TENG Gait Sensor Dataset for Machine Learning Classification

📋 Quick Summary

Item	Details
Dataset Name	TENG Gait Sensor Data (Oscilloscope + ADC)
Total Subjects	12 (S01-S08: new data, S09-S12: previous healthy baseline)
Activities	Stand, Walk, Run, Jump (84 recordings total)
Sensors	1x TENG sensor (copper/PTFE) in shoe insole
Data Sources	Oscilloscope (primary), ESP32 ADC (validation)
Labels	QOM (Quality of Movement): 1-10 scale + Demographics
Format	CSV (time-series voltage/current signals)
Collection Date	October 2025
Purpose	Multiclass gait classification + QOM regression + Bio-authentication

---

🎯 Research Goals

Primary Objectives:

1. **Activity Classification**: Detect gait states (stand/walk/run/jump) using single TENG sensor
2. **QOM Assessment**: Predict movement quality (1-10 scale) from gait patterns
3. **Bio-authentication**: Predict user weight from stride patterns (novel feature)

Expected Outcomes:

- Activity classification accuracy: **95-98%** (RF/XGBoost)
- QOM regression MAE: **<1.5/10** (XGBoost)
- Weight prediction MAE: **<8 kg** (RF)
- Feature importance analysis showing demographics matter

📁 Dataset Structure

```
User_Gait_Data_Master/
│
├── User_Data_Labelled/          ← PRIMARY DATA (Oscilloscope)
│   ├── S01STAND.csv             ← Time-series: Time(s), Voltage(V)
│   ├── S01WALK.csv
│   ├── S01WALK01.csv
│   ├── S01RUN.csv
```

```
│   ├── S01RUN01.csv
│   ├── S01JUMP.csv
│   ├── S01JUMP01.csv
│   ├── S02STAND.csv
│   └── ... (84 files total)
│
├── ADC_User_Data_Labelled_ref/     ← VALIDATION DATA (ESP32 ADC)
│   ├── S01STAND_ADC.csv            ← Time-series: Time(s), ADC_Value(0-4095)
│   ├── S01WALK_ADC.csv
│   └── ... (84 files total)
│
├── QOM/                        ← LABELS & METADATA
│   └── gait_labels_qom.csv         ← Master labels file
│
└── README.md                   ← This file
```

📊 Data Description

1. Primary Data: Oscilloscope Recordings (`User_Data_Labelled/`)

File naming convention:

```
S[01-12][ACTIVITY][TRIAL].csv

Examples:
- S01STAND.csv     → Subject 1, Standing, Trial 1
- S01WALK.csv      → Subject 1, Walking, Trial 1
- S01WALK01.csv    → Subject 1, Walking, Trial 2
- S01RUN.csv       → Subject 1, Running, Trial 1
- S01JUMP01.csv    → Subject 1, Jumping, Trial 2
```

CSV Format:

```csv
Time(s), Voltage(V)
0.000, 0.05
0.010, 0.12
0.020, 0.45
0.030, 0.68
...
```

Signal Characteristics:

- **Sampling rate**: ~100-200 Hz
- **Duration**: ~30 seconds per recording

- **Amplitude range**: 0-80V (open-circuit voltage from TENG)
- **Frequency content**: 0.5-5 Hz (gait frequencies)
- **Noise level**: Moderate (50 Hz powerline + contact noise)

Data Quality Notes:

- Stand: Low amplitude, near-zero signal (minimal foot movement)
- Walk: Periodic peaks (~1-2 Hz stride frequency)
- Run: Higher amplitude, faster frequency (~2-3 Hz)
- Jump: High amplitude spikes, variable repetition rate

---

2. Validation Data: ESP32 ADC Recordings (`ADC_User_Data_Labelled_ref/`)

Purpose: Verify oscilloscope findings with portable ADC system

File naming: Same as primary data with `_ADC` suffix

CSV Format:

```csv
Time(s), ADC_Value
0.000, 512
0.010, 1024
0.020, 2048
...
```

ADC Characteristics:

- **Resolution**: 12-bit (0-4095)
- **Reference voltage**: 3.3V
- **Conversion**: ADC_Value = (Voltage / 3.3V) × 4095
- **Sampling rate**: ~100 Hz

Use Case:

- Compare with oscilloscope for signal fidelity
- Demonstrate portability (thesis Section 4.2)
- NOT primary data for ML training

---

3. Labels & Metadata (`QOM/gait_labels_qom.csv`)

Master labels file with all subject information and ratings.

