# 山东大学 计算机科学与技术 学院

# 大数据分析与实践 课程实验报告

| 学号：202300130005 | 姓名：于佳杭 | 班级：23 数据 |
|---|---|---|
| 实验题目：手机数据采集与分析实践 | | |
| 实验学时：2 | 实验日期： 2025.11.4 | |

实验步骤与内容：

一、数据集：

本实验采用 phyphox 应用作为数据收集工具，分别采集跑步、走路、乘坐电动车三种不同运动状态下的加速度计数据。数据收集过程中，phyphox 应用记录了时间及 X、Y、Z 三轴加速度原始数据，采样过程保持设备固定以确保数据稳定性，最终形成包含原始加速度信息的数据集

实验数据集包含三种运动状态对应的加速度计原始数据，数据字段涵盖时间（秒）、X 轴加速度（m/s²）、Y 轴加速度（m/s²）、Z 轴加速度（m/s²）等核心信息。通过数据清洗对原始数据进行预处理，删除缺失值并转换数据类型，同时计算得到合加速度（含重力）、去重力影响的加速度、净加速度、水平加速度及垂直加速度等衍生特征

| | Time (s) | Accelerati | Accelerati | Acceleration z (m/s^2) |
|---|---|---|---|---|
| 2 | 0.000496 | -1.16794 | 9.299624 | 10.43098 |
| 3 | 0.002488 | -1.17275 | 9.200381 | 10.46136 |
| 4 | 0.004488 | -1.18383 | 9.093006 | 10.49113 |
| 5 | 0.006488 | -1.17522 | 8.979272 | 10.49552 |
| 6 | 0.008489 | -1.13892 | 8.877585 | 10.4857 |
| 7 | 0.010511 | -1.08053 | 8.768784 | 10.46441 |
| 8 | 0.012488 | -1.01915 | 8.655575 | 10.42701 |
| 9 | 0.014488 | -0.94709 | 8.553496 | 10.37771 |
| 0 | 0.016488 | -0.83913 | 8.454904 | 10.32945 |
| 1 | 0.018488 | -0.71632 | 8.34204 | 10.27318 |
| 2 | 0.020488 | -0.57535 | 8.211261 | 10.22004 |
| 3 | 0.022488 | -0.4281 | 8.068344 | 10.16584 |
| 4 | 0.024488 | -0.26592 | 7.923859 | 10.11522 |
| 5 | 0.026489 | -0.1226 | 7.767958 | 10.07372 |

二、数据清洗：

1.缺失值处理

数据中可能包含缺失值（NaN），这会影响分析的准确性。此步骤首先统计各列的缺失值数量，然后删除包含缺失值的所有行。函数会记录清洗前后的数据量变化，让用户了解有多少数据因不完整而被移除，确保后续分析基于完整的数据进行。

```
print("缺失值统计:")
print(df.isnull().sum())

# 删除包含NaN的行
original_rows = len(df)
df_clean = df.dropna()
cleaned_rows = len(df_clean)

print(f"原始数据行数: {original_rows}")
print(f"删除缺失值后行数: {cleaned_rows}")
print(f"删除行数: {original_rows - cleaned_rows}")
```

2.数据类型转换

加速度数据需要是数值类型才能进行计算分析。此步骤确保所有列都转换为数值类型，使用 pandas 的 to_numeric 函数进行转换，并将转换失败的值设为 NaN。转换完成后，再次清理可能产生的 NaN 值，并验证最终的数据类型，确保所有数据都是适合计算的数值格式。

```
print(df_clean.dtypes)

# 确保所有列都是数值类型
for col in df_clean.columns:
    df_clean[col] = pd.to_numeric(df_clean[col], errors='coerce')

print("\n转换后数据类型:")
print(df_clean.dtypes)

# 再次删除转换产生的NaN
df_clean = df_clean.dropna()
```

3. 加速度特征计算

基于三轴加速度原始数据计算多种有意义的特征。包括合加速度（含重力影响）、去重力合加速度（减去 9.8 m/s² 的重力加速度）、净加速度（合加速度减重力）、水平面加速度（X-Y 平面合成）和垂直加速度（Z 轴）。同时利用峰值检测算法识别步伐特征，计算步频和步幅周期，为步态分析提供关键指标。

```
if all(col in df_clean.columns for col in ['acc_x', 'acc_y', 'acc_z']):
    # 计算合加速度（含重力）
    df_clean['acc_magnitude'] = np.sqrt(
        df_clean['acc_x'] ** 2 +
        df_clean['acc_y'] ** 2 +
        df_clean['acc_z'] ** 2
    )

    # 计算去重力影响的加速度（假设重力主要在Z轴）
    gravity = 9.8  # m/s²
    df_clean['acc_magnitude_no_gravity'] = np.sqrt(
        df_clean['acc_x'] ** 2 +
        df_clean['acc_y'] ** 2 +
        (df_clean['acc_z'] - gravity) ** 2
    )

    # 计算净加速度（合加速度减去重力）
    df_clean['net_acceleration'] = df_clean['acc_magnitude'] - gravity

    # 计算水平加速度（X-Y平面）
    df_clean['acc_horizontal'] = np.sqrt(df_clean['acc_x'] ** 2 + df_clean['acc_y'] ** 2)

    # 计算垂直加速度（Z轴，含重力）
    df_clean['acc_vertical'] = df_clean['acc_z']
```

