Project Report: WISDM-51 Sensor Data Analysis

By Joao Lucas Veras and Sania Bandekar

1. Dataset Overview

The WISDM-51 dataset contains a total of 4,804,403 records of triaxial accelerometer data from wearable sensors, with the columns:

- Subject-ID
- Activity Label
- Timestamp
- X-Accel
- Y-Accel
- Z-Accel

This dataset contains a total memory usage of 219.9 MB, but for us to utilize the entire dataset, we reduced the memory to 96.2 MB by altering the data types from 64 bits to 32 bits. Along with this, we also determined the average sampling rate of the dataset to be 1,118.6 Hz, which indicates there is a high-frequency data collection.

2. Methodology

Data Segmentation:

- Windowing: Divided data into 3-minute intervals or 180 seconds
    - Samples per window: 201,350
    - Total segments: 23 windows from 4,804,403 samples

Feature Extraction:

- Time-Domain Features
    1. Statistical Measures: Mean, Standard Deviation, Variance, Min, Max, Range, Median, Interquartile Range
    2. Signal Characteristics: RMS, Zero-Crossing Rate, Skewness, Kurtosis, Autocorrelation
    3. Peak Analysis: Peak Count, Peak Amplitude
    4. Energy Metrics: Signal Energy, Signal Magnitude Area (SMA)
- Frequency-Domain Features
    1. Utilized Fast Fourier Transform (FFT)
        a. Spectral Centroid, Spectral Entropy
        b. Spectral Energy, Dominant Frequency, Frequency Variance, Spectral Flatness
        c. Peak Frequency, Bandwidth,
        d. FFT Coefficients
        e. Power Spectral Density (PSD)

3. Experimental Results

Time-Domain Feature Summary:

X-Axis Results

Feature Type	**Mean**	**Standard Deviation**
X-Accel Mean	0.1416	.0081
X-Accel Std	5.8497	.0240
X-Accel Var	34.2199	.02812
X-Accel Min	-65.5213	5.3666
X-Accel Max	76.4085	3.1311
X-Accel Range	141.8298	6.5020
X-Accel Median	.0945	.0112
X-Accel IQR	8.3676	.0132
X-Accel RMS	5.8514	.0241
X-Accel Zero-Crossing Rate	98,322.96	201.40

Y-Axis Results

Feature Type	**Mean**	**Standard Deviation**
Y-Accel Mean	-1.6755	0.0527
Y-Accel Std	6.8767	0.0359
Y-Accel Var	47.3084	0.2544
Y-Accel Min	-77.5684	5.1785
Y-Accel Max	48.0645	3.1286
Y-Accel Range	125.6329	6.4345
Y-Accel Median	-1.6627	0.0117
Y-Accel IQR	10.3071	0.0129
Y-Accel RMS	6.8769	0.0358
Y-Accel Zero-Crossing Rate	91,532.61	203.29

Z-Axis Results

Feature Type	Mean	Standard Deviation
Z-Accel Mean	0.3598	0.0268
Z-Accel Std	5.1371	0.0078
Z-Accel Var	26.3939	0.1981
Z-Accel Min	-45.6654	5.1606
Z-Accel Max	56.2614	3.0836
Z-Accel Range	101.9268	6.3742
Z-Accel Median	0.3169	0.0115
Z-Accel IQR	6.9817	0.0156
Z-Accel RMS	5.1371	0.0078
Z-Accel Zero-Crossing Rate	98,893.83	194.65

Data Distribution Insights:

- Accelerometer Ranges:
  - X-Axis: 141.83
  - Y-Axis: 125.63
  - Z-Axis: 101.93
- Skewness: All axes showed near-zero skew
  - $|Skex| < 0.2$, which shows there was a symmetric distribution

4. Insights

- High-Frequency Data Challenges:
  - The 1,118.6 Hz sampling rate generated approximately 201,000 samples/windows, which required us to do efficient memory management.
    - Utilize data-type optimization to reduce memory by 56%
- Feature Robustness:
  - There were low autocorrelation values ranging from 0.02 to 0.03
  - This suggests that there was minimal temporal dependence between consecutive samples
  - Found consistent SMA values across segments, which imply stable motion intensity
- Limitations:
  - Class imbalance