CSV Format:

```
csv

subject_id,height_cm,weight_kg,gender,activity,trial,qom,file_path
S01,175,70,M,stand,1,8,S01STAND.csv
S01,175,70,M,walk,1,7,S01WALK.csv
S01,175,70,M,walk,2,7,S01WALK01.csv
...
```

Column Descriptions:

Column	Type	Description	Example
subject_id	String	Subject identifier	S01, S02, ... S12
height_cm	Integer	Height in centimeters	175
weight_kg	Integer	Weight in kilograms	70
gender	String	Gender (M/F)	M
activity	String	Activity type	stand, walk, run, jump
trial	Integer	Trial number (1 or 2)	1
qom	Integer	Quality of Movement (1-10 scale)	7
file_path	String	Corresponding data file	S01STAND.csv

QOM Rating Scale:

```
1-3:  POOR          - Obvious gait abnormalities
4-6:  NEEDS ATTENTION - Moderate quality, monitor
7-10: GOOD          - Normal/excellent gait patterns
```

Subject Groups:

- **S01-S08**: Newly collected data (varied QOM ratings)
- **S09-S12**: Previous baseline data (healthy subjects, QOM=8)

🔬 Data Collection Protocol

Hardware Setup:

1. **TENG Sensor**: Copper/PTFE triboelectric layers (6 layers stacked)
2. **Placement**: Front of shoe insole (captures heel strike + push-off)

3. **Signal conditioning**: 10MΩ series resistor, Butterworth low-pass filter
4. **Acquisition**:
    - Primary: Rigol RTB2004 oscilloscope (high precision)
    - Validation: ESP32 12-bit ADC (portable system)

Data Collection:

- **Environment**: Indoor laboratory, flat surface
- **Instructions**: "Perform [activity] naturally for ~30 seconds"
- **Activities**:
    - Stand: Static standing, minimal movement
    - Walk: Normal walking pace (~1-2 steps/sec)
    - Run: Jogging pace (~2-3 steps/sec)
    - Jump: Vertical jumps, 10 repetitions × 2 trials
- **Repetitions**: 1-2 trials per activity per subject
- **Total duration**: ~7 min per subject (including setup)

QOM Rating Methodology:

- **Rater**: Researcher with sports background (athlete-level assessment)
- **Criteria**: Visual observation of:
    - Stride symmetry (left vs right)
    - Balance and stability
    - Movement smoothness
    - Landing technique (for jump)
    - Gait pattern consistency
- **Confidence**: High for stand/walk, medium for run/jump
- **Validation**: To be cross-checked by physiotherapist (ground truth)

🤖 Machine Learning Pipeline

Recommended Workflow:

Step 1: Feature Extraction

Extract from each CSV file:

Time-Domain Features:

- Peak amplitude (max voltage)
- Peak-to-peak range
- RMS (root mean square)
- Signal energy
- Zero-crossing rate
- Stride duration (peak-to-peak time)
- Number of peaks (stride count)

Frequency-Domain Features:

- Dominant frequency (FFT)
- Power spectral density
- Spectral entropy
- Frequency bandwidth

Statistical Features:

- Mean, median, std
- Skewness, kurtosis
- 25th/75th percentiles

Demographic Features:

- Height (cm)
- Weight (kg)
- BMI = weight / (height/100)²
- Gender (encoded as 0/1)

Step 2: Model Training

Task A: Activity Classification (4 classes)

```
python
```

```
# Target: activity (stand/walk/run/jump)
# Models: Random Forest + XGBoost
# Expected accuracy: 95-98%

from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier

X = features  # Extracted features
y = labels['activity']

rf_model = RandomForestClassifier(n_estimators=200)
xgb_model = XGBClassifier(n_estimators=200)

# Train, evaluate, compare
```

Task B: QOM Regression (1-10 scale)

```
python

# Target: qom (movement quality)
# Model: XGBoost Regressor
# Expected MAE: <1.5/10

from xgboost import XGBRegressor

X = features
y = labels['qom']

qom_model = XGBRegressor(n_estimators=100, max_depth=4)
```

Task C: Weight Prediction (Bio-auth)

```
python

# Target: weight_kg (for user identification)
# Model: Random Forest Regressor
# Expected MAE: <8 kg

from sklearn.ensemble import RandomForestRegressor

X = features  # Gait features only
y = labels['weight_kg']

weight_model = RandomForestRegressor(n_estimators=200)
```

Step 3: Evaluation

- Activity: Confusion matrix, F1-score per class
- QOM: MAE, R^2, prediction scatter plot
- Weight: MAE, feature importance (stride duration matters most)

📈 Expected Results (Reference)

Based on TENG gait paper (Li et al., 2023, Nano Energy):

Task	Metric	Expected	Notes
Activity Classification	Accuracy	95-98%	Stand easiest, run/jump harder
QOM Regression	MAE	0.8-1.5/10	Demographics improve accuracy
Weight Prediction	MAE	5-10 kg	Novel bio-auth feature

Feature Importance (Predicted):

1. Stride duration: ~18% (temporal)
2. Peak amplitude: ~14% (intensity)
3. Weight: ~12% (demographics)
4. Height/BMI: ~10% (demographics)
5. Dominant frequency: ~9% (spectral)

⚠️ Known Limitations

1. **Sample size**: 12 subjects (small, need 50+ for clinical validity)
2. **Demographics**: Mostly healthy young adults (limited variability)
3. **Single sensor**: Paper uses 8 sensors (more spatial info)
4. **QOM rating**: Single rater (needs physiotherapist validation)
5. **Environment**: Controlled lab (real-world deployment untested)
6. **Activities**: Only 4 classes (missing stairs, sit-to-stand, etc.)