4. 步频计算：

使用三轴加速度计数据（已去除重力分量）的合成幅度来计算用户的步频。首先通过 SciPy 的 find_peaks 函数检测加速度幅度中的峰值，这些峰值对应行走时的脚步落地时刻。检测时设置最小峰值高度为 0.5、相邻峰值最小间隔为 10 个采样点以过滤噪声。若检测到多个峰值，则根据时间戳计算相邻峰值的时间间隔，并求平均值得到平均步幅周期。最后将平均步幅周期转换为步频（步/分钟）并输出结果，包括检测到的步数、平均步频和步幅周期。

```
acc_magnitude = df_clean['acc_magnitude_no_gravity'].values
from scipy.signal import find_peaks

try:
    # 找到所有峰值
    peaks, _ = find_peaks(acc_magnitude, height=0.5, distance=10)  # 最小高度0.5，最小间隔10个采样点

    if len(peaks) > 1:
        # 计算步频
        if 'time' in df_clean.columns:
            time_values = df_clean['time'].values
            peak_times = time_values[peaks]
            step_intervals = np.diff(peak_times)
            avg_step_interval = np.mean(step_intervals)

            if avg_step_interval > 0:
                steps_per_second = 1 / avg_step_interval
                steps_per_minute = steps_per_second * 60
                print(f"检测到 {len(peaks)} 个步伐峰值")
                print(f"平均步频: {steps_per_minute:.1f} 步/分钟")
                print(f"平均步幅周期: {avg_step_interval:.3f} 秒")
```

4.运动类型初步判断

基于计算得到的加速度特征，此步骤对运动类型进行初步分类。通过分析去重力加速度的平均值和波动程度，将运动分为静止、慢速步行、正常步行、快走/慢跑和跑步等不同强度等级。这为用户提供了一个直观的参考，了解所分析数据对应的运动强度。

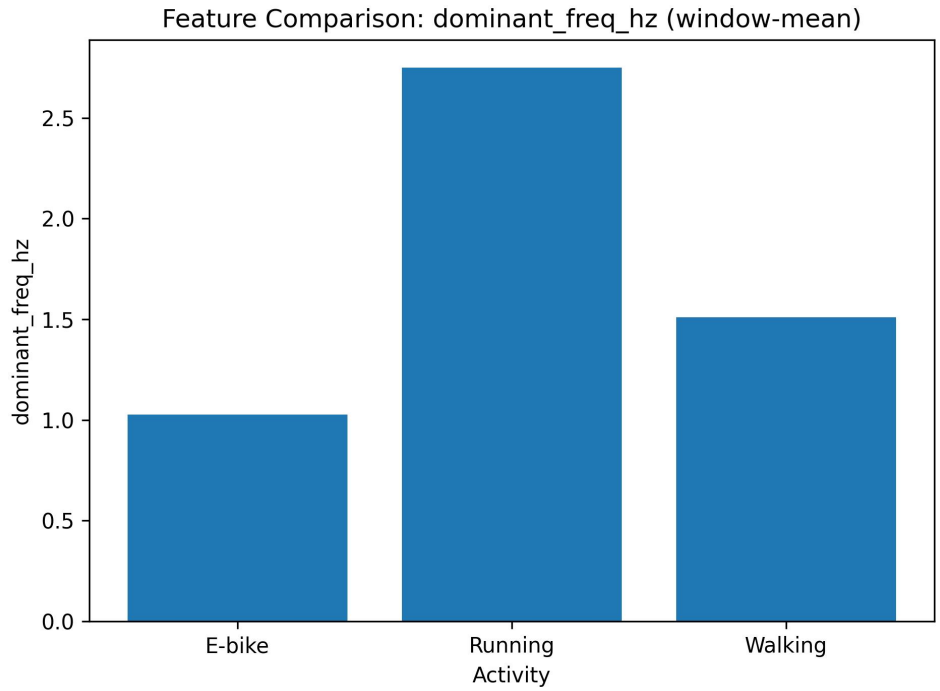
```python
avg_net_acc = df_clean['acc_magnitude_no_gravity'].mean()
acc_std = df_clean['acc_magnitude_no_gravity'].std()

if avg_net_acc < 0.3:
    print(f"  → 可能为静止状态（平均去重力加速度: {avg_net_acc:.3f} m/s²)")
    print(f"     加速度波动较小（标准差: {acc_std:.3f})")
elif avg_net_acc < 1.0:
    if acc_std < 0.5:
        print(f"  → 可能为慢速步行（平均去重力加速度: {avg_net_acc:.3f} m/s²)")
        print(f"     加速度波动适中（标准差: {acc_std:.3f})")
    else:
        print(f"  → 可能为正常步行（平均去重力加速度: {avg_net_acc:.3f} m/s²)")
        print(f"     加速度波动明显（标准差: {acc_std:.3f})")
elif avg_net_acc < 2.5:
    print(f"  → 可能为快速步行或慢跑（平均去重力加速度: {avg_net_acc:.3f} m/s²)")
    print(f"     加速度波动较大（标准差: {acc_std:.3f})")
else:
    print(f"  → 可能为跑步（平均去重力加速度: {avg_net_acc:.3f} m/s²)")
    print(f"     加速度波动很大（标准差: {acc_std:.3f})")
```

