Analyse automatique des stades du sommeil à partir des voies électrophysiologiques et cardiorespiratoires

Jade Vanbuis

DOCTORAT BRETAGNE LOIRE SCIENCES POUR L'INGENIEUR

Le Mans Université

THESE DE DOCTORAT DE

LE MANS UNIVERSITE

ECOLE DOCTORALE N° 602
Sciences pour l'Ingénieur
Spécialité : *Acoustique*

Par
Jade VANBUIS

Analyse automatique des stades du sommeil à partir des voies électrophysiologiques et cardio-respiratoires

Thèse présentée et soutenue à ANGERS, le 19/02/2021
Unité de recherche : LAUM (UMR CNRS 6613)
Thèse N° : 2020LEMA1004

Rapporteurs avant soutenance :

Anne HUMEAU-HEURTIER	Professeure des Universités, LARIS
Andrea PINNA	Maître de Conférences des Universités, LIP6

Composition du Jury :

Président :	Régine LE BOUQUIN JEANNES	Professeure des Universités, LTSI
Examinateurs :	Guillaume BAFFET	Chercheur, société CIDELEC (anciennement)
	Mathieu FEUILLOY	Enseignant-Chercheur, ESEO
	Alexandre GRAMFORT	Directeur de Recherches, INRIA
	Philippe MICHEAU	Professeur des Universités, GAUS
Dir. de thèse :	Jean-Marc GIRAULT	Enseignant-Chercheur, ESEO

Titre : Analyse automatique des stades du sommeil à partir des voies électrophysiologiques et cardio-respiratoires

Mots clés : lecture du sommeil, classification, aide au diagnostic, troubles du sommeil, polysomnographie, polygraphie ventilatoire

Résumé : Le diagnostic des troubles du sommeil repose sur l'analyse de différents signaux enregistrés lors d'un examen du sommeil. Cette analyse est réalisée par un spécialiste du sommeil qui étudie la ventilation et, selon l'outil de diagnostic, la succession des stades de sommeil. Cette dernière tâche est particulièrement chronophage et complexe. Trois algorithmes d'aide au diagnostic et dédiés à cette tâche sont présentés.
Le premier permet la classification éveil/sommeil lors de l'utilisation d'un nouvel outil de diagnostic. Il en découle la possibilité pour le médecin de diagnostiquer précisément le syndrome d'apnées du sommeil et à moindre coût.
Le deuxième, fondé sur les voies électrophysiologiques, permet d'obtenir une classification de tous les stades de sommeil à partir de l'outil de diagnostic le plus complet. Il a été implémenté en considérant les limitations à l'utilisation d'un tel algorithme en routine clinique. L'architecture de cet algorithme reproduit ainsi le processus de classification réalisé manuellement par les médecins. Une fonction de seuillage auto-adaptatif a aussi été mise en place afin de fournir une classification patient-dépendante. Les résultats obtenus sont comparables avec ceux des médecins.
Le troisième algorithme, fondé sur les voies cardio-respiratoires, permet de classifier les stades de sommeil à partir d'un outil de diagnostic très utilisé mais pour lequel il n'est normalement pas possible d'étudier les stades de sommeil. La tâche est complexe, mais les résultats obtenus sont satisfaisants vis-à-vis de la littérature.
Les trois algorithmes, destinés aux différents outils de diagnostic, permettront d'aider les spécialistes à analyser le sommeil.

Title: Automatic sleep stage analysis from electrophysiological and cardio-respiratory channels

Keywords: sleep scoring, classification, diagnosis support tools, sleep-disordered breathing, polysomnography, home sleep apnea testing

Abstract: The diagnostic of sleep-disordered breathing requires the analysis of various signals obtained while recording sleep. The analysis is carried by a sleep specialist, which studies the patient's ventilation and, depending on the diagnostic tool used for the record, sleep stages. Sleep stage scoring is a complex and time-consuming task. Three diagnosis support algorithms dedicated to this task are presented in this thesis.
The first one provides a wakefulness versus sleep classification, designed for a new diagnostic tool. It results in the ability to make a precise diagnosis of sleep apnea syndrome, at low cost.
The second algorithm, based on electrophysiological channels, provides a full sleep stage classification while using the most complete diagnosis tool. It was implemented considering the known limitations for the use of algorithms in clinical practice. Its architecture thus reproduces the manual scoring process. A self-adaptative thresholding function was also implemented to provide a patient-dependent classification. The obtained results are comparable with the ones from sleep experts.
The third algorithm, based on cardio-respiratory channels, provides a sleep stage classification while using a diagnostic tool that is insufficient for a manual sleep scoring, yet still highly used. The task is challenging but the obtained results are satisfying compared to literature.
All three algorithms, which were designed for various diagnostic tools, will help sleep experts analyzing sleep.

Remerciements

Dans un premier temps je remercie l'ensemble de mon équipe d'encadrement, avec laquelle j'ai passé de très bons moments. Je suis tout particulièrement heureuse des très bonnes relations que nous avons tous pu nouer, et qui m'ont permis d'évoluer dans un cadre extrêmement bienveillant à chaque instant. J'ai conscience de la chance que j'ai pu avoir de travailler dans une telle équipe.

Dans l'ordre de nos rencontres, je commence ainsi par remercier Guillaume Baffet. Nous nous sommes rencontrés plus d'un an avant le début de la thèse, lors de mon stage technique dans la société CIDELEC. Tu as été mon maître de stage et l'es resté durant mon année de contrat de professionalisation qui a suivi. Durant cette année, j'ai pu apprendre à te connaître et je n'en n'ai que de très bons souvenirs. Tu m'as très vite fait confiance et donné une grande autonomie, tout en restant très attentif et présent. Tu as su me donner du courage lorsqu'il m'en fallait, m'a appris à décrypter l'aspect positif d'évènements qui ne l'étaient pas forcément, m'as soutenue lorsque j'en avais besoin, et même lorsque je ne le savais pas. C'est auprès de toi que j'ai découvert le monde du travail, et que je me suis façonnée à ce métier de chercheur. Je pense que je n'aurais pas pu mieux tomber. En plus de toutes tes qualités humaines, tu m'as également appris à mieux considérer l'utile, nécessitant parfois d'être un peu moins dans la rigueur, m'a aidé à prendre plus de hauteur, à lever le nez du guidon. Je me rends compte de la chance que j'ai eu de travailler auprès de toi, tu as été un manager exceptionnel. C'est grâce à toi que je me suis lancée dans l'aventure de la thèse. Je savais que je pouvais te faire confiance, et je ne le regrette pas. Je te remercie également, malgré ton départ de la société CIDELEC, d'être resté présent, et d'avoir gardé contact alors que rien ne t'y obligeait. Je suis très heureuse d'avoir pu continuer à apprendre à te connaître d'une manière moins formelle. Je te remercie, Guillaume, d'être devenu un ami.

Je remercie ensuite Mathieu Feuilloy, qui a tout d'abord été mon professeur principal durant mes semestres d'option à l'ESEO. Je savais que j'avais beaucoup de chance de t'avoir dans mon encadrement, avant même le début de la thèse. Tu avais déjà été extrêmement pertinent lors de nos échanges sur le projet de fin d'études. Je ne compte plus les très bons conseils et les pistes vers lesquelles je me suis orientée après l'une de nos discussions. Tu as été l'un de mes soutiens scientifiques durant cette thèse et je retiendrai ta pertinence. Je savais également que je pourrais facilement me concerter avec toi, et je tiens également à te remercier pour ton accessibilité et ta sympathie. J'ai particulièrement apprécié nos « points cafés » durant lesquels tu prends des nouvelles de mon état d'esprit, de mes ressentis. Ces points m'ont été précieux et je sais très bien qu'ils ont largement impacté, de manière positive, ces années de thèse. Je te remercie, Mathieu, de ta gentillesse et salue particulièrement ta pertinence.

Ensuite, je remercie Alain Le Duff, qui a été mon directeur de thèse durant un an. Je suis heureuse d'avoir pu garder contact avec toi, qui a *switché* de travail et est parti vivre au Canada. J'apprécie que nous puissions *jaser* de ta nouvelle vie là-bas, c'est *tiguidou laï laï. À la prochaine chicane*, Alain, et merci à toi.

Pour finir avec mon équipe d'encadrement, je remercie Jean-Marc Girault, arrivé en cours de thèse en remplacement d'Alain. Tu as été un directeur de thèse formidable. J'ai particulièrement apprécié la délicatesse avec laquelle tu t'es intégré dans l'équipe. Très attentif à la dynamique déjà en place, tu as su sans aucune difficulté prendre place dans l'équipe d'encadrement, sans jamais précipiter les choses ou t'imposer. Tu as réellement considéré le travail réalisé avant ton arrivé, et tu t'es même investi très rapidement dans ce travail. J'ai apprécié particulièrement ton expertise, qui complète celle déjà présente dans l'équipe, et tes qualités de management. Merci à toi de m'avoir accompagnée en congrès, j'en garde d'excellents souvenirs. Je te remercie également pour ta curiosité et toute l'attention que tu as pu porter au domaine du diagnostic

du sommeil. Tu es vite devenu quelqu'un sur qui compter pour défendre les intérêts de mon travail de thèse, ta diplomatie est indéniable. Merci aussi d'avoir pris le temps d'apprendre à me connaître en tant que personne, et non seulement en tant que doctorante. Merci beaucoup, Jean-Marc, pour ton dynamisme et ton soutien.

Dans un second temps, je remercie la société CIDELEC, qui est à l'origine de cette thèse et m'a fait confiance du début à la fin. Je vous remercie pour ce sujet passionnant, qui aura su me motiver tout au long de la thèse. J'adresse des remerciements chaleureux à Cédric Freycenon, président de la société et manager du service Recherche, pour ta gentillesse, ta confiance, ta reconnaissance, et ton accueil dès les premiers jours. Je te remercie également, Cédric, de me permettre de travailler sur l'industrialisation du travail réalisé dans cette thèse, et de me faire confiance pour de nouveaux projets.
Je remercie également l'ensemble du pôle Recherche, qui n'a cessée de grandir ces dernières années. Un grand merci au pôle Développement Produit, avec qui j'ai travaillé conjointement pour l'industrialisation d'HypnoLighT, et qui permet à mon travail d'aboutir réellement, ce qui est très important pour moi. Merci au pôle Commercial et à tous les ingénieurs produits avec qui j'ai pu partager de bons moments lors des journées angevines ou des congrès médicaux. Merci, enfin, à tous les autres collègues des pôles Opérationnel, Ressources et Relations Transversales, Qualité et Affaires Réglementaires, et Marketing et Communication, que j'ai pu côtoyer souvent en pause et qui ont tous participé d'une manière ou d'une autre à mon équilibre lors de cette thèse. Je suis heureuse de pouvoir continuer mon travail auprès de vous pour cette après-thèse.

Je remercie également le LAUM, mon laboratoire de rattachement, ainsi que mes membres de CSI, Frédéric Ossant et Philippe Micheau, que j'ai eu le plaisir de rencontrer durant ma thèse. Ce fut un plaisir de partager quelques discussions autour d'un repas avec vous.
Je remercie Anne Humeau-Heurtier et Andrea Pinna d'avoir accepté la fonction de rapporteur de thèse, Régine Le Bouquin Jeannès de présider le jury, et Alexandre Gramfort, de prendre part à cette thèse en tant qu'examinateur.

Je remercie ensuite Nicole Meslier et Frédéric Gagnadoux. Malgré qu'ils ne fassent ni partie de l'ESEO, ni de la société CIDELEC, je les considère réellement comme des collègues. C'est durant mon contrat de professionalisation, ainsi que cette thèse, que j'ai réellement pu vous côtoyer. Je suis ravie de travailler à vos côtés. Je garde d'excellents souvenirs du DAFKA, pendant lequel nous avons pu partager de nombreux moments. Vous êtes deux personnes très inspirantes, chacune pour des raisons différentes, et je me rends compte de la chance que j'ai de pouvoir travailler auprès de vous. C'est grâce à vous que j'ai pu, d'année en année, mieux m'investir dans le secteur médical et que j'ai l'opportunité aujourd'hui de participer à plusieurs formations. Je tiens particulièrement à vous remercier pour toutes les fois où vous m'avez introduit à l'un de vos pairs. Je vous remercie, Nicole et Frédéric, pour votre bienveillance et votre gentillesse.
J'en profite également pour remercier Julien Godey, Laetitia Moreno et Marion Vincent, les techniciens du sommeil du CHU d'Angers, pour leur travail sans lequel ce travail de thèse n'aurait pas été possible.

Je tiens à grandement remercier l'ensemble des collègues de l'ESEO, et particulièrement Fabien Chhel et Richard Perdriau, pour leur extrême gentillesse durant toutes ces années. Merci à Roberto Longo et Guy Plantier, pour votre aide précieuse durant cette thèse. Merci aussi aux étudiants avec lesquels j'ai pu travailler : Corentin Lequet, Lucile Quillien, Maureen Manche, mais aussi Alice Bleines, Alexandre Cavel, Camille Csanki, Laetitia Dominjon, Anwar Gasri, Mohammed Hussein Khalife, Maëlys Montantin, Hamza Ouazzani et Paul Pierrat. Et, bien sûr, je remercie tout particulièrement les doctorants. J'ai passé de très bons moments avec vous lors des pauses et également en dehors du travail. Merci à Romain Cormerais, Valentin Besnard, Nathalie Freycenon, Romain Feron, et, mes trois piliers tout au long de cette thèse : Théo Richard, Lucile Riaboff et Margaux Blanchard.
Je te remercie grandement, Théo, pour ton accueil fabuleux et ta sympathie. J'ai adoré tous ces repas avec toi, ces cafés, nos discussions, nos rires. Tu es une personne formidable ! J'espère que tu pourras encore m'apprendre des tas de choses sur toutes tes plantes, que l'on parlera gadgets connectés encore de nombreuses fois, et que l'on pourra se raconter les derniers exploits (ou les dernières bêtises) de nos chats !

Un grand merci à toi, Lucile, avec qui j'ai pu échanger sur tout et sur rien. Sur l'aspect professionnel, merci pour tous tes conseils, et ton aide très précieuse pour la rédaction des articles. Tu es une personne très inspirante et je ne doute pas que tu sauras mettre à profit toutes tes grandes qualités pour réaliser un travail impressionnant lors de ton poste à Dublin. Mais, plus important encore, je te remercie mille fois pour tous ces moments partagés en-dehors du travail. Merci d'être venue partager ma souffrance au crossfit maintes et maintes fois, alors que tu fais dix mille autres activités sportives à côté. Merci aussi pour tous ces repas à la guinguette (au lieu d'aller au crossfit d'ailleurs). Merci pour ton humour à la Blanche Gardin, tu as rendu ces années de thèse bien plus marrantes qu'elles n'auraient pu l'être.
Pour finir, merci à toi, Margaux, pour ton humour et la bonne humeur constante qui t'accompagne. Je suis très heureuse d'avoir pu apprendre à te connaître. Je te remercie particulièrement pour cette relation si spéciale de collègue-amie, qui était nécessaire puisque tu réalises aussi ta thèse au sein de l'entreprise CIDELEC. Je suis persuadée qu'une relation telle que cela n'est pas évidente à créer, et que cela a été possible grâce à ton caractère et ta manière d'être. Je te remercie donc pour ta sincérité et l'équilibre que nous avons pu mettre en place sur un plan professionnel, mais, plus encore, pour tous les instants que nous avons partagés en dehors du travail. Merci pour tous les sourires que tu as pu me donner, et pour ton écoute et ta sincérité. J'ai rarement rencontré quelqu'un d'aussi positive, si agréable à vivre. Vivement que l'on aille de nouveau tester tous les petits restaurants d'Angers, acheter des plantes ou faire du shopping. Tu as contribué, et continue aujourd'hui de contribuer, à mon équilibre. Je ne changerai de collègue-amie pour rien au monde. C'est un vrai plaisir d'évoluer auprès de toi. J'espère du fond du cœur pouvoir continuer de le faire très longtemps.

Enfin, je tiens à remercier tous mes proches, qui m'ont soutenue à chaque instant de cette thèse. Merci tout d'abord aux copains, et particulièrement à Nicouillu, Agathe, Lolotte, Rémy, Cyrillou, Cécile et Leti, Tijette, Didi et Clémentine, Juju et Diane, Clémence, Quentin et Oriane. Merci à vous pour tous ces bons instants passés durant cette thèse et à vos petites attentions.
Merci ensuite à toute la famille. Une grosse pensée à vous, Papi DD et Bà Nội. Un merci aussi à Mamée, Papé, Jojo et Camille pour votre soutien au Botswana, lorsque j'ai appris mon premier refus d'article. Merci Michaël, Shinta, Jess, Mathieu, Capue, papa et maman, pour avoir été présents tout au long de ces années. Merci particulièrement à vous, papa et maman, qui êtes au moment où j'écris ces lignes, en train de relire le présent manuscrit pour y déceler les coquilles et fautes d'orthographe.

Je remercie pour finir ceux qui passent, à cause de (ou grâce à ?) la pandémie mondiale, toute la journée auprès de moi. Merci les chats, de m'accepter sur votre territoire toute la journée. Je sais que parfois ça vous embête, même si je vous fais plus souvent des gratouilles.
Et, enfin, merci à toi, Scot, mon conjoint et mon meilleur ami, pour l'immensurable aide et soutien que tu m'as apporté chaque jour de cette thèse, sans exception. Merci pour la niaque que tu m'as redonnée dans les moments plus difficiles, merci pour les heures entières que tu as passé à m'écouter raconter ma journée, merci pour les (autres) heures entières que tu as passé à me rassurer et me remonter le moral, ou au contraire à m'écouter laisser exploser ma joie, merci pour les heures entières (d'autres encore) que tu as passé à lire et relire mes articles et le présent manuscrit. Tout simplement un énorme merci, Scot, d'avoir été aussi présent pour moi, durant ces 3 dernières années, mais aussi les 7 précédentes.

Table des matières

Acronymes et symboles

Symboles | A | C | D | E | F | I | J | K | L | M | P | R | S | T | V

Symboles

Acc taux d'accord

P_e proportion d'accord aléatoire (concordance attendue dans l'hypothèse où l'analyse manuelle et automatique sont totalement indépendantes)

P_o proportion d'accord observée

Se Sensibilité

Sp Spécificité

κ Kappa de Cohen

k-fold CV validation croisée à k blocs, de l'anglais *k-fold Cross Validation*

A

AASM *American Academy of Sleep Medicine*

ACP Analyse en Composantes Principales

ADL Analyse Discriminante Linéaire

ADQ Analyse Discriminante Quadratique

C

CNN *Convolutional Neural Network*

CPAP *Continuous Positive Airway Pressure* en anglais

D

DL Deep Learning

E

EA Éveil Agité

ECG électrocardiogramme

EEG électroencéphalogramme

EMG électromyogramme

EOG électro-oculogramme

EYF Éveil Yeux Fermés

EYO Éveil Yeux Ouverts

F

FN Faux Négatif

FP Faux Positif

I

IA Intelligence Artificielle

IAH Index d'Apnées Hypopnées

ICM Interface Cerveau-Machine
IMC Indice de Masse Corporelle
IND Interface Neuronale Directe

J
JDL *Joint Directors of Laboratories*

K
kNN méthode des k plus proches voisins

L
LOOCV validation croisée un contre tous, de l'anglais *Leave-One-Out Cross Validation*
LSTM *Long Short Term Memory*

M
ML Machine Learning
MLP *Multilayer Perceptron*
MMC Modèle de Markov à états Cachés
MOL Mouvement Oculaire Lent
MOR Mouvement Oculaire Rapide

P
PPC Pression Positive Continue
PSG polysomnographie
PV Polygraphie Ventilatoire

R
RF forêts d'arbres décisionnels ou forêts aléatoires, de l'anglais *Random Forest*
RNN *Recurrent Neural Network*

S
SAHS Syndrome d'Apnées Hypopnées du Sommeil
SE *Sleep Efficiency*
SFFS *Sequential Floating Forward Selection*
SFS *Sequential Forward Selection*
SL Sommeil Lent ou regroupement du N1, N2 et N3
SOL *Sleep Onset Latency*
SVM Séparateur à Vaste Marge ou machine à vecteur de support

T
TE Temps d'Enregistrement
TILE Test Itératif de Latence à l'Endormissement
TST Temps de Sommeil Total

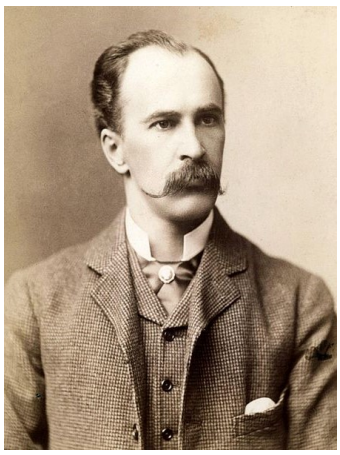
V
VN Vrai Négatif
VP Vrai Positif
VPN Valeur Prédictive Négative
VPP Valeur Prédictive Positive

Introduction

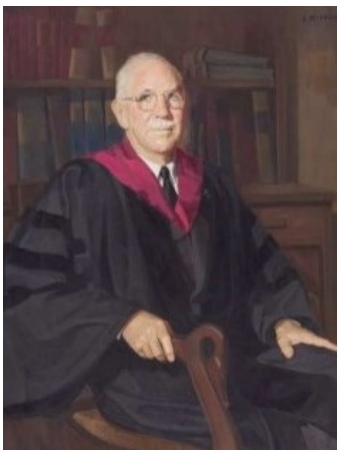
1 Motivations cliniques

Le sommeil est une fonction vitale chez l'Homme qui passe, en moyenne, un tiers de sa vie à dormir. C'est en effet durant le sommeil que des fonctions telles que l'apprentissage, la mémorisation et l'adaptation sont assurées.
Ces dernières années, l'étude du sommeil et des dysfonctionnements qui lui sont liés s'est développée et de nombreux troubles du sommeil ont été identifiés. Leur origine est généralement due à un trouble ventilatoire, comme le Syndrome d'Apnées Hypopnées du Sommeil (SAHS), ou à un dysfonctionnement cérébral, comme l'insomnie ou la narcolepsie. Cependant, avec la découverte de nouvelles pathologies liées au sommeil, les médecins provenant de spécialités de plus en plus variées s'y intéressent. Auparavant, les troubles du sommeil, pourtant connus, n'étaient que rarement détectés et traités. Aujourd'hui, il est ainsi possible de se faire diagnostiquer dans des laboratoires de sommeil, ou même chez son pneumologue, cardiologue, neurologue ou médecin généraliste.

Parmi les différents troubles du sommeil, le SAHS est l'un des plus connus. Ce syndrome a été décrit pour la première fois par W. Osler en 1919. Ce syndrome fut ensuite décrit par C.S. Burwell, sous le nom de « syndrome de Pickwick » en 1956, mais incorrectement associé à un surpoids et à un taux trop élevé de dioxide de carbone dans le sang (ou hypercapnie). C'est seulement en 1972 que C. Guilleminault distingue du syndrome de Pickwick un syndrome apparaissant chez des patients ne souffrant pas nécessairement de surpoids et/ou d'hypercapnie, mais souffrant d'interruptions complètes ou partielles de la respiration pendant le sommeil.



(a) W. Osler



(b) C.S. Burwell

(c) C. Guilleminault

Figure I1 – Portraits de W. Osler, C.S. Burwell et C. Guilleminault, à l'origine de la découverte du Syndrome d'Apnées Hypopnées du Sommeil (SAHS).

La prévalence des troubles liés au sommeil varie selon les populations étudiées. On estime qu'approximativement un tiers de la population adulte serait touchée (Croft, 2017; Heinzer *et al.*, 2015). Cette prévalence aurait augmenté durant ces deux dernières décennies (Peppard *et al.*, 2013).

Senaratna *et al.* (2017) ont étudié la prévalence du SAHS dans la population générale. Si l'estimation précise de cette prévalence reste compliquée (entre 9 % et 38 % selon les différentes études citées), il est démontré qu'elle est supérieure chez les hommes, qu'elle augmente avec l'âge et qu'elle est également liée à l'Indice de Masse Corporelle (IMC) (Franklin et Lindberg, 2015; Senaratna *et al.*, 2017).
Les patients atteints du SAHS souffrent généralement de nombreux symptômes :

— de jour : somnolence, grande fatigue, troubles de la mémoire et de l'attention, baisse des performances, troubles de l'humeur, troubles du comportement, troubles de l'acuité auditive, maux de tête matinaux ;

— de nuit : ronflements, arrêts respiratoires, sommeil agité, nombreux réveils pour uriner, réveils en sursaut, baisse de la libido, sueurs nocturnes, essoufflement nocturne, salivation excessive.

Ces symptômes, non spécifiques et fréquents dans la population générale, alarment peu le patient qui ne se soumet souvent que très tardivement à un examen du sommeil.

Le diagnostic repose sur la considération de ces symptômes et d'enregistrements dont la mise en place est complexe. Ainsi, le nombre d'infrastructures permettant de réaliser un examen du sommeil reste limité. Une étude réalisée par Flemons *et al.* (2004) a pointé un manque de ressources pour le diagnostic et traitement des troubles du sommeil dans cinq pays (Royaume-Uni, Belgique, Australie, États-Unis et Canada). En conséquence, le délai pour pouvoir réaliser un examen du sommeil est souvent supérieur à plusieurs mois (Flemons *et al.*, 2004; Stewart *et al.*, 2015).

