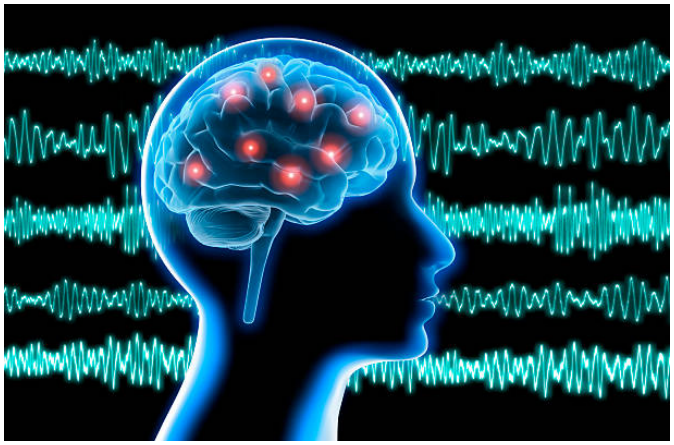
NTNU ENSTA INSTITUT POLYTECHNIQUE DE PARIS

ENSTA — Formation ingénieur
Stage d'assistant ingénieur — 2ème année

Classification automatique des stades de sommeil et détection du contenu onirique à partir de l'EEG



Rapport de stage présenté par
Alexandre Pinard

Organisme d'accueil : **Norwegian University of Science and Technology (NTNU)**
Lieu : Trondheim, Norvège
Tuteur de stage : Professeure Marta Molinas

Période du stage : 07/05/2026 – 10/09/2026

Résumé

Ce stage, réalisé au département *Engineering Cybernetics* de la Norwegian University of Science and Technology (NTNU), porte sur la classification automatique des stades de sommeil à partir d'un EEG faiblement invasif, en collaboration avec le *IIIS Human Sleep Lab* de l'université de Tsukuba, et s'inscrit dans la continuité du projet de recherche *Dream Decoder* du laboratoire, qui vise à décoder automatiquement le contenu mental du sommeil. Le travail est structuré en trois niveaux de difficulté croissante : la classification du stade de sommeil, la détection de la présence d'un rêve, puis la caractérisation de sa tonalité émotionnelle.

Sur le premier niveau, un pipeline CNN + LSTM entraîné sur un signal EEG mono-canal (jeu de données Sleep-EDF Expanded, 153 sujets) atteint un Kappa de Cohen de 0,83, un niveau d'accord excellent, validant la faisabilité d'un scoring automatique de haute qualité sur un montage réduit. Sur le deuxième niveau, la détection de rêve est étudiée sur trois jeux de données (*Multiple Awakenings*, Tononi, Zhang & Wamsley) avec une méthodologie PSD + XGBoost et sélection de canaux par permutation ; un Kappa de l'ordre de 0,80 est obtenu en validation croisée classique, mais s'effondre à une valeur proche de zéro en validation *Leave-One-Subject-Out* (LOSO), révélant une fuite de sujet systématique que ce stage documente et quantifie sur l'ensemble des jeux de données testés.

Sur le troisième niveau, le plus difficile, une classification de la colère rapportée en rêve (jeu de données REM_Turku), fondée sur trois features de puissance gamma frontale sélectionnées par corrélation et correction pour comparaisons multiples (FDR), atteint 61 % d'exactitude en LOSO et se révèle statistiquement significative par test de permutation ($p = 0{,}048$) – la seule des vingt émotions testées à passer ce contrôle. Cette réussite ciblée contraste avec l'absence de signal significatif pour une classification plus générale de la valence émotionnelle et pour la réplication de la corrélation historique entre asymétrie frontale alpha et colère (Sikka et al., 2019) : le signal EEG de l'émotion du rêve semble exister, mais de façon spécifique, non générale.

Au-delà des résultats chiffrés, ce stage documente et corrige plusieurs pièges méthodologiques classiques de l'apprentissage automatique appliqué à l'EEG sur petits échantillons : fuite de sujet entre les ensembles d'entraînement et de test, artefacts non physiologiques confondus avec le signal d'intérêt, erreurs de labellisation, et seuils de décision optimisés de façon optimiste. Cette rigueur méthodologique, systématiquement appliquée aux trois parties du rapport, est ce qui permet de distinguer, parmi les résultats obtenus, ceux qui sont réellement défendables de ceux qui ne le sont pas.

Mots-clés : EEG, sommeil, rêve, émotion, apprentissage automatique, validation croisée, fuite de données, asymétrie frontale alpha.

Abstract

This internship, carried out at the Department of Engineering Cybernetics of the Norwegian University of Science and Technology (NTNU), focuses on automatic sleep stage classification from minimally invasive EEG, in collaboration with the IIIS Human Sleep Lab at the University of Tsukuba, and builds on the host laboratory's *Dream Decoder* research project, which aims to automatically decode the mental content of sleep. The work is organized into three levels of increasing difficulty: sleep stage classification, dream presence detection, and dream emotional tone characterization.

At the first level, a CNN + LSTM pipeline trained on single-channel EEG (Sleep-EDF Expanded dataset, 153 subjects) reaches a Cohen's Kappa of 0.83, an excellent level of agreement, validating the feasibility of high-quality automatic sleep scoring on a reduced montage. At the second level, dream detection is studied across three datasets (*Multiple Awakenings*, Tononi, Zhang & Wamsley) using a PSD + XGBoost pipeline with permutation-based channel selection ; a Kappa around 0.80 is obtained under standard cross-validation but collapses to near zero under Leave-One-Subject-Out (LOSO) validation, revealing a systematic subject-level leakage that this internship documents and quantifies across every dataset tested.

At the third and most difficult level, classifying anger reported in dreams (REM_Turku dataset), based on three frontal gamma-power features selected through correlation analysis with multiple-comparison correction (FDR), reaches 61% LOSO accuracy and proves statistically significant under a permutation test ($p = 0.048$) – the only one of twenty tested emotions to pass this check. This targeted success contrasts with the absence of significant signal for a broader positive / negative valence classification, and with the failure to replicate the historical correlation between frontal alpha asymmetry and anger (Sikka et al., 2019): EEG signatures of dream emotion appear to exist, but specifically rather than generally.

Beyond the numerical results, this internship documents and corrects several classic methodological pitfalls of EEG-based machine learning on small samples: subject-level leakage between training and test sets, non-physiological artefacts confused with the signal of interest, labelling errors, and optimistically tuned decision thresholds. This methodological rigor, applied consistently across all three parts of the report, is what makes it possible to separate genuinely defensible results from those that are not.

Keywords: EEG, sleep, dreaming, emotion, machine learning, cross-validation, data leakage, frontal alpha asymmetry.

Table des matières

Résumé i

Abstract ii

1 Contexte 1
- 1.1 Le laboratoire d'accueil . . . 1
- 1.2 Le projet de recherche *Dream Decoder* . . . 1
- 1.3 Positionnement du stage . . . 1

2 Problématique, objectifs et enjeux 2
- 2.1 Le sujet du stage . . . 2
- 2.2 Une problématique à trois niveaux de difficulté croissante . . . 2
- 2.3 Enjeux . . . 3

3 Le cerveau, le sommeil et le rêve : éléments de neuroscience 4
- 3.1 Anatomie cérébrale . . . 4
- 3.2 Le sommeil et son architecture . . . 5
- 3.3 Le rêve : neurophysiologie et enjeux cliniques . . . 6
- 3.4 L'électroencéphalographie . . . 7

4 Cadre scientifique et méthodologique 9
- 4.1 Recueil du contenu mental : le paradigme du réveil provoqué . . . 9
- 4.2 Analyse spectrale de l'EEG . . . 9
- 4.3 Asymétrie frontale alpha et affect du rêve . . . 10
- 4.4 Modèles de classification utilisés . . . 10
- 4.5 Déséquilibre des classes et SMOTE . . . 10
- 4.6 Fuite de données et validation croisée . . . 11
- 4.7 Métriques d'évaluation . . . 11
- 4.8 Contrôle statistique de la significativité . . . 12

5 Description de l'activité et des résultats 13
- 5.1 Partie 1 — Classification des stades de sommeil . . . 14
  - 5.1.1 Méthodologie . . . 14
  - 5.1.2 Résultats . . . 14
  - 5.1.3 Bilan . . . 14
- 5.2 Partie 2 — Détection de rêve . . . 16
  - 5.2.1 Multiple Awakenings : identification et correction d'une erreur de labellisation . . . 16
  - 5.2.2 Réplication fidèle sur Tononi et Zhang & Wamsley . . . 17
- 5.3 Partie 3 — Détection de l'émotion du rêve . . . 19
  - 5.3.1 Premiers essais et alerte méthodologique . . . 19
  - 5.3.2 Construction itérative du pipeline . . . 19

5.3.3 Réplication fidèle d'un travail antérieur 20
5.3.4 Prototype d'apprentissage de représentation 21
5.3.5 Un signal significatif pour la colère : sélection par corrélation et validation par permutation 21
5.3.6 Robustesse statistique du pipeline sur la valence générale 24
5.3.7 Réplication de l'analyse de corrélation de Sikka et al. 25
5.4 Synthèse comparative des résultats 26

6 Conclusion **27**
6.1 Réalisations au regard des enjeux identifiés 27
6.2 Apports du stage pour ma formation et mon projet professionnel 27
6.3 Perspectives 28

Bibliographie **29**

Appendices **30**

Chapitre 1

Contexte

1.1 Le laboratoire d'accueil

Le stage s'est déroulé au sein du département *Engineering Cybernetics* de la Norwegian University of Science and Technology (NTNU), à Trondheim, sous la responsabilité de Professeure Marta Molinas, sur la période 07/05/2026 – 10/09/2026. L'équipe EEG de ce département développe des méthodes de traitement du signal et d'apprentissage automatique appliquées à l'électroencéphalographie, avec un accent particulier sur les interfaces cerveau-machine (BCI) et le décodage automatique de l'activité de sommeil.

Le sujet du stage s'inscrit dans une collaboration entre cette équipe et le *IIIS Human Sleep Lab* de l'université de Tsukuba (Japon), qui met à disposition les données de sommeil support du travail de classification des stades de sommeil (chapitre 5, partie 1).

1.2 Le projet de recherche *Dream Decoder*

Le stage s'inscrit également dans la continuité du projet *Dream Decoder*, porté par la même équipe, qui vise à développer des méthodes de décodage automatique du contenu des rêves à partir de l'EEG de sommeil, et a donné lieu à plusieurs travaux de référence directement mobilisés pendant le stage :

- la thèse de master de Packiyanathan et Torvestad [3], portant sur la détection binaire rêve / absence de rêve (*Dream / No Dream*) à partir de l'EEG, avec une méthodologie PSD + XGBoost validée sur les jeux de données Tononi et Zhang & Wamsley ;
- la thèse de master de Beslagic [1], *Dream Catcher*, qui étend cette approche au décodage du contenu des rêves en sommeil NREM en vue d'interfaces cerveau-machine portables.

Ces deux travaux fournissent le code source, les conventions méthodologiques (choix des bandes de fréquence, du classifieur, du protocole de validation croisée) et les jeux de données de référence sur lesquels s'appuie directement la partie 2 de ce stage (chapitre 5, section 5.2).

Le stage s'inscrit enfin dans le cadre plus large de la base de données *DREAM* (*Dream EEG and Mentation*, Gott et al. [2]), une collection standardisée de plus de vingt jeux de données EEG de sommeil associés à des rapports de rêves, totalisant 505 participants et 2643 réveils, dont sont issus l'ensemble des jeux de données de rêve et d'émotion utilisés durant le stage (Multiple Awakenings, Tononi, Zhang & Wamsley, REM_Turku).