三、可视化：

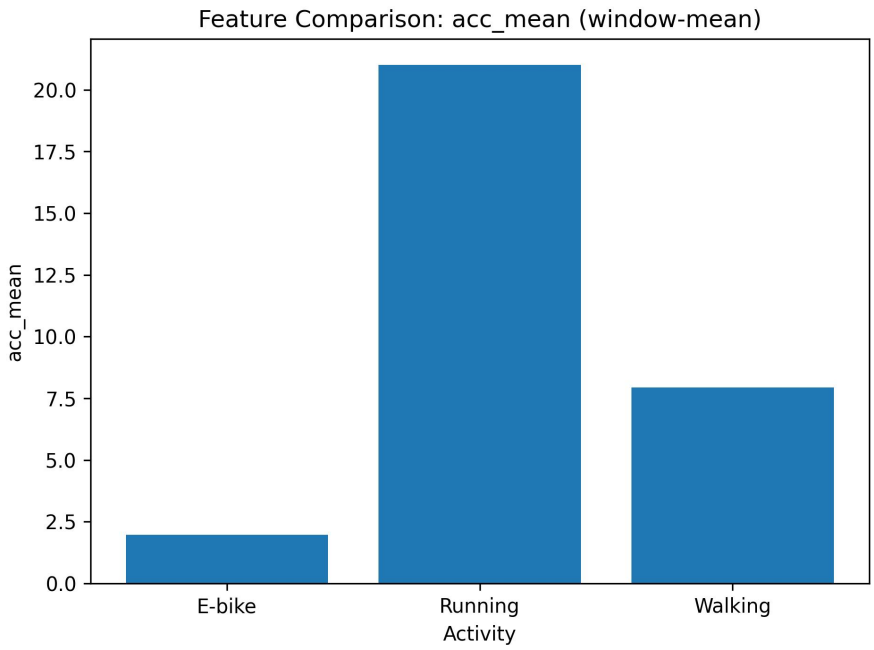
1.主频特征（dominant_freq_hz）

主频代表加速度信号的主要振动频率，跑步的主频（约 2.7Hz）远高于步行（约 1.5Hz）和电动车（约 1.0Hz）



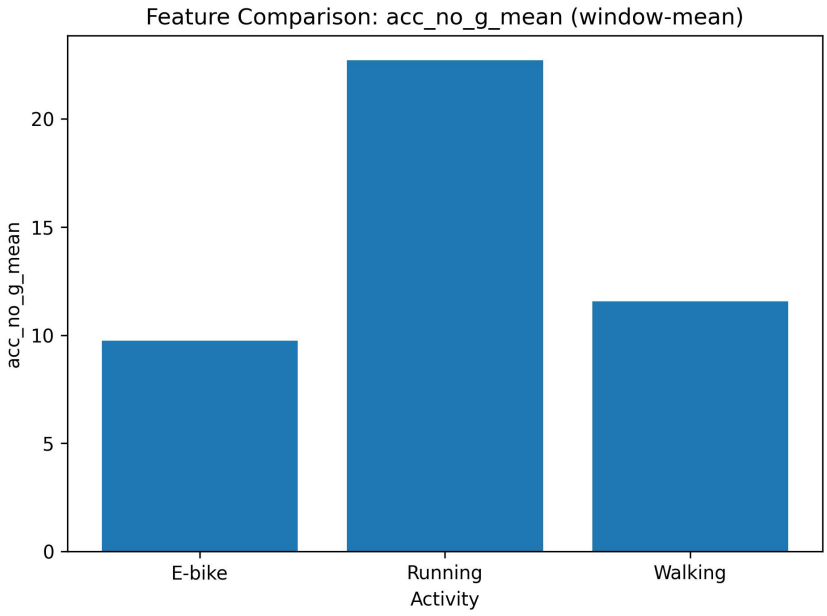

2.净加速度均值（acc_mean）

净加速度是合加速度减去重力后的结果：跑步的净加速度均值（约 10m/s²）显著为正，说明跑步时身体的加速度波动远大于重力；步行的净加速度均值略负，接近重力影响的抵消状态；电动车的净加速度均值（约 - 7.5m/s²）为负且绝对值大，反映其运动状态下加速度受重力的 “抵消效应” 更明显，整体更平稳。