Différentes solutions pour pallier ce délai ont donc été mises en place ces dernières années :

— côté électronique (voir la Figure 1), des appareils permettant un examen ambulatoire ont été développés. En passant l'examen directement à son domicile, le patient désengorge les chambres de diagnostic du sommeil et permet d'augmenter le nombre d'examens réalisés en simultané ;

— côté logiciel (voir la Figure I3), de nombreux outils informatiques permettant d'accompagner les médecins lors de leur diagnostic ont été proposés. Ils permettent ainsi aux médecins de réduire le temps alloué à l'étude des examens du sommeil.

Figure I2 – Appareil de diagnostic ambulatoire (de la société CIDELEC).

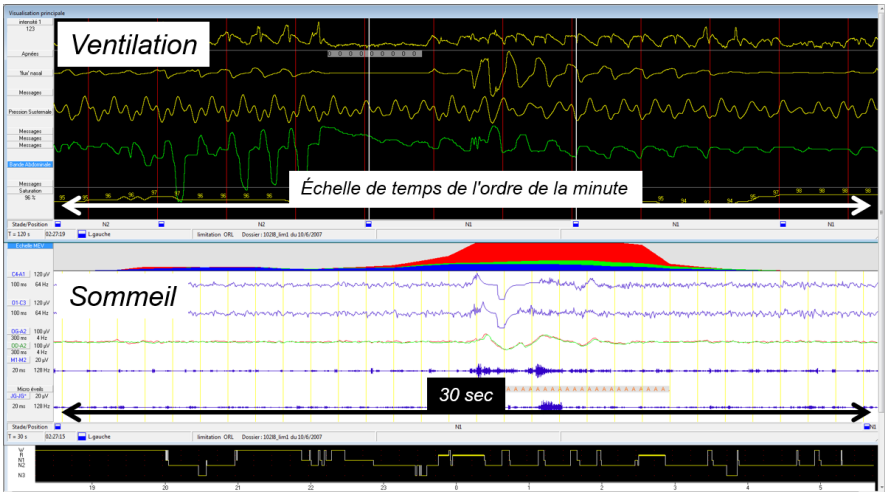


Figure I3 – Logiciel d'analyse du sommeil (de la société CIDELEC).

2 Objectif du projet de recherche

Le travail de Recherche dans lequel s'inscrit cette thèse consiste à développer des outils informatiques permettant d'accompagner les médecins lors du diagnostic du sommeil. Les outils ainsi créés ont pour objectif d'être industrialisés et proposés aux médecins par le biais de la société CIDELEC (Sainte-Gemmes-sur-Loire, FRANCE).

CIDELEC est une société impliquée dans le domaine du sommeil, et plus particulièrement du diagnostic du sommeil. Créée dans les années quatre-vingt, elle propose des appareils permettant l'enregistrement de plusieurs signaux physiologiques chez le patient. Ces signaux, mis en parallèle avec les symptômes du patient, sont étudiés par le médecin qui peut alors émettre un diagnostic. En plus de ces appareils, CIDELEC fournit aussi un logiciel permettant de les exploiter plus simplement et facilitant le suivi du patient au sein de laboratoires du sommeil par exemple.

Cette thèse s'inscrit dans un environnement multidisciplinaire, regroupant l'école d'ingénieurs Angevine ESEO, le LAUM UMR CNRS 6613, la société CIDELEC et le Laboratoire du Sommeil du CHU d'Angers.

L'objectif a été d'apporter des connaissances plus poussées non pas sur l'aspect ventilatoire, déjà très maîtrisé par CIDELEC, mais sur les stades de sommeil.

3 Organisation du manuscrit de thèse

Cette thèse débute par une présentation des modalités du sommeil et de son diagnostic (voir Chapitre A). Ici, seront expliqués les différents éléments qui composent le sommeil avec des détails sur les étapes nécessaires à son diagnostic.
Ensuite, une revue de l'état de l'art de l'analyse des stades de sommeil est présentée (voir Chapitre B). Elle met en lumière les différentes approches méthodologiques fréquemment utilisées.

Le reste de la thèse, qui regroupe toutes mes contributions, est structuré en trois chapitres.
Dans le premier chapitre (chapitre C), un nouvel outil permettant le diagnostic du sommeil est présenté. Cet outil est une version améliorée de la polygraphie ventilatoire, l'un des appareils de diagnostic les plus utilisés. Un travail a donc été réalisé pour permettre aux médecins d'obtenir des informations quant à l'état de somnolence (éveil ou sommeil) du patient lors d'un enregistrement effectué avec cet outil. Pour aller plus loin dans la caractérisation des troubles du sommeil, un second travail a été effectué avec pour objectif l'identification de certains événements ventilatoires.
Les deux chapitres suivants sont dédiés à l'analyse automatique du sommeil dans le cas de l'utilisation des outils (déjà existants) couramment utilisés pour le diagnostic du sommeil. Le deuxième chapitre (chapitre D) traitera ainsi de l'analyse du sommeil lors de la polysomnographie, l'enregistrement le plus complet. Une méthode prenant en considération les limitations empêchant l'utilisation des algorithmes d'analyse automatique du sommeil par les spécialistes en routine clinique est présentée. Elle a pour objectif d'aider le spécialiste lors de sa lecture du sommeil, pouvant ainsi lui faire gagner du temps et faciliter grandement son travail. L'impact clinique de la méthode sur le diagnostic est également évalué.
Le troisième chapitre (chapitre E) présente une analyse du sommeil lors d'un enregistrement utilisant un outil très répandu mais moins complet, à savoir la polygraphie ventilatoire. Cet outil, du fait qu'il soit moins complet, possède l'avantage d'être moins coûteux, et plus facile à installer et à étudier. Cependant, pour ce type d'enregistrement, le spécialiste n'est pas en mesure de lire le sommeil. Une nouvelle méthode sera décrite et, à l'image des études précédentes, l'impact de la méthode sur le diagnostic du patient sera évalué.
Mes contributions sont ainsi composées de différentes analyses automatiques des stades de sommeil, pensées pour répondre aux différentes problématiques rencontrées par les spécialistes du domaine, selon l'outil de diagnostic employé.

Pour conclure, le dernier chapitre (chapitre F) de cette thèse présentera une discussion générale du travail réalisé ainsi que les différentes perspectives de recherche futures.

Bibliographie

Croft, J. B. (2017). CDC's Public Health Surveillance of Sleep Health.

Flemons, W. W., Douglas, N. J., Kuna, S. T., Rodenstein, D. O. et Wheatley, J. (2004). Access to Diagnosis and Treatment of Patients with Suspected Sleep Apnea. *American Journal of Respiratory and Critical Care Medicine*, 169(6):668–672.

Franklin, K. A. et Lindberg, E. (2015). Obstructive sleep apnea is a common disorder in the population— a review on the epidemiology of sleep apnea. *Journal of Thoracic Disease*, 7(8):1311–1322.

Heinzer, R., Vat, S., Marques-Vidal, P., Marti-Soler, H., Andries, D., Tobback, N., Mooser, V., Preisig, M., Malhotra, A., Waeber, G., Vollenweider, P., Tafti, M. et Haba-Rubio, J. (2015). Prevalence of sleep-disordered breathing in the general population : the HypnoLaus study. *The Lancet Respiratory Medicine*, 3(4):310–318.

Peppard, P. E., Young, T., Barnet, J. H., Palta, M., Hagen, E. W. et Hla, K. M. (2013). Increased Prevalence of Sleep-Disordered Breathing in Adults. *American Journal of Epidemiology*, 177(9):1006–1014.

Senaratna, C. V., Perret, J. L., Lodge, C. J., Lowe, A. J., Campbell, B. E., Matheson, M. C., Hamilton, G. S. et Dharmage, S. C. (2017). Prevalence of obstructive sleep apnea in the general population : A systematic review. *Sleep Medicine Reviews*, 34:70–81.

Stewart, S. A., Skomro, R., Reid, J., Penz, E., Fenton, M., Gjevre, J. et Cotton, D. (2015). Improvement in Obstructive Sleep Apnea Diagnosis and Management Wait Times : A Retrospective Analysis of a Home Management Pathway for Obstructive Sleep Apnea. *Canadian Respiratory Journal*, 22(3):167–170.

Chapitre A

Généralités sur le sommeil et son diagnostic

A.1 Le sommeil

Le sommeil est constitué de plusieurs phases appelées stades de sommeil. Ces stades définissent différents états de vigilance qui sont plus ou moins propices à la récupération physique, à la mémorisation, ou au rêve, par exemple. Une nuit se compose de quatre à six cycles de sommeil, chacun composé d'une alternance des différents stades de sommeil. En général, chaque cycle est une succession, dans l'ordre, de sommeil léger, lent, lent profond et paradoxal.
Le sommeil léger, ou stade **N1**, correspond à l'endormissement. Stade de transition entre l'éveil et le sommeil lent, il est donc très bref et ne représente environ que 5 % de la nuit.
Le sommeil lent, ou stade **N2**, est à l'inverse le stade le plus présent puisqu'il représente environ 50 % de la nuit.
Le sommeil lent profond, ou stade **N3**, est le stade durant lequel le sommeil du patient est le plus profond. Il est très présent dans les premiers cycles de sommeil de la nuit, mais sa proportion diminue au fur et à mesure de l'atténuation de la fatigue physique.
Le sommeil paradoxal **SP** est appelé ainsi car il est constitué d'une activité cérébrale forte (qui pourrait parfois faire penser à de l'éveil) malgré un tonus musculaire complètement aboli. À l'inverse du stade N3, la quantité de sommeil paradoxal s'allonge au fur et à mesure de la nuit. C'est durant ce stade que les rêves sont les plus précis.

A.2 Le diagnostic du sommeil

A.2.1 Les outils de diagnostic

L'identification des stades de sommeil peut permettre au spécialiste du sommeil d'identifier un trouble directement lié aux stades de sommeil (insomnie ou narcolepsie par exemple), mais également de caractériser un trouble qui serait lié à l'activité ventilatoire du patient (comme un SAHS spécifique au sommeil paradoxal). En réalité, il existe différents outils permettant un diagnostic plus ou moins précis des troubles associés au sommeil. Ces outils ont été classifiés en différentes catégories :

Type I : La polysomnographie (PSG) réalisée en laboratoire du sommeil et supervisée par un expert du sommeil. Les signaux électrophysiologiques et cardio-respiratoires y sont enregistrés.

Type II : La PSG réalisée en ambulatoire. Les signaux électrophysiologiques et cardio-respiratoires y sont enregistrés.

Type III : La Polygraphie Ventilatoire (PV) réalisée en laboratoire ou en ambulatoire. Seuls les signaux cardio-respiratoires y sont enregistrés.

Type IV : L'enregistrement en laboratoire ou en ambulatoire d'un ou deux signaux cardio-respiratoires seulement.

Nous ne présenterons que la PV et la PSG, qui sont les deux outils de diagnostic du sommeil les plus répandus et qui permettent au médecin de prescrire un traitement.

La PV

Très répandue, la PV est composée de nombreux capteurs permettant l'étude de la ventilation, grâce aux capteurs suivants (voir la Figure A1) :

- la lunette nasale, qui permet l'enregistrement du flux nasal ;
- le capteur son trachéal (technologie CIDELEC appelée PneaVoX®, qui capte les ronflements, les bruits respiratoires et la pression sus-sternale) ;
- les sangles thoracico-abdominales (ou sangles inductives) permettant l'étude des mouvements respiratoires ;
- l'oxymètre de pouls qui permet l'acquisition du rythme cardiaque, du photopléthysmogramme et de la saturation pulsée en oxygène (ou SpO2) dans le sang ;
- la luminosité via un détecteur de luminosité directement sur le polygraphe ventilatoire ;
- l'actimétrie via un actimètre directement sur le polygraphe ventilatoire.

Cet outil est donc principalement composé de capteurs ventilatoires. Grâce à ces signaux, le spécialiste du sommeil sera en mesure de détecter les arrêts complets ou partiels de la respiration (apnées et hypopnées), qui caractérisent le SAHS.

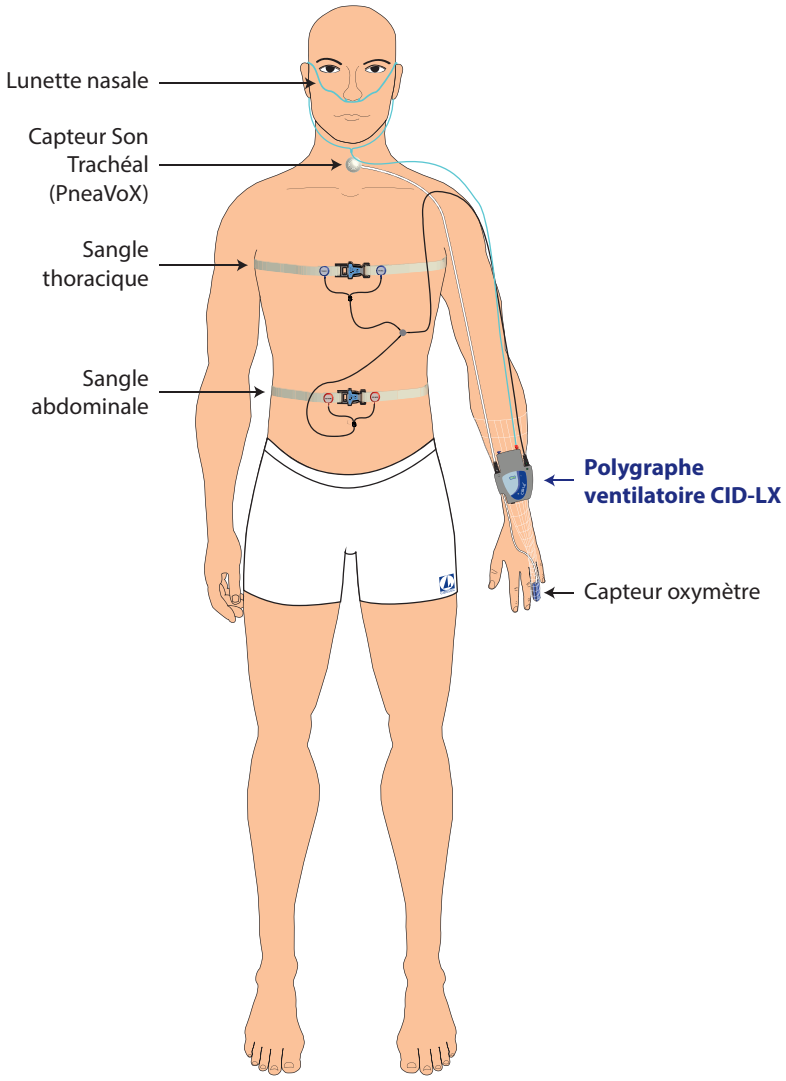


Figure A1 – Exemple de montage polygraphique, utilisant l'appareil de PV CID-LX.

La PSG

La PSG, plus complète, est considérée comme l'outil de référence (le gold standard). En plus des capteurs déjà présents en PV, elle possède en effet des capteurs dits polysomnographiques. Ainsi, le diagnostic est effectué grâce aux capteurs suivants :

- les signaux polygraphiques précédents ;
- des électroencéphalogrammes (EEGs) pour l'activité cérébrale ;
- des électro-oculogrammes (EOGs) pour l'activité oculaire ;
- un électromyogramme (EMG) mentonnier pour le tonus musculaire du menton ;
- des électromyogrammes (EMGs) jambiers pour le tonus musculaire des jambes ;
- un électrocardiogramme (ECG) pour l'activité cardiaque.

Cet outil permet aussi bien l'étude de l'activité ventilatoire du patient que l'analyse de ses stades de sommeil et de son activité cardiaque.

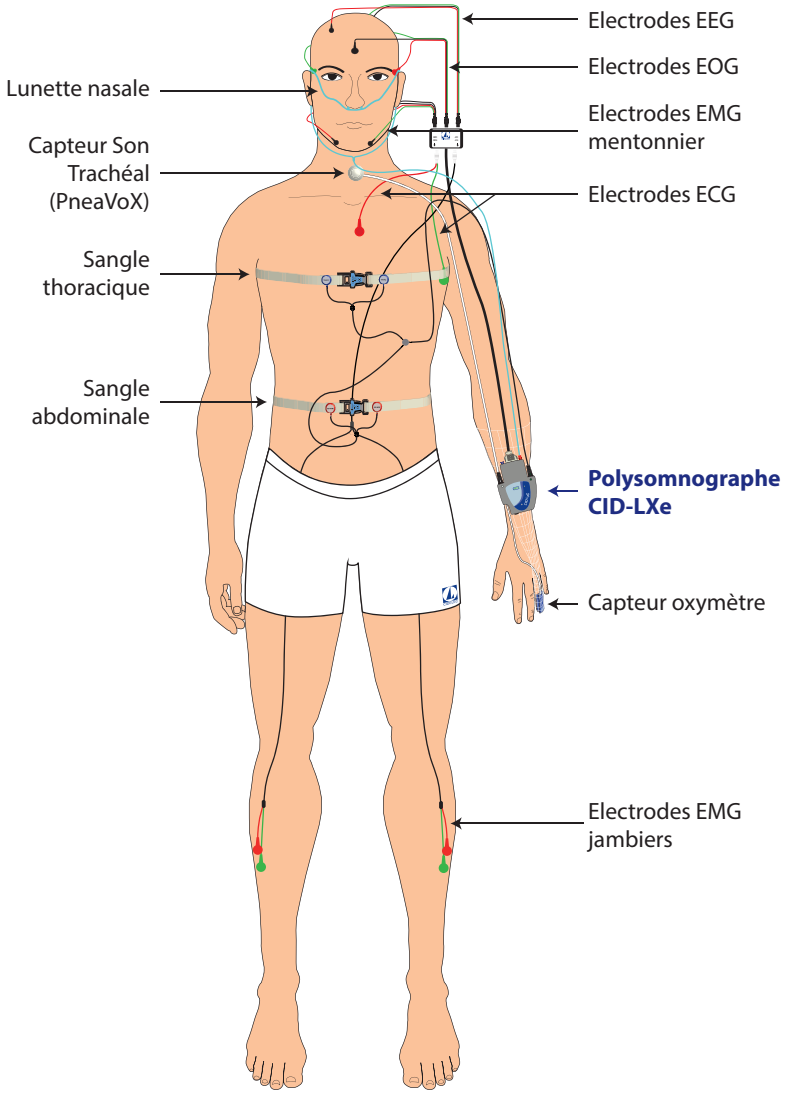


Figure A2 – Exemple de montage polysomnographique, utilisant l'appareil de PSG CID-LXe.

A.2.2 Le déroulement d'un enregistrement

Le diagnostic débute par une discussion avec le patient et un ou plusieurs tests permettant d'évaluer les symptômes du patient (fatigue diurne, manque de concentration, etc.). Selon la gravité des symptômes et les caractéristiques du patient (physiologie, âge, sexe, etc.), le spécialiste du sommeil détermine s'il effectue un enregistrement polygraphique (à l'aide d'une PV) ou

polysomnographique (à l'aide d'une PSG). En effet, de par la quantité de capteurs supérieure, la PSG est plus coûteuse et longue à mettre en place. Dans le cas d'une suspicion forte de SAHS par exemple, et malgré la plus grande précision de la PSG, le médecin pourra privilégier l'usage d'une PV à priori suffisante pour établir le diagnostic.

Les capteurs spécifiques à l'outil choisi sont ensuite installés sur le patient. Dans le cas d'un examen en ambulatoire, il est possible que le patient ait à installer certains capteurs lui-même au moment du coucher. Ce dernier passe alors une nuit complète [1] durant laquelle l'ensemble des signaux sont enregistrés.

Une analyse de l'ensemble des signaux enregistrés est ensuite nécessaire. À l'aide des signaux ventilatoires, le spécialiste du sommeil est en mesure d'identifier différents événements, comme les apnées ou hypopnées (voir Figure A3). Pour ce faire, il observe les signaux par tranche de une ou deux minutes, et repère des signes de diminution de la ventilation en suivant les recommandations des autorités françaises (SFRMS, 2010) ou américaines (Berry *et al.*, 2017). En plus d'être détectés, les événements ventilatoires sont caractérisés selon leur cause et leur effet.

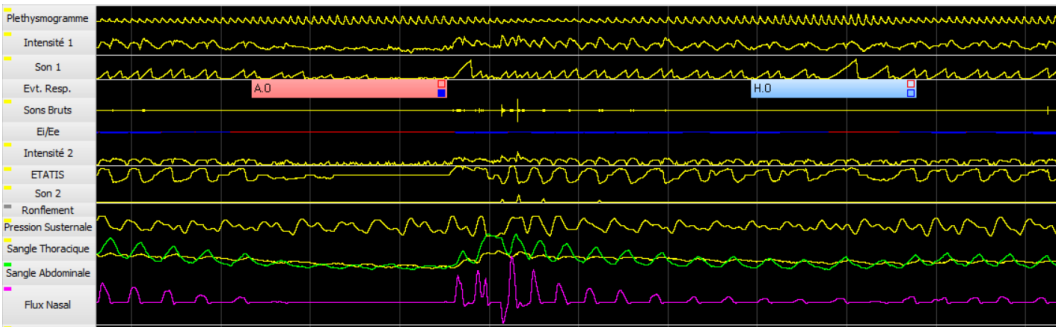


Figure A3 – Exemple d'une apnée (en rose) suivie d'une hypopnée (en bleu), observées sur l'interface du logiciel CIDELEC.

L'identification des stades de sommeil (dans le cas d'une PSG) est également effectuée manuellement, mais cette fois-ci par tranche de 30 secondes, aussi appelée **époque** (voir Figure A5). Pour chaque époque, le spécialiste identifie ainsi le stade dans lequel le patient se trouve (éveil, N1, N2, N3 ou SP, voir Section A.1).
Cette tâche est chronophage, fastidieuse et demande des connaissances spécifiques. Les recommandations de l'*American Academy of Sleep Medicine* (AASM) (Berry *et al.*, 2017) détaillent le contenu fréquentiel, temporel, ainsi que les grapho-éléments spécifiques à chaque stade de sommeil (voir Figure A4). La succession des différents stades de sommeil (un par époque) pour l'ensemble de la nuit est appelée **hypnogramme** (voir Figure A6).

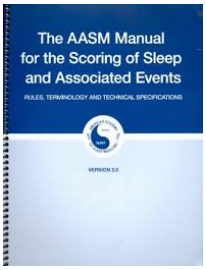
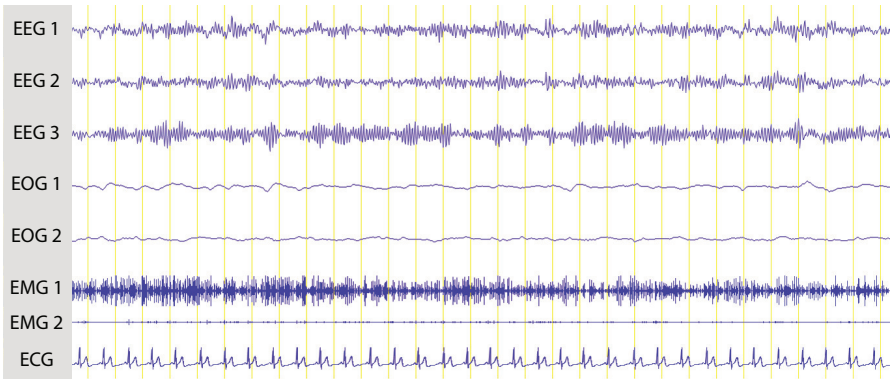


Figure A4 – Manuel de recommandations de l'AASM.

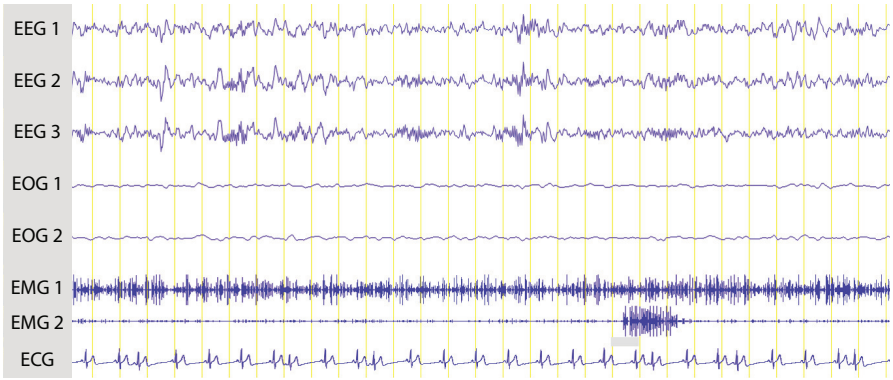
Sans rentrer dans les détails, d'autres éléments sont identifiés par le spécialiste du sommeil : diminutions importantes de la quantité d'oxygène dans le sang (appelées désaturations), très courts éveils (appelés micro-éveils), mouvements des jambes, ronflements, etc.

À partir de l'ensemble des éléments identifiés par le spécialiste du sommeil, il est ensuite possible d'effectuer un diagnostic.
Grâce à la lecture du sommeil et à l'hypnogramme, il sera ainsi possible d'identifier de possibles anomalies, et d'évaluer la fragmentation du sommeil. Par exemple, une somnolence diurne pourra être expliquée par un sommeil très fragmenté et/ou un manque de N3. Un patient qui tomberait trop vite en SP pourrait également être un patient souffrant de narcolepsie. Dans le cas d'une suspicion de SAHS, le spécialiste du sommeil calcule un index appelé Index d'Apnées Hypopnées (IAH). De manière simplifiée, cet index équivaut au nombre moyen d'événements respiratoires

1. Dans la majorité des cas, on réalise en effet un examen de nuit. Dans le cas d'une suspicion de narcolepsie, l'examen appelé Test Itératif de Latence à l'Endormissement (TILE), est effectué en journée.