1.3 Positionnement du stage

Le jeu de données REM_Turku est le seul, au sein de la base DREAM, à fournir des scores d'affect détaillés (échelle mDES) permettant, en principe, une classification de l'émotion ressentie pendant le rêve et non plus seulement de sa présence ; c'est ce qui en fait le support naturel de la partie 3 du stage, la plus avancée dans l'échelle de difficulté définie section 2.2.

Chapitre 2

Problématique, objectifs et enjeux

2.1 Le sujet du stage

Le sujet officiel du stage, tel que défini dans la convention établie entre Norwegian University of Science and Technology (NTNU), l'ENSTA Bretagne et moi-même, est la **classification automatique des stades de sommeil à partir d'un EEG faiblement invasif** (*Automatic sleep stage classification based on minimally invasive EEG*). Le scoring du sommeil est traditionnellement réalisé manuellement par des experts, qui annotent chaque époque de 30 secondes d'un enregistrement polysomnographique (PSG) selon les règles établies par l'*American Academy of Sleep Medicine* (AASM) et, historiquement, par Rechtschaffen & Kales (R&K). Cette tâche est longue et fastidieuse : automatiser tout ou partie du scoring, en particulier à partir d'un montage EEG réduit et d'un signal unique (par opposition à la polysomnographie complète, qui combine EEG, électro-oculographie et électromyographie), permettrait des solutions de suivi du sommeil beaucoup moins contraignantes qu'un enregistrement en laboratoire.

Comme présenté au chapitre 1, ce sujet prolonge les travaux de l'équipe EEG du laboratoire d'accueil et le projet Dream Decoder, qui étend la question du scoring de sommeil à la détection du contenu mental associé : la présence d'un rêve, puis sa tonalité émotionnelle.

2.2 Une problématique à trois niveaux de difficulté croissante

Ces deux axes – classification des stades de sommeil d'une part, décodage du contenu mental du dormeur d'autre part – définissent une unique question de fond, que le stage aborde par difficulté croissante : *jusqu'où le signal EEG, seul, permet-il de remonter à l'état mental du dormeur ?*

1. **Le stade de sommeil** (Éveil, N1, N2, N3, REM) est l'information la mieux définie physiologiquement : les critères de scoring AASM reposent sur des signatures EEG relativement explicites (fuseaux de sommeil et complexes K pour le stade N2, ondes delta de forte amplitude pour le stade N3, etc.). C'est un problème de classification supervisée classique, avec un consensus fort entre experts humains (d'où l'existence de règles de scoring formalisées), ce qui en fait le point de départ naturel du stage.
2. **La présence d'un rêve** (*dream experience* vs *no experience*) est une information strictement plus difficile à inférer : elle ne correspond à aucun stade de sommeil particulier (des rêves sont rapportés aussi bien en REM qu'en NREM), n'a pas de définition EEG consensuelle comparable aux règles AASM, et ne peut être validée qu'indirectement, par le rapport verbal du dormeur à son réveil – une mesure elle-même sujette à l'oubli et à la reconstruction rétrospective.
3. **La tonalité émotionnelle du rêve** (positive, négative, ou une émotion spécifique comme la colère) ajoute un niveau de difficulté supplémentaire : au-delà de savoir *si*

un contenu mental a eu lieu, il s'agit d'en caractériser une propriété fine et subjective, sur un sous-ensemble encore plus restreint d'enregistrements (seuls les rêves rapportés peuvent être notés émotionnellement), avec un fondement neurophysiologique (l'asymétrie frontale alpha [5], section 4.3) beaucoup moins consensuel que les règles de scoring du sommeil.

Le corps du rapport (chapitre 5) suit exactement cette progression : partie 1 pour le stade de sommeil, partie 2 pour la détection de rêve, partie 3 pour l'émotion. Chaque partie réutilise l'infrastructure méthodologique (extraction de features spectrales, validation croisée, contrôle de la fuite de sujet) validée à l'étape précédente, tout en l'adaptant à une tâche dont le signal est structurellement plus ténu et les données plus rares.

2.3 Enjeux

Enjeu technologique et clinique. Un pipeline de scoring automatique fiable, fonctionnant sur un montage EEG réduit, répond à un besoin concret : réduire la charge de travail des experts du sommeil et ouvrir la voie à des dispositifs de suivi portables, utilisables hors du laboratoire. Étendre ce pipeline à la détection de rêve puis à son contenu émotionnel a un intérêt clinique direct pour le suivi de troubles où l'activité onirique joue un rôle diagnostique ou pronostique – cauchemars récurrents, trouble de stress post-traumatique, dépression [1].

Enjeu scientifique. Chacun des trois niveaux ci-dessus correspond à une question de recherche ouverte, avec un état de l'art propre : le scoring automatique du sommeil est un domaine mature où l'enjeu porte sur la réduction du nombre de capteurs sans perte de performance ; la détection de rêve et sa classification émotionnelle sont des domaines beaucoup plus jeunes, où la question de savoir si un signal exploitable existe seulement reste, pour partie, ouverte.

Enjeu méthodologique. Les trois tâches partagent une même contrainte : les jeux de données disponibles sont de petite taille (quelques dizaines de sujets), ce qui rend l'évaluation particulièrement sensible aux biais classiques de l'apprentissage automatique appliqué à l'EEG (fuite de sujet entre les ensembles d'entraînement et de test, déséquilibre des classes, seuils de décision optimisés a posteriori). Ce risque croît avec le niveau de difficulté : il est faible pour le scoring de sommeil (données abondantes, signal net), et devient central pour la détection d'émotion (peu d'exemples, signal ténu). Une part importante du travail réalisé porte donc sur la construction d'une évaluation à laquelle on peut faire confiance à chacun des trois niveaux, et pas seulement sur la performance brute obtenue.

Chapitre 3

Le cerveau, le sommeil et le rêve : éléments de neuroscience

Ce chapitre présente les bases biologiques nécessaires à la compréhension du travail décrit au chapitre 5. Il s'appuie notamment sur le chapitre théorique des thèses de référence du laboratoire [1, 3].

3.1 Anatomie cérébrale

Le cerveau humain, avec le tronc cérébral et le cervelet, constitue l'essentiel du système nerveux central. Il est composé de milliards de neurones communiquant par synapses au moyen de signaux électriques [1]. Le cerveau (cerveau proprement dit, ou *cerebrum*) est la plus grande division encéphalique ; sa couche externe, le cortex cérébral (substance grise), est organisée en quatre lobes fonctionnels :

- le **lobe frontal**, siège des fonctions exécutives (planification, prise de décision) et du cortex moteur ;
- le **lobe pariétal**, qui traite l'information somatosensorielle, c'est-à-dire les sensations issues du corps (toucher, pression, température, douleur, position des membres) ;
- le **lobe temporal**, impliqué dans l'audition, la mémoire et la compréhension du langage ;
- le **lobe occipital**, siège du cortex visuel.

Les deux hémisphères cérébraux, bien que morphologiquement proches, ne sont ni parfaitement symétriques ni fonctionnellement équivalents : l'hémisphère gauche est typiquement dominant pour le langage et le traitement séquentiel, le droit pour le traitement visuo-spatial et émotionnel [1]. Cette latéralisation fonctionnelle est directement à l'origine de la mesure d'asymétrie frontale utilisée au chapitre 5 (section 4.3).

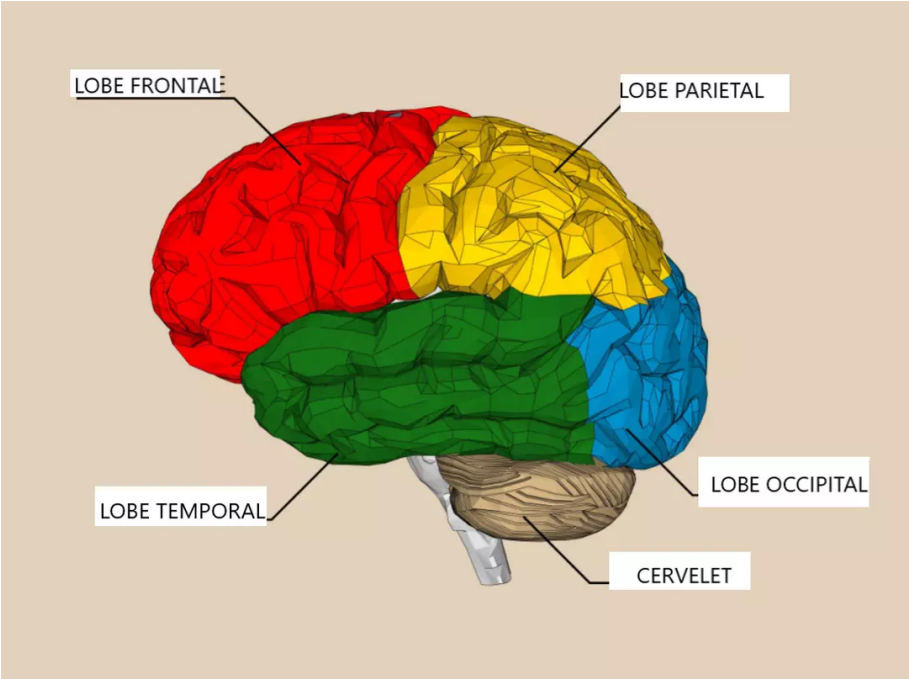


Figure 3.1 – Les quatre lobes du cortex cérébral et le cervelet.

3.2 Le sommeil et son architecture

Le sommeil occupe environ un tiers de la vie humaine et est régulé conjointement par le rythme circadien et la pression homéostatique de sommeil [1]. La polysomnographie (PSG), qui combine EEG, électro-oculographie (EOG) et électromyographie (EMG) sur des époques de 30 secondes, en reste la méthode de référence. Les règles de l'*American Academy of Sleep Medicine* (AASM) distinguent cinq stades :

Éveil EEG de fréquence mixte, clignements fréquents, tonus EMG mentonnier élevé ; rythme alpha dominant sur les régions occipitales, yeux fermés.

N1 Stade le plus léger ; les rythmes alpha cèdent la place à une activité de basse amplitude et de fréquence mixte, avec des mouvements oculaires lents. Représente 2 à 5 % du temps de sommeil total.

N2 Identifié par les fuseaux de sommeil (sleep spindles : bouffées d'oscillations de 11–16 Hz durant au moins 0,5 s) et les complexes K (K-complexes : onde biphasique ample, Figure 3.2). Le tonus musculaire continue de décroître. Stade le plus représenté (45–55 % du sommeil).

N3 Sommeil lent profond, défini par plus de 20 % de l'époque occupée par des ondes delta de forte amplitude ($\geq 75\,\mu V$, 0.5–2 Hz). Associé à la récupération physiologique (ralentissement cardiaque et respiratoire, clairance métabolique gliolymphatique).

REM *Rapid Eye Movement* : EEG de basse amplitude et fréquence mixte, proche de l'éveil, avec bouffées de mouvements oculaires rapides et quasi-atonie musculaire. Ondes en dents de scie (*sawtooth*, 2–6 Hz) fréquentes sur les dérivations centrales. Représente 20–25 % du sommeil et est associé aux rêves les plus vivaces et à la consolidation de la mémoire émotionnelle.

Le sommeil progresse cycliquement à travers ces stades, chaque cycle durant 90 à 110 minutes et se répétant 4 à 6 fois par nuit ; la première moitié de la nuit est dominée par le sommeil lent profond (N3), la seconde par le sommeil REM [1]. C'est cette architecture cyclique, et non

une succession aléatoire de stades, que le pipeline de la partie 1 (section 5.1) doit apprendre à reconnaître époque par époque.

