

Explainable AI for Sleep Staging

A Use Case with PhysioEx

Guido Gagliardi

ESAT-STADIUS, KU Leuven

June 24, 2026



KU LEUVEN

What is PhysioEx?

Open-source Python library for **sleep staging** from EEG:

- **12 datasets** with a uniform API
- **13 pretrained models** on HuggingFace
- **Explainability** methods built in
- CLI + Python API

```
pip install physioex
```



<https://github.com/guidogagl/physioex>

The Sleep Staging Problem

Score every **30-second epoch** of overnight EEG into 5 stages:

Stage	Pattern	Dominant freq.
Wake (W)	Alpha rhythm, eye blinks	8–13 Hz
N1	Theta waves, vertex sharps	4–8 Hz
N2	Sleep spindles, K-complexes	11–16 Hz
N3	Slow delta waves	0.5–4 Hz
REM	Low-voltage mixed, REMs	Mixed

- A clinician scores ~ 1000 epochs per night
- Inter-rater agreement: only **82%**
- Can we automate this **and explain** what the model sees?

Data: Two Representations

Raw Waveform — Sleep-EDF

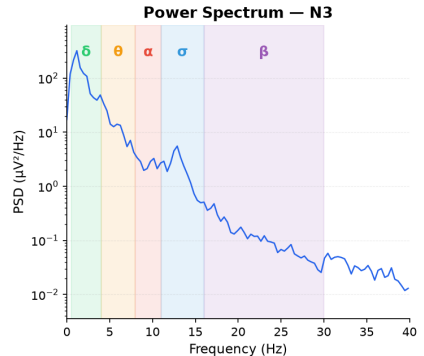
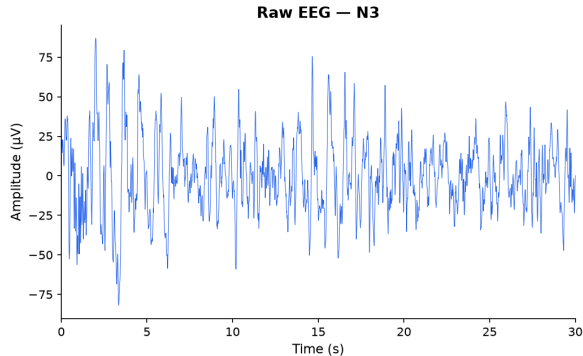
```
from physioex.data.datasets import
    SleepEDFDataset,
)
ds = SleepEDFDataset(
    channels=["EEG"],
    pipelines="raw",
    sequence_length=20,
)
# shape: (20, 3000)
# 20 epochs x 30s x 100 Hz
```

Spectrogram — MASS SS03

```
from physioex.data.datasets import
    MASSDataset,
)
ds = MASSDataset(
    cohort=3,
    channels=["EEG"],
    pipelines="seqsleepnet",
    sequence_length=20,
)
# shape: (20, 29, 129)
# 20 epochs x time x freq
```

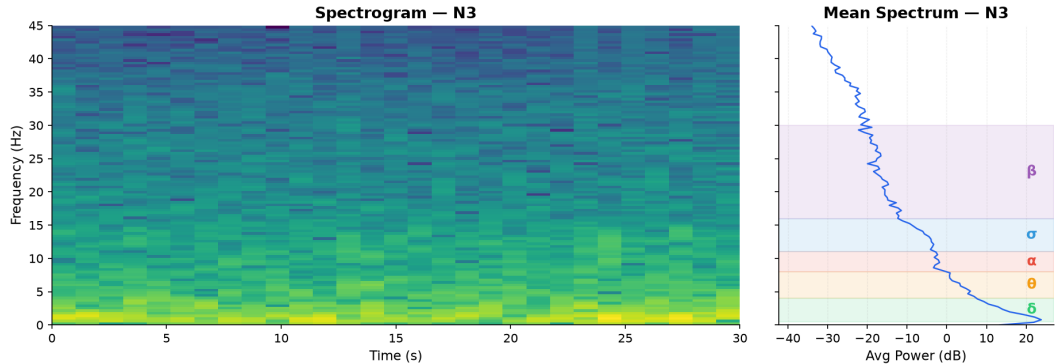
Same API, different preprocessing → different model families

Raw EEG — N3 (Deep Sleep)



Left: 30-second raw EEG epoch — large, slow oscillations typical of N3.
Right: Power spectral density — dominant power in the δ band (0.5–4 Hz).