(a)



(b)

Figure A5 – Exemple de signaux typiques d'époques en : (a) éveil calme, yeux fermés et (b) stade N2.

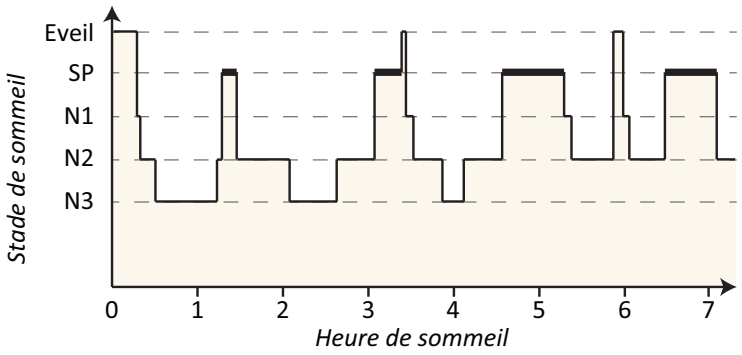


Figure A6 – Exemple d'hypnogramme.

(apnées et hypopnées) par heure, permettant ainsi d'estimer la gravité du SAHS. Il est aussi possible que les événements respiratoires apparaissent particulièrement lorsque le patient est en SP, on parle alors de REM-SAHS (*REM* est le terme anglais pour SP). Couplé avec les symptômes et les caractéristiques du patient, le spécialiste du sommeil pourra ainsi proposer un traitement adapté.

A.3 En résumé

Le diagnostic du sommeil peut être effectué à l'aide de différents outils. On retiendra la PV, permettant l'étude de la ventilation, et la PSG, permettant en plus l'étude des stades de sommeil et de l'activité cardiaque.
La PSG est plus complète mais plus complexe et chronophage. En effet, en dehors de la pose

des capteurs, il est nécessaire d'étudier les signaux enregistrés. Cette étude est effectuée par le spécialiste du sommeil qui visualise les différents signaux pour y reconnaître des événements ventilatoires et, dans le cas d'une PSG, les stades de sommeil (on dit qu'il effectue une lecture du sommeil). Cette lecture du sommeil, effectuée manuellement, aboutit à l'obtention d'un hypnogramme. Ce dernier permet de visualiser, par époques de 30 secondes, le stade de sommeil dans lequel le patient se trouve (éveil, N1, N2, N3 ou SP).
En couplant les informations obtenues avec les symptômes et les caractéristiques du patient, le médecin est en mesure d'effectuer son diagnostic et de proposer un traitement adapté.

Bibliographie

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., lloyd, R. M., Quan, S. F., Troester, M. M. et Vaughn, B. V. (2017). *The AASM Manual for the Scoring of Sleep and Associated Events : Rules, Terminology and Technical Specifications*. Numéro 2.4 de American Academy of Sleep Medicine. Darien IL.

SFRMS (2010). *Recommandations pour la pratique clinique du syndrome d'apnées hypopnées obstructives du sommeil de l'adulte*, volume 27.

Chapitre B

Revue de la littérature

B.1 Introduction

Afin de lire manuellement le sommeil, le médecin suit les recommandations de l'AASM (Berry *et al.*, 2017) et détermine, pour chaque époque, si le patient est éveillé, en sommeil léger, lent, lent profond ou paradoxal. Cette tâche requiert beaucoup de temps. En effet, un enregistrement durant en moyenne 8 heures, il faut étudier, pour chacune des 960 époques (environ) constituant chaque enregistrement, l'ensemble des voies électrophysiologiques enregistrées.
Voici une version très simplifiée des recommandations appliquées pour la lecture du sommeil :

Stade Éveil : Présence dans l'époque d'au moins 50 % d'ondes dites alpha et/ou de clignements des yeux, de Mouvements Oculaires Rapides (MORs) associés à un tonus musculaire mentonnier normal ou haut ou des mouvements oculaires de type lecture. Les ondes alpha sont les ondes EEG dont la fréquence est comprise entre 8 et 13 Hz (voir Figure B1). On recherche donc ici une activité EEG principale entre 8 et 13 Hz ainsi que de l'activité EOG pendant plus de la moitié de l'époque.

Stade N1 : Cette étape étant la transition entre l'éveil et le sommeil lent, on y observe un ralentissement de l'activité EEG. Les ondes alpha disparaissent donc et laissent place à des ondes theta de faible amplitude (voir Figure B1). On peut y observer des vertex tranchants et des Mouvements Oculaires Lents (MOLs).

Stade N2 : On reconnaît ici des complexes K et des fuseaux (voir Figure B2), des graphoéléments reconnaissables et spécifiques au stade N2.

Stade N3 : Le passage en N3 se fait lorsque l'on a au moins 20 % d'ondes lentes (dont la fréquence est comprise entre 0,5 et 2 Hz et l'amplitude est supérieure à 75 μV).

Stade SP : On observe souvent des MORs dans ce stade. Le tonus musculaire mentonnier est cependant le plus bas. On observe aussi des ondes en dents de scie et l'activité EEG est mixte avec une faible amplitude sans complexes K ni fuseaux.

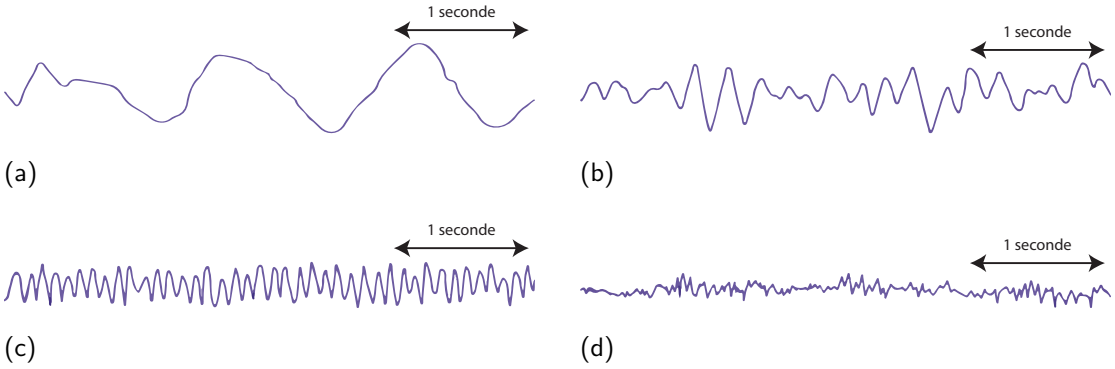


Figure B1 – Bandes de fréquences des ondes des EEGs utilisées pour la lecture du sommeil : a) ondes delta (0-3,99 Hz), b) ondes theta (4-7,99 Hz), c) ondes alpha (8-13 Hz) et d) ondes beta (>13 Hz).

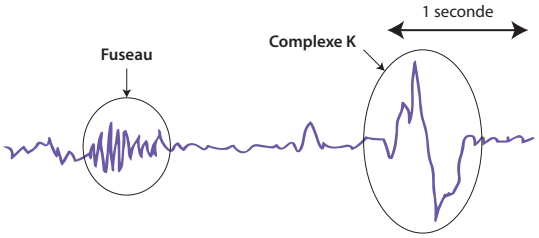


Figure B2 – Fuseau suivi d'un complexe K, visualisés sur une voie de type EEG.

En réalité, de nombreux autres éléments sont détaillés dans les recommandations AASM. On y trouve également des indications concernant les transitions possibles ou non entre les stades de sommeil, et des éléments concernant des cas spécifiques auxquels peut être confronté le lecteur. En effet, les signaux sont très différents d'un patient à l'autre, et la reconnaissance des ondes EEG et des grapho-éléments propres au patient est laborieuse. Pour cette raison, un spécialiste de la lecture du sommeil passe tout de même entre 30 minutes et 1 heure 30 minutes pour lire le sommeil. On comprend donc aisément qu'un outil d'analyse automatique du sommeil serait d'une grande aide aux médecins.

B.2 Détection automatique des stades de sommeil

B.2.1 Utilisation

Dans la majorité des cas, les analyses automatiques des stades de sommeil sont imaginées pour être exécutées après l'enregistrement des signaux.
Certaines analyses, dites entièrement automatiques, ne nécessitent pas l'intervention de l'Homme pour fonctionner. D'autres, dites semi-automatiques, nécessitent par exemple la lecture manuelle de quelques époques. L'algorithme utilise ensuite les époques identifiées manuellement pour classifier automatiquement les autres. L'avantage d'une approche semi-automatique est qu'elle permet une identification des stades adaptée à chaque enregistrement.
Il est toujours recommandé qu'il y ait une correction manuelle du spécialiste du sommeil en aval. L'utilisation à l'aveugle des analyses reste fortement déconseillée (pour différentes raisons explicitées ci-après, Section B.2.3).

B.2.2 Signaux utilisés

Les signaux utilisés par l'analyse automatique varient grandement d'une étude à l'autre. Ils dépendent principalement de l'objectif visé, mais aussi de la qualité des signaux disponibles. Les recommandations AASM spécifient les résolutions, les fréquences d'échantillonnage et les bandes passantes nécessaires pour l'analyse du sommeil. Bien souvent, c'est le constructeur de l'outil d'acquisition des signaux qui s'assure de respecter ces recommandations, et il n'est pas nécessaire, pour une analyse manuelle, d'effectuer des traitements supplémentaires.

Dans le cas où l'objectif est d'aider le spécialiste du sommeil lors de la lecture du sommeil (donc en PSG), l'ensemble des signaux polysomnographiques peut être utilisé. La majorité des études utilisent les signaux électrophysiologiques (EEGs, EMGs et EOGs), qui sont ceux utilisés lors d'une lecture manuelle.

Dans le cas où l'objectif est de proposer une analyse du sommeil en PV ou pour un système de diagnostic simplifié (de Type IV par exemple), une analyse à partir d'un ou de plusieurs signaux cardio-respiratoires est proposée. En effet, même si ces signaux ne sont pas directement liés à la profondeur du sommeil, ils sont tout de même influencés par celle-ci. Par exemple, le rythme de la respiration et le rythme cardiaque diminuent lors de l'endormissement. Il est très rare que l'analyse des stades de sommeil à partir de ces signaux soit suffisamment performante pour permettre une classification en 5 stades (éveil/N1/N2/N3/SP). Généralement, certains stades sont combinés. Les classifications en 4 stades (éveil/N1-N2/N3/SP), 3 stades (éveil/NREM [1]/SP) ou

1. NREM = N1, N2 et N3

encore 2 stades (éveil/sommeil ou SP/reste) sont donc assez répandues. Malgré qu'elles soient moins précises, ces classifications réduites apportent tout de même au médecin des informations qui sont normalement indisponibles pour ce type de système. Elles lui permettent ainsi d'effectuer un diagnostic plus précis.

Pour finir, certaines études proposent des systèmes composés d'une ou plusieurs voies EEG ou EOG seulement. Ces systèmes, pensés pour le dépistage, sont souvent destinés à être vendus directement au public. Ils permettent ainsi à l'utilisateur de monitorer lui-même son sommeil.

B.2.3 Problématiques et difficultés rencontrées

Les altérations des signaux

Les signaux utilisés pour la lecture du sommeil sont parfois altérés, rendant celle-ci délicate. Différents facteurs peuvent en être à l'origine :

- la pose des capteurs, essentielle pour l'enregistrement de signaux de qualité. Elle nécessite une fois encore un savoir-faire des spécialistes du sommeil. Une pâte conductive est ainsi utilisée pour permettre un enregistrement des signaux de bonne qualité. Cependant, il peut arriver que la conductivité soit altérée (transpiration excessive, fils tirés, etc.), résultant parfois en époques microvoltées. Ces époques ont un contenu temporel et spectral très peu présent qui rend la lecture extrêmement complexe ;
- la présence de certains syndromes ou dysfonctionnements qui impactent les signaux électrophysiologiques (EEGs : épilepsie, *alpha-delta sleep* ; EMGs : bruxisme, syndrome des jambes sans repos) ;
- la prise de certains traitements, comme les médicaments psychotropes qui impactent les EEGs.

Il existe aussi de nombreux artéfacts pouvant perturber les signaux. Ces artéfacts, en influençant localement le contenu temporel et fréquentiel, altèrent les informations utilisées pour la lecture du sommeil. On distingue les artéfacts extrinsèques, provenant d'une source extérieure au patient, et les artéfacts intrinsèques qui, à l'inverse, sont directement liés au patient. Marella (2012) présente les principaux artéfacts influant sur les EEGs, mais certains d'entre eux peuvent aussi influer sur les EOGs et les EMGs. Une liste des principaux artéfacts ainsi que leur effet sur les voies électrophysiologiques et leur prise en compte dans différentes études est disponible en Annexe page 39.

La variabilité des signaux

Pour toutes ces raisons, il existe une variabilité conséquente des signaux entre les patients. De ce fait, la lecture du sommeil est complexifiée.
Ainsi, la lecture manuelle d'un même enregistrement par deux spécialistes n'est pas rigoureusement identique. Le taux d'accord entre deux médecins (ou taux inter-scorer) n'atteint donc jamais 100 %. En fait, un même médecin ne saura pas non plus lire un même enregistrement deux fois de la même manière.
Lucey *et al.* (2016) ont effectué des évaluations mensuelles dans le cadre du programme de fiabilité inter-scores de l'AASM et ont obtenu une moyenne de 92 % d'accord avec le gold standard. Avec plus de 2500 lecteurs et 1800 époques provenant de 9 enregistrements différents, l'étude de Rosenberg et Van Hout (2013), menée par l'AASM, montre que le taux d'accord total avec la majorité n'atteint qu'approximativement 83 %. Un taux d'accord de 90 % est donc un taux excellent, et 80 % reste un taux largement acceptable.
Cette variabilité des signaux inter-patients crée une méfiance des médecins vis-à-vis des analyses automatiques. Il est donc essentiel de tester les modèles développés sur de nombreux enregistrements provenant de patients aux pathologies multiples.

Difficultés relatives à l'identification de chaque stade

À titre indicatif, et pour mieux appréhender l'impact des complexités présentées jusqu'ici, nous avons relevé les principaux obstacles rencontrés dans la littérature pour l'identification de chaque stade de sommeil individuellement :

Stade Éveil : Sa proportion varie fortement d'un patient à l'autre. En effet, elle dépend de nombreux facteurs comme les horaires habituels de coucher et de lever, l'âge, la consommation de caféine, d'alcool, la pratique d'exercice physique avant le coucher, la température et la luminosité de l'environnement, et bien sûr le besoin propre à chacun de sommeil. Certains syndromes (insomnie) ou phénomènes (*First Night Effect*), impactent également la proportion d'éveil.

> **First Night Effect (Agnew *et al.*, 1966) :**
> Il a été démontré que lors de la première nuit d'enregistrement, le sommeil du patient n'est pas entièrement représentatif de son sommeil habituel. En effet, le possible changement d'environnement et le stress dû aux nombreux capteurs et à l'enregistrement en lui même modifient la composition du sommeil. Agnew *et al.* (1966) montre que la proportion d'éveil est augmentée au contraire des stades N1 et SP.

De plus, l'éveil est un stade irrégulier. En effet, les signaux capturés lors d'un éveil agité, calme yeux ouverts ou encore calme yeux fermés sont très différents. Cependant, la détection automatique de l'éveil à partir des voies électrophysiologiques se fait généralement sans distinction entre ces trois types d'éveil. Les caractéristiques propres à chacun de ces types pouvant être très différentes, cela peut compliquer la détection automatique. Pour pallier cela, certaines études ne prennent pas en compte les époques durant lesquelles le patient est trop agité. Lucey *et al.* (2016) mettent ainsi de côté 9 % des époques enregistrées parce que l'on y trouve des mouvements dus à l'éveil. Koley et Dey (2012) rejettent, quant à eux, les époques durant lesquelles on trouve une amplitude suggérant des mouvements musculaires ou des mouvements oculaires sur les EEGs. D'autres études comme celles de Popovic *et al.* (2014) ou Charbonnier *et al.* (2011) présentent une approche comportant une étape de détection des artéfacts, qui leur permet ensuite d'éliminer certaines époques.

Stade N1 : Ce stade possède le plus petit taux d'accord inter-scorer (63 % selon l'étude de Rosenberg et Van Hout 2013). En effet, les désaccords entre les lecteurs sont le plus souvent situés aux transitions d'un état à l'autre, où il est parfois difficile de s'assurer visuellement que l'on est vraiment dans un stade et non l'autre. Le stade N1 étant bref et peu présent, il n'est donc pas surprenant qu'il soit particulièrement complexe à lire. La quasi-totalité des erreurs du stade N1 sont d'ailleurs des désaccords avec l'éveil ou le stade N2.
En ce qui concerne les analyses automatiques, c'est également le stade qui possède souvent le taux d'erreur le plus important. Pour pallier cela, certaines études comme celle de Berthomier *et al.* (2007) n'utilisent que les époques lues identiquement par plusieurs lecteurs.

Stade N2 : Ce sont des grapho-éléments tels que les complexes K ou les fuseaux (illustrés Figure B2) qui définissent le passage en N2. Cependant, ces grapho-éléments ne sont pas nécessairement présents dans chaque époque de ce stade. Ainsi, la difficulté liée à l'identification de ce stade est double : il faut détecter les grapho-éléments mais également considérer leur présence dans les époques à proximité.
Si la détection automatique des grapho-éléments reste assez rare, Mahvash Mohammadi *et al.* (2016) démontrent son importance en passant d'un taux d'accord d'environ 72 % à 84 % à l'aide des informations fréquentielles localisées (temps-fréquence) pour la détection de ces grapho-éléments.
La prise en compte des époques à proximité (ou du « contexte ») se limite souvent à un lissage final de l'hypnogramme.
Rosenberg et Van Hout (2013) indiquent également que la première époque de N2 est l'une des époques qui compte le plus d'erreurs (67.7%). La raison supposée serait la non-reconnaissance du premier complexe K ou fuseau.

Stade N3 : D'après l'étude AASM de Rosenberg et Van Hout (2013), les erreurs liées au stade N3 seraient des époques lues comme étant du N2. La difficulté à détecter correctement la fréquence des ondes lentes, leur amplitude et leur durée serait la cause de grand nombre de

ces erreurs. La solution proposée est d'ailleurs ici l'utilisation d'une analyse automatique de l'activité des ondes lentes.

Stade SP : La composition des ondes du stade SP étant proche de celle de l'éveil ou du N1, il est ici nécessaire d'utiliser d'autres signaux, et tout particulièrement l'EMG mentonnier et les EOGs. Il est en effet précisé dans les recommandations AASM que l'activité EEG en SP, de faible amplitude et de fréquence mixte, ressemble à celle du stade N1. De plus, il est commun d'observer en SP des ondes alpha, normalement associées à l'éveil. Cependant, ces ondes ont généralement une fréquence plus faible (1-2 Hz d'après les recommandations) que les ondes alpha visualisées lors de l'éveil. La détection de la fréquence spécifique aux ondes alpha de l'éveil pourrait donc permettre d'assurer une meilleure distinction éveil/N1/SP. Certaines études regroupent également les stades éveil, N1 et SP.
Une difficulté supplémentaire relevée par Rosenberg et Van Hout (2013) est la transition entre le stade N2 et le stade SP, qui serait l'époque avec le taux d'accord le plus faible (55.4%). En cause, une règle de transition complexe qui consiste à effectuer un retour en arrière (corriger les époques antérieures contingentes) après l'identification de la première époque de SP dite « certaine ».

On retiendra qu'en plus des problématiques liées à la complexité de la lecture en elle-même, il est courant que les signaux subissent des altérations. En réalité, les signaux et leur contenu (fréquentiel et temporel) sont donc assez éloignés de la théorie, rendant la lecture compliquée. En pratique, les spécialistes du sommeil visualisent rapidement toutes les époques avant de commencer la lecture. Cette étape leur permet de se familiariser avec les signaux propres à chaque patient et de connaître leurs spécificités.

B.2.4 Algorithmique

Bien souvent, les stades de sommeil sont identifiés par époque de 30 secondes, de la même manière qu'une lecture manuelle. L'étude de Stephansen *et al.* (2018) parue dans Nature présente des classifications automatiques en utilisant des segments de longueurs variables (5, 10, 15 et 30 secondes). Les meilleurs résultats ont été obtenus pour les segments de 30 secondes, malgré des performances très proches pour les segments plus courts.

Les approches développées peuvent avoir recours à des algorithmes issus d'intelligence artificielle plus ou moins élaborés.

Intelligence artificielle, Machine Learning et Deep Learning :
L'Intelligence Artificielle (IA) est née dans les années 50. Elle regroupe « l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence » (Larousse). Tout processus ou algorithme ayant pour objectif de réaliser une tâche qui nécessite de l'« intelligence » fait donc partie du domaine de l'IA.
Le Machine Learning (ML), qui a émergé dans les années 80, regroupe les processus capables d'utiliser, en plus, des approches mathématiques et statistiques pour « apprendre » à réaliser cette tâche. Il n'est ainsi plus nécessaire de programmer le processus pour réaliser la tâche. Il réalise lui-même son « apprentissage » à partir des observations qui lui sont fournies en entrée. Le réseau de neurones artificiels, par exemple, construit son apprentissage en créant des connexions neuronales, comme peut le faire un cerveau humain. Le ML est donc une sous-partie de l'IA.
Le Deep Learning (DL), qui a émergé dans les années 2010, est quant à lui une sous-partie du ML. Les algorithmes de DL peuvent être considérés comme une extension des réseaux de neurones artificiels. Ils sont composés de nombreuses couches de neurones qui communiquent les unes avec les autres. Grâce à leur architecture dite « profonde », ils sont plus puissants. Les algorithmes de DL sont ainsi capables, à partir de signaux bruts, d'« apprendre » à réaliser des tâches très complexes.

Dans le cas de l'analyse du sommeil, certaines méthodes sont implémentées de telle sorte que l'algorithme reproduise et imite, à l'aide de règles logiques, l'analyse manuelle des stades de sommeil. Ces algorithmes font donc partie de l'IA, sans être ni du ML, ni du DL. Dans la suite de ce manuscrit, le terme « approches basées sur les connaissances des experts » sera utilisé pour

y faire référence. L'analyse manuelle étant complexe, ces méthodes sont peu répandues et obtiennent généralement des résultats décevants. Elles ne seront donc pas détaillées par la suite.
D'autres méthodes permettent à l'analyse automatique de générer ses propres règles en utilisant des raisonnements mathématiques et statistiques, à partir d'un jeu de données bien structuré. Bien plus répandues, elles obtiennent des résultats tout à fait satisfaisants. Ces méthodes font donc partie du ML, mais pas du DL. Dans la suite de ce manuscrit, afin d'éviter toute confusion, on considérera que les méthodes dites « de ML » font référence aux méthodes de ML sans englober les méthodes de DL.
Les méthodes de Deep Learning (DL), qui peuvent quant à elles être utilisées à partir de données brutes et sans être guidées par l'Homme, obtiennent des résultats souvent épatants. Cependant, elles soulèvent des inquiétudes de par leur opacité. En effet, elles sont souvent considérées comme des boites noires et sont difficilement acceptées par les spécialistes du sommeil. Elles peuvent également être vulnérables à la présence de certaines perturbations, qui peuvent leurrer l'algorithme (Dalvi *et al.*, 2004; Lab, sd).
Pour finir, il existe des approches dites hybrides, qui combinent le ML ou le DL avec des connaissances des experts.

Approches de Machine Learning

Les analyses automatiques des stades de sommeil utilisant du ML ont fait l'objet de nombreuses études depuis la fin des années 90. En 2000, Penzel et Conradt détaillent les principes de l'analyse automatique du sommeil (Penzel et Conradt, 2000). Ils concluent que l'utilisation du ML est possible, mais soulignent la nécessité d'une relecture manuelle. Vingts ans plus tard, malgré de nombreux logiciels permettant l'analyse automatique, leur utilisation reste controversée au sein de la communauté médicale.

Le fonctionnement général des analyses basées sur du ML reste le même. Ces méthodes utilisent une lecture manuelle, réalisée en amont par un spécialiste du sommeil, pour permettre à l'algorithme de réaliser son apprentissage. Utilisant une référence, l'apprentissage est dit supervisé.
La structure algorithmique peut être décomposée de la sorte :

1. **Pré-processing** : Dans un premier temps, les signaux d'entrée sont nettoyés et préparés. Des filtrages peuvent être utilisés pour réduire le niveau de bruit. La normalisation des signaux ainsi que la reconnaissance et la suppression de certains artéfacts peuvent être réalisées pour faciliter l'apprentissage futur. Dans certaines études, des méthodes de séparation de source sont utilisées pour isoler les informations cérébrales, oculaires et cardiaques, parfois visibles sur différents signaux électrophysiologiques à la fois.

