Why We Chose the MHEALTH 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.

**NEXT STEPS:**

1. Validation of Proxy Injury Risk Labels

Given the absence of true injury annotations in the MHEALTH dataset, we developed a proxy labeling strategy based on scientifically supported biomechanical and physiological indicators of injury risk. These labels simulate real-world risk scenarios and allow us to frame our project as a supervised learning task.

Proxy Injury Risk Indicators Used:

| **Proxy Indicator** | **Rationale** |
| --- | --- |
| High Impact Acceleration | Detected using sudden spikes in total chest acceleration (>3.5g); may reflect unsafe landings, falls, or jerky movement |
| Fatigue Signals | Based on unusually high heart rate (ECG) during low-intensity activities such as walking; suggests poor recovery or cardiovascular strain |
| Repetitive Stress | Extended duration of high-load activities like jogging or jumping with limited rest; simulates overuse injury risk |
| (Optional) Postural Instability | Intended to capture unstable body movement during transitions using gyroscope data; often a precursor to ligament injuries |

These proxy labels serve as a foundation for model training. As part of the next step, we will:

• Refine threshold values using visual and statistical validation.

• Ensure balanced class distribution by addressing overrepresented “no risk” segments.

• Potentially expand labeling logic by incorporating additional features (e.g., gyroscopic change rates or symmetry analysis).

2. Segmentation into Time-Series Windows

To support training of deep learning models (e.g., LSTMs), the time-series data will be segmented into overlapping windows:

• Each window will contain a fixed number of time steps (e.g., 100, corresponding to 2 seconds at 50Hz sampling rate).

• A label will be assigned to each window based on the dominant or terminal proxy risk label.

• This format—(samples, timesteps, features)—preserves temporal patterns and is essential for sequential modeling.

3. Model Development and Training

We will implement and compare both traditional and sequential models.

A. Traditional Models (Baselines)

• Random Forest, Logistic Regression, and SVM will be trained on aggregated or flattened features.

• These will serve as baseline models for performance benchmarking.

B. Sequential Deep Learning Models

• LSTM (Long Short-Term Memory) networks will be trained on the windowed time-series data.

• (Optional) 1D Convolutional Neural Networks (CNNs) may also be explored to detect short-term local motion patterns.

• These models are expected to better capture injury precursors by analyzing sequences over time.

4. Evaluation and Comparison

All models will be evaluated using standard metrics:

• Accuracy, Precision, Recall, F1-score

• Confusion Matrix for class-wise performance

• Early Warning Capability (for sequential models) to assess how early high-risk movement can be detected

We will perform a comparative analysis to determine the effectiveness of sequential modeling versus traditional approaches for injury risk detection.

5. Visualization and Interpretability

We plan to include:

• Plots of sensor trends during high-risk vs no-risk windows

• Performance comparison charts across all models

• Interpretability tools such as feature importance (Random Forest) or attention heatmaps (for LSTM, if implemented)

6. Documentation and Final Reporting

The final phase will involve:

• Detailed documentation of methods, model design choices, and findings

• Integration of results, visuals, and analysis into the final report

• Preparation of a slide deck for presentation, emphasizing practical implications and recommendations for real-world injury risk monitoring