About The Dataset

For our project titled "Comparative Study of Approaches for Injury Risk Prediction in Athletes", the MHEALTH (Mobile Health) dataset is particularly well-suited due to the following reasons:

1. Rich Multisensor Time-Series Data

The dataset includes high-frequency sensor data collected from multiple body locations (chest, wrist, and ankle). This is crucial for our project because athletic injuries often relate to biomechanical movement patterns and physical load, which are best captured through detailed sensor readings.

2. Activity-Based Labeling

The dataset contains clearly labeled physical activities, such as walking, running, cycling, jumping, and crouching. These activities are directly relevant to sports training and athletic performance. By analyzing patterns during these activities, we can infer potential injury risk scenarios.

3. Sequential Nature of Data

Since the dataset is collected as continuous time-series data, it enables the use of sequential deep learning models (like LSTMs or GRUs) that are ideal for learning temporal dependencies—exactly what is needed for injury risk prediction based on evolving movement patterns.

4. Realistic Simulation of Wearable Data

The data was collected using wearable sensors at a sampling rate of 50Hz, which reflects realistic setups in sports and health monitoring. This aligns well with our goal to build a model that can be practically integrated into wearable systems for athletes.

5. Open-Access and Well-Documented

The dataset is part of the UCI Machine Learning Repository, making it accessible, reliable, and already widely cited in academic research. This ensures reproducibility and allows benchmarking against related work.

What the MHEALTH Dataset Contains

- Subjects: 10 volunteers (8 male, 2 female), aged 20–35, performed a series of physical activities while wearing sensors.
- Sensor Modalities:
  - Accelerometer, gyroscope, and magnetometer at the ankle and wrist
  - Accelerometer and ECG at the chest
- Sampling Frequency: 50 Hz (i.e., 50 samples per second)
- Total Columns: 24 features + 1 activity label + 1 subject ID

Activity Labels

There are 12 defined activity labels:

1. Standing still
2. Sitting and relaxing
3. Lying down
4. Walking
5. Climbing stairs
6. Waist bends forward
7. Frontal elevation of arms
8. Knees bending (crouching)
9. Cycling
10. Jogging
11. Running
12. Jumping front and back

Additionally, the dataset includes a significant number of rows labeled as `0`, which likely represent idle periods or unlabeled rest phases.

Relevance to Our Project

- By analyzing sequential patterns in the sensor data during these activities, we aim to identify pre-injury risk indicators such as:

  - Abrupt changes in gait or limb coordination

  - Sudden spikes in physical exertion

  - Asymmetrical movement patterns

- We will compare sequential models (e.g., LSTM) with traditional models (e.g., Random Forest) on this data to assess their relative performance in predicting such risk patterns.