2. **Évaluation des caractéristiques** : Les signaux qui découlent du pré-processing sont ensuite décomposés en plusieurs vecteurs de données ayant un sens mathématique ou statistique. Par exemple, il est possible de simplement évaluer l'amplitude, la moyenne ou l'écart-type d'une voie EEG pour chaque époque. Chaque vecteur constitue une caractéristique. Il existe de nombreuses caractéristiques possibles. On les décrit souvent comme faisant partie du domaine statistique (comme la médiane), temporel (comme l'amplitude), fréquentiel (comme la quantité d'ondes alpha), non-linéaire (comme l'entropie approximative) ou comme étant lié à une forme particulière, tel qu'un complexe ECG.

3. **Réduction de la dimensionnalité** : Il est possible qu'après leur évaluation, un nombre conséquent de caractéristiques (et donc de dimensions) soit obtenu. Si ces nombreuses caractéristiques garantissent en quelque sorte une transcription complète des informations contenues dans le signal d'origine, elles peuvent constituer un obstacle pour l'apprentissage du modèle. En effet, ce dernier est plus rapide et efficace lorsqu'il n'est pas confronté à de la redondance d'information, ou à des caractéristiques non pertinentes pour la tâche qu'il doit apprendre à effectuer. Il est donc nécessaire de réduire la dimensionnalité du problème à l'aide de méthodes variées. On obtient alors un ensemble de caractéristiques choisi spécifiquement pour le problème à résoudre.
   Attention cependant à la malédiction de la dimensionnalité. En effet, plus le nombre de caractéristiques d'origine est important, et plus les stratégies de réduction de la dimensionnalité sont difficiles à mettre en place. Avec un nombre de dimensions élevé, l'information

est éparpillée dans des caractéristiques qui diffèrent de bien des façons. Il est alors très complexe d'identifier des caractéristiques qui seraient plus pertinentes que les autres, et donc de réduire le nombre de caractéristiques à retenir pour la classification.

4. **Classification** : À partir de l'ensemble des caractéristiques, le classifieur s'entraîne à identifier les stades de sommeil. Dans le cas d'un apprentissage supervisé, il utilise l'hypnogramme lu manuellement comme étant sa référence. Différents algorithmes de classification peuvent être testés et comparés.
5. **Enregistrement des modèles entraînés** : Le modèle de classification entraîné est enregistré afin d'être réutilisé en aval.
6. **Évaluation des performances** : Les performances du modèle sont évaluées en comparant les stades de sommeil identifiés manuellement et automatiquement. Pour cela, on utilise généralement un jeu de données inconnu pour le modèle, n'ayant pas servi à l'apprentissage.

En fonction des performances du modèle, les différentes étapes sont répétées jusqu'à l'obtention d'un modèle final. C'est ce modèle qui sera proposé aux spécialistes du sommeil.

Dans le domaine du diagnostic du sommeil, certaines méthodes de réduction de la dimensionnalité, de classification et d'évaluation des performances ressortent :

La réduction de la dimensionnalité est très peu utilisée dans le domaine du diagnostic du sommeil. En effet, dans de nombreux cas, un nombre peu important de caractéristiques est évalué. Ces caractéristiques sont d'ailleurs souvent les mêmes, puisqu'elles traduisent assez clairement des éléments médicaux.
Il est possible de débuter la réduction de la dimensionnalité par des tests permettant la comparaison des caractéristiques entre elles. Ainsi, des caractéristiques quantitatives (c'est le plus souvent le cas) similaires (et donc redondantes) peuvent être identifiées à l'aide de la corrélation de Pearson ou de Spearman (tests paramétriques et non paramétriques respectivement).
Des méthodes de sélection de caractéristiques peuvent ensuite être implémentées. Elles servent à définir le sous-ensemble le plus pertinent parmi l'ensemble des caractéristiques de départ. Pour cela, deux approches sont possibles :

— les approches « *filters* », qui consistent à supprimer des caractéristiques jugées individuellement comme étant les moins bonnes (mesures de distance, de corrélation, etc.) ;
— les approches « *wrappers* », qui consistent à construire un ensemble de caractéristiques jugé comme étant le meilleur. Ces approches peuvent être exhaustives (B&B - Clausen 1999), séquentielles (SFS ou SFFS - Pudil *et al.* 1994) ou non déterministes (algorithmes génétiques - Goldberg 1989).

Les études de Charbonnier *et al.* (2011); Fonseca *et al.* (2015); Lajnef *et al.* (2015) font partie des rares études dans lesquelles une étape de réduction de la dimensionnalité est utilisée. Ils utilisent une approche « *wrapper* » séquentielle nommée *Sequential Forward Selection* (SFS). SFS est une heuristique, qui fonctionne par approches successives. À la première itération, elle sélectionne la caractéristique permettant d'obtenir les meilleures performances. Cela constitue le premier ensemble. Ensuite, à chaque itération, elle teste l'ajout de chaque caractéristique restante à l'ensemble précédent, et sélectionne la caractéristique pour laquelle le nouvel ensemble génère les meilleures performances. En étudiant l'évolution des performances obtenues en fonction du nombre de caractéristiques, il est ainsi possible de choisir le sous-ensemble de caractéristiques le plus optimal. Avec à l'origine 142 caractéristiques, Fonseca *et al.* (2015) observent une performance optimale pour un nombre de 105 caractéristiques seulement. Cependant, l'évolution des performances entre 80 et 105 caractéristiques étant minime, ils choisissent de se limiter au nombre de 80 caractéristiques. Les résultats finaux, qu'ils évaluent à partir de cet ensemble optimisé, sont supérieurs à ceux qu'ils auraient obtenus à partir des 142 caractéristiques initiales. La méthode SFS a cependant l'inconvénient de ne pas permettre la suppression de caractéristiques devenues obsolètes ou redondantes après l'ajout d'une nouvelle caractéristique. La méthode *Sequential Floating Forward Selection* (SFFS), qui fonctionne sur le même principe, permet de supprimer une caractéristique précédemment sélectionnée mais devenue peu significative dans l'ensemble actuel.
Pour finir, des méthodes dites d'extraction de caractéristiques peuvent être utilisées. Ces méthodes servent à construire de nouvelles caractéristiques en combinant les caractéristiques initiales. Dans le domaine du diagnostic du sommeil, ces approches sont peu utilisées. En effet, les

caractéristiques obtenues étant des combinaisons (linéaires ou non) des caractéristiques initiales, elles ajoutent une forme d'opacité. Cependant, l'utilisation d'une méthode d'extraction de caractéristiques nommée Analyse en Composantes Principales (ACP) et ses variantes ressortent tout de même, non pas pour l'analyse des stades de sommeil mais pour la suppression d'artéfacts (voir les études de Enshaeifar *et al.* 2016; Hapuarachchi 2006; McCurry 2017; Nguyen *et al.* 2012). En bref, l'ACP (Hotelling, 1933) permet de résumer l'information contenue dans un ensemble de données en un nombre de caractéristiques synthétiques constituées d'une combinaison linéaire des caractéristiques originelles. Ces caractéristiques synthétiques sont appelées « composantes principales ».

En conclusion :
Malgré le fait que les méthodes de réduction de la dimensionnalité ne soient que rarement utilisées pour l'analyse automatique des stades de sommeil, les méthodes de comparaison de caractéristiques et celles de sélection de caractéristiques ont tout intérêt à être testées. En effet, elle permettent l'obtention d'un ensemble de caractéristiques optimal pour l'application, en limitant la redondance et les caractéristiques non pertinentes.

Parmi les méthodes de **classification** existantes, certaines ressortent plus que d'autres dans le cas de l'analyse automatique des stades de sommeil :

— le Séparateur à Vaste Marge ou machine à vecteur de support (SVM) est l'un des classifieurs les plus utilisés. Dans le cas de l'analyse du sommeil, le SVM (Cortes et Vapnik, 1995) consiste à cartographier chaque époque dans un espace, de manière à ce que les époques de différents stade de sommeil soient séparées par une marge la plus vaste possible. Les époques à classifier sont ensuite placées dans ce même espace. Selon leur position, elles sont rattachées au stade le plus proche. Cette méthode est très répandue pour notre application (Aktaruzzaman *et al.* 2017; Enshaeifar *et al.* 2016; Fehrmann 2013; Koley et Dey 2012; Lajnef *et al.* 2015; Lewicke *et al.* 2008; Mahvash Mohammadi *et al.* 2016; Yılmaz *et al.* 2010) ;

— l'Analyse Discriminante Linéaire (ADL) est également un classifieur régulièrement utilisé. Dans le cas de l'analyse du sommeil, l'ADL (Cohen *et al.*, 2013) est utilisée pour trouver une combinaison linéaire de caractéristiques qui permet de caractériser ou séparer les stades de sommeil du patient de manière optimale (voir les études de Ebrahimi *et al.* 2013; Long *et al.* 2014, 2015; Redmond *et al.* 2007). L'Analyse Discriminante Quadratique (ADQ), étroitement liée à l'ADL, peut également être utilisée (voir études de Ebrahimi *et al.* 2013; Redmond *et al.* 2007; Yılmaz *et al.* 2010) ;

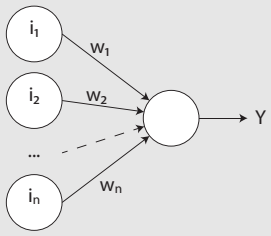
— le Modèle de Markov à états Cachés (MMC) est un système Bayésien. Contrairement à une chaîne de Markov, le MMC (Baum et Pietrie, 1966) est un automate pour lequel les états sont inconnus de l'utilisateur. Dans notre cas, cet automate permet, à partir des caractéristiques de chaque époque, d'estimer sa probabilité d'être dans chaque stade de sommeil (voir les études de Doroshenkov *et al.* 2007; Mendez *et al.* 2009, 2010) ;

— la méthode des k plus proches voisins (kNN) utilise quant-à elle une mesure de distance. La méthode kNN (Altman, 1992) utilise pour cela un espace créé à partir des différentes caractéristiques. Les époques sont classifiées selon le stade majoritaire parmi les k (nombre à définir) époques les plus proches, du point de vue de la mesure de distance (voir les études de Malaekah 2016; Yılmaz *et al.* 2010) ;

— les réseaux de neurones artificiels ont un fonctionnement inspiré des neurones biologiques (du cerveau humain). Il existe de nombreuses sortes de réseaux de neurones :

  • le perceptron peut-être considéré comme le réseau neuronal le plus simple. Il est constitué d'un unique neurone auquel toutes les entrées sont connectées, et il ne possède qu'une sortie ;

Fonctionnement du perceptron :



Le perceptron somme ses entrées après les avoir pondérées.

$$Y = f(\sum_{a=1}^{n} i_a \times w_a)$$

C'est à partir du résultat obtenu que la sortie est évaluée, en utilisant une fonction d'activation f (par exemple la fonction heaviside, linéaire ou sigmoïde).

Figure B3 – Schéma d'un perceptron.

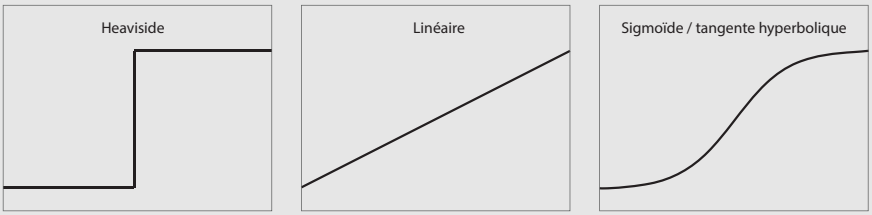


Figure B4 – Principales fonctions d'activation.

- le perceptron multicouche (Rumelhart *et al.*, 1986), ou *Multilayer Perceptron* (MLP), est composé de plusieurs perceptrons simples, agencés entre eux. Comme précédemment, une couche d'entrée reçoit les données. Cependant, cette fois-ci, une ou plusieurs couches cachées précèdent la couche de sortie.

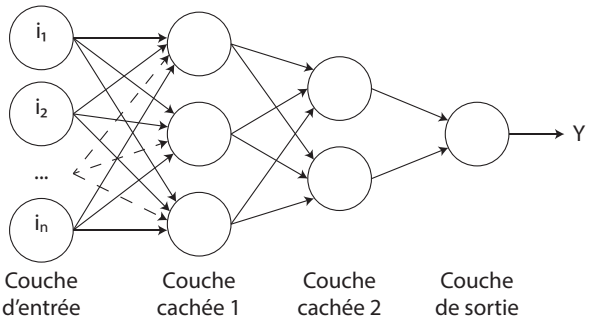


Figure B5 – Schéma d'un MLP à deux couches cachées de 3 et 2 neurones, respectivement.

Le MLP est généralement utilisé avec rétropropagation du gradient de l'erreur, qui permet de propager l'erreur de la dernière couche vers la première afin d'ajuster les poids (notamment ceux des couches cachées) en fonction de l'erreur finale obtenue. Charbonnier *et al.* (2011); Garcia-Molina *et al.* (2012); Lewicke *et al.* (2008) ont utilisé des MLPs pour analyser le sommeil ;

- le *Recurrent Neural Network* (RNN), qui possède, au contraire du MLP, une temporalité le rendant particulièrement adapté aux données séquentielles. Son fonctionnement lui permet donc de traiter en entrée non pas un vecteur, mais une séquence de vecteurs. Cependant, dans sa version classique, le RNN est difficile à entraîner (du fait d'un phénomène appelé vanishing gradient), et possède une mémoire relativement courte. Parmi différentes variantes de RNN créées pour remédier à ces problèmes, on retiendra le *Long Short Term Memory* (LSTM) (Hochreiter et Schmidhuber, 1997). Avec leur habilité à identifier des patterns pour faire des propositions plus sensées, les LSTMs atteignent d'excellents résultats dans les domaines de la traduction automatique, de la modélisation linguistique et du traitement multilingue des langues. Dans le cas de l'analyse automatique du sommeil, ils s'avèrent intéressants puisque la notion de temporalité est importante pour la création de l'hypnogramme (voir l'étude de Hsu *et al.* 2013) ;

- les *Convolutional Neural Networks* (CNNs), à l'inverse, sont réputés pour les applications pour lesquelles la dimension spatiale est importante, comme la reconnaissance d'images. Les données d'entrée sont analysées par régions qui se chevauchent (appelées champs récepteurs). Comparé à un MLP, l'analyse est donc organisée selon la disposition géographique des données d'entrées. Dans le cas de l'analyse automatique du sommeil, ils s'avèrent intéressants puisque l'identification de certains stades nécessite l'identification de motifs particuliers (voir l'étude de Sors *et al.* 2018).

— pour finir, les méthodes combinant plusieurs classifieurs sont également très répandues, telsque le boosting, voting et les forêts d'arbres décisionnels ou forêts aléatoires, de l'anglais *Random Forest* (RF). Les RF (Ho, 1995) sont constituées d'une multitude d'arbres de décision [2], entraînés à prédire, dans notre cas, le stade de sommeil de chaque époque. Chaque arbre est entraîné à partir d'un sous-groupe de données d'apprentissage formé de manière aléatoire, et d'une partie seulement des caractéristiques, sélectionnées de manière aléatoire. Le choix final est réalisé par vote majoritaire entre les différents arbres. Les études de Fraiwan *et al.* 2012; Xiao *et al.* 2013; Zokaeinikoo 2016 ont utilisé des RF.

Dans une étude parue en 2007, Lotte *et al.* ont comparé différentes méthodes d'apprentissage avec pour problématique l'étude des voies EEGs pour la création d'une Interface Neuronale Directe (IND) ou d'une Interface Cerveau-Machine (ICM). Des recommandations ont été proposées pour le choix du classifieur. L'objectif n'étant pas exactement le même que pour notre application, les conclusions en terme de supériorité des méthodes par rapport aux autres n'ont pas été admises pour le présent travail.

En conclusion :

Il existe un nombre conséquent d'approches permettant la classification automatique des stades de sommeil à partir d'un ensemble de caractéristiques. Il semblerait qu'aucune méthode n'ait démontré une supériorité nette sur les autres dans le cas de l'analyse automatique des stades de sommeil. Il est donc d'usage de sélectionner plusieurs méthodes à tester, selon nos objectifs.

L'**évaluation des performances** se fait nécessairement à partir de données n'ayant pas été utilisées lors de l'apprentissage. En amont, on partitionne donc les données pour avoir un groupe d'apprentissage (utilisé pour entraîner le modèle), de validation (facultatif, utilisé pour contrôler l'ajustement les paramètres du modèle comme par exemple des poids, etc.) et de test (utilisé pour évaluer le modèle entraîné).

Il est très important, dans le cas de notre application, de ne pas utiliser les époques d'un même enregistrement pour l'apprentissage - plus validation si nécessaire - et le test. En effet, pour un enregistrement unique, les caractéristiques des époques d'un même stade sont très proches voire similaires. Évaluées à partir d'époques semblables à celles ayant servi à l'apprentissage, les performances sont surestimées de manière importante. La plupart des études partitionnent correctement leurs données en considérant les enregistrements au lieu des époques directement.

L'échantillonnage peut être effectué de différentes manières, selon la quantité de données, l'objectif recherché, etc.

Un échantillonnage relativement simple mais couramment utilisé consiste à utiliser 2/3 des données pour l'apprentissage - plus validation si nécessaire - et 1/3 restant pour le test (voir les études de Fraiwan *et al.* 2012; Lewicke *et al.* 2008; Xiao *et al.* 2013).

La validation croisée est également très répandue. On retrouve principalement les deux approches suivante dans la littérature :

— la validation croisée à k blocs, de l'anglais *k-fold Cross Validation* (*k-fold CV*) consiste à séparer les enregistrement en k blocs. Successivement, chaque bloc est sélectionné comme étant l'ensemble de test, et le reste (donc les k-1 blocs restants) comme étant l'ensemble d'apprentissage. Au final, k modèles sont donc entraînés et testés, et le meilleur est retenu. Les études de Charbonnier *et al.* (2011); Fehrmann (2013); Fonseca *et al.* (2015); Koley et Dey (2012); Lajnef *et al.* (2015); Long *et al.* (2015); Mahvash Mohammadi *et al.* (2016); Malaekah (2016); Yılmaz *et al.* (2010); Zokaeinikoo (2016) utilisent cette approche ;

— la validation croisée un contre tous, de l'anglais *Leave-One-Out Cross Validation* (LOOCV) consiste à considérer chaque enregistrement comme étant l'ensemble de test, et tous les

2. Un arbre de décision est un graphe constitué de noeuds, de branches et de feuilles. Un noeud correspond à un test logique appliqué à une caractéristique (par exemple, « L'amplitude de l'EEG est-elle supérieure à 75 μV ? »), et qui est à l'origine de deux branches (« Oui » ou « Non »). Chaque branche aboutit à un autre noeud, ou bien à une feuille. Dans notre cas, la valeur des feuilles correspond à l'un des stade de sommeil.

autres comme étant l'ensemble d'apprentissage. Cela revient à effectuer une *k-fold CV* avec k égal au nombre d'enregistrements total. De la même manière, le modèle ayant obtenu les meilleures performances est retenu.
Dans le cas de l'analyse du sommeil, certains désavantages peuvent être liés à l'utilisation de cette méthode. En effet, comme expliqué précédemment, les signaux sont extrêmement variables d'un patient à l'autre. Or, le choix du meilleur modèle repose sur les performances obtenues à partir d'un seul enregistrement. Dans le cas d'un enregistrement de test provenant d'un patient sans troubles du sommeil, avec peu de transitions et peu d'altérations des signaux, le modèle entraîné aura de grandes chances d'obtenir de bonnes performances, au contraire du cas précédent. Les études de Aktaruzzaman *et al.* (2017); Garcia-Molina *et al.* (2012); Long *et al.* (2014); Mendez *et al.* (2009, 2010); Redmond *et al.* (2007) utilisent cette approche.

En conclusion :
Il est nécessaire d'évaluer la méthode à partir de données provenant de patients n'ayant pas été utilisés lors de l'apprentissage. On parle de données d'apprentissage et de données de test. Selon la quantité d'enregistrements à disposition, il est possible de partitionner les données de différentes manières.

Approches de Deep Learning

Les études utilisant le DL sont bien plus récentes. La Figure B6, extraite de l'étude de Roy *et al.* (2019), présente d'ailleurs le nombre d'études portées sur le traitement des EEGs en Deep Learning entre les années 2010 et 2018.

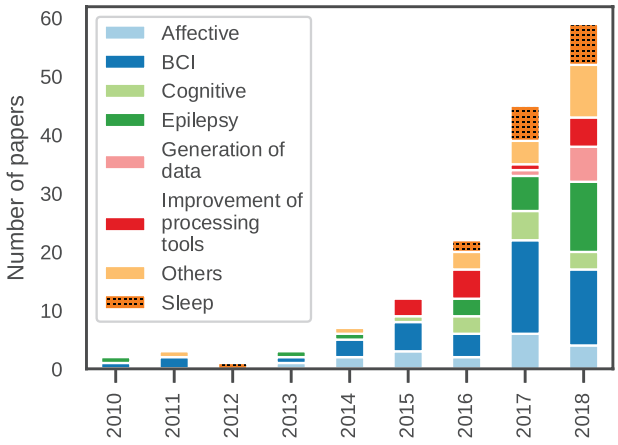


Figure B6 – Figure extraite de l'étude de Roy *et al.* (2019), dénombrant les publications sur le traitement des EEGs en DL, par domaine et par an. Différentes catégories ont été identifiées, parmi lesquelles le sommeil (en orange).

Les analyses basées sur du DL peuvent être considérées comme une extension des réseaux de neurones artificiels. Cette fois-ci, il n'est pas nécessaire de réaliser les étapes de pré-processing, d'évaluation des caractéristiques et de réduction de la dimensionnalité, telles que précédemment définies. En effet, ces étapes sont directement intégrées dans les algorithmes d'apprentissage supervisé de DL, plus puissants que ceux de ML, qui peuvent recevoir en entrée des signaux bruts (ou simplement normalisés).
La structure algorithmique est donc constituée d'un ensemble de couches agencées de telle sorte à réaliser l'apprentissage. Par exemple, les études de Biswal *et al.* (2018); Bresch *et al.* (2018); Stephansen *et al.* (2018); Zhang *et al.* (2019) présentent des approches combinant un CNN (permettant d'identifier des éléments pertinents pour la classification) et un RNN ou LSTM (permettant de générer la séquence des stades de sommeil, soit l'hypnogramme). Avec un raisonnement similaire, Patanaik *et al.* (2018) combinent un CNN avec un MLP prenant en compte l'époque à classifier mais également les cinq époques précédentes et suivantes. Dans leur étude, Dong *et al.* (2018) combinent un MLP avec un LSTM.

En conclusion :
Les approches de DL englobent les étapes de pré-processing, d'évaluation des caractéristiques, de réduction de la dimensionnalité et de classification. Elles peuvent recevoir en entrée un signal brut directement. On retiendra que, dans le cas de l'analyse du sommeil, les structures algorithmiques mises en place combinent souvent un CNN, particulièrement efficace pour les problèmes ayant une dimension spatiale, et un RNN (ou LSTM), qui excelle dans les problèmes ayant une dimension temporelle. Cela s'explique par le fait que la classification des stades de sommeil nécessite à la fois l'identification d'éléments remarquables dans les signaux (ondes particulières, motifs), et la connaissance des époques précédentes et suivantes.

En ce qui concerne l'**évaluation des performances**, les méthodes sont principalement les mêmes qu'en ML. On notera cependant que l'ensemble de validation (qui était rarement utilisé en ML), est souvent nécessaire dans le cas du DL, pour le paramétrage des réseaux neuronaux.

Approches hybrides

Récemment, l'article de synthèse de Fiorillo *et al.* (2019) a présenté les principaux obstacles à l'utilisation de la lecture automatique du sommeil par les cliniciens. La principale limitation identifiée a été le fonctionnement opaque des algorithmes de DL.
Aujourd'hui, un certain nombre de chercheurs tentent d'améliorer l'interprétabilité de leurs modèles Doshi-Velez et Kim (2017). En effet, les applications médicales sont souvent confrontées à une importante variabilité des signaux, ainsi qu'à un grand nombre de cas particuliers. Dans l'idéal, l'apprentissage devrait être réalisé sur un jeu de données couvrant toutes les situations possibles, ce qui est peu probable pour les applications médicales. Sachant cela, les approches dites « boites noires » ont du mal à convaincre, d'autant plus que l'obtention d'un grand nombre d'enregistrements du sommeil est parfois difficilement réalisable. Sur 154 analyses automatiques des EEGs en DL présentées dans l'article de synthèse de Roy *et al.* (2019), seules la moitié d'entre elles comprenaient plus de 13 patients. Or, sur peu d'enregistrements, le risque de surapprentissage [3] est important.
Pour toutes ces raisons, de nouvelles approches combinant les connaissances médicales avec du ML ou DL ont émergé.
L'étude de Al-Hussaini *et al.* (2019) combine un CNN avec des règles médicales et la création de prototypes représentant les époques de chaque stade. La comparaison entre les époques réelles et les prototypes permet ensuite de classifier le sommeil via un modèle facilement interprétable (arbres de décision ou régression logistiques).
L'analyse automatique présentée dans Chen (2016); Chen *et al.* (2019); Ugon (2015) repose quant à elle sur la fusion symbolique. Cette approche permet de classifier les stades de sommeil à partir de caractéristiques qualitatives permettant de décrire au mieux les recommandations AASM pour la lecture du sommeil (par exemple, les caractéristiques « *EEGstabilityStable* » ou « *EEGstabilityUnstable* »). Des règles sont ensuite utilisées pour identifier le stade de sommeil à partir de ces caractéristiques.

En conclusion :
Récemment, des méthodes dites hybrides sont développées afin d'augmenter l'interprétabilité de l'analyse automatique. Ces méthodes combinent généralement du ML ou DL avec des éléments extraits directement des connaissances des experts.

