ML - Kaggle Challenge : Beacon Biosignals Sleep Staging

Oceam Toujan & Ivan Bautista

December 2025

Abstract

The objective of this Kaggle Challenge is to predict sleep stages based on EEG recordings. There are 5 **sleep stages** : "Wake", "N1", "N2", "N3", "REM" ; and 5 EEG channels recorded on 8 patients. The EEG signal is batched in 30 **s-epochs**, and we want to label each epoch with one of the 5 sleep stages.

To that purpose, we **extracted and engineered several features** based on EEG signal processing and sleep states litterature. After reduction, we ended up with 700 **features** processed for each 30 s-epoch.

The classification task upon those 700 features is then handled by an `XGBoost Classifier`. This algorithm is inherently stochastic and the random seed introduces variance in the outputs. To cope with that bottleneck, we implemented **seed averaging strategy** with a custom wrapper. This custom wrapper initializes five similar `XGBClassifier`s but with different random seeds, and the prediction is made by averaging all classifiers' outputs.

Finally, we operate a 'manual curation' to **smoothen** high discrepancies in stage predictions that appear **physiologically unplausible** and will never be predicted by any expert (for example very short sudden awake stage within long deep sleep period).

1 Data Preprocesssing

1.1 Signal Acquisition and Segmentation

The dataset consists of multichannel EEG polysomnography recordings acquired during overnight sleep. Each recording is sampled at a frequency of $f_s = 100\,\text{Hz}$ and contains signals from five EEG channels. The continuous signals are segmented into non-overlapping epochs of 30 s, corresponding to 3000 samples per channel and per epoch.

Each epoch is associated with a single sleep stage label belonging to one of the following classes: Wake, N1, N2, N3, and REM. These labels are provided at the epoch level and are assumed to be temporally aligned with the EEG signals.

1.2 Data Loading and Integrity Checks

Raw signals and labels are stored as NumPy arrays and loaded using a set of dedicated helper functions. Several integrity checks are performed at load time : verification of the number of channels, validation of signal dimensionality, consistency between signal length and epoch duration, alignment between signal epochs and target labels. These checks ensure that all downstream processing operates on well-formed and synchronized data.

1.3 Band-pass Filtering

Prior to feature extraction, each EEG channel is filtered using a zero-phase Butterworth band-pass filter in the range $[0.5, 40]\,\text{Hz}$. From a neuro-physiological point of view, this frequency range includes most neural rythms, from delta band ($[0.5, 4]\,\text{Hz}$) to sigma one ($> 30\,\text{Hz}$), through theta ($[4, 8]\,\text{Hz}$), alpha ($[8, 13]\,\text{Hz}$) and beta ($[13, 30]\,\text{Hz}$). Furthermore, zero-phase filtering (via forward–backward filtering using `filtfilt`) is used to avoid phase distortions that could affect temporal or spectral features.

Filtering is not intended to denoise the signal aggressively, but rather to remove slow drifts and high-frequency artifacts while preserving physiologically relevant oscillatory activity in the common frequency bands (delta, theta, alpha, beta).

1.4 Temporal Sub-Segmentation

In order to capture short-term temporal dynamics within each 30 s epoch, epochs are further subdivided into 6 non-overlapping sub-epochs of 5 s. Features are first computed at the sub-epoch level and subsequently aggregated at the epoch level.

This multi-scale temporal representation increases robustness to transient artifacts and allows the model to capture within-epoch variability that is characteristic of several sleep stages.

2 Feature Extraction

Each EEG epoch is transformed into a high-dimensional feature vector combining time-domain, frequency-domain, time–frequency, nonlinear, and physiological proxy features. Feature extraction is performed channel-wise and complemented by global features capturing inter-channel relationships [1, 2].

The overall design aims to balance the three following objectives [3] :

i) physiological interpretability
ii) robustness to noise and inter-subject variability
iii) discriminative power across sleep stages

Feature engineering was deliberately designed as a multi-level representation, combining local descriptors computed on short sub-epochs, aggregation operators capturing within-epoch dynamics, and contextual features encoding temporal continuity across successive epochs. This hierarchical design mirrors both the temporal organization of sleep and the way human experts integrate local EEG patterns into global stage decisions.