Spectrogram — N3 (Deep Sleep)



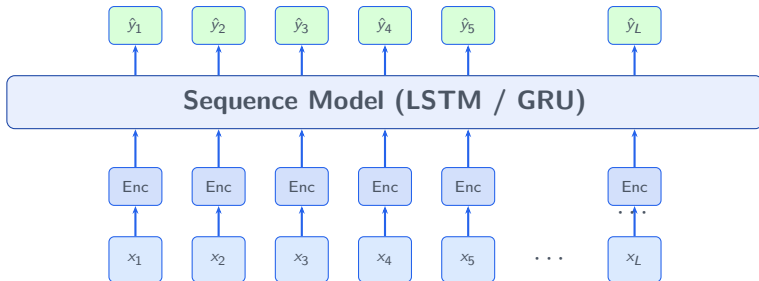
Left: Time-frequency spectrogram — bright yellow band at 0–4 Hz = strong delta power.

Right: Time-averaged power confirms δ dominates by >40 dB over higher bands.

Both representations capture the same physiology; the spectrogram makes frequency **explicit**.

Sequence-to-Sequence Sleep Staging

Sleep has **temporal structure** — models see $L = 20$ consecutive epochs (10 min):



Per-epoch encoder:

CNN (raw) or Filterbank+RNN
(spectrogram)

Sequence model:

Captures temporal context across epochs

Pretrained Models in PhysioEx

```
from physioex.models import load_from_pretrained

# Load any of 13 pretrained models with one call
tiny = load_from_pretrained("tinysleepnet-supratak", device="cuda")
seq  = load_from_pretrained("seqsleepnet-phan",      device="cuda")

# Inference: (batch, L=20 epochs, channels, ...)
logits = tiny(x_raw)          # x: (1, 20, 1, 3000)      -> (1, 20, 5)
logits = seq(x_spec)         # x: (1, 20, 1, 29, 129)    -> (1, 20, 5)
```

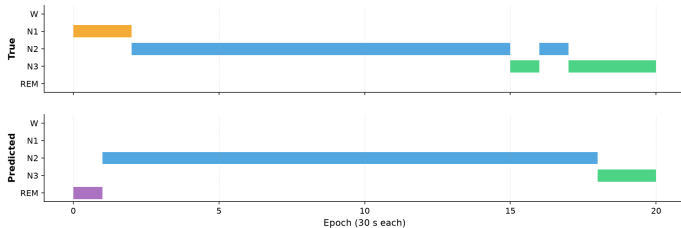
- Output: (batch, L, 5) — one 5-class probability distribution per epoch
- All 13 models share the same interface: load, call, done
- Models hosted on HuggingFace — downloaded and cached automatically

TinySleepNet — Raw EEG → Sleep Stages

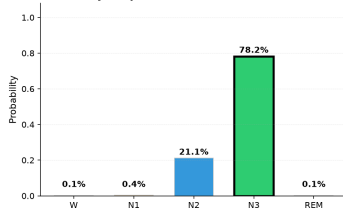
Architecture: $4 \times \text{Conv1D} (3000 \rightarrow 2048) \rightarrow \text{LSTM} (2048 \rightarrow 128) \rightarrow \text{Linear} (128 \rightarrow 5)$

Supratak & Guo, 2020

TinySleepNet on Sleep-EDF (in-domain)



TinySleepNet — Prediction Confidence



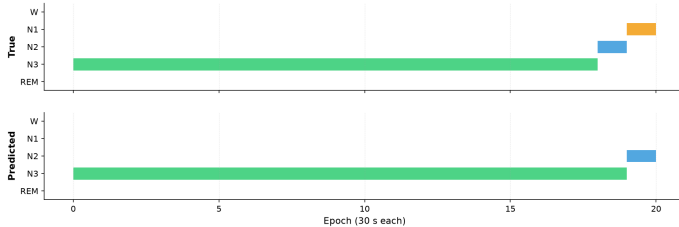
N3 epoch at **78% confidence** — correct, but not perfectly certain.