B.3 Évaluation de la classification

B.3.1 Estimation des performances de l'hypnogramme automatique

Il existe de nombreux indices statistiques permettant de quantifier la justesse de la classification automatique comparée à la référence (pour rappel, la lecture manuelle). Les performances sont évaluées sur l'ensemble de test.

3. En ML ou DL, on parle de surapprentissage (ou overfitting en anglais) lorsque le modèle apprenant (le classifieur dans notre cas) est trop adapté aux données d'apprentissage et à ses particularités, et a des difficultés à se généraliser. Dès lors, il obtient des performances excellentes sur les données d'apprentissage, mais bien plus faibles sur celles de test.

Il est possible d'évaluer les performances d'identification de chaque stade de sommeil individuellement. Dans ce cas, on considère successivement chaque stade comme étant la classe positive, et l'ensemble des autres classes comme étant la classe négative. On détermine ensuite le nombre de Vrais Positifs (VP), Vrais Négatifs (VN), Faux Positifs (FP) et Faux Négatifs (FN) pour chaque stade.

Les VP correspondent ainsi au nombre d'époques correctement identifiées dans le stade en question. Les VN correspondent aux époques correctement identifiées comme n'étant pas dans le stade en question. Les FP correspondent aux époques automatiquement identifiées comme étant dans le stade en question mais en désaccord avec la lecture de référence. Enfin, les FN correspondent aux époques qui auraient dû être identifiées dans le stade en question mais ne l'ont pas été. Les VP et VN correspondent donc aux succès, contrairement aux FP et FN.

Table B1 – Matrice de confusion dans le cas de deux états (positif et négatif).

		Auto	
		Pos	Neg
Ref	Pos	VP	FN
	Neg	FP	VN

Plusieurs indices statistiques peuvent être évaluées à partir de ces quantités :

— la Sensibilité (Se) décrit la capacité de l'analyse à identifier correctement le stade en question lorsqu'il est présent. Ainsi, la sensibilité du stade d'éveil correspond à la proportion d'époques correctement identifiées comme étant de l'éveil parmi les époques d'éveil de la référence ;

$$Se\ (\%) = \frac{VP}{VP + FN} \times 100$$

— la Spécificité (Sp) accompagne souvent la Se. Pour l'exemple précédent, la Sp du stade d'éveil correspond à la proportion d'époques correctement identifiées comme étant du sommeil parmi les époques de sommeil de la référence ;

$$Sp\ (\%) = \frac{VN}{VN + FP} \times 100$$

— la Valeur Prédictive Positive (VPP) décrit quant-à-elle la probabilité qu'une époque soit effectivement de l'éveil lorsque l'analyse automatique l'indique. En reprenant le même exemple, la VPP de l'éveil correspond à la proportion d'époques correctement identifiées comme étant de l'éveil parmi les époques d'éveil de l'analyse automatique ;

$$VPP\ (\%) = \frac{VP}{VP + FP} \times 100$$

— la Valeur Prédictive Négative (VPN) accompagne souvent la VPP. Pour l'exemple précédent, la VPN du stade d'éveil correspond à la proportion d'époques correctement identifiées comme étant du sommeil parmi les époques de sommeil de l'analyse automatique ;

$$VPN\ (\%) = \frac{VN}{VN + FN} \times 100$$

— le taux d'accord (Acc) correspond à la proportion d'époques correctement identifiées parmi l'ensemble des époques.

$$Acc\ (\%) = \frac{VP + VN}{VP + VN + FP + FN} \times 100$$

Ces indicateurs doivent donc être les plus grands possibles pour chaque stade de sommeil. Généralement, on cherche un équilibre entre la sur-détection (les FP) et la sous-détection (les FN). Pour cela, il est nécessaire de considérer plusieurs de ces indicateurs. On étudie donc souvent l'Acc avec le couple (Se ; Sp), ou le couple (VPP ; VPN).

Il est possible d'évaluer les performances de l'analyse automatique en considérant l'ensemble des stades de sommeil. Dans ce cas, il est nécessaire de créer une table de contingence (voir Table B2). Cette table, aussi appelée matrice de confusion, permet de connaître le nombre d'époques identifiées dans chaque stade, selon le stade lu par le spécialiste.

Les n_{ii}, situés sur la diagonale, correspondent aux succès. Le reste correspond aux erreurs. Une

Table B2 – Table de contingence (ou matrice de confusion) dans le cas de l'analyse automatique des stades de sommeil.

		Analyse automatique					
		Éveil	N1	N2	N3	SP	Total
Référence	Éveil	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	$n_{1.}$
	N1	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	$n_{2.}$
	N2	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	$n_{3.}$
	N3	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	$n_{4.}$
	SP	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	$n_{5.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	n

manière d'étudier de façon plus visuelle les performances de l'analyse automatique consiste à colorer la table de contingence. Le terme heatmap est généralement utilisé pour décrire ce type de représentation. Cependant, comme mentionné dans la Section A, certains stades sont bien plus présents que d'autres. Il est donc d'usage d'exprimer le nombre d'époque en pourcentage du nombre d'époques de chaque stade (du point de vue de la référence). On utilise souvent une échelle de couleur allant du bleu clair, pour les nombres faibles, au bleu foncé, pour les nombres importants. La Figure B7 présente un exemple de heatmap utilisant la palette de couleurs Matlab (généralement utilisée) « Bone ».

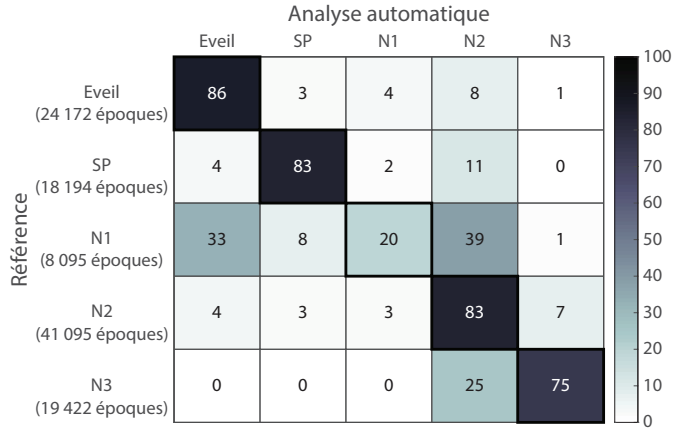


Figure B7 – Exemple de heatmap. Chaque valeur est exprimée en pourcentage (somme du chaque ligne égale à 1). On observe rapidement que le stade N1 obtient des résultats moins bons que les autres stades.

Plusieurs indices statistiques peuvent être évaluées à partir de la table de contingence :

— le taux d'accord (*Acc*) correspond à la proportion d'époques correctement identifiées parmi l'ensemble des époques ;

$$Acc\ (\%) = \frac{1}{n} \sum_{i=1}^{5} n_{ii} \times 100$$

— le Kappa de Cohen (κ) (Cohen, 1960) est sans doute l'indicateur le plus utilisé dans le domaine. En effet, il sert à mesurer l'accord entre la référence et l'analyse automatique, en prenant en considération la composante aléatoire de cet accord. Le κ est calculé à partir de la proportion d'accord observée, notée P_o, et la proportion d'accord aléatoire (concordance attendue dans l'hypothèse où l'analyse manuelle et automatique sont totalement indépendantes), notée P_e.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

On remarquera que P_o correspond au taux d'accord *Acc* (en valeur numérique au lieu du pourcentage).

$$P_o = \frac{1}{n} \sum_{i=1}^{5} n_{ii}$$

P_e, quant-à-elle, correspond à la somme des produits des effectifs marginaux divisée par le carré du nombre d'époques total.

$$P_e = \frac{1}{n^2} \sum_{i=1}^{5} n_{i.} \times n_{.i}$$

Le κ est interprété en utilisant six fourchettes :

- $\kappa < 0,0$: accord mauvais
- $0,0 \leq \kappa < 0.2$: accord faible
- $0,2 \leq \kappa < 0.4$: accord médiocre
- $0,4 \leq \kappa < 0.6$: accord modéré
- $0,6 \leq \kappa < 0.8$: accord fort
- $0,8 \leq \kappa$: accord excellent

B.3.2 Mesure de l'impact clinique

Il est également possible de comparer la lecture automatique du sommeil avec la lecture manuelle, en évaluant leur impact sur le diagnostic du patient. Pour cela, on calcule l'IAH qui découle des deux lectures. L'IAH se calcule différemment selon les informations à disposition.
Il est possible de distinguer quatre sortes d'événements respiratoires :

— l'apnée, que l'on notera par la suite A. De manière simplifiée, elle est définie comme étant une diminution de la ventilation $\geq 90\,\%$ pendant au moins 10 secondes ;

— l'hypopnée désaturante, que l'on notera par la suite H_{desat}. De manière simplifiée, elle est définie comme étant une diminution de la ventilation $< 90\,\%$ mais $\geq 30\,\%$, pendant au moins 10 secondes. Comme l'indique son nom, elle doit être suivie par une diminution significative ($\geq 3\,\%$) du taux d'oxygène dans le sang ;

— l'hypopnée micro-éveillante, que l'on notera par la suite H_{MEV}. Proche de l'hypopnée désaturante, la seule différence réside dans le fait qu'elle doit être suivie non pas d'une désaturation mais d'un micro-éveil (pour rappel, c'est un très court éveil) ;

— l'hypopnée éveillante, que l'on notera par la suite H_{Ev}. Cette fois-ci, elle est suivie d'une époque d'éveil.

Les diminutions de la ventilation et les désaturations sont observables à partir des signaux ventilatoires. Au contraire, les micro-éveils et les éveils ne sont identifiables qu'à partir des voies électrophysiologiques. Dans le cas d'une PV, les H_{MEV} et H_{Ev} sont donc manquantes. Dans le cas d'une analyse automatique des stades de sommeil ne permettant pas la détection automatique des micro-éveils, seules les H_{MEV} sont manquantes.
Un autre élément différencie le calcul de l'IAH en PV et en PSG. En effet, dans le cas de la PV, l'IAH équivaut au rapport entre le nombre d'événements ventilatoires identifiés et la durée de l'enregistrement (ou Temps d'Enregistrement, noté TE). En PSG, seuls les événements ayant eu lieu pendant le sommeil sont comptabilisés. De plus, le temps de référence est remplacé par la durée de sommeil (ou Temps de Sommeil Total, noté TST).
Le tableau B3 récapitule ces informations.

Table B3 – Estimation de l'IAH selon le type d'enregistrement réalisé.

	PV	PSG	
		sommeil automatique	sommeil manuel
Événements respiratoires	A H_{desat}	A *en sommeil*$_{auto}$ H_{desat} *en sommeil*$_{auto}$ H_{Ev} *en sommeil*$_{auto}$	A *en sommeil* H_{desat} *en sommeil* H_{Ev} *en sommeil* H_{MEV} *en sommeil*
Temps de référence	TE	TST_{auto}	TST
IAH	$\frac{A+H_{desat}}{TE}$	$\frac{A+H_{desat}+H_{Ev}}{TST_{auto}}$	$\frac{A+H_{desat}+H_{Ev}+H_{MEV}}{TST}$

Afin de comparer les IAHs obtenus par lecture manuelle et automatique, il ne s'agit donc pas uniquement de modifier la valeur du dénominateur (le temps de référence). Les événements respiratoires ne peuvent en effet être comptabilisés que s'ils apparaissent pendant une époque de sommeil.

Afin de comparer les IAHs entre eux, deux approches sont généralement utilisées :

— le Bland-Altman (Altman et Bland, 1983) est un graphe permettant de visualiser la différence des paires d'IAH (les IAHs obtenus selon 2 méthodes, par patient), en fonction de la moyenne des paires. Très répandu, il permet de se rendre compte de la tendance des erreurs réalisées. Sur ce graphe, on estime souvent la moyenne des différences et les limites basses et hautes entre lesquelles 95 % des différences se trouveront (si les différences suivent une loi normale) ;

— la courbe de corrélation, qui permet de visualiser, paire par paire, les IAHs obtenus avec une première méthode, comparées avec ceux d'une seconde méthode. On estime généralement le coefficient de corrélation de Pearson, noté r, qui renseigne sur le degré de corrélation entre les IAHs. On détaillera également l'équation du modèle de régression linéaire.

La Figure B8 présente un exemple pour lequel on ne considère que cinq enregistrements, pour lesquels les IAHs de référence seraient [2 ; 6 ; 16 ; 21 ; 37] et ceux obtenus avec l'analyse automatique seraient [3 ; 7 ; 14 ; 24 ; 32].

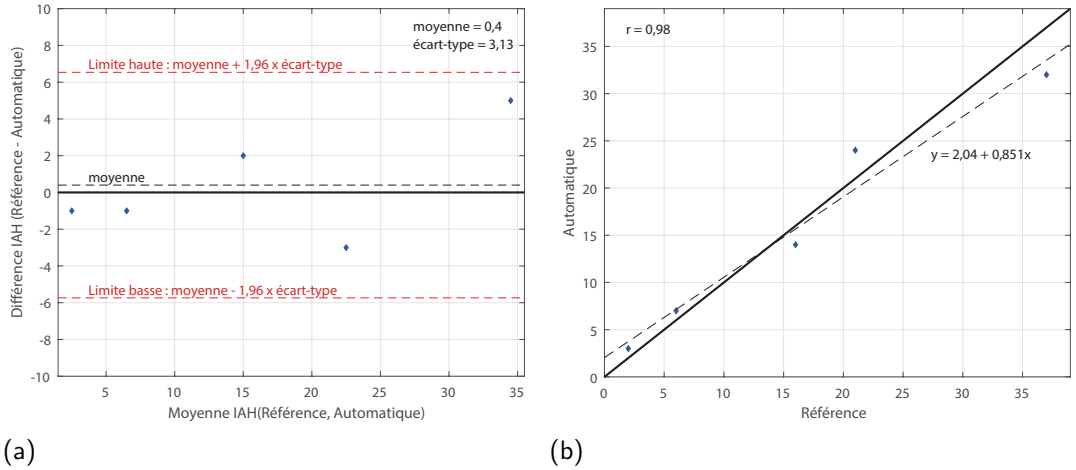


Figure B8 – Exemple de graphes permettant la comparaison d'IAHs : (a) graphe Bland-Altman et (b) courbe de corrélation. Sur les deux graphes, les résultats sont meilleurs lorsque les points sont proches de la courbe en gras (qui représente une différence de zéro entre les IAHs pour le Bland-Altman, et la fonction identité pour la courbe de corrélation).

En plus de l'IAH, certains indicateurs sont utilisés par le spécialiste du sommeil lors de son diagnostic :

— les proportions de chaque stade, notées propN1, propN2, propN3 et propSP. On regroupe souvent les stades N1 et N2 : propN1N2. La proportion d'éveil est généralement remplacée par l'efficacité du sommeil, ou *Sleep Efficiency* (SE) en anglais, qui correspond, à l'inverse, à la proportion de sommeil ;

— les latences associées à la première occurrence des principaux stades (latence du N2, latence du N3 et latence du SP). L'enregistrement commençant bien souvent à l'éveil, on définit aussi la latence d'endormissement, ou *Sleep Onset Latency* (SOL) en anglais, qui équivaut à la durée entre l'extinction des lumières et l'endormissement du patient.

Par exemple, une latence de SP très faible peut orienter, dans certains cas, vers le diagnostic d'une narcolepsie.

La sévérité du SAHS est évaluée à partir de l'IAH :

- $IAH < 5/h$: pas de SAHS

- $5/h \leq IAH < 15/h$: SAHS léger
- $15/h \leq IAH < 30/h$: SAHS modéré
- $30/h \leq IAH$: SAHS sévère

En considérant ces catégories de sévérité, on peut obtenir la matrice de confusion présentée Table B4.

Table B4 – Matrice de confusion associée à la sévérité du SAHS.

		Analyse automatique				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
Référence	Pas de SAHS	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
	SAHS léger	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
	SAHS modéré	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
	SAHS sévère	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

Il est ainsi possible, comme précédemment, d'estimer un taux d'accord (Acc) et un Kappa de Cohen (κ) pour la classification du point de vue de la sévérité du SAHS.
Il faut également savoir que les appareils permettant le traitement du SAHS ne sont pris en charge par la sécurité sociale que dans certaines conditions. D'une manière simplifiée, un patient ayant un SAHS sévère pourra être traité. Un patient avec un SAHS modéré pourra également être traité sous condition de présence d'au moins 10 micro-éveils par heure de sommeil et/ou en présence d'une maladie cardio-vasculaire grave associée. Ainsi, les erreurs de la Table B4 annotées n_{31}, n_{41}, n_{32} et n_{42} sont potentiellement graves puisqu'elles représentent des patients qui ne seront pas traités si l'analyse automatique est utilisée sans relecture. Au contraire, les erreurs annotées n_{13}, n_{14}, n_{23} et n_{24} représentent des patients qui seraient traités à tort dans le cas d'une analyse automatique sans relecture. Ces erreurs peuvent d'ailleurs apparaître lors de l'utilisation d'une PV au lieu d'une PSG. En effet, comme expliqué précédemment (voir Table B3), l'IAH calculé à partir d'une PV est différent de celui estimé à partir de la PSG. En général, l'IAH de PV est sous-estimé par rapport à celui de PSG. Parfois, lorsque l'examen réalisé en première intention est une PV et que l'IAH obtenu est insuffisant malgré les symptômes du patient, le spécialiste du sommeil prescrit un second examen du sommeil, de type PSG cette fois-ci. C'est grâce à ce second examen que certains patients peuvent être traités.

B.4 Conclusion du chapitre

L'analyse automatique du sommeil peut être réalisée à partir de différents signaux d'entrée. Le choix de ces derniers dépendent de l'objectif de l'étude. Dans le cas d'une analyse automatique en PSG, les signaux électrophysiologiques sont couramment employés, puisqu'ils sont utilisés pour la lecture manuelle du sommeil. Dans le cas d'une analyse automatique en PV, seuls les signaux cardio-respiratoires peuvent être employés. Même s'ils sont influencés par la profondeur du sommeil, ils ne permettent pas une analyse des stades de sommeil aussi précise qu'avec les signaux électrophysiologiques. Dans ce cas, il est donc courant de combiner certains stades entre eux et de ne proposer qu'une classification en 2, 3 ou 4 stades (au lieu de 5). Les quelques informations extraites de ces classifications réduites peuvent tout de même s'avérer utiles pour le clinicien lors de son diagnostic.

Cependant, de manière générale, l'analyse automatique est confrontée à plusieurs difficultés :

— l'altération fréquente des signaux (prises médicamenteuses, qualité de l'enregistrement, artéfacts, syndromes particuliers) ;

— la variabilité des signaux entre les patients, qui limite le taux d'accord inter-scorer ;

— la complexité des règles de lecture des stades de sommeil et des transitions.

De nombreuses études se sont confrontées à ces difficultés et ont proposé une méthode d'analyse automatisée des stades de sommeil. Les méthodes utilisant du ML ressortent beaucoup dans la littérature. Elles nécessitent cependant un travail d'évaluation de caractéristiques, et parfois de

réduction de la dimensionnalité. Les méthodes utilisant du DL sont plus récentes. Elle combinent souvent un CNN et un RNN afin de profiter des avantages des deux algorithmes (sur le plan spacial et temporel). Pour finir, des méthodes hybrides basées sur l'association des approches de ML ou DL avec les connaissances des experts, voient le jour. Elles permettent d'apporter une interprétabilité à l'analyse automatique que ne possèdent pas les autres méthodes.
Il est nécessaire que l'apprentissage du modèle choisi soit effectué sur des enregistrements indépendants des enregistrements de test.

L'évaluation des performances peut être réalisée sur différents niveaux, permettant d'évaluer :

— l'identification de chaque stade de sommeil individuellement, à l'aide d'indicateurs comme le taux d'accord (Acc), la sensibilité Se, la spécificité Sp, la Valeur Prédictive Positive (VPP) et la Valeur Prédictive Négative (VPN) ;
— l'estimation de l'hypnogramme (donc l'ensemble des stades), à l'aide d'indicateurs comme le taux d'accord (Acc) et le Kappa de Cohen (κ), ainsi qu'à l'aide de représentations graphiques comme une heatmap ;
— l'estimation de l'IAH résultant, à l'aide de graphes comme le Bland-Altman et la courbe de corrélation ;
— l'estimation d'indicateurs supplémentaires, à l'aide des mêmes types de graphes ;
— l'estimation de la sévérité du SAHS associé, à l'aide du taux d'accord (Acc) et du Kappa de Cohen (κ) également.

C'est en jugeant les performances d'une analyse automatique sur l'ensemble de ces niveaux que l'on pourra s'assurer de son efficacité.

Annexe : les artéfacts

On différencie deux types d'artéfacts. Les artéfacts extrinsèques apparaissent à cause d'un élément extérieur au patient. Au contraire, les artéfacts intrinsèques sont directement liés au patient.

Les artéfacts extrinsèques

On peut rencontrer trois sortes d'artéfacts extrinsèques, dont l'origine est liée :

— aux électrodes : à cause d'une perte de contact entre l'électrode et la peau, une activité lente (4-5 Hz) et continue peut perturber les signaux temporairement. Dans la littérature, il est courant de dévalider (entièrement ou en partie) les voies touchées par cet artéfact.
Il est également possible d'observer un artéfact appelé « *electrode pop* », qui est dû à la décharge spontanée du potentiel électrique. On observe alors une variation soudaine, brusque et brève.
L'utilisation d'une quantité de gel conducteur trop importante peut également faire glisser l'électrode sur la peau.
Pour finir, un encrassage des électrodes peut être à l'origine d'un pont salin générant un artéfact de faible fréquence ;

— au secteur : le bruit électrique de 50/60 Hz peut être visible sur certaines voies ;

— à l'environnement : la ventilation de la pièce, la présence d'appareils comme des intraveineuses ou même un téléphone dans la pièce peuvent influer sur les signaux.

Les artéfacts intrinsèques

On peut rencontrer cinq sortes d'artéfacts intrinsèques, dont l'origine est liée :

— aux mouvements oculaires : les mouvements oculaires sont souvent visibles sur les voies EEGs proches des cavités oculaires. On peut ainsi observer des traces de clignements, des mouvements de lecture, d'observation de l'environnement ou même des mouvements oculaires ayant lieu avec les yeux fermés.
Les clignements des yeux ont une signature bien précise très souvent visible sur les voies EEGs frontales, avec une forte amplitude et une courte durée.
Les flutter oculaires (mouvements conjugués involontaires en salves) génèrent également des répercussions sur les voies EEGs frontales, mais de plus faible amplitude et en salves.
Les mouvements de lecture ou d'observation de l'environnement sont facilement visibles également ;

— à l'activité musculaire : trois artéfacts sont ici distingués.
Le premier est l'artéfact électromyographique de surface, très présent puisqu'il est dû à l'impact d'un mouvement musculaire sur les électrodes et/ou leur fil. Il est à l'origine d'un bruit de très haute fréquence et de durée variable, apparaissant le plus souvent sur les régions temporales et frontales.
Le deuxième, l'artéfact photomyogénique, est la réponse à une stimulation photique (flash de lumière forte). Il est répétitif, de courte durée et principalement visible sur les régions frontales et périorbitales.
Pour finir, l'artéfact glossokinétique est la conséquence des mouvements de langue engendrés lors de mastication ou de déglutition. En effet, la langue agit comme un dipôle et peut donc générer des perturbations de haute amplitude sur les EEGs ;

— à l'activité cardiaque : on peut distinguer deux artéfacts cardiaques.
Les artéfacts cardiaques électriques se traduisent par l'image de l'ECG (ou du moins l'ECG appauvri, soit uniquement les pics Q, R et S ou même simplement R) sur une autre voie. Il est plus facilement visible sur les voies pour lesquelles les électrodes sont situées loin l'une de l'autre.
Les artéfacts cardiaques mécaniques sont l'image de la pulsion sanguine captée par les électrodes (cet artéfact est d'ailleurs parfois considéré comme un artéfact d'électrodes). Il est plus facilement visible sur les voies à proximité des vaisseaux sanguins, comme par exemple les tempes, les mastoïdes ou les jambiers ;

— à l'activité ventilatoire : les mouvements ventilatoires peuvent influer sur les signaux électrophysiologiques. Dans ce cas, on distingue sur les voies artéfactées un rythme tel que celui de la ventilation (soit 10-20 cycles ventilatoires par minute pour un adulte).
Les ronflements peuvent également générer une activité rapide sur les EEGs et l'EMG mentonnier ;

— à l'électrodermogramme : la transpiration est à l'origine de cet artéfact, qui est visible sous la forme d'une onde très lente (0,5-1 Hz) et durant plusieurs secondes. Cet artéfact apparaît plus souvent en N3 que durant les autres stades.

Le traitement des artéfacts

Si certains artéfacts n'ont pas d'influence sur la lecture des époques, de par leur brièveté (*electrode pop*) ou bien tout simplement car ils sont des indices permettant de faciliter la lecture des stades de sommeil (mouvements oculaires, activité musculaire), d'autres dégradent le signal (secteur, environnement, activité cardiaque, activité ventilatoire, électrodes). Certaines études retirent les artéfacts ou dévalident les sections artéfactées. Le bruit 50 Hz est systématiquement retiré électroniquement par filtrage passe-bas. Les artéfacts de mouvement et les artéfacts ECG sont parfois détectés, mais les autres artéfacts ne sont que très rarement supprimés.