3. 去重力加速度均值（acc_no_g_mean）

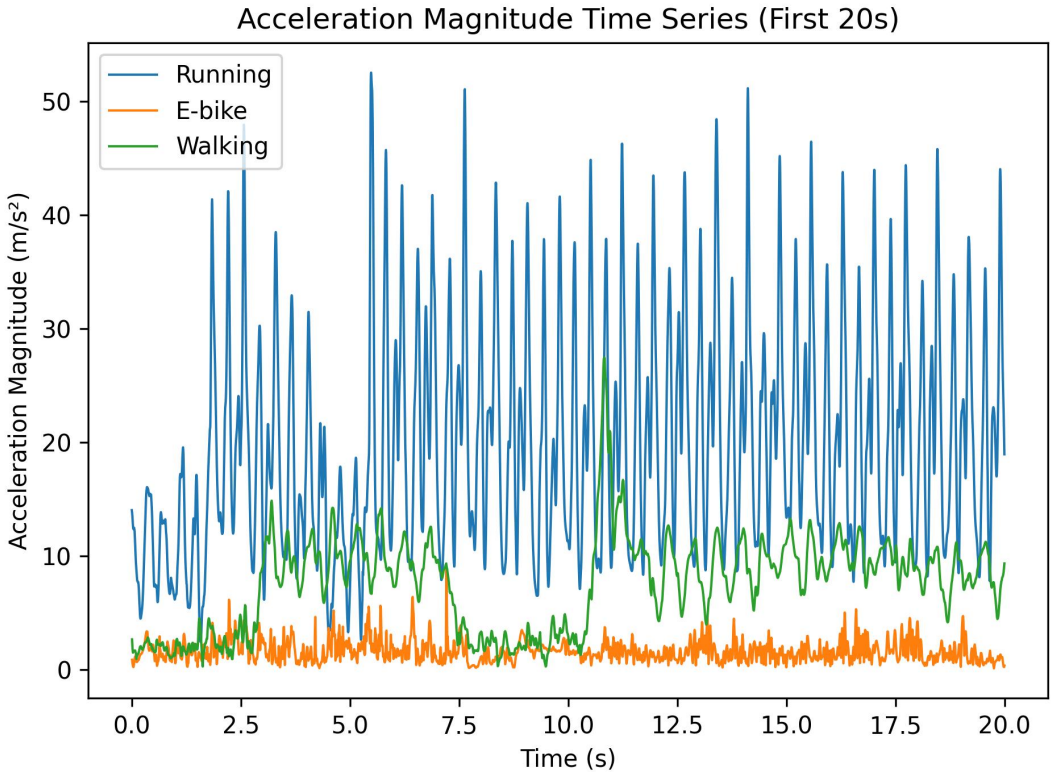
该指标剔除了地球重力的恒定影响（9.8 m/s²），专注于反映设备自身运动产生的纯动态加速度。数据显示，跑步状态下的去重力加速度均值最高（约 22 m/s²），显著高于步行状态（约 11 m/s²）和电动车出行（约 9.5 m/s²）。这一差异清晰表明：跑步时人体主动产生的加速度强度最大，步行动作次之，而乘坐电动车时虽然整体移动速度更快，但人体自身产生的动态加速度却相对平缓。此指标能更准确地反映不同运动模式下人体或设备的真实动态强度。




4. 加速度幅值时间序列（前 20 s）

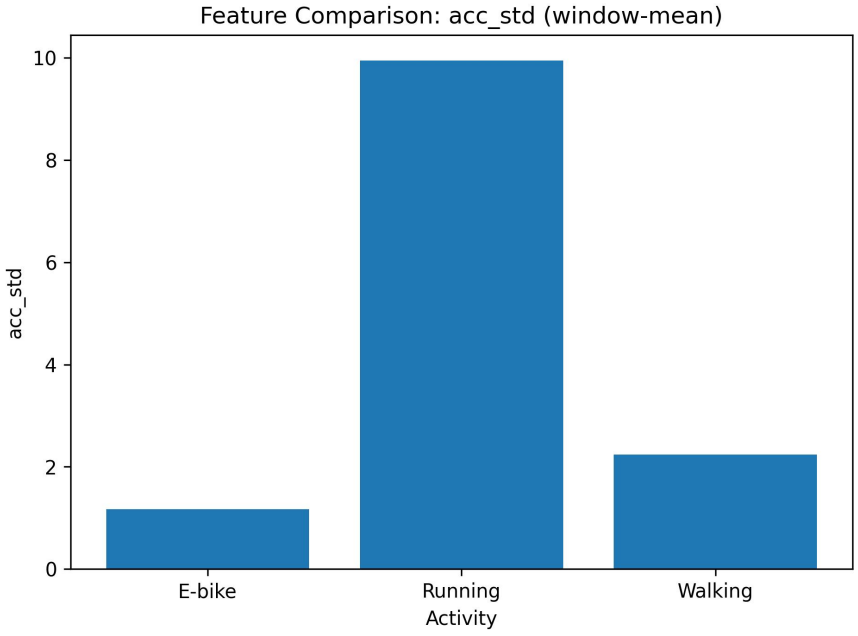
时间序列图直观呈现了三种状态的波动差异：跑步的加速度幅值波动最大（峰值超50m/s²），且波动频率高；步行的幅值波动次之（峰值约 25m/s²），周期性较明显；电