Mitigation strategies (discuss in thesis):

- Cross-validation for robust evaluation
- Compare with TENG paper (8-sensor system)
- Acknowledge as pilot study, future: clinical population
- Plan physiotherapist review for ground truth

🚀 Quick Start

1. Load Data

python

```python
import pandas as pd
import glob

# Load all oscilloscope CSVs
data_files = glob.glob('User_Data_Labelled/*.csv')
signals = {f: pd.read_csv(f) for f in data_files}

# Load labels
labels = pd.read_csv('QOM/gait_labels_qom.csv')
```

2. Extract Features

python

```python
def extract_features(signal_df):
    """Extract features from voltage-time signal"""
    voltage = signal_df['Voltage(V)'].values
    time = signal_df['Time(s)'].values

    features = {
        'peak_amplitude': voltage.max(),
        'rms': np.sqrt(np.mean(voltage**2)),
        'energy': np.sum(voltage**2),
        # ... add more features
    }
    return features

# Extract for all files
feature_df = pd.DataFrame([extract_features(signals[f]) for f in data_files])
```

3. Train Model

python

```
from sklearn.ensemble import RandomForestClassifier

X = feature_df
y = labels['activity']

model = RandomForestClassifier(n_estimators=200)
model.fit(X_train, y_train)

accuracy = model.score(X_test, y_test)
print(f"Accuracy: {accuracy*100:.1f}%")
```

📊 Data Visualization

See `plot_sample_signals.png` for example voltage-time plots showing signal characteristics for each activity.

Key observations:

- **Stand**: Near-zero signal, slight drift
- **Walk**: Periodic peaks at ~1-2 Hz (stride frequency)
- **Run**: Higher amplitude, faster frequency (~2-3 Hz)
- **Jump**: Sharp high-amplitude spikes, irregular timing

📖 References

1. **TENG Gait Paper**: Li et al. (2023). "A triboelectric gait sensor system for human activity recognition and user identification." *Nano Energy*, 112, 108473.
   - Their system: 8 sensors, 97.9% activity accuracy, 99.4% user ID accuracy
   - Our system: 1 sensor (simpler, lower cost)
2. **Similar Work**:
   - Lin et al. (2019): TENG insole for gait monitoring
   - Zhang et al. (2020): TENG smart socks with deep learning

---

📧 Contact & Support

Dataset Creator: [Your Name] **Institution**: M.Sc ESE, University of Freiburg **Date**: October 2025

Questions?

- Dataset issues: [Your Email]
- ML pipeline: See `gait_analysis.py` in package

- Feature extraction: See example code above

✅ Data Quality Checklist

Before using this dataset, verify:

- [ ] All 84 oscilloscope CSV files present in `User_Data_Labelled/`
- [ ] All 84 ADC CSV files present in `ADC_User_Data_Labelled_ref/`
- [ ] `gait_labels_qom.csv` has 84 rows (excluding header)
- [ ] No missing values in height/weight columns (except S02-S08 if not filled)
- [ ] QOM ratings present for all recordings
- [ ] CSV files load without errors
- [ ] Signal amplitude reasonable (0-80V for oscilloscope, 0-4095 for ADC)

Data Validation Script: See `validate_dataset.py` in package

🎓 Thesis Integration

How to use this dataset in your thesis:

Chapter 3 (Methodology):

- Section 3.1: TENG sensor design (cite Li et al. paper)
- Section 3.2: Data collection protocol (use this README)
- Section 3.3: Signal processing pipeline
- Section 3.4: QOM rating methodology

Chapter 4 (Results):

- Section 4.1: Activity classification (RF vs XGBoost comparison)
- Section 4.2: QOM regression (MAE, R^2, feature importance)
- Section 4.3: Weight prediction (novel bio-auth feature)
- Section 4.4: Confusion matrices, error analysis

Chapter 5 (Discussion):

- Compare 1-sensor vs 8-sensor system (Li et al.)
- Discuss demographic features importance
- Limitations: sample size, single rater, healthy subjects
- Future work: clinical validation, more sensors, real-world deployment

---

Version: 1.0 (October 26, 2025)

Last Updated: October 26, 2025

📦 What's in the ZIP?

```
User_Gait_Data_Master.zip
├── User_Data_Labelled/       (84 oscilloscope CSVs)
├── ADC_User_Data_Labelled_ref/ (84 ADC CSVs)
├── QOM/
│   └── gait_labels_qom.csv
├── README.md                 (this file)
├── plot_sample_signals.png   (visualization)
├── validate_dataset.py       (data check script)
└── gait_analysis.py          (ML pipeline starter)
```

Ready to use! Unzip, validate, and start training your models. 🚀