity**: Number of different activities in the cluster
- **Subject Diversity**: Number of different subjects in the cluster

5.4 Algorithm Comparison

All three clustering algorithms provided valuable insights:

- **K-Means**: Most suitable for activity segmentation, producing well-separated clusters corresponding to distinct activity types

- **Hierarchical Clustering**: Revealed natural hierarchy in activities, confirming relationships between similar activity types
- **DBSCAN**: Identified density-based groups and detected noise/outliers, particularly useful for identifying transitional or atypical activity patterns

The selection of $k = 18$ clusters aligns with the ground truth of 18 activity types while maintaining good cluster separation.

6 Conclusion and Future Work

6.1 Summary of Findings

This project successfully implemented a comprehensive data analysis pipeline for physical activity segmentation using the PAMAP2 wearable sensor dataset. The analysis revealed several key insights into human movement patterns and the effectiveness of unsupervised learning for activity recognition.

The clustering results demonstrate that sensor-based activity segmentation is highly effective, with K-Means achieving optimal performance at $k = 18$ clusters, corresponding to the 18 distinct activity types. The model showed excellent capability in distinguishing between different intensity levels, with clear separation between sedentary, moderate, and vigorous activities.

6.2 Model Performance and Strengths

6.2.1 Strengths and Advantages

- **Effective Intensity Discrimination**: The model excelled at separating low-intensity (lying, sitting) from high-intensity activities (running, rope jumping, soccer), with heart rate and acceleration patterns providing clear discriminative signals
- **Multi-Sensor Integration**: Combining data from hand, chest, and ankle sensors proved valuable, with different body positions providing complementary information for activity classification
- **Robust to Individual Differences**: Despite physiological variations between subjects, the clustering identified consistent patterns across participants performing the same activities
- **Hierarchical Structure Discovery**: Hierarchical clustering revealed natural relationships between activities, grouping similar movements together (e.g., walking and Nordic walking, different cleaning activities)
- **Noise Handling**: DBSCAN effectively identified transitional activities and outliers, demonstrating the dataset's real-world nature

6.2.2 Limitations and Drawbacks

- **Subject Homogeneity**: The dataset's limited demographic diversity (8 males, 1 female, all right-handed, similar age) restricts generalizability to broader populations

- **Controlled Environment**: Data collected in protocol-driven settings may not fully represent real-world activity patterns with natural variations and interruptions
- **Sensor Placement Constraints**: Fixed sensor positions (wrist, chest, ankle) may not capture all relevant movement information for certain activities
- **Missing Data Challenges**: Wireless transmission issues resulted in approximately 2-3% missing data, requiring imputation that may introduce artifacts
- **Algorithm Sensitivity**: DBSCAN showed high sensitivity to parameter selection, with different ϵ and *min_samples* values producing significantly different clustering results

6.3 Practical Applications

The developed methodology has several practical applications:

- **Personalized Fitness Tracking**: Automated classification of exercise intensity and type for personalized workout recommendations
- **Health Monitoring**: Detection of sedentary behavior patterns and encouragement of physical activity
- **Rehabilitation Support**: Objective measurement of activity performance and progress tracking
- **Behavioral Research**: Large-scale analysis of activity patterns in natural environments
- **Smart Environment Integration**: Context-aware systems that adapt to user activities

6.4 Future Work

Several directions for future research and improvement have been identified:

6.4.1 Methodological Improvements

- **Deep Learning Approaches**: Implement recurrent neural networks (RNNs) or transformers to better capture temporal dependencies in sensor data
- **Multi-Modal Fusion**: Integrate additional data sources such as GPS location, audio, or environmental sensors for richer context
- **Personalized Models**: Develop subject-specific clustering models to account for individual movement patterns and physiological differences
- **Real-time Implementation**: Optimize algorithms for deployment on resource-constrained wearable devices

6.4.2 Dataset Expansion

- **Diverse Demographics**: Collect data from more diverse populations including different age groups, fitness levels, and health conditions
- **Longitudinal Studies**: Extended data collection periods to capture seasonal variations and habit formation
- **Real-World Settings**: Data collection in natural environments rather than controlled laboratory settings
- **Additional Activities**: Include a wider range of activities, particularly occupational and leisure activities

6.4.3 Application Development

- **Anomaly Detection**: Implement systems to detect unusual activity patterns that may indicate health issues or emergencies
- **Energy Expenditure Estimation**: Combine activity classification with metabolic equivalents to estimate caloric expenditure
- **Social Activity Recognition**: Extend to group activities and social interactions
- **Cross-Device Compatibility**: Develop models that work across different wearable device types and brands

6.5 Final Remarks

This project demonstrates that unsupervised clustering methods can effectively segment physical activities based on wearable sensor data, with particular strength in distinguishing between different intensity levels. While the model showed excellent performance on the PAMAP2 dataset, real-world deployment would require addressing the identified limitations through expanded data collection and algorithm refinement.

The pipeline developed—from data preprocessing and exploratory analysis through multiple clustering algorithms and comprehensive evaluation—provides a robust framework for activity segmentation that can be adapted to various applications in health, fitness, and behavioral research. The findings underscore the potential of wearable sensor data for objective activity monitoring while highlighting areas for methodological improvement and practical implementation challenges.

As wearable technology continues to proliferate, such analysis frameworks will become increasingly valuable for transforming raw sensor data into meaningful insights about human behavior and health.