动车的幅值波动最小（峰值仅约 5m/s²），整体处于低波动水平，体现了不同运动状态下加速度的 “剧烈程度” 差异。



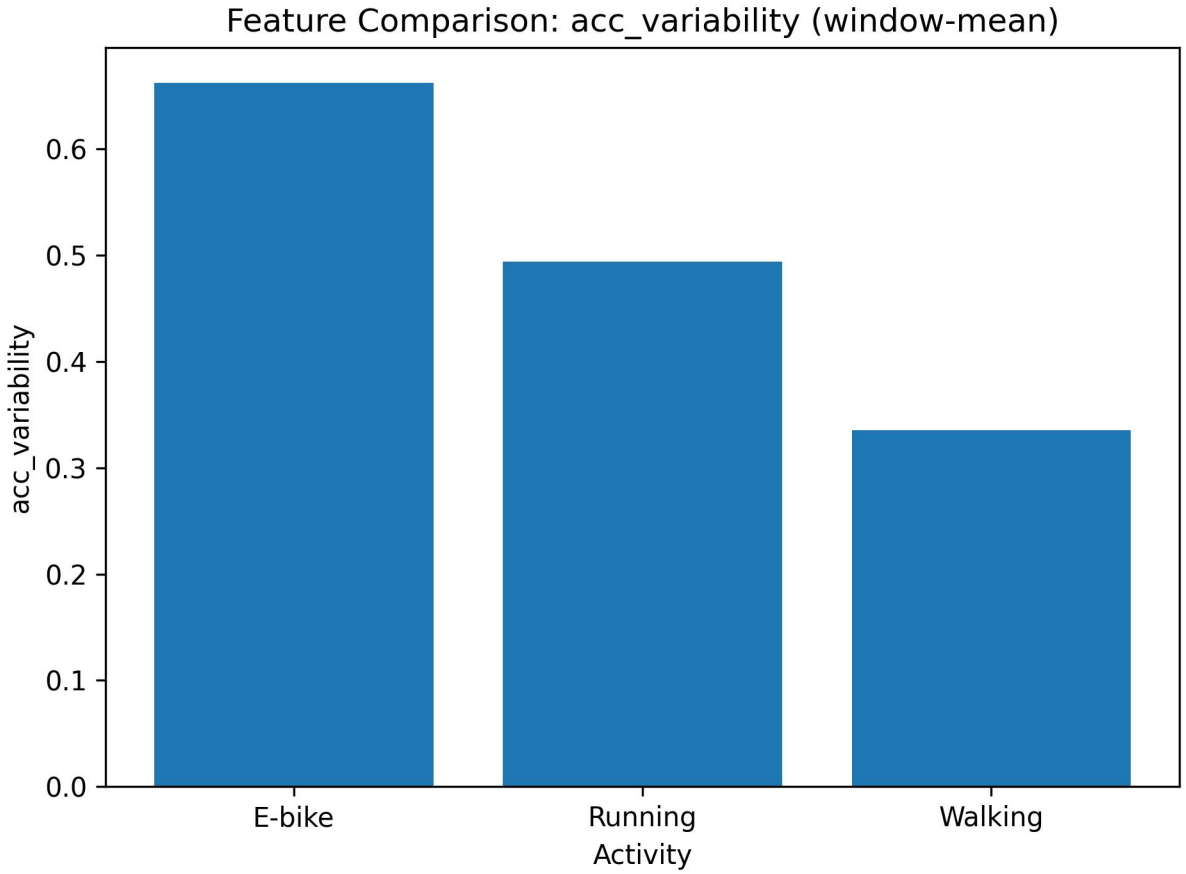

5. 加速度标准差（acc_std）

标准差代表加速度的波动程度：跑步的标准差（约 10）远高于步行（约 2）和电动车（约 1），说明跑步时加速度的 “不稳定程度” 最强；电动车的标准差最小，对应其运动状态更平稳。