2.1 Time-Domain Features

For each sub-epoch and each channel, several time-domain descriptors are computed, including standard deviation, zero-crossing rate, signal sharpness (based on first-order differences) and Teager–Kaiser energy operator (TKEO) [4].

These features capture signal amplitude, variability, and abrupt temporal changes. Wake and REM stages typically exhibit higher variability and sharp transitions, whereas deep sleep (N3) is characterized by smoother, high-amplitude oscillations.

Time-domain descriptors have been widely used in automatic sleep staging due to their robustness and low sensitivity to spectral estimation artifacts. In particular, the Teager–Kaiser energy operator has been shown to be sensitive to rapid energy fluctuations and transient events, making it suitable for detecting micro-arousals and abrupt EEG changes [5].

2.2 Frequency-Domain Features

Spectral features are extracted using Welch's method [6] and, in selected cases, a multitaper power spectral density estimator to improve robustness.

The following spectral descriptors are computed [3, 7]:

i) relative band power in the delta, theta, alpha, sigma, and beta bands
ii) ratios between key bands (e.g. alpha/theta, delta/theta, slow/fast)
iii) spectral slope
iv) peak alpha frequency
v) spectral edge frequencies (SEF50 and SEF95)
vi) spectral centroid and spectral spread

These features reflect well-known electrophysiological markers of sleep. For instance, delta power dominates deep sleep, sigma activity is associated with sleep spindles during N2, and alpha–theta dynamics play a key role during transitions and REM sleep.

Relative spectral power and band ratios constitute the backbone of most classical and modern sleep staging systems. Delta dominance is a defining marker of slow-wave sleep (N3), sigma-band activity reflects spindle density during N2, while alpha and theta activity characterize wakefulness, drowsiness, and REM sleep. Spectral slope and edge frequencies further capture global shifts in spectral organization that occur across sleep stages [2].

2.3 Time–Frequency and Burst-Related Features

Using the analytic signal obtained via the Hilbert transform, envelope-based features are computed in the alpha and sigma bands : maximum envelope amplitude, burst duration above adaptive thresholds and coefficient of variation of the envelope.

These features are particularly relevant for detecting sleep spindles and transient oscillatory events, which are strong indicators of specific sleep stages. Sleep spindles are a defining hallmark of stage N2, while transient alpha activity is frequently associated with micro-arousals and wake intrusions. Envelope statistics and burst duration metrics provide a compact and noise-robust representation of these phenomena [8, 9].

2.4 Wavelet-Based Features

Discrete wavelet transforms are applied to sub-epochs in order to capture transient oscillations and localized energy patterns. Energy coefficients in the delta and sigma ranges are retained as features [10].

Wavelet-based representations complement classical Fourier analysis by providing improved sensitivity to short-lived events such as K-complexes and spindles.

Wavelet-based energy features enable multi-resolution analysis of EEG signals, offering improved sensitivity to non-stationary and transient events such as K-complexes and slow waves. Unlike Fourier-based methods, wavelets preserve both temporal and frequency localization, which is particularly advantageous for sleep EEG analysis [4].

2.5 Nonlinear and Complexity Measures

To characterize signal irregularity and complexity, several nonlinear descriptors are extracted [11] :

i) spectral entropy
ii) spectral flatness
iii) kurtosis in selected frequency bands
iv) asymmetry measures
v) a spectral chaos index based on inter-band power variability

These features tend to increase during wakefulness and REM sleep, reflecting more complex and less regular brain activity.

Nonlinear and complexity-based features aim to quantify signal irregularity and state instability. Increased entropy, spectral flattening, and inter-band variability are typically associated with wakefulness, REM sleep, and micro-arousals, whereas stable sleep stages exhibit more structured spectral profiles. The proposed spectral chaos index captures this phenomenon by measuring the dispersion of relative band powers within a sub-epoch [12].

2.6 Inter-Channel and Global Features

Beyond channel-wise descriptors, global features are computed to capture spatial structure and synchrony [13] :

i) spectral coherence between selected channel pairs
ii) global slow-wave correlation
iii) fronto–occipital gradients in alpha and delta activity

Such features reflect large-scale coordination of brain activity, which is a hallmark of sleep depth and stage transitions.

