Feature Engineering Summary – Finger Tapping Dataset

Purpose (Why feature engineering was necessary)

The raw dataset consisted of **1800-sample amplitude–time signals** per hand recording. Machine learning models cannot learn effectively from raw time-series at this scale while remaining interpretable.

Feature engineering was therefore required to:

- compress raw signals into **meaningful summaries**
- preserve **clinically relevant motor patterns**
- produce a dataset that is **robust, interpretable, and suitable for classification**

Overall Approach (What was done)

Each hand recording was converted into a **single feature vector** describing:

- tapping speed
- rhythm regularity
- amplitude stability
- motor fatigue
- hesitation behaviour

Left and right hands were treated as **independent samples**, as motor impairment can be asymmetric.

Step-by-Step Feature Engineering Process

1. Data loading & validation

- Loaded left and right hand datasets separately.
- Converted amplitude and time columns from strings to lists.
- Verified:
    - all recordings contained **exactly 1800 samples**
    - no missing or truncated signals

- Result: dataset was structurally consistent.

2. Label verification (issue identified & fixed)

Problem

- The UPDRS column was initially assumed to contain 0–4 severity scores.
- Exploratory checks showed only values **0 and 1**, collapsing the dataset into a single class when thresholded.

Identification

- Value-count checks showed no UPDRS values above 1.
- Cross-checking with the original rating sheet revealed:
    - raw rater scores (0–4) had already been averaged and **binarised upstream**.

Fix

- Removed all UPDRS thresholding logic.
- Used updrs_left / updrs_right directly as **binary motor-impairment labels**.
- Retained ID-based labels only as a sanity check.

3. Signal preprocessing

- Applied light Savitzky–Golay smoothing.
- Reduced noise while preserving:
    - hesitations
    - rhythm variability
    - fatigue-related amplitude changes

4. Peak detection (tap segmentation)

- Used **adaptive peak detection** rather than fixed thresholds.
- Peak prominence scaled to each signal's amplitude range.

- Minimum inter-tap distance enforced physiological plausibility.

This ensured consistent tap detection across different impairment levels.

5. Padding issue discovered & resolved

Problem

- Some time vectors contained repeated values due to signal padding.
- This caused errors when estimating timing from raw time differences.

Resolution

- Signals were **not discarded**.
- Inter-tap intervals were computed using:
    - peak index differences
    - known fixed sampling duration (30s / 1800 samples)
- This avoided padding artefacts while preserving correct timing behaviour.

6. Feature extraction

For each hand recording, the following features were computed:

- **Speed**: number of taps
- **Amplitude**: mean peak amplitude, amplitude variability
- **Fatigue**: amplitude decrement (early vs late taps)
- **Rhythm**: mean inter-tap interval, variability, coefficient of variation
- **Hesitation**: count and proportion of prolonged pauses

Each recording was represented by **one feature vector (9 features)**.

7. Feature table construction

- Built separate feature tables for left and right hands.
- Each table includes:

    - extracted features
    - patient ID
    - hand indicator
    - binary target label
- Final shape: **66 samples × 13 columns per hand**.

8. Validation

- Compared feature distributions between classes using boxplots.
- Observed expected trends:
    - reduced tap count in impaired group
    - higher amplitude decrement
    - increased rhythm irregularity
    - lower mean peak amplitude

This confirmed the features are **informative and clinically aligned**.

Final Outcome

- Feature engineering completed successfully.
- Labels verified and correctly interpreted.
- All encountered issues were identified, diagnosed, and resolved.
- Final datasets are:
    - interpretable
    - robust to artefacts
    - suitable for machine learning.

What's Next (How this feeds into modelling)

The engineered feature tables can now be used directly for modelling:

- Combine left and right feature tables or model them separately.
- Apply feature scaling (e.g. standardisation).
- Train baseline classifiers (e.g. logistic regression, random forest).
- Evaluate performance using cross-validation and class-balanced metrics.

No further preprocessing or feature engineering is required before modelling.

Feature Set:

Feature Name	**What it Measures**	**Why This Feature Was Chosen**	**How It's Used in Modelling**
num_taps	Total number of detected taps in the recording	Tapping speed is a core indicator of motor impairment; PD patients typically tap more slowly	Strong global discriminator between impaired and non-impaired samples
mean_peak_amp	Average amplitude of detected taps	Reduced amplitude reflects bradykinesia and reduced motor output	Separates strong, confident movement from weak execution
std_peak_amp	Variability of tap amplitudes	Motor instability leads to inconsistent tap strength	Captures irregular motor control beyond average amplitude
amp_decrement	Difference between early and late tap amplitudes (fatigue effect)	Explicitly assessed in UPDRS finger-tapping tasks	Key indicator of motor fatigue and progressive impairment
mean_iti	Mean inter-tap interval (average time between taps)	Slower movements result in longer intervals	Complements tap count by modelling timing directly
std_iti	Variability of inter-tap intervals	PD often causes irregular rhythm rather than just slow movement	Identifies unstable timing patterns

Feature Name	What it Measures	Why This Feature Was Chosen	How It's Used in Modelling
cv_iti	Coefficient of variation of inter-tap intervals	Scale-independent measure of rhythm irregularity	Robust discriminator across different baseline speeds
num_long_pauses	Count of unusually long pauses between taps	Captures hesitations and motor interruptions	Detects discrete breakdowns in motor execution
prop_long_pauses	Proportion of long pauses relative to total taps	Normalises hesitation frequency across participants	Scale-invariant hesitation feature