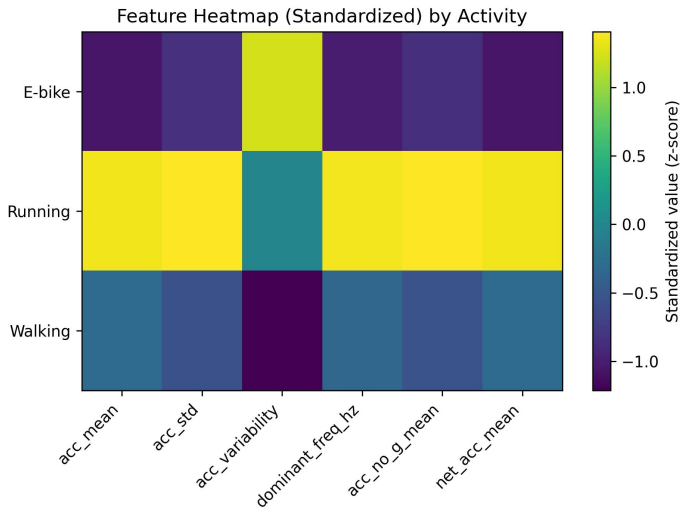
6. 加速度变异性（acc_variability）

变异性反映加速度信号的波动复杂性：电动车的变异性（约 0.65）最高，这是因为电动车运动虽整体平稳，但可能存在轻微的路面颠簸等 “无规律小波动”；跑步的变异性（约 0.5）次之，步行的变异性（约 0.33）最低，体现了步行步伐的周期性更稳定。



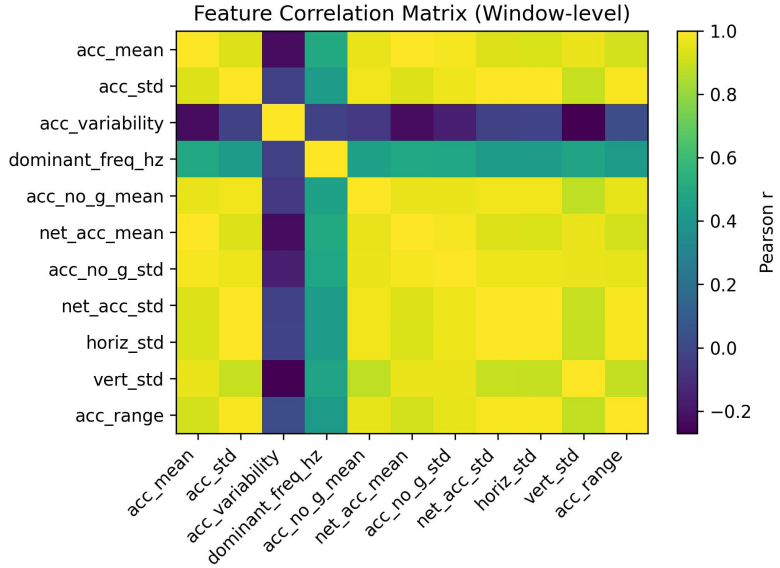

7. 特征热力图（标准化）

标准化热力图通过 Z-score 值量化各特征在不同运动状态下的相对强度：跑步的所有特征标准化值均处于高位（接近 1.0），整体呈现强特征信号；步行的各特征标准化值处于中等偏低区间（-0.5~-1.0），信号强度弱于跑步；电动车仅加速度变异性（acc_variability）的标准化值较高（约 1.0），其余特征均处于低位（接近-1.0），形成了独有的特征辨识度。



