*This document serves as a high-level overview of dataset generation. For detailed derivations, coefficient tables,
and empirical parameter justification, see* `ATHLETE_READINESS_TRANSFORMER_COMPREHENSIVE.pdf`

SYNTHETIC DATASET: ATHLETE READINESS PROJECT

OVERVIEW

This file contains the information required to reproduce the 17 features and 3 corresponding labels for all 100 athletes over 150 days. Reproduction of data can be executed using a commercial LLM (e.g. ChatGPT). Below are the formulas used to calculate next-day data.

Formulas are based on physiologically realistic relationships between physiology, sleep, training, and recovery. Each feature follows within- and between-athlete variability inspired by empirical datasets (PPG-DaLiA, Sleep-EDF).

- sinDOWt and cosDOWt are markers to enable weekly patterns to emerge by associating a combination of their values to a particular day.
- µ contributes to individuality in a particular metric (interindividual variation).
- Lag (hysteresis) and noise are also included.

FEATURES

HRV and RHR (physiology)

HRV dynamics (z-scored)

$$HRV_{z,t} = \rho_h\,HRV_{z,t-1}\;-\;a_1\,Load_{z,t-1}\;-\;a_2\,ACWR_t\;-\;a_3\,RHR_{z,t-1}$$
$$+\;b_1\,SleepEff_{z,t-1}\;+\;b_2\,SleepDur_{z,t-1}$$
$$-\;a_4\,Load_{z,t-2}\;-\;a_5\,ACWR_{t-1}$$
$$+\;e_1\,(Load_{z,t-1}\times SleepEff_{z,t-1})\;+\;\epsilon_t$$

RHR dynamics (z-scored)

$$RHR_{z,t} = \rho_r\,RHR_{z,t-1}\;+\;c_1\,Load_{z,t-1}\;+\;c_2\,ACWR_t\;+\;c_3\,(-HRV_{z,t-1})$$
$$-\;d_1\,SleepEff_{z,t-1}\;-\;d_2\,SleepDur_{z,t-1}$$
$$+\;c_4\,Load_{z,t-2}\;+\;c_5\,ACWR_{t-1}$$
$$+\;e_2\,(Load_{z,t-1}\times SleepEff_{z,t-1})\;+\;\eta_t$$

Sleep Duration and Sleep Efficiency (sleep)

Sleep duration

$$Dur_{z,t} = \rho_d\,Dur_{z,t-1}\;+\;\gamma_1\,Def_{t-1}\;+\;\gamma_2\,Load_{z,t-1}\;-\;\gamma_3\,ACWR_t$$
$$+\;\kappa_{d1}\sin DOW_t\;+\;\kappa_{d2}\cos DOW_t$$
$$+\;\gamma_4\,(Load_{z,t-1}\times Def_{t-1})\;+\;\epsilon_t^{(d)}$$

Sleep efficiency

$$Eff_{z,t} = \rho_e\,Eff_{z,t-1}\;-\;\beta_1\,Load_{z,t-1}\;-\;\beta_2\,ACWR_t\;-\;\beta_3\,Fatigue_{z,t-1}$$
$$+\;\beta_4\,Def_{t-1}\;+\;\kappa_{e1}\sin DOW_t\;+\;\kappa_{e2}\cos DOW_t$$
$$+\;\beta_5\,(Load_{z,t-1}\times Eff_{z,t-1})\;+\;\epsilon_t^{(e)}$$

Note: Def_{t-1} is an individual's sleep deficit determined by yesterday's sleep. Calculated using following formula:

$$Def_{t-1} = \max\left(0,\;\frac{TargetDur_i - Dur_{raw,t-1}}{SD_{Dur,i}}\right)$$

Rating of Perceived Exertion (training)

$$
\begin{aligned}
RPE_{z,t}^{(i)} = {}& \rho_p^{(i)} \, RPE_{z,t-1}^{(i)} \;+\; \phi_1^{(i)} \, Load_{z,t-1}^{(i)} \;+\; \phi_2^{(i)} \, ACWR_{t-1}^{(i)} \;+\; \phi_3^{(i)} \, Fatigue_{z,t-1}^{(i)} \\
& -\; \phi_4^{(i)} \, Readiness_{t-1}^{(i)} \;-\; \phi_5^{(i)} \, SleepEff_{z,t-1}^{(i)} \;-\; \phi_6^{(i)} \, SleepDur_{z,t-1}^{(i)} \\
& +\; \lambda_1^{(i)} \, \sin DOW_t \;+\; \lambda_2^{(i)} \, \cos DOW_t \\
& +\; \phi_7^{(i)}\big(Fatigue_{z,t-1}^{(i)} \times (1 - Readiness_{t-1}^{(i)})\big) \;+\; \mu_{RPE}^{(i)} \;+\; \epsilon_t^{(i,p)}
\end{aligned}
$$