SeqSleepNet — Spectrogram → Sleep Stages

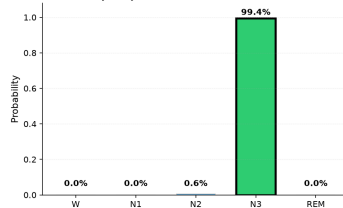
Architecture: LearnableFilterbank ($129 \rightarrow 32$) → BiLSTM+Attn → BiGRU → Linear ($128 \rightarrow 5$)

Phan et al., 2019

SeqSleepNet on MASS S503 (in-domain)



SeqSleepNet — Prediction Confidence



N3 epoch at **99.4% confidence** — the model is very sure.

Both models correctly classify **N3 (deep sleep)**.

TinySleepNet: 78%

SeqSleepNet: 99.4%

But why?

Which parts of the input drove the prediction?

Can we **validate** the explanation against known physiology?

Integrated Gradients — Intuition

Imagine a **path** from silence to the actual signal:

1. Start from a **baseline** (silence / noise floor)
2. Gradually interpolate toward the real input
3. At each step, measure the model's **gradient** — which features change the output most?
4. **Accumulate** these sensitivities along the path

Result: an **attribution map** (same shape as input)

Reading the Attribution

- **Red (+)** = pushes toward the predicted class
- **Blue (−)** = pushes away from it

Key Property: Completeness

Attributions sum exactly to $f(x) - f(\text{baseline})$

Nothing is “lost” or double-counted.

Integrated Gradients — Formulation

Definition (Sundararajan et al., ICML 2017):

For input x , baseline x' , and model output F :

$$\text{IG}_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

In practice — Riemann sum with $N = 64$ steps:

$$\text{IG}_i(x) \approx (x_i - x'_i) \times \frac{1}{N} \sum_{k=1}^N \frac{\partial F(x' + \frac{k}{N}(x - x'))}{\partial x_i}$$

Why not vanilla gradients?

- Gradients at a single point can be noisy / saturated
- IG satisfies **Sensitivity**: if a feature matters, it gets nonzero attribution

Key axioms:

- **Completeness**:
 $\sum_i \text{IG}_i = F(x) - F(x')$
- **Implementation Invariance**: same outputs $\Rightarrow$ same attributions

Computing IG Attributions

```
from physioex.explain.posthoc import IntegratedGradients

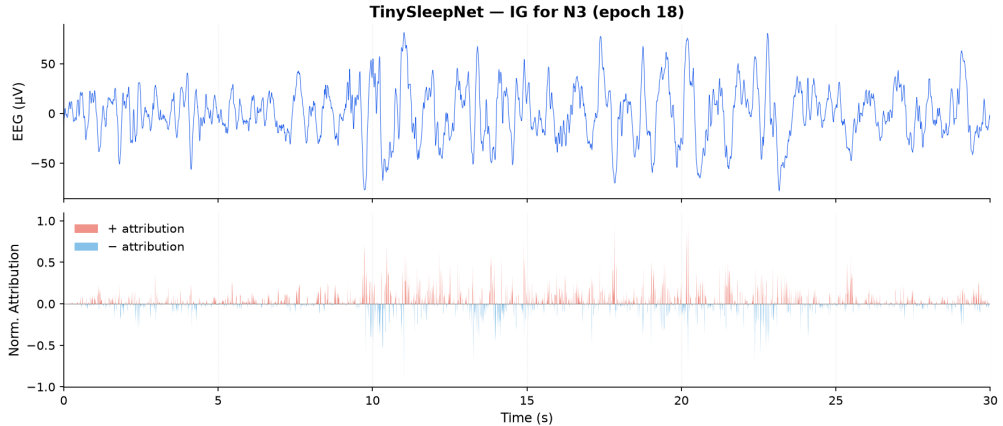
# 1. Wrap model: single epoch -> predicted-class confidence
def score_fn(x_epoch):
    logits = model(x_epoch.unsqueeze(1).unsqueeze(1))
    return logits[:, 0, :].softmax(-1)[:, pred_class]

# 2. Create IG (64 Riemann steps along baseline -> input path)
ig = IntegratedGradients(f=score_fn, steps=64, expects_batch=True)

# 3. Compute attributions (same shape as input)
attr = ig(x_epoch) # baseline = zeros (raw EEG)
attr = ig(x_epoch, baseline=noise_floor) # baseline = noise (spectrogram)
```