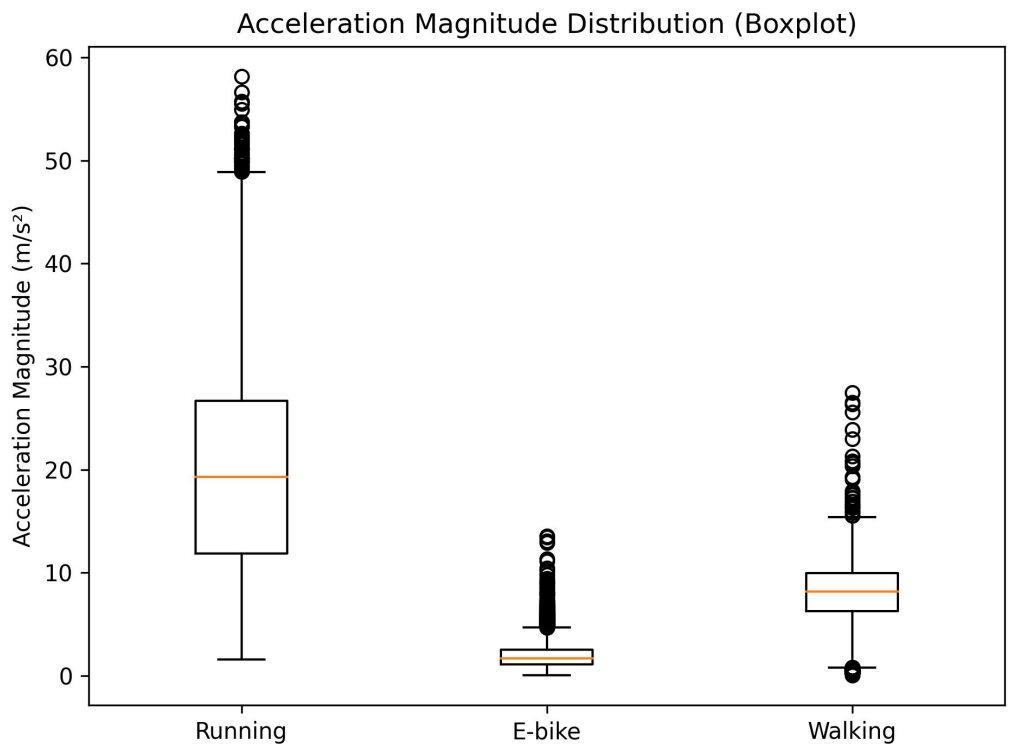

8. 特征相关矩阵

特征相关矩阵展示了各加速度特征间的线性相关程度：加速度均值（acc_mean）、去重力加速度均值（acc_no_g_mean）与净加速度均值（net_acc_mean）呈强正相关（相关系数接近 1），说明这三类特征本质上反映的是运动强度的相似维度；加速度变异性（acc_variability）与加速度均值（acc_mean）呈弱相关甚至负相关，是相对独立的特征维度；大部分特征间的正相关关系，说明这些特征可协同反映运动的“剧烈程度”。



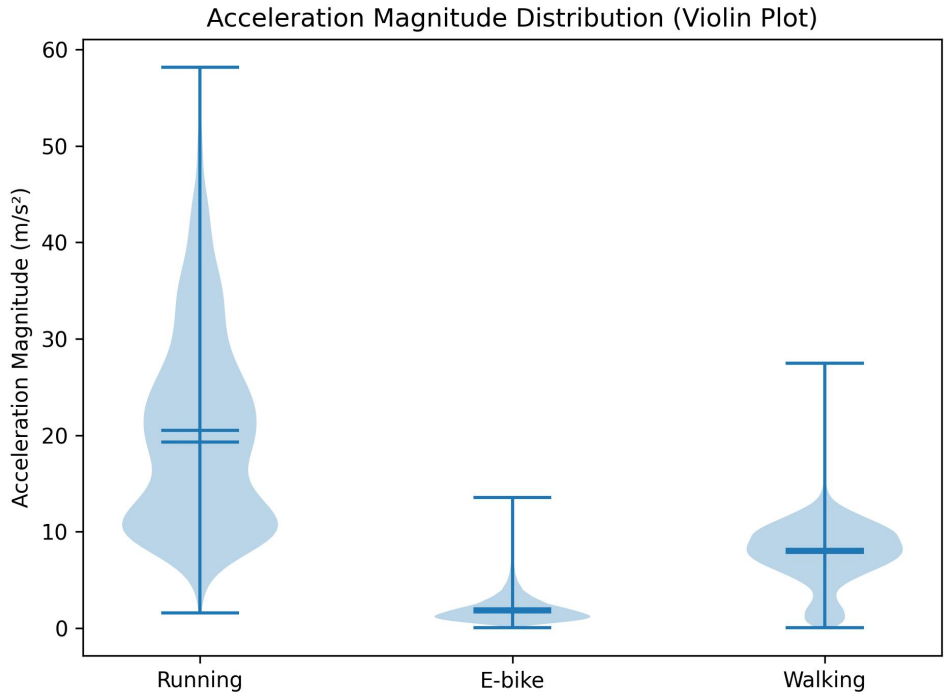

9. 加速度幅值箱线图

箱线图呈现了加速度幅值的统计分布范围：跑步的加速度幅值中位数约 20m/s²，箱体范围在 15~50m/s² 之间，且存在多个高值异常点，体现其幅值整体大、波动范围广；步行的中位数约 8m/s²，箱体范围在 5~15m/s² 之间，波动程度处于中间水平；电动车的中位数接近 0m/s²，箱体范围集中在 0~5m/s²，是三者中幅值最稳定的状态。