Sleep is a spatially organized phenomenon, characterized by increasing inter-regional synchrony with sleep depth. Coherence and correlation-based features capture large-scale coordination patterns that are not observable at the single-channel level. Fronto–occipital gradients further reflect known spatial asymmetries in alpha and delta activity across sleep stages [14].

2.7 Physiological Proxy Features

High-level physiological proxy features are extracted using established signal processing pipelines implemented in the `YASA` toolbox [15] :

i) automatic detection of sleep spindles and slow waves

ii) cross-frequency coupling between delta and sigma bands

iii) micro-arousal indicators based on beta-band envelope dynamics

iv) EMG and REM proxies derived from high-frequency activity

These proxies aim to directly quantify canonical sleep markers such as spindles and slow waves, thereby bridging the gap between raw EEG features and clinically interpretable sleep phenomena.

2.8 Feature Selection & Aggregation

To incorporate temporal context, we introduce a lagged feature representation whereby the sleep stage of epoch e is predicted using features extracted from epoch e as well as from the 1-, 5-, and 21-previous epochs. This strategy allows the model to capture sleep continuity, stage inertia, and gradual transitions, which are central characteristics of sleep architecture [16, 17].

Overly correlated features are simply discarded, allowing us to boil feature extraction down to 700 remaining meaningful features.

2.9 Illustrative Analysis

Figure 1 presents a two-dimensional PCA projection of the 700 most informative features. Although no clear linear separation between sleep stages is observed, structured regions emerge in the feature space. In particular, deep sleep (N3) and Wake tend to occupy more extreme regions, while transitional stages such as N1 exhibit strong overlap with neighboring classes.

This visualization highlights both the intrinsic difficulty of the task and the need for nonlinear classification models capable of exploiting complex feature interactions rather than simple linear boundaries.

It should be noted that PCA is used here solely as an exploratory visualization tool ; the absence of clear linear separation does not preclude strong discriminative performance when nonlinear classifiers are employed.