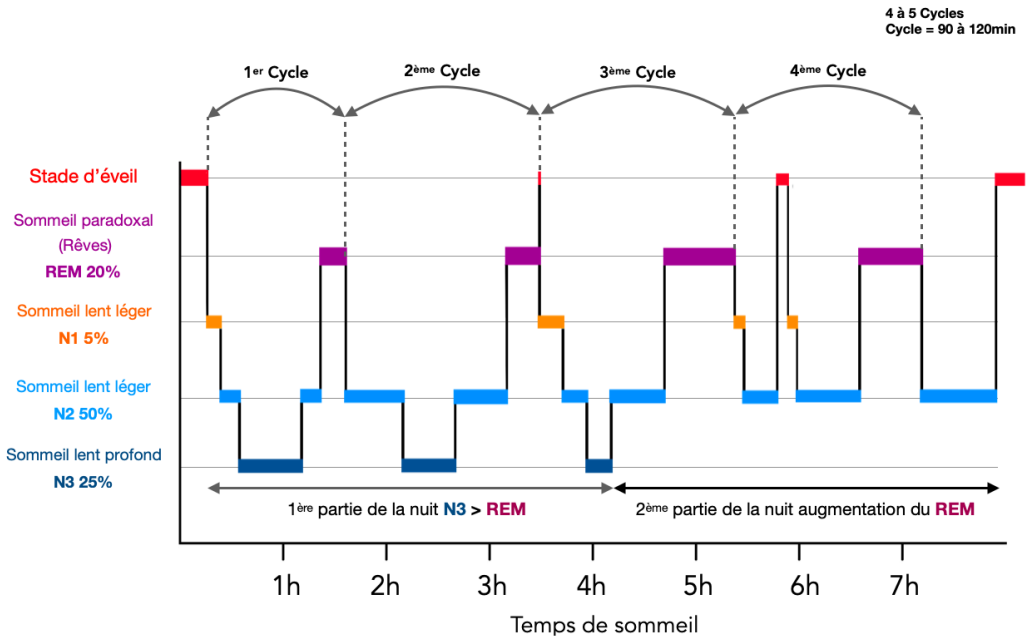


Figure 3.2 – Répartition des stades de sommeil au cours d'une nuit type, par cycle.

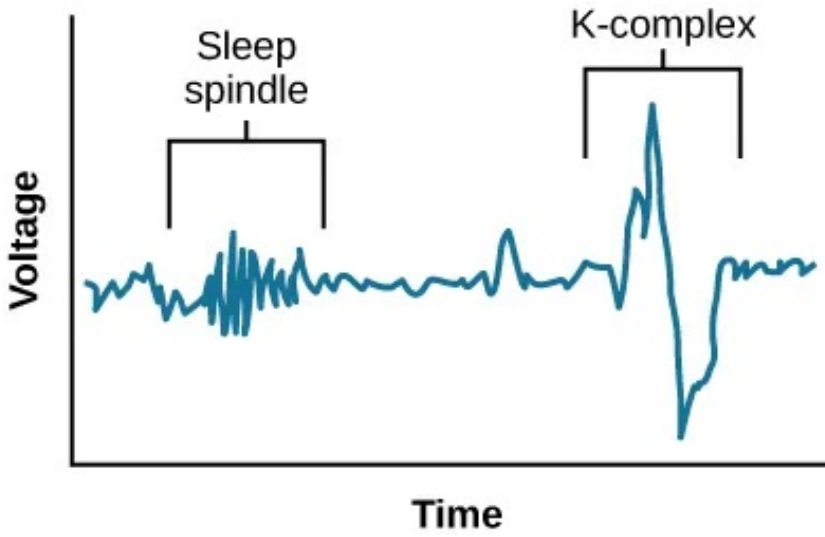


Figure 3.3 – Exemple d'un fuseau de sommeil (*sleep spindle*) et d'un complexe K (*K-complex*).

3.3 Le rêve : neurophysiologie et enjeux cliniques

Le contenu mental du sommeil demeure, malgré son universalité, incomplètement compris. Plusieurs théories tentent d'expliquer sa fonction : l'interprétation psychanalytique historique de Freud (le rêve comme reflet inconscient de conflits non résolus) ; les théories contemporaines de consolidation mnésique, qui associent la réactivation des patterns neuronaux hippocampiques pendant le sommeil à la consolidation de mémoires épisodiques et spatiales ; et l'hypothèse, proposée par Francis Crick, d'un mécanisme d'*apprentissage inversé* éliminant les associations neuronales redondantes [1]. Aucune de ces théories n'est aujourd'hui pleinement tranchée, mais toutes s'accordent sur le rôle cognitif et émotionnel central du rêve.

Le contenu et la probabilité de rappel des rêves diffèrent nettement entre REM et NREM. En REM, 80 à 90 % des réveils provoqués donnent lieu à un rapport de rêve, typiquement vivace, riche en imagerie visuelle et émotionnellement marqué, formant des récits narratifs peu liés aux événements immédiats de la veille. En NREM, en particulier au stade N2, le taux de rappel avoisine 50 % ; les rêves y sont plus courts, plus conceptuels, moins chargés émotionnellement, et reflètent davantage les préoccupations récentes du dormeur [1]. Cette asymétrie REM / NREM justifie a posteriori le choix des jeux de données mobilisés aux parties 2 et 3 du chapitre 5 : le sommeil REM pour la détection d'émotion (REM_Turku, Zhang & Wamsley), le NREM pour l'étude complémentaire de la détection de rêve (Tononi).

Au-delà de l'intérêt fondamental, la caractérisation automatique du contenu du rêve a une portée clinique directe. Jusqu'à 63 % des patients souffrant de trouble de stress post-traumatique rapportent des troubles du sommeil incluant insomnie et cauchemars ; les personnes souffrant d'insomnie ont un risque de développer une dépression multiplié par deux ; schizophrénie, trouble bipolaire et risque suicidaire ont tous été associés aux troubles du sommeil et à la présence de cauchemars [1]. Un outil capable de caractériser automatiquement, à partir du seul EEG, la survenue et la tonalité émotionnelle des rêves constituerait potentiellement un indicateur de suivi précoce pour ces populations.

3.4 L'électroencéphalographie

L'EEG mesure, au moyen d'électrodes de surface, la différence de potentiel générée par l'activité synchrone de larges populations de neurones pyramidaux corticaux : lorsque ces neurones, dont les dendrites sont orientées perpendiculairement à la surface corticale, déchargent en phase, leurs courants post-synaptiques s'additionnent en dipôles dont le champ électrique traverse le crâne et le cuir chevelu et devient mesurable en surface [1]. Chaque trace EEG représente ainsi l'agrégat temporel de milliers de potentiels post-synaptiques sous l'électrode. Cette technique, dont les prémices remontent à la découverte de l'activité électrique cérébrale par Richard Caton en 1875 et à sa première mesure non invasive chez l'humain par Hans Berger dans les années 1920, reste aujourd'hui la méthode de référence pour l'étude du sommeil, en raison de sa résolution temporelle fine et de son caractère non invasif.

L'EEG est cependant également sensible à des signaux non neuronaux : mouvements oculaires (EOG), activité musculaire (EMG), pouls cardiaque (ECG) et bruit électrique environnemental. Cette sensibilité aux artefacts, documentée dans la littérature de référence [1], n'est pas qu'une remarque théorique : elle est directement à l'origine de l'un des pièges méthodologiques rencontrés au chapitre 5 (section 5.3.1), où un premier classifieur d'émotion s'est révélé apprendre à reconnaître l'intensité des mouvements oculaires plutôt que le contenu émotionnel du rêve.

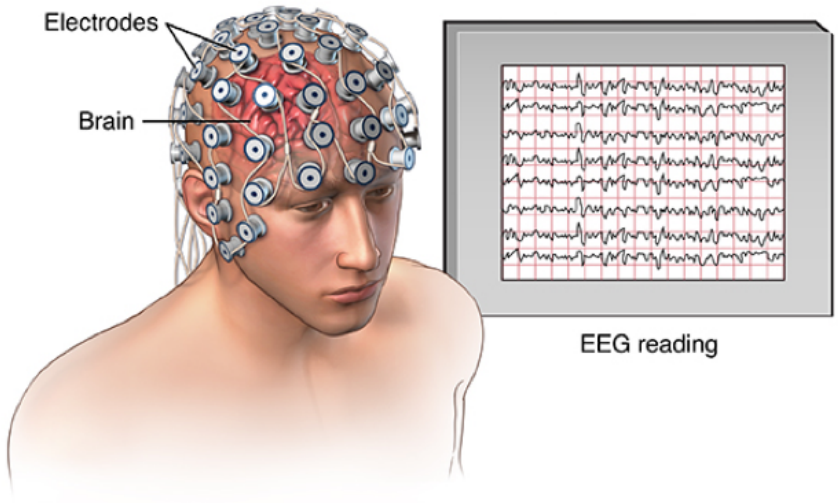


Figure 3.4 – Principe d'acquisition de l'EEG : électrodes de surface et tracé obtenu.

Chapitre 4

Cadre scientifique et méthodologique

Ce chapitre rassemble les outils, formules et notions méthodologiques mobilisés dans le traitement du signal et l'apprentissage automatique appliqués à l'EEG, en cohérence avec le formalisme employé dans les thèses de référence du laboratoire [1, 3]. Le chapitre 3 a présenté le *quoi* (cerveau, sommeil, rêve) ; ce chapitre présente le *comment* (formules et outils utilisés pour l'extraire du signal et évaluer les modèles).

4.1 Recueil du contenu mental : le paradigme du réveil provoqué

La méthode expérimentale standard pour associer un signal EEG à un contenu mental consiste à réveiller le participant à un moment contrôlé du sommeil (*awakening paradigm*) et à recueillir immédiatement un rapport verbal ou un questionnaire structuré sur l'expérience vécue. Le jeu de données REM_Turku, mobilisé pour la détection d'émotion (partie 3), utilise l'échelle mDES (*modified Differential Emotions Scale*), qui recueille dix scores d'affect positif (PA1–PA10) et dix scores d'affect négatif (NA1–NA10), notés par le participant immédiatement après le réveil.

4.2 Analyse spectrale de l'EEG

Le signal EEG est classiquement décomposé en bandes de fréquence associées à différents états fonctionnels (Table 4.1).

Table 4.1 – Bandes de fréquence EEG utilisées dans ce rapport

Bande	**Fréquence**
Delta (δ)	0.5–4 Hz
Thêta (θ)	4–8 Hz
Alpha (α)	8–12 Hz
Bêta (β)	12–30 Hz
Gamma (γ)	$>$ 30 Hz

La densité spectrale de puissance (*Power Spectral Density*, PSD) d'un signal quantifie la répartition de sa puissance en fonction de la fréquence. L'estimateur de base, le périodogramme, se calcule à partir de la transformée de Fourier discrète (TFD) du signal $x(n)$ de N échantillons :

$$X_k = \sum_{n=0}^{N-1} x(n)\, e^{-i2\pi k n N^{-1}}, \qquad PSD(k) = \frac{1}{N}\left|\hat{X}(k)\right|^2 \tag{4.1}$$

Cet estimateur est cependant bruité pour un signal non stationnaire comme l'EEG. La méthode de Welch, retenue dans l'ensemble des pipelines du stage, réduit cette variance en segmentant le signal en fenêtres courtes et recouvrantes, en calculant le périodogramme de chaque segment, puis en moyennant les résultats – au prix d'une résolution fréquentielle légèrement réduite.

Au-delà des puissances de bande brutes, plusieurs features dérivées ont été mobilisées : la puissance relative (ratio d'une bande sur la puissance totale, moins sensible aux variations d'impédance inter-sujets), l'entropie spectrale (mesure de la platitude du spectre), la dimension fractale de Higuchi (mesure de la complexité temporelle du signal), et la cohérence spectrale entre canaux (mesure de synchronisation inter-régions à une fréquence donnée).