10. 加速度幅值小提琴图

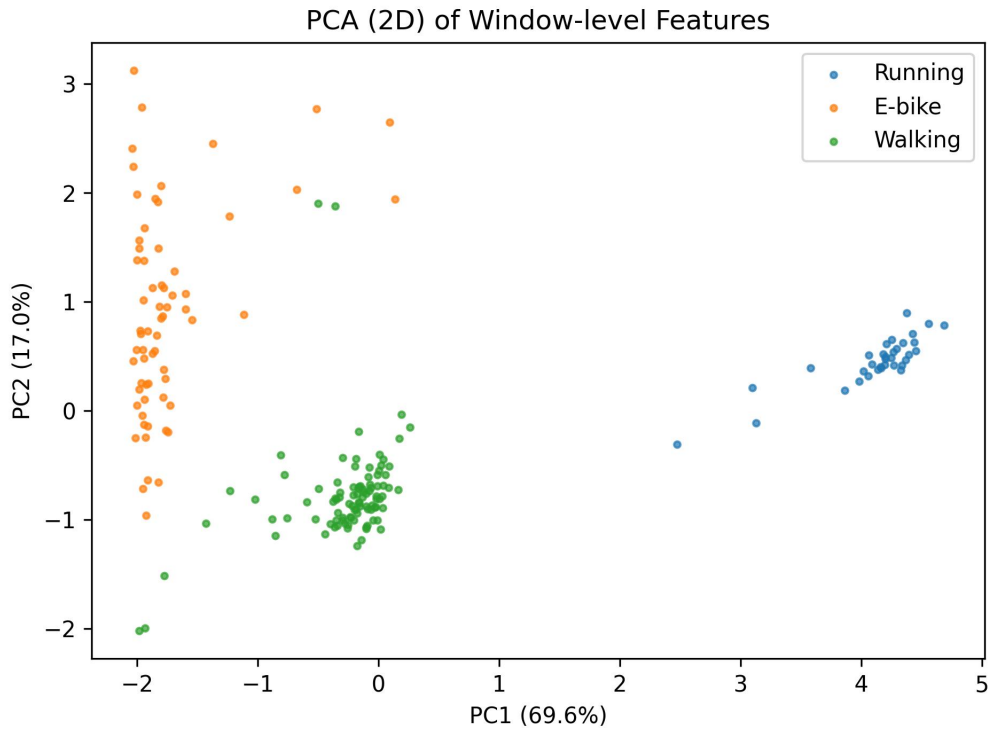
小提琴图结合概率密度分布展示加速度幅值的分布形态：跑步的幅值分布呈“双峰+长尾”形态，说明跑步时既有集中的中等幅值区间，也频繁出现高幅值波动；步行的幅值分布为单峰且集中在中等值区，波动规律更稳定；电动车的幅值分布为单峰且集中在低值区，体现其运动状态的平稳性。




11. PCA 降维散点图

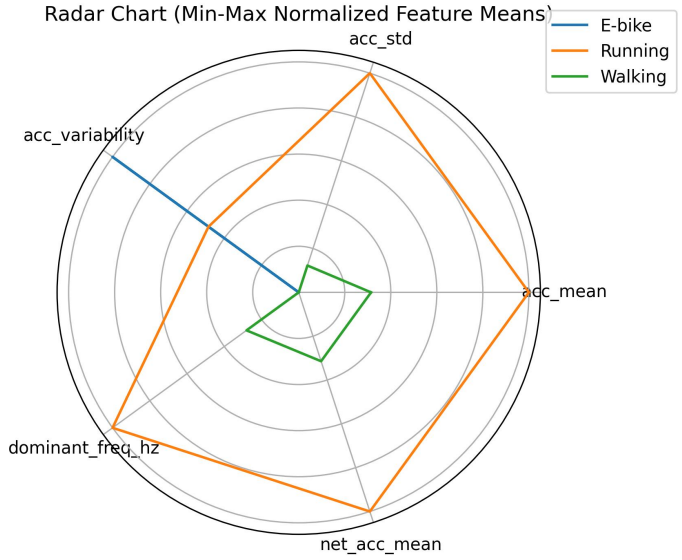
PCA 降维将高维特征压缩为两个主成分（PC1 解释 69.6%方差，PC2 解释 17.0%方差）：

跑步样本集中在 PC1 高值区（右侧），电动车样本集中在 PC2 高值区（左上方），步行样本集中在 PC2 低值区（左下方），三类样本形成明显的聚类效果，说明加速度特征具备极强的区分度，可有效划分不同运动状态。




12. 雷达图（归一化特征均值）

雷达图展示了三种运动状态的归一化特征均值轮廓：跑步的所有特征归一化值均接近最大值，轮廓最为饱满，体现其各维度特征均处于最高水平；电动车仅加速度变异性（acc_variability）的归一化值较高，其余特征均处于低位；步行的各特征归一化值处于中间区间，轮廓紧凑且均衡，进一步验证了三类运动状态的特征差异具有全局区分性。



结论分析与体会：

跑步、步行、电动车三种运动状态的加速度特征差异显著且可量化：跑步表现为高加速度均值（约 20m/s²）、高标准差（约 10）、高主频（约 2.7Hz），净加速度显著为正，幅值波动剧烈；步行各项特征均处于中间水平，波动周期性明显；电动车则以低加速度均值、低波动、低主频为特点，仅加速度变异性相对较高，整体运动状态最平稳。

所选加速度特征具备极强的区分能力与实用价值：运动强度类特征（均值、去重力均值等）、波动特性类特征（标准差、主频等）及变异性特征形成互补，高维特征无明显冗余，PCA 降维后三类样本聚类清晰，累计可解释 86.6% 的方差，为精准识别提供了坚实基础。

基于加速度数据的运动状态自动识别具备高可行性，核心特征组合可直接应用于实际场景。无论是健身监测、交通出行识别还是运动能耗估算，均可通过特征阈值判断或简单分类模型实现高效识别，相关量化标准也能为个性化运动方案制定提供数据支撑。