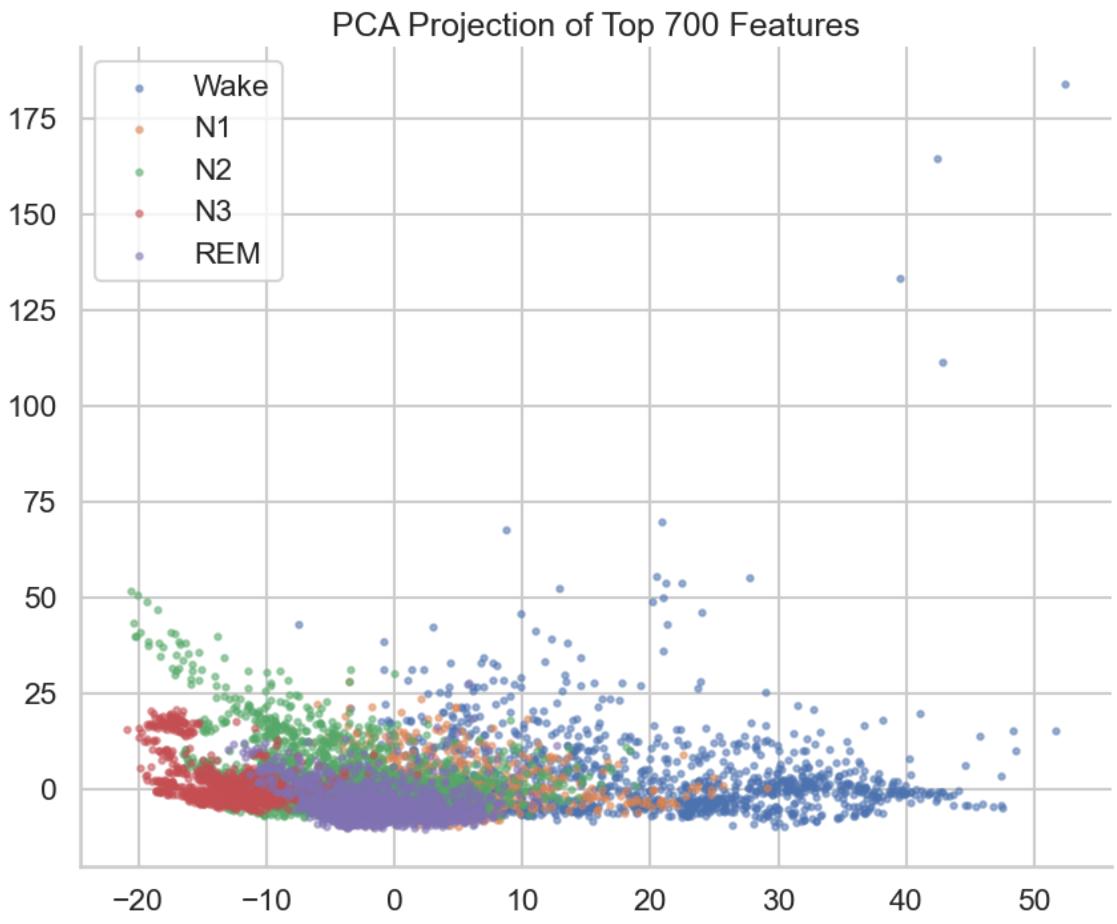


Figure 1: PCA-Analysis on Top-700 Features

3 Model Description

3.1 Choice of Model

The classification task is addressed using a gradient boosting model based on decision trees `XGBoost`. Gradient boosting constructs an ensemble of weak learners in a sequential manner, where each new tree corrects the errors of the previous ensemble.

This model class is particularly well suited to EEG-based sleep staging due to its ability to handle high-dimensional feature spaces, model complex nonlinear interactions and accommodate heterogeneous feature distributions.

3.2 Multiclass Formulation and Regularization

The model is trained in a multiclass setting with five output classes corresponding to the sleep stages. Regularization is enforced through constraints on tree depth, subsampling of observations and features, and explicit L_1 and L_2 penalties.

These mechanisms reduce overfitting and improve generalization across subjects, which is critical given the strong inter-individual variability of sleep EEG signals.

3.3 Model Wrapper & Random Seed Sensitivity

Since `XGBoost` is a stochastic algorithm, there is a risk of a 'Random Seed Bootleneck' : the initialisation of the random generator itself can introduce some variance in the results.

To cope with that problem, we wrapped the `XGBClassifier` in a custom wrapper `'SeededXGBClassifier'` to allow us to train simultaneously 5-differently-seeded models (random seed = 42, 43, 44, 45, 46).

The schematic pipeline is very straighforward. We trained a single `XGBClassifier` and found its optimal hyperparameters using small cross-validation protocol (see below). Then, we generated inside our custom wrapper five `XGBClassifiers` with identical hyperparameters but different random seeds. Finally, `SeededXGBClassifier`'s predictions are claimed as the average of the five `XGBClassifiers`' outputs.

3.4 Manual Smoothing

Given the `'SeededXGBClassifier'`'s predictions, we applied a final step of "Human-Manual Smoothing" : for example, experts will never predict the following sequence "deep sleep - deep sleep - awake - deep sleep" because it lacks continuity to be bio-physiologically plausible. Such sequences are thus being smoothened is a 'post-processing' fashion.