4.3 Asymétrie frontale alpha et affect du rêve

L'asymétrie frontale alpha (*Frontal Alpha Asymmetry*, FAA) est définie comme la différence entre le logarithme de la puissance alpha de l'hémisphère droit et celui de l'hémisphère gauche, typiquement mesurée entre les électrodes F4 et F3 :

$$\mathrm{FAA} = \log P_\alpha(F4) - \log P_\alpha(F3) \tag{4.2}$$

Une puissance alpha élevée traduisant une moindre activation corticale, une FAA positive indique une activation relative de l'hémisphère gauche.

Sikka et al. [5] ont montré, sur le jeu de données REM_Turku ($n = 17$ sujets), que la FAA mesurée en sommeil REM est corrélée positivement à l'intensité de la colère rapportée dans le rêve ($R^2 = 0{,}604$, $p = 0{,}006$), en cohérence avec le modèle de valence hémisphérique de l'émotion (approche/retrait). La valence est la dimension d'une émotion qui indique si elle est agréable (positive) ou désagréable (négative), par opposition à l'arousal, qui mesure son intensité. Ce résultat constitue la référence directe de la partie 3 du stage (section 5.3).

4.4 Modèles de classification utilisés

XGBoost (*eXtreme Gradient Boosting*) construit un ensemble d'arbres de décision peu profonds de façon séquentielle, chaque nouvel arbre corrigeant l'erreur résiduelle de l'ensemble précédent, avec régularisation L_1/L_2 sur les poids des feuilles. Choisi pour sa robustesse sur des jeux de features tabulaires de petite taille.

Régression logistique modélise directement la probabilité d'une issue binaire par une combinaison linéaire des features passée dans une fonction sigmoïde.

SVM linéaire recherche l'hyperplan séparateur de marge maximale entre les classes dans l'espace des features.

Random Forest agrège les prédictions d'un grand nombre d'arbres de décision non élagués, chacun entraîné sur un sous-échantillonnage aléatoire des observations et des features, pour réduire la variance par rapport à un arbre unique.

Tous les modèles ci-dessus sont entraînés sur des features tabulaires extraites manuellement. Une approche par apprentissage profond (auto-encodeur convolutif 1D pour l'émotion, CNN + LSTM pour le stade de sommeil) a également été mise en œuvre et est décrite directement dans les sections correspondantes du chapitre 5.

4.5 Déséquilibre des classes et SMOTE

Les classes minoritaires (rêve négatif, absence de rêve selon le jeu de données) sont systématiquement sous-représentées, avec un déséquilibre pouvant atteindre un exemple minoritaire

pour cinq à six exemples majoritaires. La technique SMOTE (*Synthetic Minority Oversampling Technique*), qui génère des exemples synthétiques de la classe minoritaire par interpolation dans l'espace des features, est utilisée systématiquement, en veillant à ne l'appliquer qu'à l'intérieur de chaque pli d'entraînement pour ne pas contaminer les données de test.

4.6 Fuite de données et validation croisée

Le risque méthodologique central de ce type d'étude est la *fuite de sujet* : lorsque plusieurs observations (époques, réveils) d'un même sujet se retrouvent réparties entre l'ensemble d'entraînement et l'ensemble de test, le modèle peut apprendre à reconnaître l'identité ou les caractéristiques individuelles du sujet plutôt que le phénomène ciblé, ce qui gonfle artificiellement la performance apparente sans se traduire par une quelconque erreur de code visible.

La parade standard consiste à comparer deux protocoles de validation croisée : une validation en k plis classique (ici 5 plis), qui autorise ce type de fuite si le regroupement par sujet n'est pas respecté, et la validation *Leave-One-Subject-Out* (LOSO), dans laquelle chaque pli de test correspond à l'ensemble des observations d'un unique sujet totalement absent de l'entraînement. L'écart entre les deux métriques renseigne directement sur l'ampleur de la fuite.

Le choix de systématiquement rapporter également le Kappa en 5 plis a une seconde justification, propre à ce stage : c'est la métrique utilisée dans les thèses de référence [1, 3], ce qui permet une comparaison directe des résultats obtenus à ceux de la littérature, tout en la complétant systématiquement par le LOSO pour en vérifier la validité.

4.7 Métriques d'évaluation

Soit TP, TN, FP, FN respectivement les vrais positifs, vrais négatifs, faux positifs et faux négatifs (Figure 4.1).

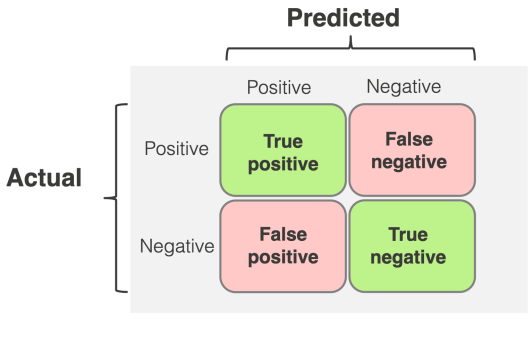


Figure 4.1 – Matrice de confusion et définition de TP, FN, FP, TN.

Exactitude (*accuracy*). Proportion de prédictions correctes ; trompeuse en cas de déséquilibre des classes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

Précision et rappel. La précision mesure la proportion de prédictions positives réellement correctes ; le rappel, la proportion de positifs réels correctement identifiés.

$$\text{Précision} = \frac{TP}{TP + FP}, \qquad \text{Rappel} = \frac{TP}{TP + FN} \tag{4.4}$$

Score F1. Moyenne harmonique de la précision et du rappel, préférée à l'exactitude lorsque les classes sont déséquilibrées.

$$F1 = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{4.5}$$

Kappa de Cohen. Mesure l'accord observé entre prédiction et réalité, corrigé de l'accord attendu par hasard :

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.6}$$

où p_o est la proportion d'accord observée (l'exactitude) et p_e la proportion d'accord attendue sous l'hypothèse d'une association purement aléatoire entre prédictions et labels, compte tenu de leurs distributions marginales respectives. κ varie entre -1 (désaccord complet) et 1 (accord parfait), 0 correspondant à un étiquetage aléatoire. C'est la métrique principale retenue dans ce rapport : contrairement à l'exactitude, un modèle qui prédit toujours la classe majoritaire obtient un Kappa nul, quel que soit le déséquilibre des classes.

AUROC. Aire sous la courbe ROC (taux de vrais positifs contre taux de faux positifs, pour tous les seuils de décision possibles), quantifiant la capacité de séparation du modèle indépendamment d'un seuil fixé.

4.8 Contrôle statistique de la significativité

Une métrique ponctuelle (Kappa, F1, accuracy) calculée sur un unique passage d'une validation croisée reste une estimation, dont la variance peut être élevée sur des échantillons de moins de trente sujets. Deux outils complémentaires, utilisés au chapitre 5 chaque fois que la taille de l'échantillon le justifie, permettent de quantifier cette incertitude :

- l'**intervalle de confiance bootstrap** : ré-échantillonnage avec remise au niveau sujet (pas au niveau des observations individuelles, pour respecter la structure de dépendance de la validation LOSO), recalcul de la métrique sur chaque ré-échantillon, et lecture des percentiles de la distribution obtenue ;
- le **test de permutation** : mélange aléatoire des labels (détruisant toute relation réelle entre les features et le label), ré-exécution du protocole complet, et comparaison de la métrique observée à la distribution ainsi obtenue sous l'hypothèse nulle. La proportion de permutations égalant ou dépassant la métrique observée constitue la valeur p.

La valeur p est la probabilité d'obtenir, sous l'hypothèse nulle (absence de relation réelle entre les features et le label), un résultat au moins aussi bon que celui observé. Dans un test de permutation, elle s'estime directement : c'est la proportion de permutations dont la métrique égale ou dépasse la métrique observée. Par convention, on retient le seuil de 0,05 : si $p < 0{,}05$, l'hypothèse nulle est rejetée et le résultat est dit statistiquement significatif, ce qui signifie qu'on accepte d'attribuer au hasard, au plus, un résultat sur vingt. La valeur p ne mesure ni la probabilité que le résultat soit « vrai », ni l'ampleur de l'effet, et un résultat proche du seuil reste fragile.

Chapitre 5

Description de l'activité et des résultats

Conformément à la problématique posée au chapitre 2, ce chapitre est organisé selon les trois niveaux de difficulté croissante annoncés section 2.2 : la classification des stades de sommeil (partie 1), la détection de rêve (partie 2), puis la détection de l'émotion du rêve (partie 3). Dans la mesure du possible, chaque partie cherche à réutiliser et adapter l'infrastructure méthodologique de la précédente (validation LOSO, contrôle de la fuite de sujet – chapitre 4). Cette continuité n'est cependant pas systématique : le passage de la partie 1 à la partie 2, en particulier, correspond aussi à un changement d'approche (d'un apprentissage profond sur signal brut à un pipeline de features spectrales manuelles), rendu nécessaire par la nature différente de la tâche et par la méthodologie de référence disponible pour la détection de rêve. Chaque partie fait par ailleurs face à un signal structurellement plus ténu et à des données plus rares que la précédente.

Le tableau 5.1 résume les jeux de données, méthodes et résultats principaux de chacune des trois parties.

Table 5.1 – Vue d'ensemble des trois parties du stage

Partie	**Tâche**	**Jeu(x) de données**	**Résultat clé**
1	Stade de sommeil (5 classes)	Sleep-EDF Expanded (153 sujets)	CNN + LSTM fine-tuné : Kappa LOSO = 0,83
2	Détection de rêve (binaire)	Multiple Awakenings, Tononi, Zhang & Wamsley	Kappa 5-fold jusqu'à 0,828
3	Émotion du rêve (colère, valence)	REM_Turku	Colère : accuracy = 61 %, p = 0,048 (significatif) ; valence générale : non significatif

La progression de la partie 1 à la partie 3 n'est pas qu'organisationnelle : elle correspond à une dégradation progressive et attendue du rapport signal/bruit disponible. Le stade de sommeil est prédit avec un excellent niveau d'accord (partie 1) ; la détection de rêve, plus subtile, résiste bien en validation croisée classique mais s'effondre en LOSO sur les trois jeux de données testés (partie 2), révélant une fuite de sujet massive ; l'émotion du rêve, enfin, ne livre un signal statistiquement défendable que sur une cible et un jeu de features spécifiques, identifiés après un travail de sélection et de validation rigoureux (partie 3).

5.1 Partie 1 — Classification des stades de sommeil

5.1.1 Méthodologie

Conformément au sujet officiel du stage (chapitre 2), cette première partie vise à classifier automatiquement les cinq stades de sommeil (Éveil, N1, N2, N3, REM) à partir d'un signal EEG réduit à un **unique canal**, en cohérence avec l'objectif de solutions de suivi du sommeil minimalement invasives. Le travail est mené sur le jeu de données Sleep-EDF Expanded (PhysioNet), 153 sujets, en réservant un ensemble de sujets totalement indépendant pour l'évaluation finale (108 sujets en entraînement, 27 à 31 en test selon les expériences, aucun chevauchement de sujet – conformément au principe de la section 4.6).

Quatre familles de modèles ont été comparées, par complexité croissante :

1. **Random Forest** sur features spectrales (2 classes puis 5 classes) ;
2. **XGBoost** sur les mêmes features (5 classes) ;
3. **CNN 1D** directement sur le signal brut mono-canal (3000 échantillons par époque de 30 s), sans extraction manuelle de features ;
4. **CNN + LSTM**, où l'encodeur convolutif du CNN précédent extrait une représentation par époque, puis un LSTM bidirectionnel modélise la dépendance temporelle entre époques successives d'une même nuit, avant un fine-tuning de bout en bout.