Table B5 – Revue de la littérature : traitement des artéfacts.

Auteurs	Voies	Artéfact(s)	Méthode(s)	Élément(s)	Utilisation
Hapuarachchi (2006)	EEG EOG	Transpiration	Modification des bornes des ondes delta	Informations fréquentielles	Suppression
		Artéfacts biologiques ou intermittents	WDA (débruitage à l'aide des ondelettes) ou ICA (Analyse en composantes indépendantes)	Corrélation et amplitude	Suppression
Charbonnier *et al.* (2011)	EEG EOG EMG	Hyper ou micro voltage	Seuillage	Amplitudes crête et crête-à-crête	Dévalidation totale ou partielle des époques
		Trop hautes et basses fréquences	Seuillage	Amplitude après passe-bande et SEF95 (95ème centile Spectral Edge Frequency)	Dévalidation totale ou partielle des époques
		ECG	Seuillage	Dérivées, amplitude et interquartile	Dévalidation totale ou partielle des époques
		Mouvements	Seuillage	Amplitude, variance et médiane de la variance	Dévalidation totale ou partielle des époques
Popovic *et al.* (2014)	EEG	Oculaire	Filtre médian	Informations fréquentielles	Suppression de l'artéfact
		Mouvements, cardio-respiratoires, EMG	Seuillage	Amplitude pic-à-pic, pente, moyennage	Dévalidation des époques
McCurry (2017)	EEG	Mouvement	Seuillage	Amplitude	Dévalidation des époques

Bibliographie

Agnew, H. W., Webb, W. B. et Williams, R. L. (1966). The First Night Effect : An Eeg Studyof Sleep. *Psychophysiology*, 2(3):263–266.

Aktaruzzaman, M., Rivolta, M. W., Karmacharya, R., Scarabottolo, N., Pugnetti, L., Garegnani, M., Bovi, G., Scalera, G., Ferrarin, M. et Sassi, R. (2017). Performance comparison between wrist and chest actigraphy in combination with heart rate variability for sleep classification. *Computers in Biology and Medicine*, 89:212–221.

Al-Hussaini, I., Xiao, C., Westover, M. B. et Sun, J. (2019). SLEEPER : interpretable Sleep staging via Prototypes from Expert Rules. *Machine Learning for Healthcare*, 106:1–18.

Altman, D. G. et Bland, J. M. (1983). Measurement in Medicine : The Analysis of Method Comparison Studies. *The Statistician*, 32(3):307.

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46(3):175–185.

Baum, L. E. et Pietrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, pages 1554–1563.

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., lloyd, R. M., Quan, S. F., Troester, M. M. et Vaughn, B. V. (2017). *The AASM Manual for the Scoring of Sleep and Associated Events : Rules, Terminology and Technical Specifications*. Numéro 2.4 de American Academy of Sleep Medicine. Darien IL.

Berthomier, C., Drouot, X., Herman-Stoïca, M., Berthomier, P., Prado, J., Bokar-Thire, D., Benoit, O., Mattout, J. et D'Ortho, M.-P. (2007). Automatic Analysis of Single-Channel Sleep EEG : Validation in Healthy Individuals. *Sleep*, 30(11):1587–1595.

Biswal, S., Sun, H., Goparaju, B., Westover, M. B., Sun, J. et Bianchi, M. T. (2018). Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association*, 25(12):1643–1650.

Bresch, E., Grossekathöfer, U. et Garcia-Molina, G. (2018). Recurrent Deep Neural Networks for Real-Time Sleep Stage Classification From Single Channel EEG. *Frontiers in Computational Neuroscience*, 12:85.

Charbonnier, S., Zoubek, L., Lesecq, S. et Chapotot, F. (2011). Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging. *Computers in Biology and Medicine*, 41(6):380–389.

Chen, C. (2016). *An e-health system for personalized automatic sleep stages classification*. Thèse de doctorat, Université Pierre et Marie Curie - Paris VI.

Chen, C., Ugon, A., Sun, C., Chen, W., Philippe, C. et Pinna, A. (2019). Towards a Hybrid Expert System Based on Sleep Event's Threshold Dependencies for Automated Personalized Sleep Staging by Combining Symbolic Fusion and Differential Evolution Algorithm. *IEEE Access*, 7:1775–1792.

Clausen, J. (1999). Branch and Bound Algorithms - Principles and Examples. page 30.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, J., Cohen, P., West, S. et Aiken, L. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Taylor & Francis.

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dalvi, N., Domingos, P., Mausam, Sanghai, S. et Verma, D. (2004). Adversarial classification. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 99–108, New York, NY, USA. Association for Computing Machinery.

Dong, H., Supratak, A., Pan, W., Wu, C., Matthews, P. M. et Guo, Y. (2018). Mixed Neural Network Approach for Temporal Sleep Stage Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):324–333. arXiv : 1610.06421.

Doroshenkov, L. G., Konyshev, V. A. et Selishchev, S. V. (2007). Classification of human sleep stages based on EEG processing using hidden Markov models. *Biomedical Engineering*, 41(1):25–28.

Doshi-Velez, F. et Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv :1702.08608 [cs, stat]*. arXiv : 1702.08608.

Ebrahimi, F., Setarehdan, S.-K., Ayala-Moyeda, J. et Nazeran, H. (2013). Automatic sleep staging using empirical mode decomposition, discrete wavelet transform, time-domain, and nonlinear dynamics features of heart rate variability signals. *Computer Methods and Programs in Biomedicine*, 112(1):47–57.

Enshaeifar, S., Kouchaki, S., Took, C. C. et Sanei, S. (2016). Quaternion Singular Spectrum Analysis of Electroencephalogram With Application in Sleep Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(1):57–67.

Fehrmann, E. A. (2013). *Automated sleep classification using the new sleep stage standards*. Rochester Institute of Technology.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L. et Faraci, F. D. (2019). Automated sleep scoring : A review of the latest approaches. *Sleep Medicine Reviews*, 48.

Fonseca, P., Long, X., Radha, M., Haakma, R., Aarts, R. M. et Rolink, J. (2015). Sleep stage classification with ECG and respiratory effort. *Physiological Measurement*, 36(10):2027–2040.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H. et Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, 108(1):10–19.

Garcia-Molina, G., Abtahi, F. et Lagares-Lemos, M. (2012). Automated NREM sleep staging using the Electro-oculogram : A pilot study. *In Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2255–2258. IEEE.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., USA, 1st édition.

Hapuarachchi, P. (2006). *Feature selection and artifact removal in sleep stage classification*. Thèse de doctorat, University of Waterloo.

Ho, T. K. (1995). Random Decision Forests.

Hochreiter, S. et Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9:1735–1780.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables Into Principal Components. page 25.

Hsu, Y.-L., Yang, Y.-T., Wang, J.-S. et Hsu, C.-Y. (2013). Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing*, 104:105–114.

Koley, B. et Dey, D. (2012). An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine*, 42(12):1186–1195.

Lab, M. (s.d.). A Brief Introduction to Adversarial Examples.

Lajnef, T., Chaibi, S., Ruby, P., Aguera, P.-E., Eichenlaub, J.-B., Samet, M., Kachouri, A. et Jerbi, K. (2015). Learning machines and sleeping brains : Automatic sleep stage classification using decision-tree multi-class support vector machines. *Journal of Neuroscience Methods*, 250:94–105.

Larousse, d. (s.d.). Encyclopédie Larousse en ligne - intelligence artificielle.

Lewicke, A., Sazonov, E., Corwin, M., Neuman, M. et Schuckers, S. (2008). Sleep Versus Wake Classification From Heart Rate Variability Using Computational Intelligence : Consideration of Rejection in Classification Models. *IEEE Transactions on Biomedical Engineering*, 55(1):108–118.

Long, X., Fonseca, P., Foussier, J., Haakma, R. et Aarts, R. M. (2014). Sleep and Wake Classification With Actigraphy and Respiratory Effort Using Dynamic Warping. *IEEE Journal of Biomedical and Health Informatics*, 18(4):1272–1284.

Long, X., Yang, J., Weysen, T., Haakma, R., Foussier, J., Fonseca, P. et Aarts, R. M. (2015). Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging. *Physiological Measurement*, 36(3):625–625.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F. et Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 4(2):R1–R13.

Lucey, B. P., Mcleland, J. S., Toedebusch, C. D., Boyd, J., Morris, J. C., Landsness, E. C., Yamada, K. et Holtzman, D. M. (2016). Comparison of a single-channel EEG sleep study to polysomnography. *Journal of Sleep Research*, 25(6):625–635.

Mahvash Mohammadi, S., Kouchaki, S., Ghavami, M. et Sanei, S. (2016). Improving time–frequency domain sleep EEG classification via singular spectrum analysis. *Journal of Neuroscience Methods*, 273:96–106.

Malaekah, A. (2016). *Automated sleep stage detection and classification of sleep disorders*. Thèse de doctorat, Electrical and Computer Engineering College of Science, Engineering and Health RMIT University.

Marella, S. (2012). EEG Artifacts.

McCurry, M. (2017). *Automatic Spectral-Temporal Modality Based EEG Sleep Staging*. Thèse de doctorat, Georgia Institute of Technology.

Mendez, M., Matteucci, M., Cerutti, S., Aletti, F. et Bianchi, A. (2009). Sleep staging classification based on HRV : Time-variant analysis. *In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 9–12, Minneapolis, MN. IEEE.

Mendez, M. O., Matteucci, M., Castronovo, V., Strambi, L. F., Cerutti, S. et Bianchi, A. M. (2010). Sleep staging from Heart Rate Variability : time-varying spectral features and Hidden Markov Models. *International Journal of Biomedical Engineering and Technology*, 3(3/4):246.

Nguyen, H.-A. T., Musson, J., Li, F., Wang, W., Zhang, G., Xu, R., Richey, C., Schnell, T., McKenzie, F. D. et Li, J. (2012). EOG artifact removal using a wavelet neural network. *Neurocomputing*, 97:374–389.

Patanaik, A., Ong, J. L., Gooley, J. J., Ancoli-Israel, S. et Chee, M. W. L. (2018). An end-to-end framework for real-time automatic sleep stage classification. *Sleep*, 41(5).

Penzel, T. et Conradt, R. (2000). Computer based sleep recording and analysis. *Sleep Medicine Reviews*, 4(2):131–148.

Popovic, D., Khoo, M. et Westbrook, P. (2014). Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead : validation in healthy adults. *Journal of Sleep Research*, 23(2):211–221.

Pudil, P., Novovičová, J. et Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125.

Redmond, S. J., de Chazal, P., O'Brien, C., Ryan, S., McNicholas, W. T. et Heneghan, C. (2007). Sleep staging using cardiorespiratory signals. *Somnologie - Schlafforschung und Schlafmedizin*, 11(4):245–256.

Rosenberg, R. S. et Van Hout, S. (2013). The American Academy of Sleep Medicine Inter-scorer Reliability Program : Sleep Stage Scoring. *Journal of Clinical Sleep Medicine*, (9).

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H. et Faubert, J. (2019). Deep learning-based electroencephalography analysis : a systematic review. *Journal of Neural Engineering*, 16(5).

Rumelhart, D. E., Hintont, G. E. et Williams, R. J. (1986). Learning representations by back-propagating errors. page 4.

Sors, A., Bonnet, S., Mirek, S., Vercueil, L. et Payen, J.-F. (2018). A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42:107–114.

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y. L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S. C., Kim, T. W., Pizza, F., Plazzi, G., Vandi, S., Antelmi, E., Perrin, D., Kuna, S. T., Schweitzer, P. K., Kushida, C., Peppard, P. E., Sorensen, H. B. D., Jennum, P. et Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, 9(1):5229.

Ugon, A. (2015). *Fusion Symbolique et Données Polysomnographiques*. Thèse de doctorat.

Xiao, M., Yan, H., Song, J., Yang, Y. et Yang, X. (2013). Sleep stages classification based on heart rate variability and random forest. *Biomedical Signal Processing and Control*, 8(6):624–633.

Yilmaz, B., Asyali, M. H., Arikan, E., Yetkin, S. et Özgen, F. (2010). Sleep stage and obstructive apneaic epoch classification using single-lead ECG. *Biomedical engineering online*, 9(1):39.

Zhang, L., Fabbri, D., Upender, R. et Kent, D. (2019). Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks. *Sleep*, 42(11):zsz159.

Zokaeinikoo, M. (2016). *Automatic Sleep Stages Classification*. Thèse de doctorat.

Chapitre C

Projet HypnoLighT

C.1 Objectif et contexte

Ce projet a débuté avant le début de la thèse, par le biais d'un contrat de professionalisation (septembre 2016 - septembre 2017). Il a permis la compréhension des enjeux et des difficultés dans le domaine général de l'analyse automatique du sommeil. Quelques développements ont été réalisés sur ce projet lors de la première année de thèse. L'outil créé au sein de ce projet est maintenant en phase d'industrialisation. Un travail conjoint avec d'autres pôles de CIDELEC et particulièrement le Bureau d'Étude est donc réalisé depuis ces 3 dernières années. La sortie d'HypnoLighT est prévue en 2021.

Dans ce projet, l'objectif est d'identifier uniquement l'éveil (classification éveil/sommeil). Cette détection est réalisée à partir des signaux d'un nouveau type d'enregistrement, constitué d'une PV améliorée par l'ajout d'une unique voie EEG. Comparé à la PSG, cet enregistrement reste toujours plus rapide à mettre en place et à lire, moins coûteux et plus facilement réalisable en ambulatoire. Cependant, l'identification de l'éveil permet d'obtenir un IAH plus précis qu'en PV, et donc de faciliter le diagnostic.

C.2 Problématiques identifiées

L'identification de l'éveil est normalement réalisée en PSG à l'aide de différents signaux :

— l'EEG occipital (à l'arrière du crâne) pour visualiser les ondes alpha ;

— les voies EOG pour visualiser les clignements de yeux, les mouvements de lecture ou les Mouvement Oculaire Rapide (MOR) ;

— l'EMG du menton pour visualiser le tonus musculaire [1].

Dans le cas de ce projet, seule une voie EEG est mise à disposition.

La première problématique a donc été le choix de cette voie EEG.
Il existe différentes dérivations EEG possibles. Celles-ci sont définies selon la position des deux électrodes entre lesquelles la différence de potentiel électrique est évaluée. Le positionnement des électrodes est référencé selon le système international 10-20. Les positions sont ainsi identifiées à l'aide d'une lettre, indiquant la région du cerveau observée (Pg : orifice pharyngé, Fp : préfrontale, F : frontale, C : centrale, T : temporale, P : pariétale, O : occipitale et A ou M : mastoïde) et d'un chiffre, indiquant l'hémisphère exploré (pair : droit, impair : gauche, z : centre).

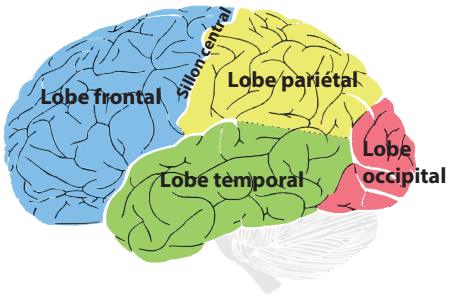


Figure C1 – Architecture du cerveau.

1. Le muscle du menton est assez représentatif du niveau de vigilance. C'est en effet l'un des derniers muscles à se relâcher lors de l'endormissement.

Les recommandations AASM conseillent l'enregistrement de trois EEGs permettant l'enregistrement de l'activité frontale (F4-M1 ou F3-M2), centrale (C4-M1 ou C3-M2) et occipitale O2-M1 ou O1-M2). Parfois, l'activité pré-frontale (Fp2-M1 ou Fp1-M2) est enregistrée à la place de l'activité frontale.
Dans la littérature, de nombreuses études traitant de l'identification de l'éveil à partir d'une unique voie EEG choisissent une voie centrale (Berthomier *et al.*, 2007; Fraiwan *et al.*, 2012; Koley et Dey, 2012; Sors *et al.*, 2018).
Cependant, d'autres études choisissent une voie EEG pré-frontale sur laquelle les mouvements oculaires importants peuvent être visibles (de par la proximité du capteur avec les yeux) ou directement les EOGs (Bresch *et al.*, 2018; Dong *et al.*, 2018; Popovic *et al.*, 2014; Virkkala *et al.*, 2008). Ces voies ont l'avantage d'être également plus faciles à poser puisqu'elles ne sont pas situées sur le cuir chevelu. Dans leur étude, Hsu *et al.* (2013) utilisent la voie Fpz-Cz.

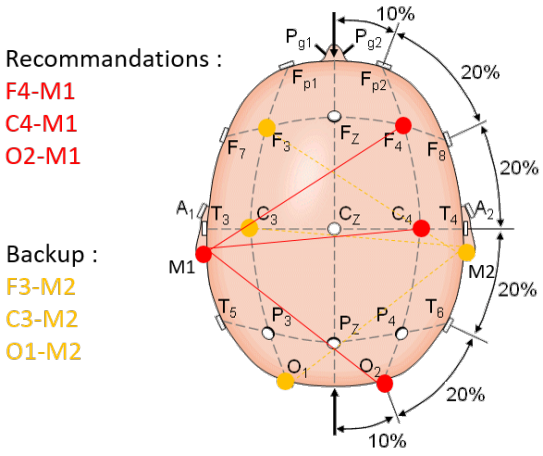


Figure C2 – Placement des électrodes recommandé par l'AASM.

Une autre problématique rencontrée a été la variabilité de la fréquence des ondes alpha, utilisées pour l'identification de l'éveil.
Premièrement, si la définition stricte des ondes alpha est une onde EEG dont la fréquence est située entre 8 Hz et 13 Hz, en réalité les ondes alpha à l'éveil d'un patient peuvent aussi bien être situées entre 7 Hz et 9 Hz qu'entre 12 Hz et 14 Hz.
Deuxièmement, il n'est pas rare d'observer en SP des ondes ressemblant aux ondes alpha à l'éveil, dont la fréquence n'est qu'1 ou 2 Hz inférieure. Si ces ondes ont une fréquence comprise entre 8 Hz et 13 Hz, elles sont techniquement considérées comme étant des ondes alpha, et une grande proportion d'entre elles indiquerait alors de l'éveil. En réalité, ce n'est pas le cas et le spécialiste du sommeil différencie bien ces ondes des ondes alpha spécifiques à l'éveil.
Il est donc nécessaire que l'algorithme identifie pour chaque enregistrement la fréquence des ondes alpha spécifiques à l'éveil pour adapter la classification qui en résulte.

C.3 Stratégie

Afin de visualiser les mouvements oculaires en plus des ondes alpha, nous avons choisi d'utiliser la voie Fp2-A1 [2]. Le tonus musculaire est estimé à partir de l'actimétrie, puisque l'EMG du menton n'est pas disponible.
Différents types d'éveil ont été définis. Ils permettent de prendre en compte les différents éléments utiles pour l'identification de l'éveil :

Éveil Agité (EA) : Il est souvent visible en début ou fin d'enregistrement, lorsque le patient est totalement réveillé. On peut y observer des mouvements sur l'actimètre ou les sangles inductives (abdominales et thoraciques). La ventilation et le rythme cardiaque ne sont pas réguliers, et les ondes des EEGs sont mixtes et irrégulières. Il est possible d'y observer des mouvements oculaires comme des clignements de yeux et des mouvements de lecture sur les EOGs. Sur certains patients, les mouvements oculaires sont même assez puissants pour être visibles sur les EEGs. Le tonus musculaire est élevé et la lumière est parfois allumée.

Éveil Yeux Ouverts (EYO) : Le patient est au repos mais il n'est pas en phase d'endormissement. Ce type d'éveil est moins anarchique que l'éveil agité. On peut observer des clignements des yeux et des mouvements oculaires type lecture sur les EOGs, et parfois même les EEGs. Le tonus musculaire reste haut mais le rythme cardiaque et la ventilation se stabilisent. L'information des EEGs reste assez mixte et irrégulière.

Éveil Yeux Fermés (EYF) : Le patient est dans un état de somnolence. Le patient a ici les yeux fermés et les EEGs montrent la plupart du temps des ondes alpha bien présentes. Le tonus

2. Les recommandations françaises préfèrent A1 et A2 aux électrodes M1 et M2 qui sont parfois très sensibles à l'activité cardiaque (proximité avec la veine jugulaire).

musculaire reste élevé mais peut diminuer légèrement. Le rythme cardiaque et la ventilation se stabilisent également.

Ces types d'éveil ont été identifiés automatiquement de manière individuelle, mais l'évaluation des résultats a été faite en les combinant de manière à pouvoir les comparer avec la lecture manuelle.

La méthode a été entraînée et testée sur 61 et 99 enregistrements polysomnographiques, respectivement. Ces enregistrements proviennent de patients sains ou avec troubles du sommeil et pour lesquels le sommeil a été enregistré et lu manuellement par des spécialistes du sommeil du Laboratoire du Sommeil du CHU d'Angers, en suivant les recommandations AASM.

L'impact de la méthode sur le diagnostic des patients a également été estimé grâce à l'évaluation du Temps de Sommeil Total (TST) et de l'IAH. Des graphiques de Bland-Altman ont été utilisés pour montrer l'apport de la méthode sur l'IAH, et les sévérités de SAHS qui en découlent ont été estimées et comparées avec celles obtenues en PSG.

Dans la suite de ce chapitre, l'approche choisie est décrite en détail, et l'impact possible de ce travail sur le diagnostic du SAHS est évalué. Pour aller plus loin, un travail de détection automatique des hypopnées micro-éveillantes est ensuite présenté. Son apport sur le diagnostic du SAHS dans le cas d'une PV améliorée par l'ajout d'une unique voie EEG et la détection automatique de l'éveil est estimé.

C.3.1 Approche et impact clinique

Ce travail a été présenté en détail par le biais d'un article clinique, publié dans le Journal of Sleep Research (Sabil *et al.*, 2019).

Résumé traduit

Le Temps de Sommeil Total (TST) est estimé lors d'un examen de polysomnographie (PSG). Il est utilisé pour le calcul de l'Index d'Apnées Hypopnées (IAH). Dans le cas d'un examen de Polygraphie Ventilatoire (PV), le manque des signaux électrophysiologiques empêche l'estimation du TST. La conséquence est que l'IAH obtenu est souvent inférieur à celui résultant d'une PSG. L'objectif de cette étude est d'évaluer la précision et la fiabilité d'un nouvel outil permettant la détection automatique éveil/sommeil, et fonctionnant à partir de la combinaison d'une unique voie électroencéphalographique (EEG), de l'actimétrie et des signaux disponibles en PV.
L'étude a été réalisée sur 160 enregistrements PSG de patients investigués pour suspicion de Syndrome d'Apnées Hypopnées du Sommeil (SAHS). Chaque PSG a été enregistrée et lue manuellement selon les règles de l'*American Academy of Sleep Medicine*. L'algorithme d'identification automatique éveil/sommeil repose sur une seule voie EEG (FP2-A1), et sur l'analyse de la variabilité des signaux de PV (flux nasal, ronflements, actimétrie, luminosité et sangles inductives). Des seuils de détection optimaux ont été estimés pour chaque signal à l'aide d'un ensemble de données d'entraînement. Les classifications automatique et manuelle ont ensuite été comparées époque par époque.
Le Kappa de Cohen est de $0,74 \pm 0,18$ et indique un accord substantiel entre les deux classifications. La sensibilité, la spécificité et les valeurs prédictives positives et négatives pour la détection de l'éveil sont respectivement de $76,51\,\% \pm 21,67\,\%$, $95,48\,\% \pm 5,27\,\%$, $81,84\,\% \pm 15,42\,\%$ et $93,85\,\% \pm 6,23\,\%$. Par rapport à l'IAH obtenu en PV, l'IAH estimé a augmenté de 22,12 %. Grâce à cela, on estime que la sévérité du SAHS de 27 patients aurait été sous-estimée en PV mais correctement estimée avec le système ici-présenté.
L'identification automatique éveil/sommeil à l'aide d'une unique voie EEG combinée aux signaux de PV est une méthode fiable pour l'estimation du TST et permet l'obtention d'un IAH plus précis qu'en PV.

Aide lexicale

- *HSAT* est le terme utilisé en anglais pour PV [a]
- *Obstructive Sleep Apnea (OSA) syndrome* est le terme utilisé en anglais pour SAHS [b]
- *Total Recording Time (TRT)* : Temps d'Enregistrement (TE)
- *REM sleep* ou *R stage* : Sommeil Paradoxal (SP)
- *NREM* : Sommeil Lent ou regroupement du N1, N2 et N3 (SL)
- *Respiratory Inductance Plethysmography (RIP) belts* : sangles inductives
- *actigraphy* ou *actimetry* : actimétrie
- *arousal* : micro-éveil

a. Dans les publications internationales, le terme « *Home Sleep Apnea Testing* (HSAT) » est communément utilisé pour mentionner la PV car cette dernière est fréquemment réalisée en ambulatoire.

b. Les évènements ventilatoires peuvent être d'origine obstructive (dus à l'obstruction des voies aériennes supérieures), centrale (d'origine neurologique) ou mixte (dont l'origine est d'abord centrale puis devient obstructive). En France, on parle du Syndrome d'Apnées Hypopnées du Sommeil en général, mais il est également possible de distinguer le Syndrome d'Apnées-Hypopnées Obstructives du Sommeil (SAHOS), le Syndrome d'Apnées Obstructives du Sommeil (SAOS) ou le Syndrome d'Apnées Centrales du Sommeil (SACS). Dans les publications internationales, et malgré l'existence de termes équivalents, c'est généralement le terme « *Obstructive Sleep Apnea (OSA) syndrome* », qui est utilisé (même si l'origine obstructive n'a pas d'importance pour l'étude en question).