Fatigue and Soreness (recovery)

$$
\begin{aligned}
Fatigue_{z,t}^{(i)} = {}& \rho_f^{(i)} \, Fatigue_{z,t-1}^{(i)} \;+\; \psi_1^{(i)} \, Load_{z,t-1}^{(i)} \;+\; \psi_2^{(i)} \, ACWR_{t-1}^{(i)} \;-\; \psi_3^{(i)} \, SleepDur_{z,t-1}^{(i)} \\
& -\; \psi_4^{(i)} \, SleepEff_{z,t-1}^{(i)} \;-\; \psi_5^{(i)} \, Readiness_{t-1}^{(i)} \;+\; \psi_6^{(i)} \, RHR_{z,t-1}^{(i)} \;-\; \psi_7^{(i)} \, HRV_{z,t-1}^{(i)} \\
& +\; \psi_8^{(i)} \, \sin DOW_t \;+\; \psi_9^{(i)} \, \cos DOW_t \\
& +\; \psi_{10}^{(i)}\big(Load_{z,t-1}^{(i)} \times (1 - Readiness_{t-1}^{(i)})\big) \;+\; \mu_{Fatigue}^{(i)} \;+\; \epsilon_t^{(i,f)}
\end{aligned}
$$

$$
\begin{aligned}
Soreness_{z,t}^{(i)} = {}& \rho_s^{(i)} \, Soreness_{z,t-1}^{(i)} \;+\; \omega_1^{(i)} \, Load_{z,t-1}^{(i)} \;+\; \omega_2^{(i)} \, ACWR_{t-1}^{(i)} \;+\; \omega_3^{(i)} \, Fatigue_{z,t-1}^{(i)} \\
& -\; \omega_4^{(i)} \, SleepDur_{z,t-1}^{(i)} \;-\; \omega_5^{(i)} \, SleepEff_{z,t-1}^{(i)} \;-\; \omega_6^{(i)} \, Readiness_{t-1}^{(i)} \\
& +\; \omega_7^{(i)} \, \sin DOW_t \;+\; \omega_8^{(i)} \, \cos DOW_t \\
& +\; \mu_{Soreness}^{(i)} \;+\; \epsilon_t^{(i,s)}
\end{aligned}
$$

Note that the remaining 10 features are derived from the above features (e.g. InternalLoad is simply RPE × SessDuration).

The **remaining 10 features**:
- HRV Delta *(relative to past 7-day baseline)*
- RHR Delta *(relative to past 7-day baseline)*
- Session Duration
- Internal Load
- Acute Load
- Chronic Load
- ACWR
- DOW_sin
- DOW_cos
- Previous CRI *(yesterday)*

LABELS

Continuous Readiness Index (CRI)

$$
Index_t = \sigma\{ +w_1 \, HRV_z - w_2 \, RHR_z + w_3 \, SleepEff_z - w_4 \, ACWR + w_5 \, (HRV_z \times SleepEff_z) - w_6 \, (ACWR \times RHR_z) + lags + \epsilon \}
$$

The CRI will produce a value between 0 and 1, where 0 signals significant fatigue and 1 signals optimal physiological state.

The **remaining 2 labels** are derived from CRI:

- ReadyBin: 0 if CRI < 0.5; 1 if CRI ≥ 0.5
- ReadyOrd: 0 if 0 ≤ CRI < 0.2 | 1 if 0.2 ≤ CRI < 0.4 | 2 if 0.4 ≤ CRI < 0.6 | 3 if 0.6 ≤ CRI < 0.8 | 4 if 0.8 ≤ CRI ≤ 1

FILE CONTENTS

- `readiness_timeseries.csv` - Full simulated dataset (100 × 150 = 15,00 rows)
- `schema.json` - Column names, data types, and units
- `normalization_stats.json` - Athlete baselines and within-athlete SDs
- `simulation_config.json` - Simulation parameters and random seeds
- `splits_loao.json` - Leave-One-Athlete-Out (LOAO) cross-validation splits