- `baseline=None` defaults to zeros — silence for raw waveforms
- For spectrograms: noise floor = mean power in bins > 40 Hz (≈ -35 dB)
- Attribution has the **same shape** as the input: (1, 3000) or (1, 29, 129)

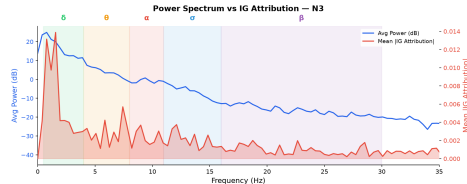
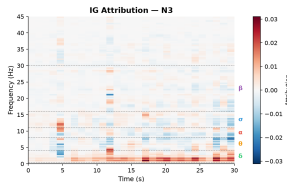
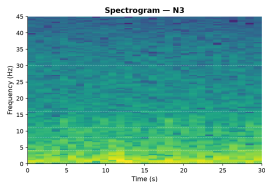
IG on TinySleepNet — Raw Signal



What we see: attribution peaks align with the large slow oscillations in the EEG.

Limitation: in the time domain we see *where* the model looks, but not *at which frequency* — we can't confirm it's delta.

IG on SeqSleepNet — Spectrogram



Left: spectrogram + IG heatmap — attribution concentrates in the lowest frequency bins.

Right: overlay of mean attribution vs. power spectrum — the δ band (0.5–4 Hz) drives the prediction.

The spectrogram representation lets us **validate** that the model uses the right frequency bands.

The Baseline Choice

IG attributions are *relative to a baseline* — the choice matters:

Model	Baseline	Interpretation
TinySleepNet	Zeros (silence)	“What changes from silence to this signal?”
SeqSleepNet	Noise floor (−35 dB)	“What is above the recording noise?”
<i>Alternatives:</i>		
	Random noise	“What is non-random in this signal?”
	Mean training signal	“What deviates from the average?”
	Contrastive epoch	“What differs from another stage?”

Why noise floor for spectrograms?

The signal is bandpass-filtered at 0.3–40 Hz, so bins above 40 Hz contain only noise. Setting the baseline to their mean value (≈ -35 dB) means: “explain what the model sees *above the noise*.”

The **Completeness** axiom guarantees: $\sum_i IG_i = F(x) - F(\text{baseline})$.

Wrap-Up

What we did

1. Examined N3 in two representations; confirmed delta dominance via PSD
2. Ran two seq-to-seq models that correctly classify N3
3. Applied Integrated Gradients to both; explanations match known physiology
4. Showed that spectrogram attributions are directly interpretable in frequency

Key Insight

Input representation

affects both model performance
and explanation interpretability.

Raw signal → temporal attribution
Spectrogram → frequency attribution

PhysioEx also supports

- SpectralGradients — freq-band attribution
- CSD — concept-level explanations
- 9 foundation models

Thank You

Questions?

Guido Gagliardi

guido.gagliardi@kuleuven.be
ESAT-STADIUS, KU Leuven

GitHub <https://github.com/guidogagl/physioex>

PyPI `pip install physioex`

Models huggingface.co/4rooms/physioex



KU LEUVEN