We can report here the importance of each hyperparameter of the `XGBClassifier`

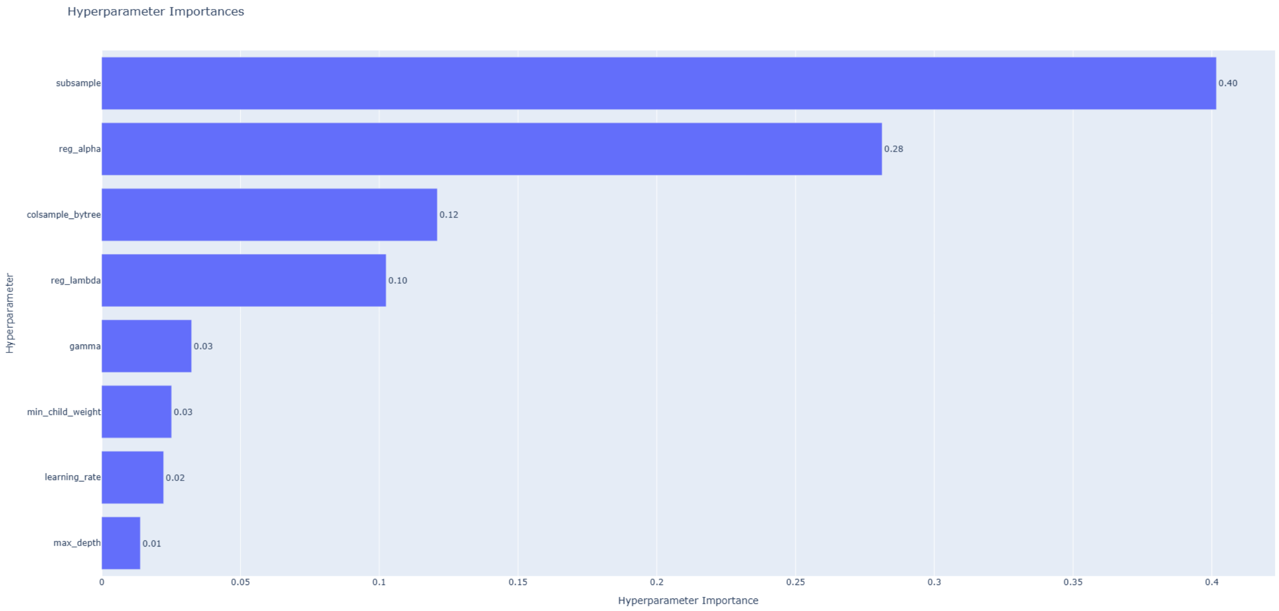


Figure 2: Parameter Importances

Figure 2 reports the relative importance of hyperparameters with respect to the optimization objective. `subsample` and `reg_alpha` appears dominant.

4 Cross-Validation Protocol

Hyperparameter optimization is performed using the Optuna framework, which automates the search for optimal model configurations. For each trial, Optuna samples a candidate set of XGBoost hyperparameters, trains the corresponding model, and evaluates its performance using subject-wise cross-validation. The resulting macro-averaged F1 score is used as the optimization objective. Based on the outcomes of previous trials, Optuna adaptively guides the search toward promising regions of the hyperparameter space, allowing efficient exploration while limiting computational cost.

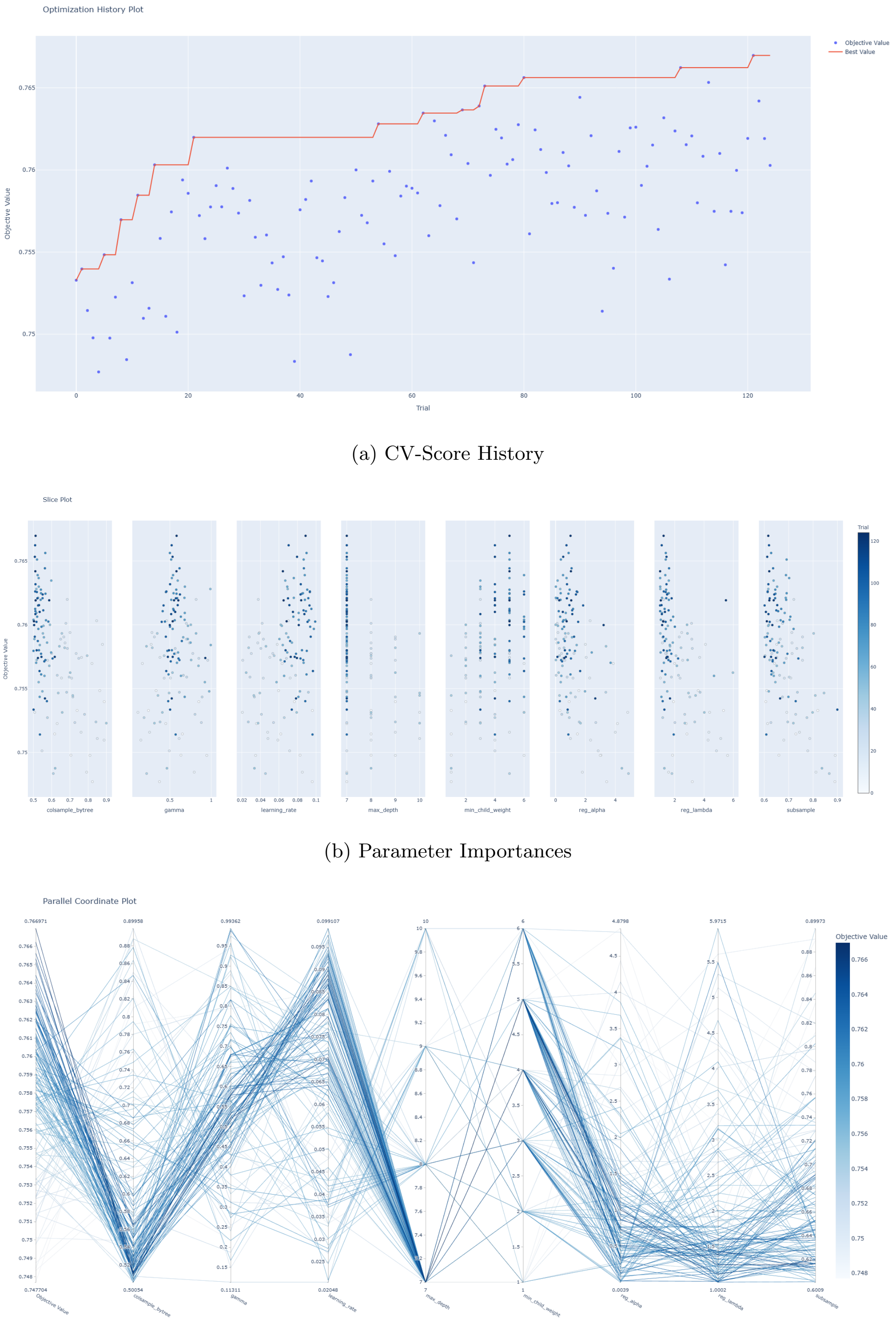
Formally, a Group K-Fold cross-validation scheme is used, ensuring that no data from a given subject appears simultaneously in the training and validation sets. This protocol simulates a realistic deployment scenario in which the model is applied to unseen individuals.