5.1.2 Résultats

La figure 5.1 synthétise la progression du Kappa selon le modèle utilisé.

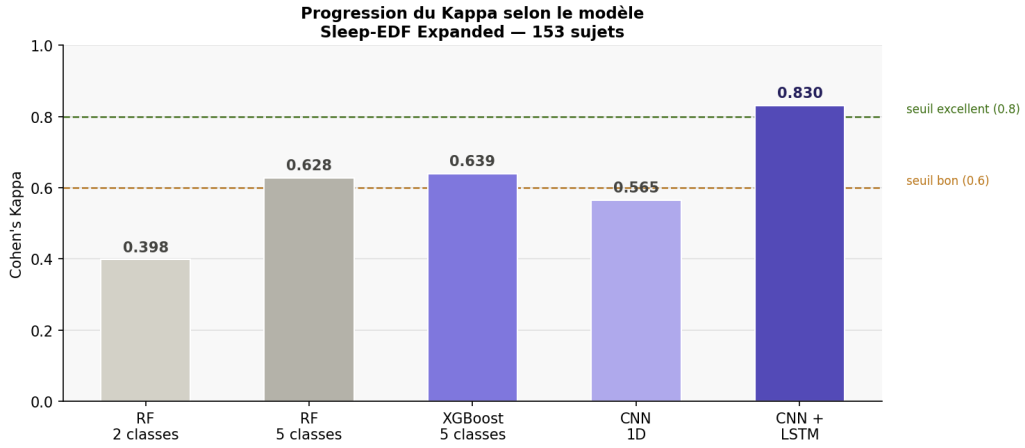


Figure 5.1 – Progression du Kappa selon le modèle, stades de sommeil, Sleep-EDF Expanded (153 sujets).

Le passage d'un classifieur classique sur features manuelles (Random Forest, Kappa = 0,63 en 5 classes) à un modèle exploitant directement le signal brut et sa structure temporelle (CNN + LSTM, Kappa = 0,83) apporte un gain net et franchit le seuil d'accord « excellent » ($\kappa > 0{,}8$, section 4.7). Le CNN 1D seul, sans modélisation temporelle inter-époques, plafonne à 0,57 : c'est bien l'ajout du LSTM – qui exploite le fait que les stades de sommeil ne se succèdent pas aléatoirement mais suivent l'architecture cyclique décrite section 2.2 – qui explique le gain final, non l'architecture convolutive seule.

5.1.3 Bilan

Ce premier résultat valide, sur le sujet officiellement défini par la convention de stage, la faisabilité d'un scoring automatique de haute qualité à partir d'un signal EEG unique. Il constitue

la brique méthodologique de base (prétraitement du signal, entraînement d'un encodeur convolutif, validation par sujets disjoints) réutilisée et adaptée aux parties 2 et 3, où la tâche cible – la détection de rêve, puis de son émotion – est nettement plus difficile à isoler du signal EEG que le stade de sommeil.

5.2 Partie 2 — Détection de rêve

Après le stade de sommeil (partie 1), le niveau de difficulté suivant est de détecter si un rêve a eu lieu, indépendamment de son contenu. Cette question est traitée sur trois jeux de données indépendants : *Multiple Awakenings*, Tononi et Zhang & Wamsley (ZW).

5.2.1 Multiple Awakenings : identification et correction d'une erreur de labellisation

Le jeu de données *Multiple Awakenings* (Scarpelli et al. [4], Sapienza Università di Roma, 28 canaux EEG, 250 Hz, 456 réveils, 19 sujets) a été choisi pour sa compatibilité de montage avec REM_Turku. L'examen du fichier de métadonnées utilisé par le pipeline existant a révélé que les 455 enregistrements exploitables étaient tous étiquetés « rêve » : la colonne source `Experience` utilisait les codes 2 = rêve, 0 = absence de rêve, −2 = incertain, mais le script de construction des métadonnées avait mappé toute valeur non nulle vers 1. Le classifieur entraîné sur ces données n'avait donc aucun exemple négatif issu de *Multiple Awakenings* ; les seuls exemples « absence de rêve » provenaient d'un jeu de données de sommeil sans protocole de rêve et avec un matériel EEG différent, si bien que le modèle apprenait probablement à distinguer les dispositifs d'acquisition plutôt que le phénomène cognitif visé.

Après correction des labels à partir du fichier source (314 rêve, 121 absence de rêve, 20 incertains exclus) et filtrage au sommeil REM, 154 enregistrements exploitables restaient (125 rêve, 29 absence de rêve). Le pipeline PSD + XGBoost (filtrage 0,5–30 Hz, segmentation 2 s, 19 canaux communs avec REM_Turku, features spectrales – section 4.2) a été exécuté en 5 plis et en LOSO :

Table 5.2 – Détection de rêve, *Multiple Awakenings* (REM uniquement, 154 enregistrements)

Protocole	Exactitude	Kappa	Spécificité (absence de rêve)
5-fold CV	68,5 %	+0,20	40,5 %
LOSO (19 sujets)	67,1 %	+0,10	26,5 %

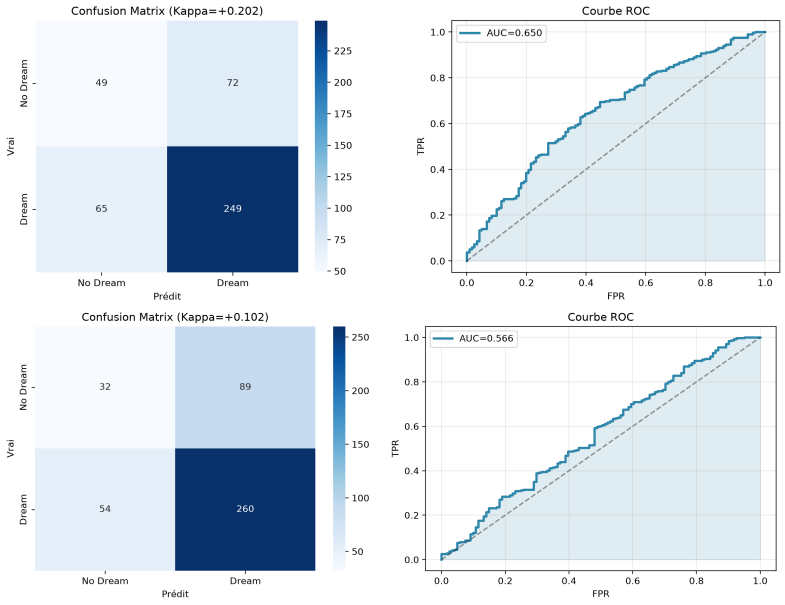


Figure 5.2 – Multiple Awakenings – comparaison 5-fold (haut) et LOSO (bas).

L'écart entre le Kappa 5-fold (+0,20) et le Kappa LOSO (+0,10) illustre directement la fuite

de sujet décrite section 4.6. La variabilité inter-sujets en LOSO est en outre considérable (Kappa allant de +0,65 pour le sujet le mieux classé à −0,45 pour le moins bien classé), ce qui suggère que les corrélats EEG du rêve varient fortement d'un individu à l'autre.

5.2.2 Réplication fidèle sur Tononi et Zhang & Wamsley

Pour consolider ce résultat, la méthodologie de référence du laboratoire [3] a été reproduite aussi fidèlement que possible sur les jeux de données Tononi et Zhang & Wamsley (ZW), avec deux adaptations : l'exclusion des rêves « DEWR » (*Dream Experience Without Recall*, rapports ambigus), et une sélection de canaux par permutation reproduisant l'algorithme d'importance des thèses de référence – le nombre de canaux conservés étant ré-optimisé à chaque pli de la validation LOSO pour éviter toute fuite d'information dans le choix des canaux eux-mêmes. L'inspection des données a par ailleurs montré que le jeu de données Tononi ne contient aucun stade REM (100 % des enregistrements sont en stade N2) ; l'analyse a donc été conduite en NREM pour Tononi et en REM pour ZW, conformément à la nature réelle de chaque jeu de données.

Table 5.3 – Détection de rêve, méthodologie répliquée avec sélection de canaux

Jeu de données	Stade	Canaux retenus	Kappa 5-fold	Kappa LOSO
Tononi (hors DEWR)	NREM (N2)	50	+0,792	−0,040
Zhang & Wamsley	REM	58	+0,824	−0,065

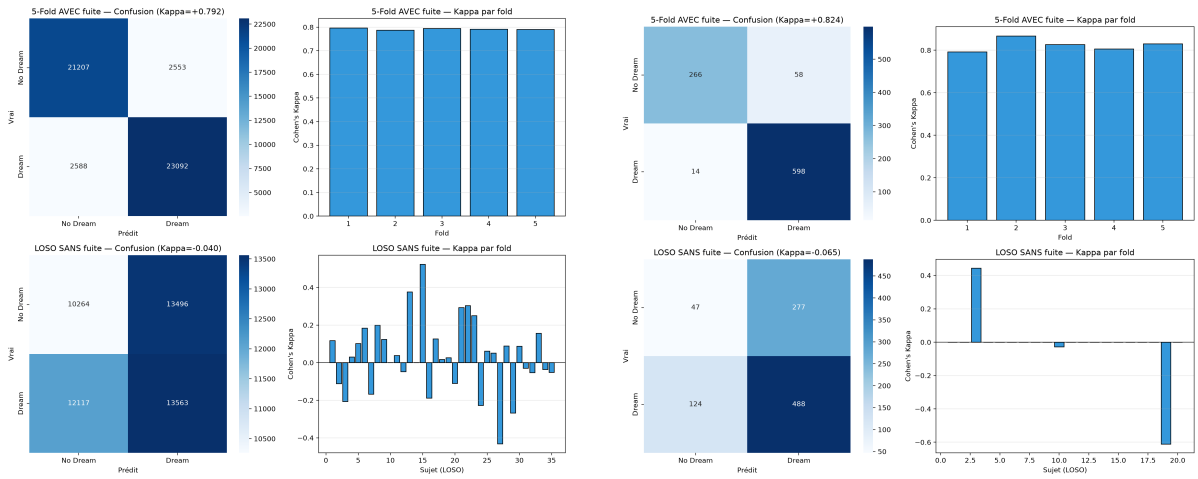


Figure 5.3 – Résultats de détection de rêve – Tononi (gauche) et Zhang & Wamsley (droite).

Le résultat 5-fold (Kappa ≈ 0,80) est comparable à celui rapporté dans la littérature de référence – notamment à celui obtenu par Beslagic [1] sur le même jeu de données Tononi (Kappa ≈ 0,96 en validation aléatoire contre ≈ 0,21 en LOSO) –, ce qui confirme la fidélité de la réplication et situe nos résultats dans la fourchette attendue. Le résultat LOSO, en revanche, s'effondre à une valeur proche de zéro, voire légèrement négative, sur les deux jeux de données. Cet écart – observé de façon reproductible sur trois jeux de données indépendants au total avec *Multiple Awakenings* (et sur un cinquième dans la littérature de référence) – constitue l'un des enseignements méthodologiques les plus robustes du stage : **la performance en validation croisée classique de ce type de pipeline provient essentiellement de la reconnaissance de caractéristiques propres à chaque sujet, et non de la reconnaissance du phénomène cognitif ciblé**. La sélection de canaux par permutation (Figure 5.4) reste néanmoins utile : elle identifie les régions les plus informatives selon le critère 5-fold et permet une

comparaison directe avec la topographie attendue (Figure 5.5), même si sa validité sur nouveau sujet reste à démontrer.