        - We did not incorporate activity labels as they limited our insights into behavioral patterns
    - Lack of computer processing did not allow us to take the dataset as it was originally

5. Conclusion

Key Findings

1. Efficient Data Loading and Pre-processing

- The dataset comprising over 4.8 million records with six columns: Subject-id, Activity Label, Timestamp, and three accelerometer axes (X, Y, Z).
- Data types were optimized by converting columns to more memory-efficient formats (e.g., int32, float32, and category), reducing memory usage from approximately 220 MB to 96 MB.

2. High-Frequency Sensor Data

- The calculated average sampling rate was approximately 1118.6 Hz which is very high-frequency accelerometer recordings.

3. Comprehensive Feature Extraction

- Extensive time-domain features were extracted for each window and axis, including mean, standard deviation, variance, min, max, range, median, interquartile range, RMS, zero-crossing rate, skewness, kurtosis, autocorrelation, peak count, peak amplitude, energy, and signal magnitude area (SMA).
- The resulting feature set for each window comprised 49 features, facilitating detailed characterization of the sensor data.

Lessons Learnt

1. Data Type Optimization is Crucial

- Converting data types significantly reduced memory usage, which is essential when handling large-scale sensor datasets.

2. Windowing Strategy Impacts Feature Quality

- Segmenting data into fixed-length windows (e.g., 3 minutes) ensures consistent input for feature extraction and modelling, but the choice of window size should be informed by the nature of the activities and the application context.

3. Rich Feature Sets Enhance Modelling Potential

- Extracting a wide range of time-domain features enables more nuanced analysis and improves the potential for accurate activity recognition or anomaly detection.

Potential Improvements

1. Advanced Feature Extraction Using Deep Learning

- Incorporating deep learning models, particularly convolutional neural networks (CNNs), can automate and enhance feature extraction from raw sensor data. CNNs

can learn multi-scale features and maintain data independence, leading to stronger generalization and adaptability across different sensor modalities and tasks. Hybrid models combining CNNs with recurrent architectures (e.g., BiLSTM or GRU) further improve the capture of spatial and temporal dependencies in activity recognition.

2. Energy-Efficient and Distributed Query Processing

- Sensor nodes often have limited power and computational resources. Future systems should focus on intelligent in-network data reduction, such as local aggregation and synopsis computation, to minimize energy consumption and network traffic. Adopting database-inspired, declarative query approaches can further abstract and optimize data retrieval and aggregation across distributed sensor networks.

3. Leveraging Semi-Supervised and Attention-Based Learning

- Semi-supervised learning approaches can exploit both labelled and unlabelled data, improving model robustness and scalability in scenarios where labelled data is scarce. Additionally, attention-based models (such as transformers) are emerging as powerful tools for modelling long-range dependencies in sequential sensor data, outperforming traditional recurrent models in many human activity recognition (HAR) tasks.

Future Research Directions

1. Real-Time, On-Device Analytics

- Research should focus on developing lightweight, real-time analytics pipelines that can operate directly on edge devices, enabling immediate feedback and reducing reliance on cloud infrastructure.

2. Multi-Modal and Context-Aware Sensing

- Integrating data from diverse sensors (e.g., accelerometers, gyroscopes, magnetometers, WiFi signals) and contextual information can enhance recognition accuracy and enable new applications in healthcare, smart environments, and surveillance.

3. Privacy-Preserving Sensor Data Processing

- As sensor deployments become ubiquitous, privacy-preserving techniques (such as federated learning or differential privacy) will be critical to protect user data while still enabling large-scale analytics and model training.

4. Benchmarking and Dataset Expansion

- Continued development and public release of diverse, well-annotated datasets are essential for benchmarking new algorithms and ensuring progress in the field.

5. Adaptive and Transferable Models

- Future research should explore models that can adapt to new activities, sensor configurations, and unseen environments with minimal retraining, leveraging advances in transfer learning and domain adaptation.