Received: 13 August 2018 | Revised: 17 October 2018 | Accepted: 22 October 2018

DOI: 10.1111/jsr.12795

REGULAR RESEARCH PAPER

Journal of Sleep Research

Automatic identification of sleep and wakefulness using single-channel EEG and respiratory polygraphy signals for the diagnosis of obstructive sleep apnea

AbdelKebir Sabil[1,*] | Jade Vanbuis[2,*] | Guillaume Baffet[1] | Mathieu Feuilloy[2,3] | Marc Le Vaillant[4] | Nicole Meslier[5,6] | Frédéric Gagnadoux[5,6]

[1]Recherche et Développement, CIDELEC, Angers, France

[2]Ecole Supérieure d'Electronique de l'Ouest, Angers, France

[3]Laboratoire d'Acoustique, Université du Maine, Le Mans, France

[4]Institut de Recherche en, Santé Respiratoire des Pays de la Loire, Beaucouzé, France

[5]Département de Pneumologie, Centre Hospitalier Universitaire, Angers, France

[6]INSERM, UMR 1063, Université d'Angers, Angers, France

Correspondence
AbdelKebir Sabil, Recherche et Développement, CIDELEC, Angers, France.
Email: ksabil99@hotmail.com

Funding Information
The study was financially supported by an unrestricted grant by CIDELEC, France.

Abstract

Polysomnography (PSG) is necessary for the accurate estimation of total sleep time (TST) and the calculation of the apnea–hypopnea index (AHI). In type III home sleep apnea testing (HSAT), TST is overestimated because of the lack of electrophysiological sleep recordings. The aim of this study was to evaluate the accuracy and reliability of a novel automated sleep/wake scoring algorithm combining a single electroencephalogram (EEG) channel with actimetry and HSAT signals. The study included 160 patients investigated by PSG for suspected obstructive sleep apnea (OSA). Each PSG was recorded and scored manually using American Academy of Sleep Medicine (AASM) rules. The automatic sleep/wake-scoring algorithm was based on a single-channel EEG (FP2-A1) and the variability analysis of HSAT signals (airflow, snoring, actimetry, light and respiratory inductive plethysmography). Optimal detection thresholds were derived for each signal using a training set. Automatic and manual scorings were then compared epoch by epoch considering two states (sleep and wake). Cohen's kappa coefficient between the manual scoring and the proposed automatic algorithm was substantial, 0.74 ± 0.18, in separating wakefulness and sleep. The sensitivity, specificity and the positive and negative predictive values for the detection of wakefulness were 76.51% ± 21.67%, 95.48% ± 5.27%, 81.84% ± 15.42% and 93.85% ± 6.23% respectively. Compared with HSAT signals alone, AHI increased by 22.12% and 27 patients changed categories of OSA severity with the automatic sleep/wake-scoring algorithm. Automatic sleep/wake detection using a single-channel EEG combined with HSAT signals was a reliable method for TST estimation and improved AHI calculation compared with HSAT.

KEYWORDS
automatic scoring algorithm, home polygraphy, OSA subjects, single-channel EEG

1 | INTRODUCTION

Obstructive sleep apnea (OSA) is a highly prevalent disease characterized by recurrent episodes of partial or complete obstruction of the upper airway during sleep. Most recent estimates of OSA prevalence suggest that 6% of women and 13% of men have clinically significant OSA (Heinzer et al., 2015; Peppard et al., 2013). The diagnosis of OSA requires accurate measurement of breathing during sleep. The reference standard for OSA management is in-laboratory polysomnography (PSG), but this method is expensive and time

*AbdelKebir Sabil and Jade Vanbuis contributed equally to this publication.

consuming. Home sleep apnea testing (HSAT) is an accepted (Kapur et al., 2017) and less costly (Corral et al., 2017) alternative in adult patients presenting with signs and symptoms indicating an increased risk of moderate to severe disease and no significant comorbidities. However, measurement error is expected in HSAT, compared with PSG, as standard sleep staging channels are not available, resulting in the use of total recording time (TRT) rather than total sleep time (TST) to define the denominator of the apnea–hypopnea index (AHI). Therefore, the AHI is the ratio of the total number of detected events over TRT for HSAT and over TST for PSG. This overestimation of sleep duration with calculation of the AHI is called the "dilution effect". In a recent observational study including 11,049 patients from the multicentric European Sleep Apnea Cohort (ESADA), patients evaluated by HSAT had a 30% lower AHI on average, compared with patients evaluated by PSG. The proportion of patients with an AHI ≥ 15/hr, indicating moderate-to-severe OSA, was 64% in the subjects who underwent PSG and 47% in the subjects who underwent HSAT (Malhotra et al., 2013). In a recent multicentric, randomised controlled trial including 430 patients investigated for suspicion of OSA, the median (interquartile range) AHI was 20.9/hr (33.4/hr) in the HSAT group and 28.5/hr (43.3/hr) in the PSG group (Corral et al., 2017).

In many healthcare systems, the AHI is a major criterion for the reimbursement of OSA treatments (Escourrou et al., 2015). For instance, the French health insurance system requires an AHI ≥ 15/hr combined with OSA-related symptoms for the reimbursement of continuous positive airway pressure (CPAP) or mandibular advancement device (MAD) therapies for OSA. Spanish sleep physicians recommend CPAP for patients with an AHI ≥ 5/hr, when associated with symptoms or previous cardiovascular diseases, and for patients with an AHI ≥ 30/hr with less severe symptoms (Corral et al., 2017). By underestimating AHI, HSAT could compromise the prescription of CPAP or MAD and/or lead to the prescription of a PSG in patients with borderline AHI. In a recent randomised controlled trial (Corral et al., 2017), performed in 12 tertiary hospitals in Spain, CPAP was prescribed in 15% fewer patients using HSAT compared with PSG.

In the late 1990s, Ehlert et al. introduced one of the first single-channel electroencephalogram (EEG) (FP1-FP2) algorithms, the QUISI system, developed for ambulatory EEG recording device with three electrodes placed on the forehead. This system offered an automatic sleep stage classification based on neural networks (Ehlert et al., 1998). Recently, automated scoring software has been largely developed for use with a single channel of EEG data. These studies have evaluated algorithms that perform identification of simple sleep/wake (Kaplan, Wang, Loparo, Kelly, & Bootzin, 2014) or all sleep stages (Berthomier et al., 2007; Koley & Dey, 2012; Malhotra et al., 2013). Different single-lead EEG channels were used in these studies (A1–A2, CZ–PZ or C4–A1 EEG) (Berthomier et al., 2007; Kaplan et al., 2014; Koley & Dey, 2012). The results suggest that automated scoring algorithms could potentially replace the visual scoring of sleep/wake with as little as one channel of EEG data in both healthy individuals and those with sleep disorders.

In the absence of EEG data during HSAT recording, indirect interpretation of sleep and wake from actigraphy-based systems is often used. Acquired accelerometer data are processed to compute conventional sleep/wake statistics such as TST, which could then be used to calculate a more accurate AHI for HSAT (Fietze et al., 2015; Marino et al., 2013; Sadeh, 2011) . However, actigraphy by itself is limited in terms of accuracy regarding sleep/wake detection because of the potentially inconsistent relation between sleep and patient motion. In a study comparing actigraphy with PSG in detecting sleep, the actigraphy sensitivity was 96.5%, whereas specificity was very poor and only 32.9% (Marino et al., 2013). This study showed that although actigraphy may be a good tool to identify sleep, it is not a reliable method to detect wakefulness. Cardiorespiratory signals from HSAT have also been used for automatic sleep staging classification (Fonseca et al., 2015; Long, Foussier, Fonseca, Haakma, & Aarts, 2013; Redmond & Heneghan, 2006).

The aim of this study was to evaluate the accuracy and reliability of a novel automated scoring algorithm combining a single EEG channel with actimetry and HSAT signals, for sleep/wake identification and OSA diagnosis.

2 | MATERIALS AND METHODS

2.1 | Study population

This study was conducted on the *Institut de Recherche en Santé Respiratoire des Pays de la Loire [IRSR]* sleep cohort. Since 15 May 2007, consecutive patients aged ≥18 years who were investigated for suspected OSA in seven centres from the west of France have been eligible for inclusion in the IRSR sleep cohort (Gagnadoux et al., 2011). Approval was obtained from the University of Angers ethics committee and the "*Comité Consultative sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé* (C.C.T.I.R.S.) (07.207bis). The database is anonymous and complies with the restrictive requirements of the "*Commission Nationale Informatique et Liberté* (C.N.I.L.) and the French information technology and personal data protection authority. All patients included in the *IRSR* sleep cohort have given their written informed consent.

Two hundred and twenty-seven patients from the *IRSR* sleep cohort investigated by PSG at the Angers University Hospital between June 2016 and December 2017 were assessed for eligibility. Sixty-seven patients were excluded because of psychotropic drug use potentially affecting electroencephalographic (EEG) data (n = 56) and unreliable EEG data (n = 11). Thus, the final study sample size included 160 patients arbitrarily split between the development group (61 patients) and the validation group (99 patients).

2.2 | Measurements, questionnaires and sleep studies

Each patient enrolled in the *IRSR* sleep cohort completed surveys including anthropometric data, smoking habits and medical history.

Excessive daytime sleepiness was evaluated by the Epworth sleepiness scale (Johns, 1991).

All patients underwent a full-night laboratory PSG using the CID102L8D PSG system (Cidelec, France). Recorded data included all electrophysiological signals for sleep evaluation as well as airflow by thermistor and nasal pressure (NP), respiratory inductance plethysmography (RIP) belts, pulse oximetry, body position, limb movements, actigraphy and light. In addition to the laboratory routine, tracheal sounds (TS) and suprasternal pressure (SSP) signals using the PneaVoX® (Cidelec, France) technology were recorded. Placed on the sternal notch, the PneaVoX detects snoring, breathing sound and respiratory effort (Amaddeo et al., 2016; Glos et al., 2018; Meslier et al., 2002). However, only the snoring sound and the suprasternal signals were used in our algorithm for the detection of sleep epochs with snoring or with obstructive events. The EEG channel selected for the single-lead EEG analysis in this study was the FP2-A1 derivation.

2.3 | Automatic sleep/wake scoring algorithm

The algorithm was implemented using Matlab R2016a (Mathworks, Massachusetts, USA). The automatic sleep/wake scoring algorithm (Figure 1), HypnoLighT, was based on a combination of fast Fourier transform (FFT) analysis of a single-lead EEG (FP2-A1) and the variability analysis of HSAT signals (airflow, actimetry, light, snoring, SSP and RIP belts). HypnoLighT operated on a single epoch at a time, so the input to the algorithm consists of 30-s blocks of data. Optimal detection thresholds were derived for each signal using a training set with random recordings of 61 patients (development group). Recordings from the remaining 99 patients were used as a validation group for the developed algorithm. The HypnoLighT algorithm detected wakefulness and sleep, by means of a preprocessing step, followed by five detection steps.

2.3.1 | Pre-processing step: alpha frequencies' analysis

Spectral analysis of the FP2-A1 EEG derivation was first applied on the entire signal to define the alpha wave frequency range specific to the patient. Instead of using a unique 8–13-Hz range, HypnoLighT adapted the detection to each patient's alpha frequency range. This was computed for each patient based on their data alone, and therefore it considered the EEG signal differences among individuals. This pre-processing step was necessary to avoid misclassifications caused by alpha frequency on the edge of the frequency limits or patients with alpha-delta sleep. Alpha rhythms in alpha-delta sleep are usually 1–2 Hz lower than waking alpha (Hauri & Hawkins, 1973). In addition, this step improved the algorithm in differentiating between rapid eye movement (REM) sleep and wake stages. The American Academy of Sleep Medicine (AASM) recommendations indicate that alpha frequency in REM sleep often is 1–2 Hz slower than during wakefulness (Berry et al., 2012).

2.3.2 | Detection steps

For each 30-s block of data, FFT analysis was applied to the FP2-A1 derivation signal and the alpha rhythm was detected. Wake stage was identified when an epoch contained more than 50% of alpha waves. For this step, a first threshold is used to determine whether each recorded second is composed of an alpha wave. A second threshold is then applied to define as 'awake' the epochs in which alpha waves are present for more than 50% of the epoch's time. The eye-blinks pattern was also recognized based on the signal magnitude and its shape and was used to detect wake with open eyes. The agitated wake detection was based on detection of body movements using RIP belts or actimetry, combined with the presence of less than 50% of alpha waves. Whereas the first three steps detected if the patient was awake, the fourth step verified if the patient was asleep while the light was ON. The light signal analysis and the detection of slow waves in the EEG signal were added to enhance the wake-stage detection and to improve the algorithm sensitivity. Finally, a sleep detection step using FP2-A1 derivation, airflow, RIP belts, snoring and SSP was applied to improve the obtained sleep/wake classification. This step detected the presence of slow waves, snoring or obstructive/mixed apneas as well as obstructive hypopneas. In the last step of the algorithm, the presence of snoring or obstructive events is used to detect sleep. Although snoring is used for the characterization of hypopneas as obstructive, the suprasternal pressure is used as a supplement sensor to the RIP belts for a reliable apnea characterization. This fifth step was used to detect sleep and correct the false detections of the first four steps in order to increase wakefulness detection specificity. The detection algorithm is summarized in Figure 1.

2.4 | Sleep scoring and AHI calculation

Sleep and respiratory events were scored visually, according to AASM 2012 guidelines, by three qualified technicians and validated by two somnologists from the Sleep Laboratory (F.G. and N.M.). Epochs with poor-quality signals (1,578/178,052, 0.89%) were not included in the analysis, whereas epochs with the usual EEG artefacts if any (cardiac, sweating, electrode pops, major body movements, etc.) were kept in the analysis but they were neither visually detected nor quantified. However, epochs where any of the polygraphy signals used in the algorithm were missing (disconnected electrodes, or a very weak or saturated signal) or corrupted by noise were excluded from the analysis.

An apnea was defined as a complete cessation of airflow and a hypopnea was defined as at least 30% decrease in the nasal pressure signal combined with either ≥3% arterial oxygen desaturation or an arousal, both lasting at least 10 s (Berry et al., 2012). For each recording, three AHIs were calculated. (i) AHI_{PSG} was based on PSG scoring with respect to the total sleep time (TST) and included hypopneas with no desaturation terminated by an arousal. (ii) AHI_{HSAT} was based on HSAT channels (airflow, actimetry, light, snoring, SSP and RIP belts) using TRT instead of TST. Only hypopneas with ≥3% arterial oxygen desaturation were included in AHI_{HSAT}. (iii)

$AHI_{HypnoLighT}$ was calculated using the TST as detected by the HypnoLighT algorithm, $TST_{HypnoLighT}$, and the number of respiratory events as detected on HSAT channels. Figure 2 illustrates the signals recorded and analyzed for the three AHI estimation approaches. To evaluate the impact of the algorithm on the diagnosis, AHI_{PSG}, AHI_{HSAT} and AHIHypnoLighT were compared.

2.5 | Statistical analysis

The *t* test analyses in Table 1 were performed using the SAS/STAT package 2002–2003 (SAS Institute, Cary, NC, USA). The results presented in Table 2 are obtained using calculation programs developed using Matlab R2016a software. Values are presented as

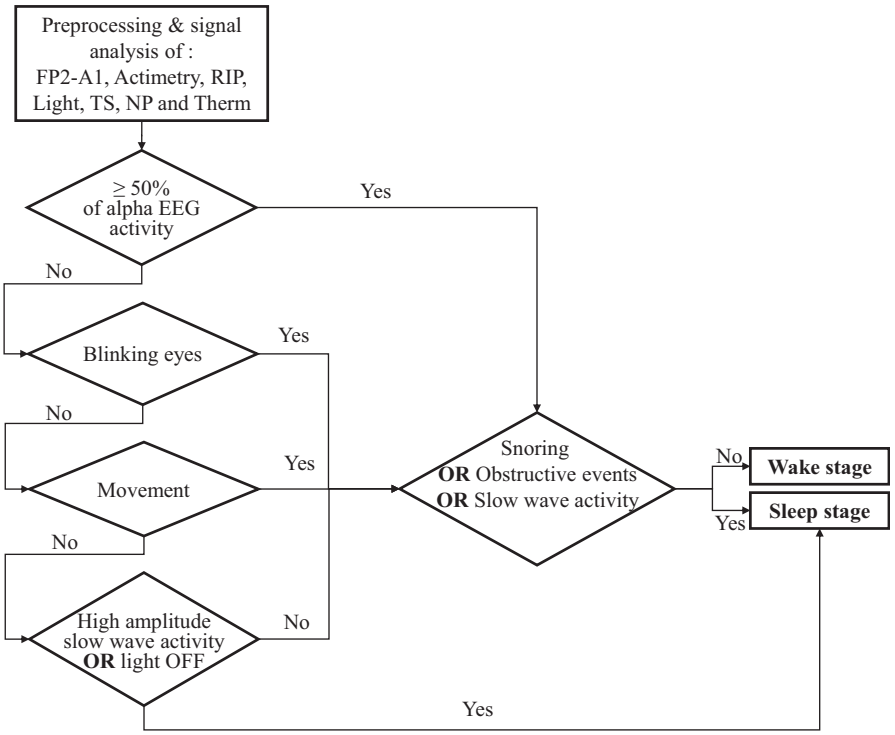


FIGURE 1 The HypnoLighT automatic sleep/wake detection algorithm. The automatic sleep/wake scoring algorithm was based on a combination of fast Fourier transform analysis of a single-lead EEG (FP2-A1) and the variability analysis of HSAT signals (airflow, actimetry, light, snoring, SSP and RIP belts). The algorithm analyzed a 30-s epoch at a time based on optimal detection thresholds that were derived for each signal using a training set with 61 random recordings. EEG, electroencephalogram; HSAT, home sleep apnea testing; SSP, suprasternal pressure; Therm, thermistor; NP, nasal pressure; TS, tracheal sounds; RIP, respiratory inductance plethysmography

	AHI_{PSG}	AHI_{HSAT}	$AHI_{HypnoLighT}$
Signals	• 3 EEGs, 2 EOGs, 1 EMG • 1 ECG • Oronasal thermistor • Nasal pressure • RIP belts • Pulse oximetry • Body position, limb movements • Actigraphy • Light • Tracheal sound (snoring)	• **No electrophysiological signals** • Oronasal thermistor • Nasal pressure • RIP belts • Pulse oximetry • Body position, limb movements • Actigraphy • Light • Tracheal sound (snoring)	• 1 EEG (FP2-A1) • Oronasal thermistor • Nasal pressure • RIP belts • Pulse oximetry • Body position, limb movements • Actigraphy • Light • Tracheal sound (snoring) • Suprasternal pressure
Information	• Wake, N1, N2, N3 and REM • TST • Apneas • Hypopneas with desaturation (H_{Desat}) • Hypopneas with arousal (H_{Ar})	• **No sleep stages** • TRT • Apneas • H_{Desat} • **NO hypopneas with arousal**	• **Sleep/wake stages** • $TST_{Hypnolight}$ • Apneas • H_{Desat} • **NO hypopneas with arousal**
	$AHI_{PSG} = \frac{\text{Apneas} + H_{Desat} + H_{Ar}}{TST_{PSG}}$	$AHI_{HSAT} = \frac{\text{Apneas} + H_{Desat}}{TRT}$	$AHI_{HypnoLight} = \frac{\text{Apneas} + H_{Desat}}{TST_{Hypnolight}}$

FIGURE 2 Illustration of AHI estimation using three different approaches. Recorded and analyzed signals, used information and AHI equations are presented for each method (PSG, HSAT and HypnoLighT). Note that the hypopneas with arousal are not taken into consideration for the estimation of AHI in either the HSAT or the HypnoLight approaches. AHI, apnea–hypopnea index; PSG, polysomnography; HSAT, home sleep apnea testing. EOG, EEG, eMG

mean ± standard deviation (*SD*). Between-group comparisons were performed using the chi-squared test for categorical variables and *t* test for continuous variables. A $p < 0.05$ was considered significant. The Cohen's kappa, sensitivity and specificity, as well as positive predictive and negative predictive values (PPV and NPV) of the automatic wakefulness detection, were calculated. A Bland–Altman plot was also used to visually evaluate the agreement between the PSG and the automatic algorithm analyses.

3 | RESULTS

Table 1 shows the characteristics of the study population. The study population consisted of typical mild-to-severe OSA patients, predominantly male, obese or overweight, and frequently presenting with systemic hypertension and metabolic comorbidities. No significant differences were observed between the development group and the validation group. The analysis included 1,76,474 epochs for the 160 patients (38.27% for the development group and 61.63% for the validation group).

3.1 | Automatic sleep/wake scoring

Automatic spectral analysis was performed on the 160 files; 34.3% of the recordings were detected with an out-of-range alpha frequency (8–13), of which 26.9% had an alpha frequency just under 8 Hz but not distinguishable with the naked eye, 2.4% had an algorithm that was wrong (micro-volt signal) because we could still see alpha, 1.2% had alpha at a frequency that was too low (so no alpha according to AASM regulations) and 3.8% did not have alpha. Thus, only 5% of our patients turned out to be without alpha according to the AASM regulations. For these patients who were not excluded from the analysis, wake detection was based mainly on other variables. This reinforces the interest in combining EEG and non-EEG signals for sleep/wake detection.

Figure 3 shows examples of detection at different steps of the HypnoLight algorithm. Drowsy with closed eyes (3a), awake with blinking eyes 3(b) and awake and agitated (3c) were the first three types of wakefulness detected. Figure 3(d) illustrates an example that validates the use of a light signal to detect wakefulness independently of the first three steps. The combination of the detection of a high-amplitude slow wave with the light signal allows differentiation between an awake stage with the light ON and a sleep stage with the light ON. The application of the first four steps increased the sensibility of the automatic scoring. Figure 3(e) illustrates the use of snoring signals to correct the false detections of the first four steps. This final step was used to increase the specificity of the algorithm to detect sleep. Finally, Figure 4 shows the result of the application of all five steps of the automatic algorithm. For this particular example, compared with the reference PSG visual scoring, the kappa score, VPP, sensitivity and specificity were 0.82, 94%, 92% and 91%, respectively.

The two-state (wake and sleep) epoch-by-epoch agreement is summarized in Table 2. The visual analysis of PSG scored 37,136 (15,597 for the development group and 21,539 for the validation group) epochs as wake, of which 28,413 (76.51%) (12,447 for the development group and 15,966 for the validation group) were correctly classified by the HypnoLight algorithm. The sensitivity of the HypnoLighT algorithm for all individuals for wake detection was

TABLE 1 Characteristics of the study population for both the development group and the validation group

	Development group (min, max)	Validation group (min, max)	*p* value
n	61	99	–
Age (years)	53.13 ± 13.09 (20.00, 80.00)	48.82 ± 13.74 (20.00, 86.00)	0.0537
Body mass index (kg/m^2)	29.28 ± 6.48 (19.96, 51.02)	29.50 ± 7.68 (17.58, 69.40)	0.8525
Female (%)	34.43	22.34	0.0982
Epworth sleepiness scale	10.33 ± 5.04 (0.00, 22.00)	10.14 ± 4.38 (2.00, 20.00)	0.8092
Apnea–hypopnea index (/hr)	21.41 ± 16.63 (1.00, 81.00)	23.55 ± 21.73 (0.00, 103.00)	0.4892
Apnea index (/hr)	5.13 ± 6.71 (0.00, 27.00)	7.43 ± 13.79 (0.00, 99.00)	0.1863
3% oxygen desaturation index (/hr)	16.28 ± 15.63 (0.00, 66.00)	19.57 ± 19.66 (0.00, 105.00)	0.2721
Arousal index (/hr)	25.88 ± 11.51 (5.00, 55.00)	29.40 ± 17.08 (6.00, 99.00)	0.1368
Total sleep time (min)	424.25 ± 79.28 (214.00, 549.00)	446.09 ± 59.81 (277.00, 569.00)	0.0773
N1-N2 sleep (%)	57.84 ± 11.81 (27.60, 93.00)	57.78 ± 9.38 (30.20, 89.20)	0.9730
N3 sleep (%)	21.82 ± 8.42 (2.60, 38.20)	20.03 ± 6.60 (2.40, 39.00)	0.1793
REM sleep (%)	20.46 ± 6.41 (3.00, 35.40)	22.18 ± 5.32 (10.80, 35.90)	0.0789
Systemic hypertension (%)	48.39	50.00	0.8906
Diabetes (%)	20.59	13.95	0.4403
Cardiovascular diseases (%)	9.84	16.30	0.2550

Data are expressed as mean ± *SD* or just percentages. *P*-values are based on *t* tests. There is no statistically significance difference in any of the parameters. When they apply, the minimum and maximum values for certain calculated variables are also given.
Min, minimum value; max, maximum value; REM, rapid eye movement; *SD*, standard deviation.