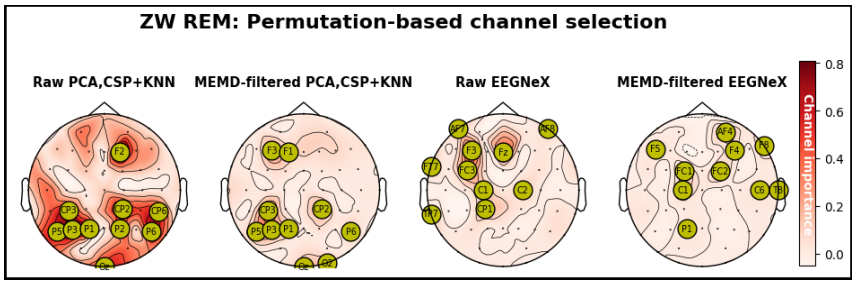


Figure 5.4 – Importance des canaux par permutation, Zhang & Wamsley (REM).

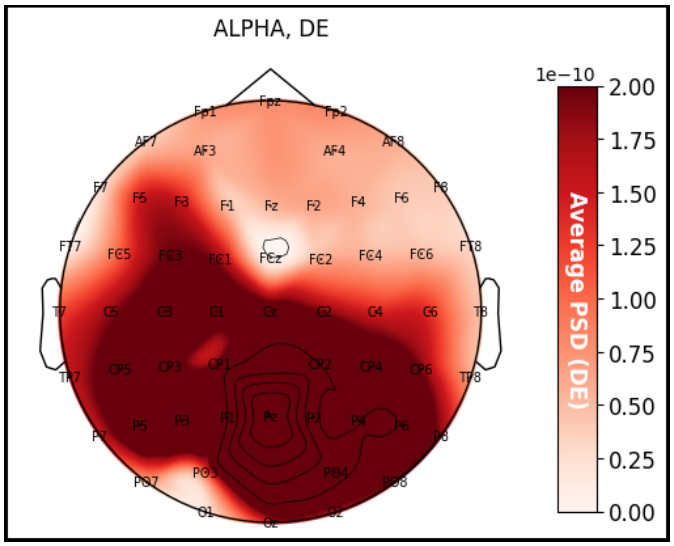


Figure 5.5 – Topographie de la puissance alpha, rêves avec contenu (DE) vs sans contenu (NE), Zhang & Wamsley.

5.3 Partie 3 — Détection de l'émotion du rêve

Le troisième et dernier niveau de difficulté, conformément à la section 2.2, consiste à caractériser non plus la présence d'un rêve mais sa tonalité émotionnelle. Cette partie, la plus développée du rapport, est conduite intégralement sur le jeu de données REM_Turku (134 réveils provoqués en REM, 18 sujets, EEG 29 canaux à 500 Hz), seul jeu de la base DREAM à fournir des scores d'affect détaillés par item (échelle mDES, section 4.2).

5.3.1 Premiers essais et alerte méthodologique

Un premier pipeline (SVM linéaire, LOSO, 8 features incluant l'énergie des signaux EOG et EMG) a obtenu une exactitude de 47 %, inférieure au hasard (63 %). L'analyse des poids du modèle a montré que la feature dominante était l'énergie EOG (mouvements oculaires) : le modèle apprenait à reconnaître l'intensité des mouvements oculaires rapides, caractéristiques du sommeil REM par définition (chapitre 3), mais sans lien causal avec l'émotion du rêve. Le retrait des canaux EOG et EMG a ramené le modèle à un comportement honnête mais peu performant ; le passage à XGBoost avec agrégation par vote majoritaire au niveau de l'enregistrement a ensuite produit un signal faible mais réel ($F1_{\text{négatif}} = 0{,}32$).

Une approche par représentation riemannienne (matrices de covariance inter-canaux, classification par distance au centre – MDM) a également été testée ; sans mécanisme de rééquilibrage des classes comparable à SMOTE, elle a obtenu des résultats inférieurs à XGBoost (MCC = +0,14 contre $\approx$ 0,32), le centre riemannien de la classe minoritaire étant estimé sur trop peu d'exemples pour être représentatif.

Ces deux essais, bien que peu performants en eux-mêmes, ont eu une valeur méthodologique déterminante : ils ont établi la nécessité de contrôler strictement la nature des features utilisées (risque de confusion avec des artefacts non neuronaux, chapitre 3) et le déséquilibre des classes, deux précautions appliquées systématiquement dans la suite de cette partie.

5.3.2 Construction itérative du pipeline

Quatre versions successives du pipeline ont ensuite été développées, chacune motivée par une hypothèse d'amélioration explicite :

V1 Époques de 30 s, bande alpha large, 6 features, XGBoost seul. Kappa LOSO = +0,066.

V2 Un seul vecteur de features par enregistrement complet, bande alpha resserrée (7–11 Hz, conforme à Sikka et al.), ajout d'une feature d'alpha temporal, comparaison de trois classifieurs. **Meilleur résultat brut sur la valence générale** : régression logistique, Kappa LOSO = +0,183.

V3 Changement de cible vers la *colère* spécifiquement (score `SR_NA1` > 0), motivé par la littérature de référence [5]. Kappa LOSO = +0,129 (XGBoost), $F1 = 0{,}472$.

V4 Ajout de features d'activation frontale (bêta, gamma), seuil de décision optimisé de façon honnête (recherche interne à chaque pli LOSO). Kappa LOSO au seuil nichée = +0,078.

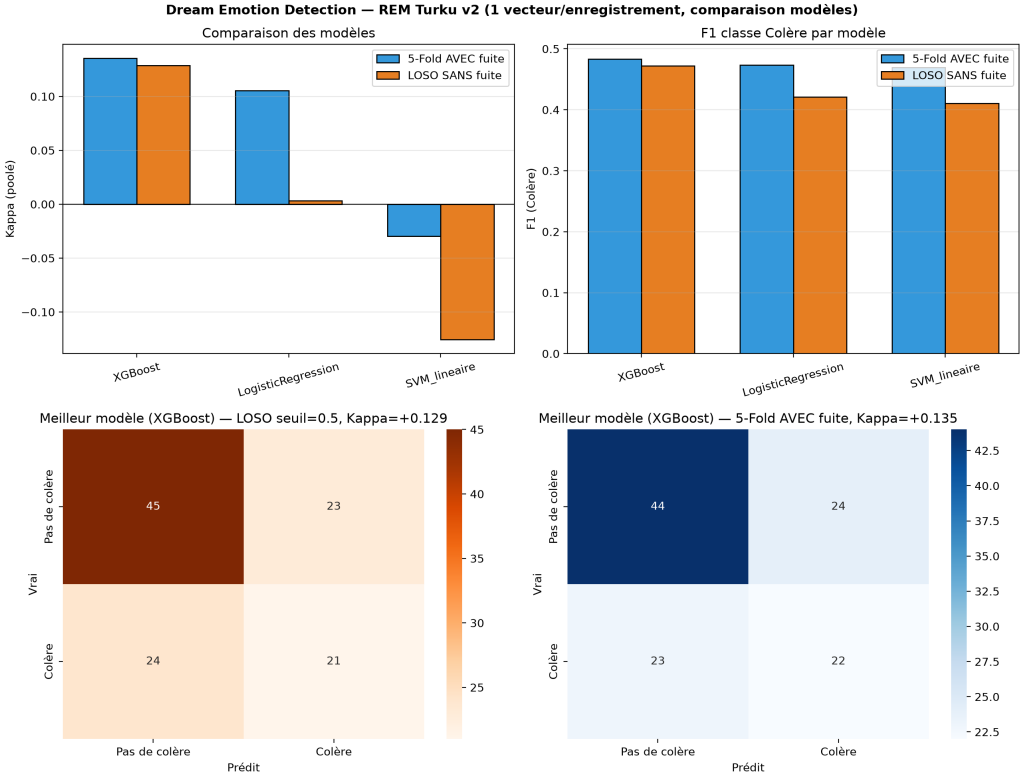


Figure 5.6 – Version 2 du pipeline – comparaison des trois classifieurs (valence générale).

5.3.3 Réplication fidèle d'un travail antérieur

Un travail antérieur mené sur ce même jeu de données (segmentation de 30 s avec recouvrement, cohérence spectrale alpha inter-hémisphérique, Random Forest) rapportait un score $F1 = 0{,}43$. Sa réplication fidèle – y compris un biais de correspondance des canaux identifié dans son code, volontairement conservé pour reproduire exactement les conditions d'obtention du résultat original – a montré qu'au seuil de décision standard, le modèle s'effondre ($F1 = 0$), tandis qu'un seuil optimisé *a posteriori* sur les données de test restitue $F1 = 0{,}389$, proche du score annoncé. Une version honnête de ce seuil (optimisé en interne à chaque pli LOSO) donne $F1 = 0{,}333$, Kappa $= +0{,}146$.

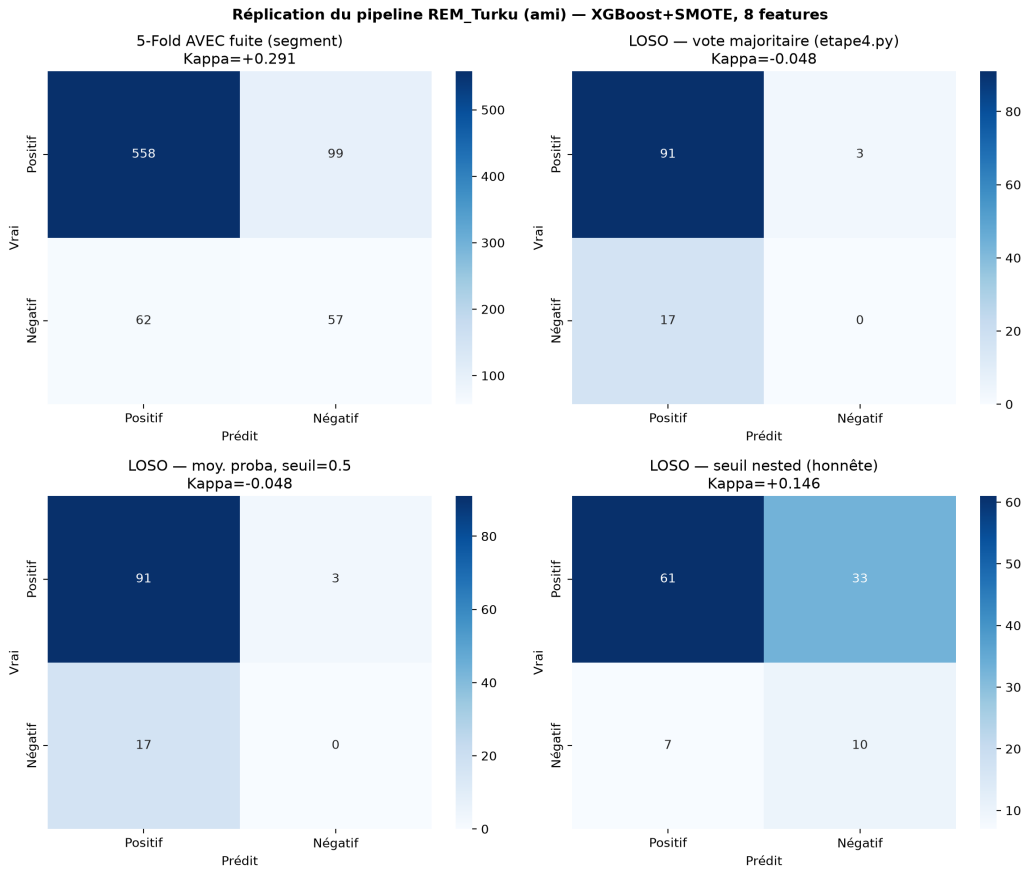


Figure 5.7 – Réplication du pipeline antérieur : effondrement au seuil standard, récupération au seuil optimisé *a posteriori*, résultat honnête au seuil nichée.