Model selection and evaluation are performed using the macro-averaged F1 score, which balances performance across all sleep stages and mitigates the impact of class imbalance.

Figure 3a shows the evolution of the objective function across hyperparameter optimization trials. Performance improves in a stepwise manner, with successive plateaux indicating progressive refinement of promising regions of the hyperparameter space. The stabilization observed in later trials suggests convergence toward a robust optimum rather than overfitting to individual folds.

Slice plots (Figure **??**) illustrate the relationship between individual hyperparameters and the validation performance. Given a vertical line (ie fixed value for the hyperparameter), vertical spread of results indicate a high variance on objective value for that hyperparameter's value ; that means that this hyperparameter has limited effect on the objective value optimization (see `max_depth` for example). On the contrary, clustered dots on y axis, with a decaying pattern over x axis highlights the necessity for such hyperparameter to be optimized (see `subsample`). An explicit link can be made with results drawn by Figure 2.

Figure 3c provides a parallel-coordinates representation of the hyperparameter configurations explored during optimization. High-performing trials concentrate within specific regions of the parameter space, revealing interactions between depth, learning rate, and subsampling ratios. This joint analysis confirms that optimal performance arises from a balanced trade-off between model expressiveness and regularization.



(a) CV-Score History

(b) Parameter Importances

(c) Parallel Score History

Figure 3: CV-Results

5 Results & Model Performances

Model performance is evaluated using cross-validated predictions. The overall retained metrics (that was imposed by the challenge) is the F1 score, whose expression is given as follows :

$$\boxed{F1 \text{ Score} = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}}$$

This metrics allows a more balanced view on model classification performance when class distribution is not homogeneous. It highlights the model's ability to discriminate between physiologically similar stages such as N1 and REM, as well as its robustness on more distinct stages such as Wake and N3.

Let us have a look at local classification performances ; starting with confusion matrix showing the per-class F1 scores :