Cette réplication retrouve, à environ 10 % près, le résultat publié, tout en suggérant que ce dernier provient probablement d'une optimisation du seuil de décision directement sur les données de test.

5.3.4 Prototype d'apprentissage de représentation

Un auto-encodeur convolutif 1D, pré-entraîné de façon auto-supervisée sur l'ensemble des fenêtres de 4 s disponibles (REM_Turku et Zhang & Wamsley confondus, $\approx 8\,800$ fenêtres), a été testé pour produire une « signature » de dimension 16 par fenêtre, en remplacement ou en complément des features manuelles. Utilisée seule, cette signature obtient un Kappa LOSO négatif ($-0{,}102$) ; combinée aux features manuelles, elle dégrade légèrement le résultat. Le volume de données de pré-entraînement s'est révélé insuffisant pour un apprentissage de représentation non supervisé fiable.

5.3.5 Un signal significatif pour la colère : sélection par corrélation et validation par permutation

Au-delà du pipeline construit itérativement (sections 5.3.1 – 5.3.2), une approche complémentaire a consisté à partir directement des corrélations entre chaque feature spectrale et chacun des 20 scores d'émotion de l'échelle mDES, avec correction de Bonferroni-FDR (*False Discovery Rate*) pour le grand nombre de tests réalisés (125 features $\times$ 20 émotions).

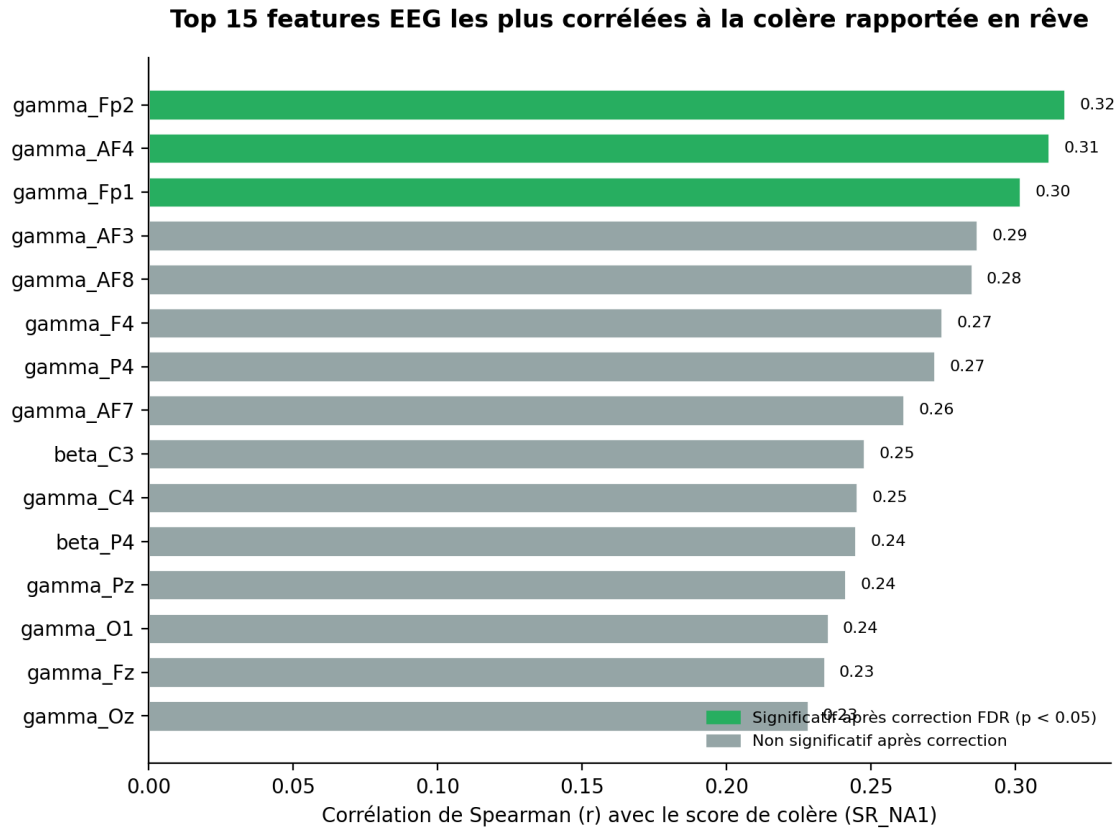


Figure 5.8 – Corrélation de Spearman entre puissance gamma frontale et intensité de colère rapportée – 3 features significatives après correction FDR.

Trois features de puissance gamma frontale (Fp1, Fp2, AF4) ressortent significativement corrélées à l'intensité de colère rapportée après correction ($p_{FDR} < 0{,}05$, Figure 5.8). Un classifieur Random Forest entraîné uniquement sur ces trois features, évalué en LOSO, atteint une **exactitude de 61,1 %** et un $F1$-macro de 0,597, contre respectivement 55,8 % et 0,550 pour un classifieur aléatoire de référence tenant compte du déséquilibre des classes. La sélection des features par corrélation et correction FDR est recalculée à l'intérieur de chaque pli d'entraînement LOSO, sans jamais utiliser le sujet de test. Un test de permutation (500 permutations des labels, ré-exécution complète du protocole LOSO) confirme que ce résultat est statistiquement significatif : $p = 0{,}048$ (Figure 5.9). Autrement dit, seules environ 5% des permutations aléatoires (24 sur 500) obtiennent un F1-macro au moins égal à celui du modèle réel (0,597), alors que la moyenne obtenue par hasard est de 0,496. Le résultat se situe toutefois juste sous le seuil conventionnel de 0,05 et appelle une confirmation sur un échantillon plus large.

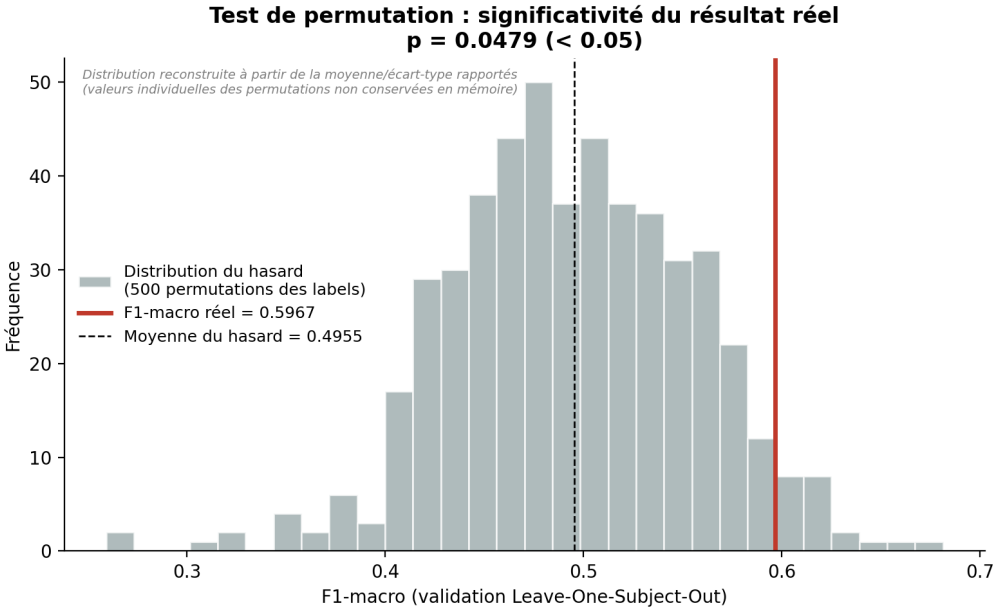


Figure 5.9 – Test de permutation pour le classifieur de colère (Random Forest, 3 features gamma frontales, LOSO) : $p = 0{,}048$.

Ce résultat gagne en solidité lorsqu'il est mis en perspective avec le reste des 20 émotions de l'échelle mDES (Figure 5.10). La fierté (*Pride*) présente pourtant *davantage* de features corrélées après correction FDR (25 contre 3 pour la colère) ; mais un test de permutation identique appliqué à sa classification échoue à atteindre la significativité ($p = 0{,}29$). La colère est ainsi la seule émotion, sur les vingt testées, dont le signal de classification résiste au test de permutation : ce n'est pas la richesse univariée d'une émotion qui prédit sa décodabilité, mais la capacité de ses corrélats à généraliser hors-échantillon.

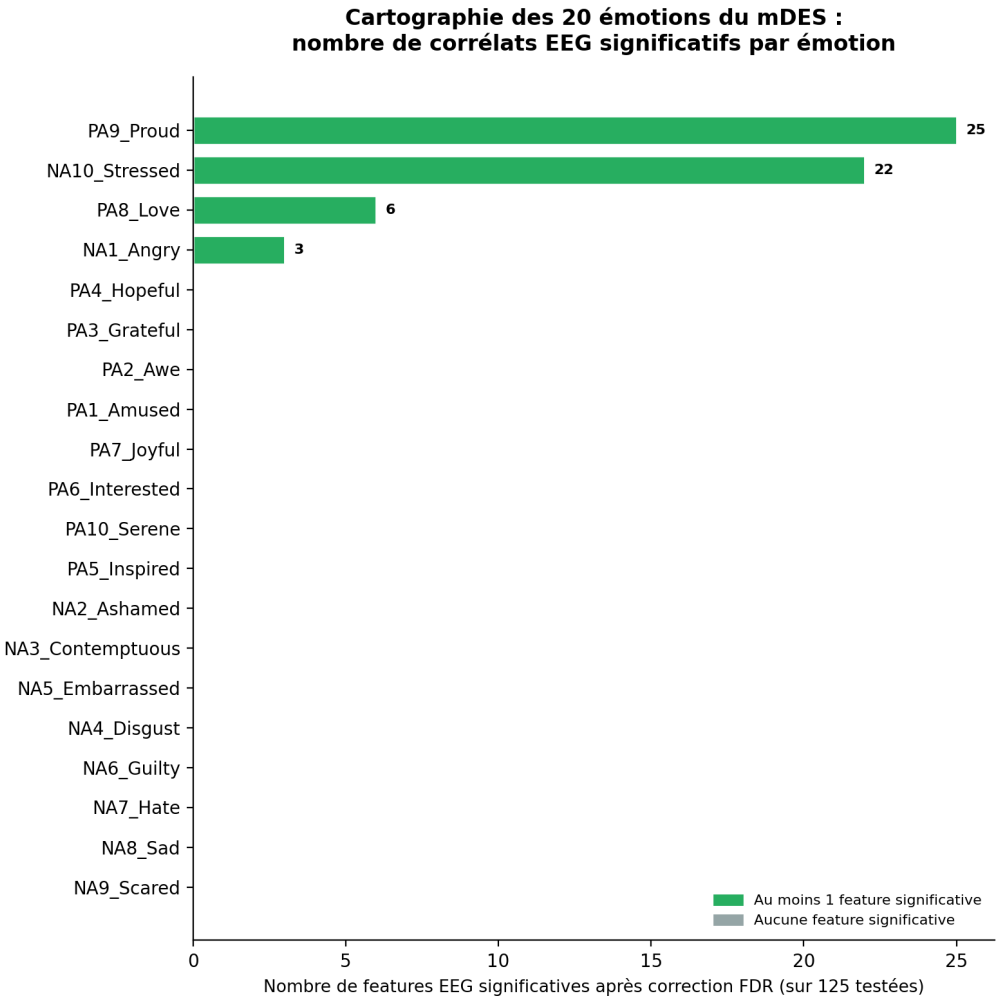


Figure 5.10 – Nombre de corrélats EEG significatifs (FDR) par émotion de l'échelle mDES – 20 émotions testées.

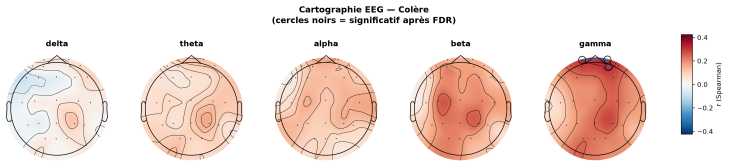


Figure 5.11 – Topographie de la puissance gamma corrélée à la colère rapportée.

C'est, sur l'ensemble du stage, le résultat le plus solidement validé pour la détection d'émotion : obtenu par sélection de features corrigée pour comparaisons multiples, évalué en LOSO, et confirmé par test de permutation.

5.3.6 Robustesse statistique du pipeline sur la valence générale

Le meilleur résultat obtenu sur la valence générale positive / négative (section 5.3.2, V2 : Kappa LOSO = +0,183) a également été soumis au même contrôle statistique (intervalle de confiance bootstrap au niveau sujet, test de permutation, section 4.8), avec le jeu de features v4 (dans le tableau suivant le Kappa étant recalculé avec le jeu v4 et sans le seuil optimisé) :

Table 5.4 – Robustesse statistique, label de valence générale, LOSO

Modèle	**Kappa observé**	**IC bootstrap 95 %**	p **(permutation)**
Régression logistique	+0,092	[−0,066 ; +0,235]	0,244
XGBoost	+0,058	[−0,105 ; +0,243]	0,304

Contrairement au résultat sur la colère (section 5.3.5), ni l'intervalle de confiance ni le test de permutation ne permettent ici de rejeter l'hypothèse nulle. La différence entre les deux résultats est informative : c'est en ciblant spécifiquement la colère, avec des features sélectionnées et corrigées pour cette cible précise (puissance gamma frontale), que le signal se distingue du hasard – une classification de la valence générale, plus large mais moins spécifique, ne le permet pas avec la puissance statistique actuellement disponible.

5.3.7 Réplication de l'analyse de corrélation de Sikka et al.

Sikka et al. [5] ont établi leur résultat fondateur par une corrélation *au niveau sujet* : la FAA moyenne en REM de chaque sujet ($n = 17$) mise en relation avec l'intensité moyenne de colère qu'il rapporte. Cette analyse, reproduite à l'identique sur les données recalculées durant le stage pour la FAA et huit autres features, ne retrouve pas le résultat publié : $R^2 = 0{,}037$, $p = 0{,}457$ pour la FAA (contre $R^2 = 0{,}604$, $p = 0{,}006$ dans la littérature), et aucune des neuf features testées ne passe le seuil de significativité après correction de Bonferroni.

Ce résultat n'est pas contradictoire avec celui de la section 5.3.5 : il montre plus précisément que c'est l'*asymétrie frontale alpha*, spécifiquement, qui ne réplique pas sur cet échantillon recalculé – tandis que la *puissance gamma frontale*, une feature différente identifiée indépendamment par sélection statistique, porte un signal qui, lui, résiste à la validation. Les deux analyses, prises ensemble, illustrent l'intérêt de ne pas se limiter à une seule feature *a priori*, aussi bien établie soit-elle dans la littérature.

5.4 Synthèse comparative des résultats

Le tableau 5.5 rassemble les résultats quantitatifs les plus significatifs des trois parties du stage.

Table 5.5 – Synthèse des résultats principaux, par partie

Partie	Tâche / jeu de données	Éval.	Résultat	Remarque
1	Stade de sommeil (Sleep-EDF)	split sujets	Kappa = 0,83	Excellent (CNN+LSTM)
2	Rêve, Multiple Awakenings	5-fold	Kappa = +0,20	
2	Rêve, Multiple Awakenings	LOSO	Kappa = +0,10	Écart ×2 vs 5-fold
2	Rêve, Tononi (NREM)	5-fold	Kappa = +0,79	
2	Rêve, Tononi (NREM)	LOSO	Kappa = −0,04	Fuite quasi-totale
2	Rêve, ZW (REM)	5-fold	Kappa = +0,82	
2	Rêve, ZW (REM)	LOSO	Kappa = −0,07	Fuite quasi-totale
3	Colère, gamma frontal (FDR+RF)	LOSO	Acc. = 61 %	$p = 0{,}048$ **(significatif)**
3	Colère, XGBoost (v3/v4)	LOSO	Kappa = +0,13	
3	Valence générale, LogReg (v2)	LOSO	Kappa = +0,18	
3	Valence générale, robustesse	LOSO	Kappa = +0,09	IC95% :[−0,07 ;+0,24], $p = 0{,}24$
3	Corrélation FAA-colère (Sikka)	sujet (n=17)	$R^2 = 0{,}04$	$p = 0{,}46$ (litt. : $R^2 = 0{,}60$)

Lecture transversale

Trois enseignements traversent les trois parties du stage, indépendamment du niveau de difficulté considéré :

1. **La difficulté croît fortement du stade de sommeil à l'émotion.** Le passage d'un Kappa « excellent » (0,83) pour le stade de sommeil à un signal ténu, détectable uniquement pour une cible spécifique et après sélection statistique rigoureuse des features, pour l'émotion, illustre concrètement la progression annoncée section 2.2 : plus l'information ciblée s'éloigne de la structure physiologique directement mesurée par l'EEG, plus le signal exploitable se raréfie.
2. **L'écart 5-fold / LOSO est systématique sur la détection de rêve.** Observé sur trois jeux de données indépendants et cohérent avec la littérature de référence [1], cet écart confirme que la variabilité inter-sujets domine largement le signal cognitif recherché pour cette tâche intermédiaire.
3. **Un signal significatif existe pour l'émotion, mais il est spécifique.** Le résultat le plus robuste de la partie 3 (colère, $p = 0{,}048$) n'a été obtenu ni par la première approche testée, ni sur la cible la plus générale (valence positive / négative), mais par une sélection de features guidée statistiquement sur une cible précise. Cela suggère que le signal EEG de l'émotion du rêve existe, mais qu'il est probablement spécifique à certaines émotions et certaines bandes de fréquence, plutôt que générale.

Chapitre 6

Conclusion

6.1 Réalisations au regard des enjeux identifiés

Le chapitre 2 posait une problématique à trois niveaux de difficulté croissante : le stade de sommeil, la détection de rêve, puis l'émotion du rêve, avec des enjeux technologique (solutions de suivi du sommeil minimalement invasives), scientifique et méthodologique (fiabilité de l'évaluation sur petits échantillons).

Sur le premier niveau, le stage atteint pleinement l'objectif fixé par la convention de stage : un pipeline CNN + LSTM entraîné sur un signal EEG mono-canal classe les cinq stades de sommeil avec un Kappa de 0,83, un niveau « excellent » au sens des critères usuels, démontrant la faisabilité d'un scoring automatique de haute qualité sur un montage réduit.

Sur le deuxième niveau, le stage apporte une contribution méthodologique nette : la mise en évidence systématique, sur trois jeux de données indépendants, d'un écart considérable entre validation croisée classique et validation *Leave-One-Subject-Out* pour la détection de rêve, ainsi que la correction de plusieurs biais concrets (bug de labellisation, confusion artefact / signal, seuil de décision optimisé sur le test) qui auraient autrement produit des résultats non reproductibles.

Sur le troisième niveau, le plus difficile, le stage aboutit à un résultat scientifique positif et rigoureusement validé : la classification de la colère rapportée en rêve, à partir de trois features de puissance gamma frontale sélectionnées statistiquement, atteint une significativité confirmée par test de permutation ($p = 0{,}048$) en évaluation LOSO. Ce résultat contraste avec l'absence de signal significatif pour une classification plus générale de la valence émotionnelle, ou pour la réplication de la corrélation FAA / colère de la littérature [5] : le signal EEG de l'émotion du rêve semble exister, mais de façon spécifique à certaines émotions et certaines bandes de fréquence, non comme une propriété générale et uniformément décodable.

Pris ensemble, ces trois niveaux répondent à l'enjeu technologique posé en introduction : le stage fournit à la fois une brique directement exploitable (scoring de sommeil sur canal unique) et un cadre d'évaluation rigoureux (protocole LOSO systématique, tests de significativité) que les travaux futurs du laboratoire pourront réutiliser pour étendre le décodage du contenu onirique.

6.2 Apports du stage pour ma formation et mon projet professionnel

Ce stage m'a avant tout permis d'approfondir concrètement le machine learning, et en particulier l'apprentissage supervisé. J'ai mené un projet de bout en bout : prétraitement de signaux, extraction de features, comparaison de modèles de complexité croissante (Random Forest, XGBoost, réseaux convolutifs, LSTM) et évaluation des résultats. J'ai surtout compris à quel point l'évaluation d'un modèle est délicate : fuite de données entre entraînement et test, déséquilibre des classes, seuils de décision optimisés sur le jeu de test, modèle qui apprend un artefact plutôt que le phénomène visé. Apprendre à détecter ces pièges, à les quantifier (validation LOSO,

intervalles de confiance, tests de permutation) et à ne retenir que les résultats défendables est la compétence la plus transférable que j'emporte, quel que soit le domaine d'application.

Le contexte, l'étude du sommeil et du rêve, était un terrain d'application stimulant, qui m'a permis de découvrir un domaine nouveau pour moi. Mais c'est surtout la démarche d'apprentissage automatique qui m'a passionné : construire un pipeline, formuler des hypothèses, les tester rigoureusement et accepter qu'un résultat soit négatif. Ce stage m'a donné envie d'approfondir cette voie, et je souhaite orienter mon projet de fin d'études, puis ma carrière, vers l'apprentissage automatique et la science des données, dans des secteurs où l'analyse de signaux ou de données complexes joue un rôle central.

6.3 Perspectives

Plusieurs pistes, identifiées mais non explorées faute de temps, pourraient prolonger ce travail :

- **Augmenter le volume de données** en intégrant d'autres jeux de la base DREAM disposant de scores d'affect, ou en contactant les auteurs de jeux de données existants pour vérifier la disponibilité de scores mDES ou équivalents non encore exploités ;
- **Reformuler la tâche en régression** plutôt qu'en classification binaire : prédire un score continu d'affect négatif exploiterait l'intégralité des enregistrements disponibles (sans exclusion liée à un seuil de pureté émotionnelle) et une corrélation significative même modeste ($r > 0{,}3$) constituerait un résultat scientifiquement exploitable ;
- **Poursuivre l'approche par personnalisation** : plutôt qu'un modèle universel, entraîner un modèle de base puis l'adapter par sujet (*few-shot*), en exploitant la variabilité inter-sujets mise en évidence dans ce rapport plutôt qu'en cherchant à s'en affranchir ;
- **Achever l'approche par espace tangent riemannien** combinée à un rééquilibrage explicite des classes, identifiée comme prometteuse mais non menée à terme (section 5.3.1).

Bibliographie

[1] Adin Beslagic. Dream catcher : Scalable EEG decoding of NREM dreams for portable brain-computer interfaces. Master's thesis, Norwegian University of Science and Technology (NTNU), Department of Engineering Cybernetics, 2025. Supervisor : Marta Molinas.

[2] Jarrod Gott et al. The dream EEG and mentation (DREAM) database. *Nature Communications*, 2025.

[3] Mithila Packiyanathan and André Torvestad. Dream decoder. Master's thesis, Norwegian University of Science and Technology (NTNU), Department of Engineering Cybernetics, 2024. Supervisor : Marta Molinas ; co-supervisor : Luis Alfredo Moctezuma.

[4] Serena Scarpelli et al. Electrophysiological correlates of dream recall during REM sleep. *Nature and Science of Sleep*, 12 :1043, 2020.

[5] Pilleriin Sikka, Antti Revonsuo, Valdas Noreika, and Katja Valli. EEG frontal alpha asymmetry and dream affect. *Journal of Neuroscience*, 39(24) :4775–4784, 2019.

Appendices