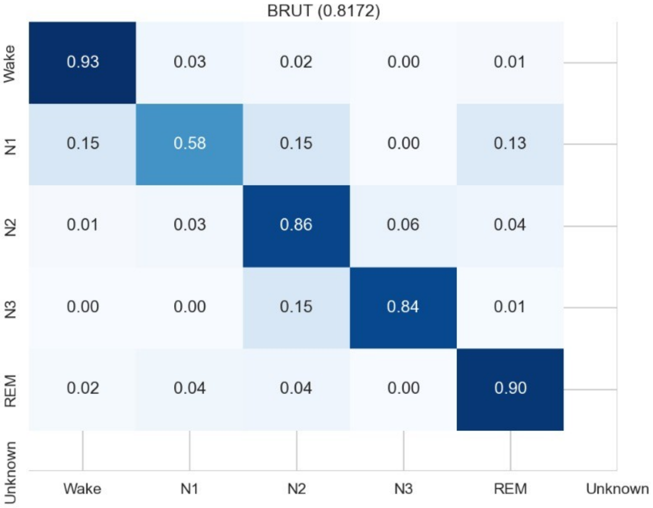
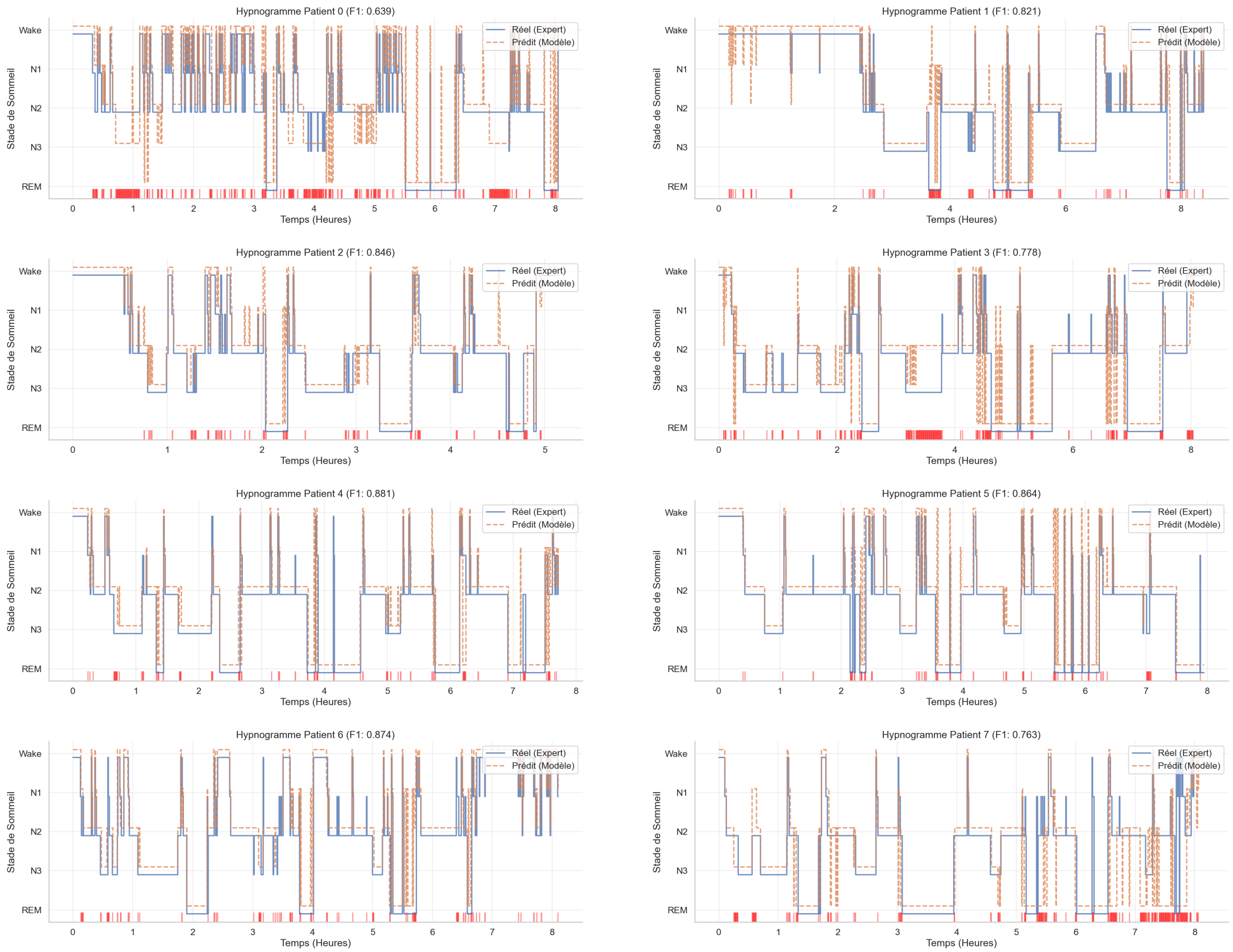


Figure 4: Per-Class F1 Scores Confusion Matrix

The most obvious comment on such matrix is that the model struggles with $N1$ sleep stage. That is something that was expected, since $N1$ stage in particular is a transient stage, making it both **scarce** and highly `non-separable`. Indeed, as previously shown by the PCA analysis, this stage highly lacked separability from others, so this is no big surprise that this is the stage for which the model performs the worst.

And finally, here are the hypnograms for the model's predictions (orange-dotted) vs expert labeling (blue-continuous) on the eight patients :


Hypnogramme Patient 0 (F1: 0.639)

Hypnogramme Patient 1 (F1: 0.821)

Hypnogramme Patient 2 (F1: 0.846)

Hypnogramme Patient 3 (F1: 0.778)

Hypnogramme Patient 4 (F1: 0.881)

Hypnogramme Patient 5 (F1: 0.864)

Hypnogramme Patient 6 (F1: 0.874)

Hypnogramme Patient 7 (F1: 0.763)

Figure 5: Hypnograms

Furthermore, the $F1$ Score are the followings :

	F1	Accuracy
Patient 0	0.6372	0.7063
Patient 1	0.8202	0.8959
Patient 2	0.8489	0.8981
Patient 3	0.7791	0.8107
Patient 4	0.8791	0.9169
Patient 5	0.8589	0.9288
Patient 6	0.8679	0.9034
Patient 7	0.7634	0.8082

Those hypnograms reveal predictions that follow well the trends and dynamics of expert labeling, but that still lack a bit of continuity. That emphatizes the difficulty with $N1$ stage classification, as well as the relevance of 'manual post-processing' to tamper with physiological-unplausible irregularities.

References

[1] A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects.* U.S. Department of Health, Education, and Welfare, 1968.

[2] American Academy of Sleep Medicine, *The AASM Manual for the Scoring of Sleep and Associated Events.* American Academy of Sleep Medicine, version 2.4 ed., 2017.

[3] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, and M. Samet, "Learning machines for automated sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 1, pp. 22–33, 2015.

[4] A. R. Hassan and M. I. H. Bhuiyan, "Automated identification of sleep stages from eeg signals using normalized power spectral density," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 12, pp. 1277–1287, 2016.

[5] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 381–384, 1990.

[6] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[7] P. Achermann and A. A. Borbély, "Sleep homeostasis and models of sleep regulation," *Brain Research Bulletin*, vol. 44, no. 5, pp. 409–421, 1997.

[8] L. De Gennaro and M. Ferrara, "Sleep spindles: An overview," *Sleep Medicine Reviews*, vol. 7, no. 5, pp. 423–440, 2003.

[9] T. Andrillon, Y. Nir, R. J. Staba, F. Ferrarelli, C. Cirelli, G. Tononi, and I. Fried, "Sleep spindles in humans: Insights from intracranial eeg," *Journal of Neuroscience*, vol. 31, no. 49, pp. 17821–17834, 2011.

[10] I. Daubechies, *Ten Lectures on Wavelets.* SIAM, 1992.

[11] E. Pereda, R. Quian Quiroga, and J. Bhattacharya, "Nonlinear multivariate analysis of neurophysiological signals," *Progress in Neurobiology*, vol. 77, no. 1–2, pp. 1–37, 2005.

[12] O. A. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schurmann, and E. Başar, "Distinguishing noise from chaos," *Physical Review Letters*, vol. 99, no. 15, p. 154102, 2007.

[13] M. Steriade, D. A. McCormick, and T. J. Sejnowski, "Thalamocortical oscillations in the sleeping and aroused brain," *Science*, vol. 262, no. 5134, pp. 679–685, 1993.

[14] P. Achermann, L. A. Finelli, and A. A. Borbély, "Regional differences in eeg slow-wave activity," *Journal of Sleep Research*, vol. 7, no. Suppl 1, pp. 38–42, 1998.

[15] R. Vallat and M. P. Walker, "An open-source toolbox for sleep analysis," *Journal of Open Source Software*, vol. 6, no. 69, p. 3434, 2021.

[16] S. Chambon, M. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 11, pp. 2379–2390, 2018.

[17] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of sleep apnea," *Sleep*, vol. 41, no. 11, 2018.