76.51% ± 21.67% (79.81% ± 19.29% for the development group and 77.12% ± 22.83% for the validation group), with a specificity of 95.48% ± 5.27% (95.83% ± 3.73% for the development group and 95.26% ± 6.04% for the validation group). The PPV and NPV were 81.84% ± 15.42% (85.19% ± 11.07% for the development group and 79.41% ± 17.46% for the validation group) and 93.85% ± 6.23% (94.05% ± 6.60% for the development group and 93.73% ± 6.03% for the validation group), respectively. Cohen's kappa agreement was 0.74 ± 0.18 (0.77 ± 0.14 for the development group and 0.71 ± 0.20 for the validation group), which falls within the range of good agreement (>0.60). Table 3 summarizes the evolution of the detection results throughout the five steps of the algorithm.

In the overall population, false wake detections by HypnoLight compared with PSG visual scoring occurred predominantly in N1 and N2 sleep stages (39.9% and 31.6%) and less frequently during N3 sleep and REM sleep stages (6.5% and 22.0%).

The average TRT for all patients was 9.27 ± 0.45 hr (9.27 ± 0.47 hr for the development group and 9.28 ± 0.44 hr for the validation group), whereas TST_{PSG} was 7.26 ± 1.12 hr (7.10 ± 1.30 hr for the development group and 7.36 ± 1.00 hr for the validation group) and the estimated $TST_{HypnoLighT}$ was 7.38 ± 1.21 hr (7.23 ± 1.29 hr for the development group and 7.48 ± 1.15 hr for the validation group). Figure 5 shows the Bland–Altman plots comparing the TST between the PSG visual scoring, TST_{PSG} and the HypnoLighT algorithm estimation, $TST_{HypnoLighT}$. For the development group, the mean difference value was −8.08 min with an *SD* of 33.8 min. The maximum difference was 156 min and the minimum difference was 1 min. For the validation group, the mean difference value was −7.24 min with an *SD* of 47.2 min. The maximum difference was 191 min and the minimum difference was 0 min.

3.2 | AHI calculation

The total number of events used for the AHI calculations was 19,797 (29.13% for development and 70.87% for test), 19,468 (29.30% for development and 70.70% for validation) and 23,789 (30.70% for development and 69.30% for validation) for HSAT, HypnoLighT and PSG, respectively. The average AHI for all patients was 20.73 ± 19.64 events/hr (17.25 ± 13.33 for development and 22.87 ± 22.47 for validation), 13.61 ± 15.62 events/hr (10.26 ± 9.27 for development and 15.67 ± 18.23 for validation) and 16.62 ± 17.97 events/hr (13.26 ± 12.06 for development and 18.69 ± 20.58 for validation) for AHI_{PSG}, AHI_{HSAT} and $AHI_{HypnoLighT}$, respectively. Figure 6 shows the Bland–Altman plots comparing AHI_{PSG}, AHI_{HSAT} and $AHI_{HypnoLighT}$ with each other. For the development group, the maximum and the minimum differences were 18.00/hr and 0/hr between AHI_{PSG} and $AHI_{HypnoLighT}$, 26.51/hr and 0.02/hr between AHI_{PSG} and AHI_{HSAT}, and 15.15/hr and 0.01/hr between AHI_{HSAT} and $AHI_{HypnoLighT}$. For the evaluation group, the maximum and the minimum differences were 23.74/hr and 0/hr between AHI_{PSG} and $AHI_{HypnoLighT}$, 26.99/hr and 0.02/hr between AHI_{PSG} and AHI_{HSAT}, and 15.48/hr and 0.01/hr between AHI_{HSAT} and $AHI_{HypnoLighT}$.

Figure 7 summarizes the distribution of AHI severity categories for PSG, HypnoLight and HSAT. Comparing AHI_{HSAT} with AHI_{PSG}, 48 patients changed categories (25 from no or mild OSA [AHI < 15] to moderate [15 ≤ AHI < 30] OSA and 23 from mild or moderate to severe [AHI > 30] OSA). Twenty-seven of these 48 patients were successfully reclassified (16 from no or mild OSA to moderate OSA and 11 from mild or moderate to severe OSA) by adding the HypnoLighT algorithm to a simple HSAT. In fact, 56.25% of patients who were classified wrongly by HSAT were classified correctly after applying the HypnoLighT algorithm.

Finally, in 61 patients with predominantly severe OSA (mean AHI_{PSG} = 35.94 ± 15.52), false detection of wake stages resulted in a loss of 329 events between HSAT and HypnoLight, including 221 central apneas, 95 central hypopneas and 13 undetermined hypopneas.

4 | DISCUSSION

Our study demonstrates that fully automated analysis of a single-lead EEG channel (FP2-A1) combined with HSAT signals provides reliable wake/sleep identification, improves AHI calculation and, thus, improves the reliability of HSAT for OSA diagnosis and severity assessment. Our findings suggest that automatic wake/sleep detection using the proposed HypnolighT algorithm could decrease by

TABLE 2 Statistical results for all patients as well as for the development group and the validation group separately

	Single EEG spectral analysis (mean ± *SD*)			HypnoLight analysis (mean ± *SD*)		
	All patients (min, max)	Development (min, max)	Validation (min, max)	All patients (min, max)	Development (min, max)	Validation (min, max)
n	160	61	99	160	61	99
Cohen's kappa	0.55 ± 0.27 (−0.10, 0.92)	0.59 ± 0.24 (0.00, 0.92)	0.52 ± 0.28 (−0.10, 0.91)	0.74 ± 0.18 (0.11, 0.94)	0.77 ± 0.14 (0.27, 0.94)	0.71 ± 0.20 (0.11, 0.91)
PPV (%)	76.28 ± 20.11 (8.27, 100.00)	81.12 ± 17.37 (25.00, 100.00)	72.75 ± 21.64 (8.27, 100.00)	81.84 ± 15.42 (22.84, 100.00)	85.19 ± 11.07 (58.06, 98.34)	79.41 ± 17.46 (22.84, 100.00)
NPV (%)	88.56 ± 9.17 (47.81, 99.48)	88.21 ± 10.25 (47.81, 99.48)	88.77 ± 8.47 (57.64, 99.20)	93.85 ± 6.23 (57.24, 99.80)	94.05 ± 6.60 (57.24, 99.80)	93.73 ± 6.03 (60.57, 99.51)
Sensitivity (%)	53.69 ± 29.94 (0.39, 99.21)	57.26 ± 26.93 (0.39, 96.59)	51.1 ± 31.10 (0.76, 99.21)	76.51 ± 21.67 (8.35, 99.21)	79.81 ± 19.29 (19.40, 98.64)	74.12 ± 22.83 (8.35, 99.21)
Specificity (%)	95.55 ± 7.08 (52.21, 100.00)	96.00 ± 4.25 (79.22, 100.00)	95.28 ± 8.36 (52.21, 100.00)	95.48 ± 5.27 (61.54, 100.00)	95.83 ± 3.73 (80.66, 99.69)	95.26 ± 6.04 (61.54, 100.00)

The results are given for the first step of the algorithm using just the spectral analysis of the single-lead EEG signal (detection of alpha wave presence ≥50%), and for the HypnoLight algorithm combining the single-lead EEG analysis with the HSAT signal analysis. Information on mean and *SD* values derived from single recordings are given for all calculated variables. The minimum and maximum values for each calculated variable are also given.
Min, minimum value; max, maximum value; PPV, positive predictive value; NPV, negative predictive value; *SD*, standard deviation.

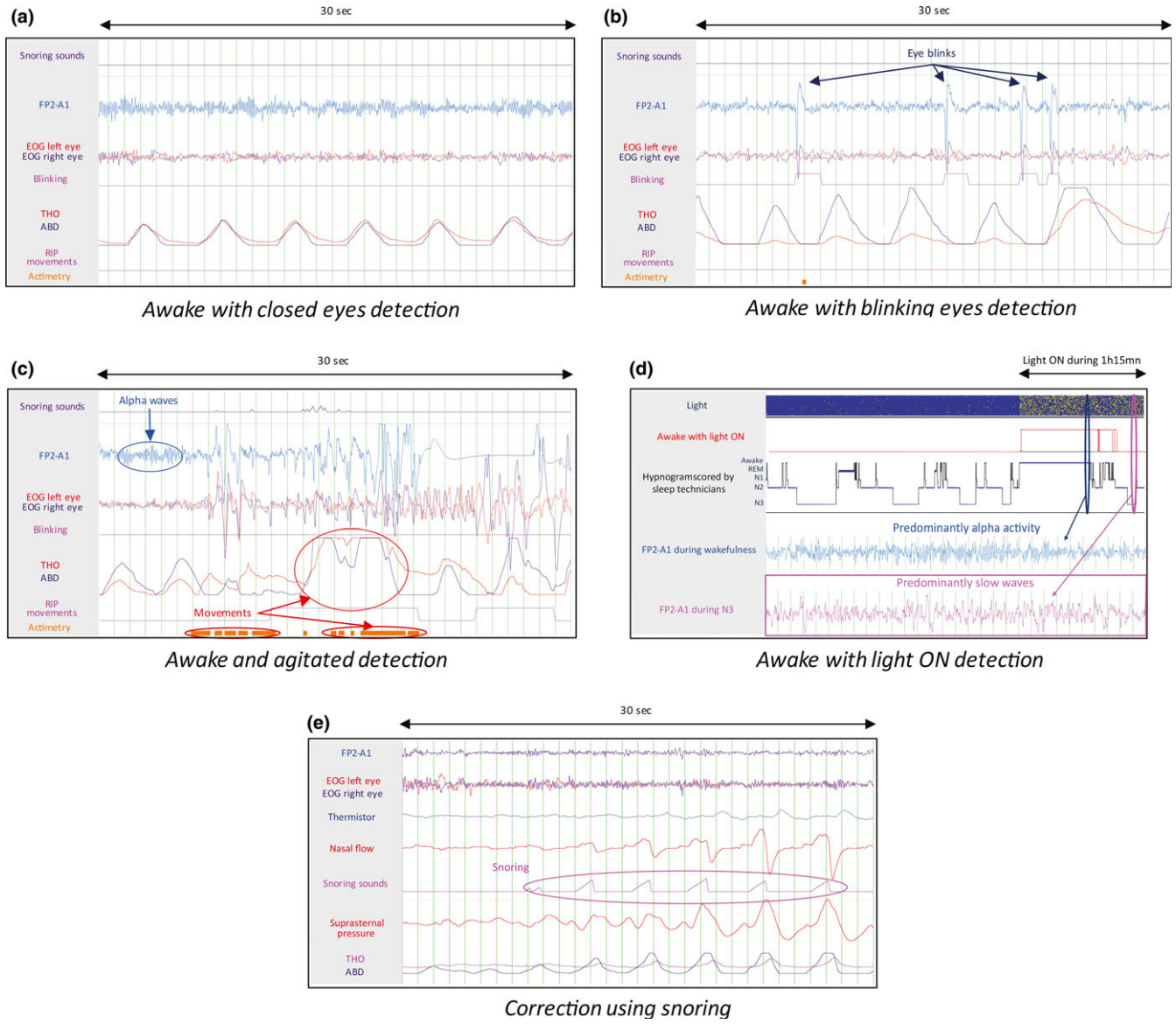


FIGURE 3 Examples of detection at different steps of the HypnoLight algorithm. The five steps are illustrated: (a) drowsy with closed eyes, (b) awake with blinking eyes, (c) awake and agitated, (d) asleep/awake with light ON detection and (e) correction using snoring events. EOG, electrooculogram; THO, thorax; ABD, abdomen; RIP, respiratory inductance plethysmography

more than 50% the number of additional PSG recordings needed in patients with borderline AHI wrongly classified by HSAT.

Previous small-sample-size studies have evaluated automated scoring software using single-channel EEG data (Fietze et al., 2015; Fraiwan, Lweesy, Khasawneh, Wenz, & Dickhaus, 2012; Garcia-Molina et al., 2015; Lucey et al., 2016; Su, Luo, Hong, Nagurka, & Yen, 2015; Zhang & Wu, 2018; Zhu, Li, & Wen, 2014). In 15 healthy adults. Berthomier et al. have compared single EEG channel (Cz-Pz) automated sleep scoring software with visual scoring of a standard PSG. Epoch-by-epoch comparison was based on classification into two states (wake/sleep), three states (wake/REM/NREM), four states (wake/REM/N1-N2/N3) or five states (wake/REM/N1/N2/N3). The overall agreements, as quantified by the kappa coefficient, were 0.82, 0.81, 0.75 and 0.72, respectively. The method achieved 98.1% sensitivity and 82.5% specificity for detecting sleep (Berthomier et al., 2007). In another study, automatic identification of sleep stages on a single EEG channel (C4–A1) produced average Cohen's kappa of 0.86 for the test subjects (n = 12) and 0.88 for the training subjects (n = 16) (Koley & Dey, 2012). In 99 patients Kaplan et al. showed that an automated single EEG channel (A1–A2) algorithm achieved an overall sensitivity for detecting sleep of 95.5%, with a specificity of 92.5%, PPV of 98.0% and NPV of 84.2% (Kaplan et al., 2014). Fietze et al. combined actigraphy with a single-lead EEG (F4-M1) in one group and added electro-oculographic (EOG) and electro-myogram (EMG) to a second group. For the first group, visual sleep stage scoring was performed only using the single-channel EEG according to Rechtschaffen & Kales, and for the second group, the scoring included EOG and EMG data as well. The study showed high

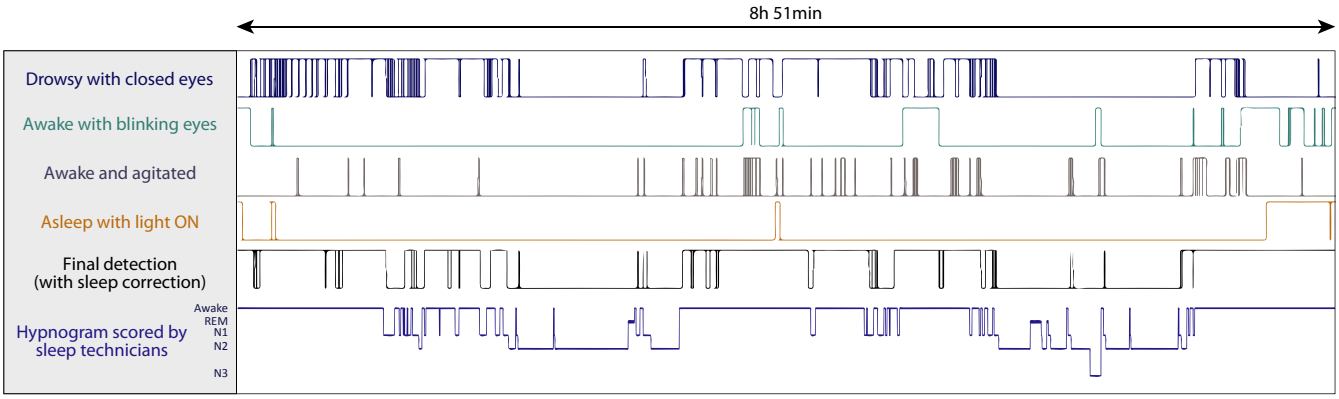


FIGURE 4 Example of the application of the sleep/awake detection algorithm to a one-night recording. The result of the detection after each step is given. The overall final sleep/wake detection using the HypnoLighT algorithm is compared with the hypnogram of the PSG reference scoring. PSG, polysomnography

TABLE 3 Evolution of the detection results throughout the steps of the algorithm

	Alpha wave activity ≥50% (min, max)	Blinking eyes (min, max)	Alpha wave activity ≤50% and movement (min, max)	High amplitude slow wave activity *or* light OFF (min, max)	Snoring *or* obstructive events *or* slow wave activity (min, max)
Cohen's kappa	0.55 ± 0.27 (−0.10, 0.92)	0.69 ± 0.23 (0.03, 0.92)	0.72 ± 0.19 (0.03, 0.92)	0.73 ± 0.19 (0.14, 0.94)	0.74 ± 0.18 (0.14, 0.94)
PPV (%)	76.28 ± 20.11 (8.27, 100.00)	78.64 ± 19.18 (16.67, 99.09)	76.44 ± 17.46 (18.51, 99.29)	75.12 ± 18.28 (18.51, 99.54)	81.84 ± 15.42 (22.84, 100.00)
NPV (%)	88.56 ± 9.17 (47.81, 99.48)	92.63 ± 7.13 (49.85, 99.60)	94.73 ± 6.50 (54.38, 99.88)	95.17 ± 6.28 (57.28, 99.88)	93.85 ± 6.23 (57.24, 99.80)
Sensitivity (%)	53.69 ± 29.94 (0.39, 99.21)	71.68 ± 26.77 (2.99, 99.60)	80.49 ± 23.19 (10.54, 99.60)	82.36 ± 22.23 (10.54, 99.60)	76.51 ± 21.67 (8.35, 99.21)
Specificity (%)	95.55 ± 7.08 (52.21, 100.00)	94.81 ± 7.07 (52.10, 99.91)	93.39 ± 8.25 (46.77, 99/88)	92.73 ± 8.47 (46.77, 99.88)	95.48 ± 5.27 (61.54, 100.00)

Data are expressed as mean ± *SD* or just percentages. The minimum and maximum values for each calculated variable are also given.
Min, minimum value; max, maximum value; PPV, positive predictive value; NPV, negative predictive value; *SD*, standard deviation.

agreement between PSG and single-lead EEG in sleep apnea patients. However, the agreement was slightly lower for the single-lead EEG by itself than when EOG and EMG data were added to the analysis (Fietze et al., 2015).

Similarly to some previous studies (Kaplan et al., 2014; Lucey et al., 2016; Wang, Loparo, Kelly, & Kaplan, 2015), we used the EEG channel FP2-A1 with the FP2 lead placed just below the hairline or right on the hairline for most of our patients. This has the advantage, especially with male patients who have lost their hair, that the lead placement may not require the use of Collodion glue and a simple snap electrode could be used. Such techniques are suitable for a variety of in-home monitoring applications. Furthermore, this placement allows us to get better quality eye movement signals for use in our algorithm.

Compared with previous studies using complex EEG automatic analysis for wake detection, the HypnoLighT algorithm uses only FFT analysis. This resulted in a high specificity (96%), but the sensitivity (54%) and Kappa coefficient (0.55) were low. This difference in performance compared with EEG data from other studies is likely to be because they were mainly obtained from healthy subjects, which was not the case for our study. However, when adding the HSAT signal analysis to the EEG data in later steps, the algorithm's performance improved and we found results similar to those in other studies. Nevertheless, the methodologies used in these studies are too heterogeneous to draw any concrete conclusions about the superiority of one algorithm over the other.

As expected, false wake detection by HypnoLight occurred predominantly during light N1 and N2 sleep stages. Indeed, differentiating wake from the N1 stage is difficult by automatic sleep analysis (Finan et al., 2016; Stepnowsky, Levendowski, Popovic, Ayappa, & Rapoport, 2013) and by visual scoring (Rosenberg & Van Hout, 2013). In a study evaluating the validity of an ambulatory EEG monitor in 14 patients, Finan et al. showed that sensitivity was low for wake and stage N1, and high for N2, N3 and REM stages. Kappa was strongest for N3 and REM stages (Finan et al., 2016). Stepnovsky et al. compared an automated sleep staging method using a bipolar electro-ocular recording with manual scoring by multiple raters. Their results showed that a single channel of forehead EEG recordings agreed with the majority of manual scoring, with the lowest agreement observed in stage N1 (Stepnowsky et al., 2013). Rosenberg et al. showed that inter-scorer agreement in stage N1 is the lowest (63%), and that almost all the errors in the N1 stage were made with wake and N2 stages (Rosenberg & Van Hout, 2013). Similarly, in our study, the differentiation between wake and N1 had the

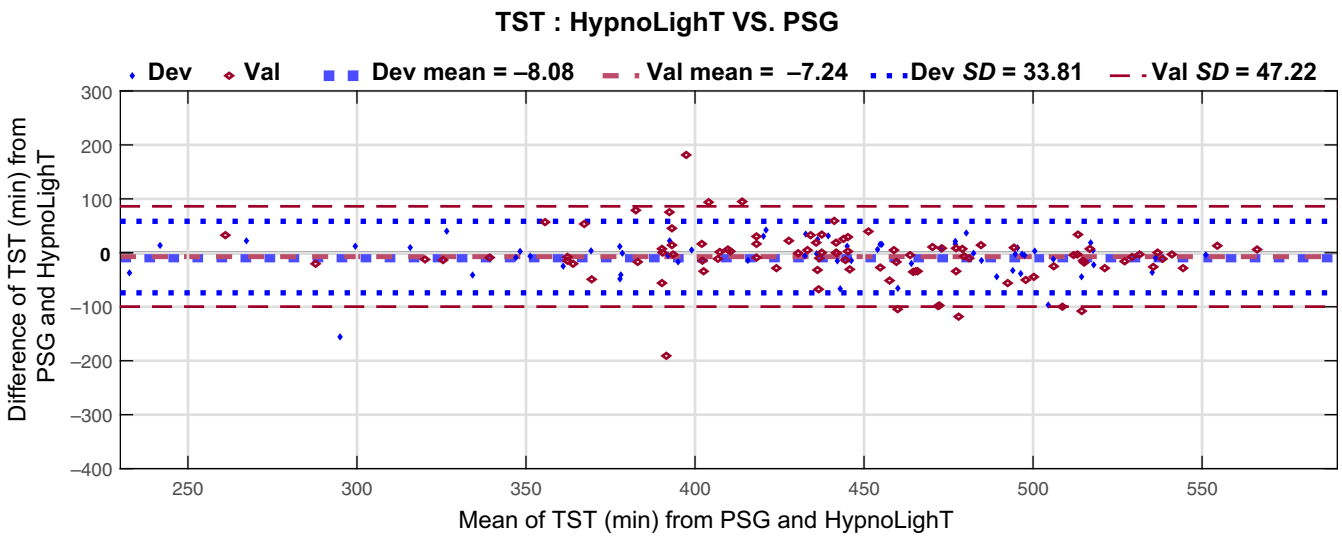


FIGURE 5 Bland–Altman plots comparing the TST between the PSG visual scoring, TST_{PSG}, and the HypnoLighT algorithm estimation, $TST_{HypnoLighT}$. Results for the development group are given in blue and the validation group in red. For the development group, the mean difference value was −8.08 min with a standard deviation of 33.81 min. For the validation group, the mean difference value was −7.24 min with a standard deviation of 47.22 min. The dashed lines in the figure indicate the mean ± 2 *SD* limits. TST, total sleep time; PSG, polysomnography; Dev, development; Val, validation; *SD*, standard deviation

highest rate of error for both the development and the validation groups. Although PSG is our "reference standard" comparator, scorers may have inaccurately estimated the wake and N1 stages.

The second highest error rate occurred in stage N2. In an additional analysis (data not shown), we found that most arousals were scored by PSG during N2 and that most of the errors associated with arousals occurred during epochs scored as N2 stage. In our study, the high occurrence of errors in the N2 stage is most likely a result of the presence of respiratory-related arousals during this stage, as false detection of a wake stage was observed predominantly in patients with severe OSA. Respiratory event-related arousals correlate significantly with AHI (Yamashiro, Suganuma, Hosaka, & Uchida, 1998) and the more severe the OSA, the more fragmented the sleep. Even manual sleep staging is less reliable when sleep is highly fragmented (Danker-Hopfe et al., 2009; Norman, Pal, Stewart, Walsleben, & Rapoport, 2000).

Despite the high agreement between $TST_{HypnoLighT}$ and TST_{PSG}, the $AHI_{HypnoLighT}$ was only increased halfway between the AHI_{PSG} and AHI_{HSAT}. This is simply because 4,261 (17.91% of total PSG events) non-desaturating hypopneas associated with arousal were scored on PSG but not used with the HypnoLighT algorithm. An automated reliable detection of arousals would allow hypopneas without desaturation to be considered for the calculation of $AHI_{HypnoLighT}$. This would enhance further the estimation of AHI and would yield more accurate AHI values compared with PSG. Although single-lead EEG may not be sufficient for estimating arousals, non-cortical indicators of arousal using cardiorespiratory signals along with other HSAT signals may increase the reliability of arousal detection without the use of a PSG. Compared with the usual HSAT signals used to evaluate sleep stages in the absence of EEG signals, our algorithm has the advantage of using TS and SSP signals, which could increase the reliability of the respiratory scoring (Sabil et al., 2018).

Furthermore, a small number of events (1.7%) were included in the HSAT and missed by the HypnoLighT algorithm. The fact that all these missed events were central shows that the algorithm was successful in scoring all epochs with obstructive events as sleep. Central events were not included in the automatic scoring of sleep stage as central events often occur during wake to sleep transition and are often observed in both wake and N1 epochs (Yamazaki, Asakura, Fujimura, Yoshida, & Matsuda, 1989). However, missing these events had no significant effect on $AHI_{HypnoLighT}$ calculation. Nevertheless, besides AHI calculation, TST automatic estimation during HSAT may be very useful in various clinical settings, such as OSA with comorbid insomnia, which occurs in between 22% and 55% of all OSA patients (Al-Jawder & Bahammam, 2012).

A limitation of this study is that we did not include patients with medical conditions or drug treatments that could affect EEG. This criterion excluded one-third of the patients in our cohort, which means that the algorithm could be unreliable for these patients. The other limitation is that EEG with artefacts was included in the analysis, which may have influenced the optimization of the algorithm. During EEG recording, several sources of artefacts exist and therefore several kinds of noise can contaminate the raw signal (Brown, 2000; Gwin, Gramann, Makeig, & Ferris, 2010). Removal of these artefacts prior to the EEG data analysis could improve the performance of the algorithm. Finally, the algorithm has been developed and tested on recordings of patients with predominantly obstructive events and may not be as accurate with patients with central events.

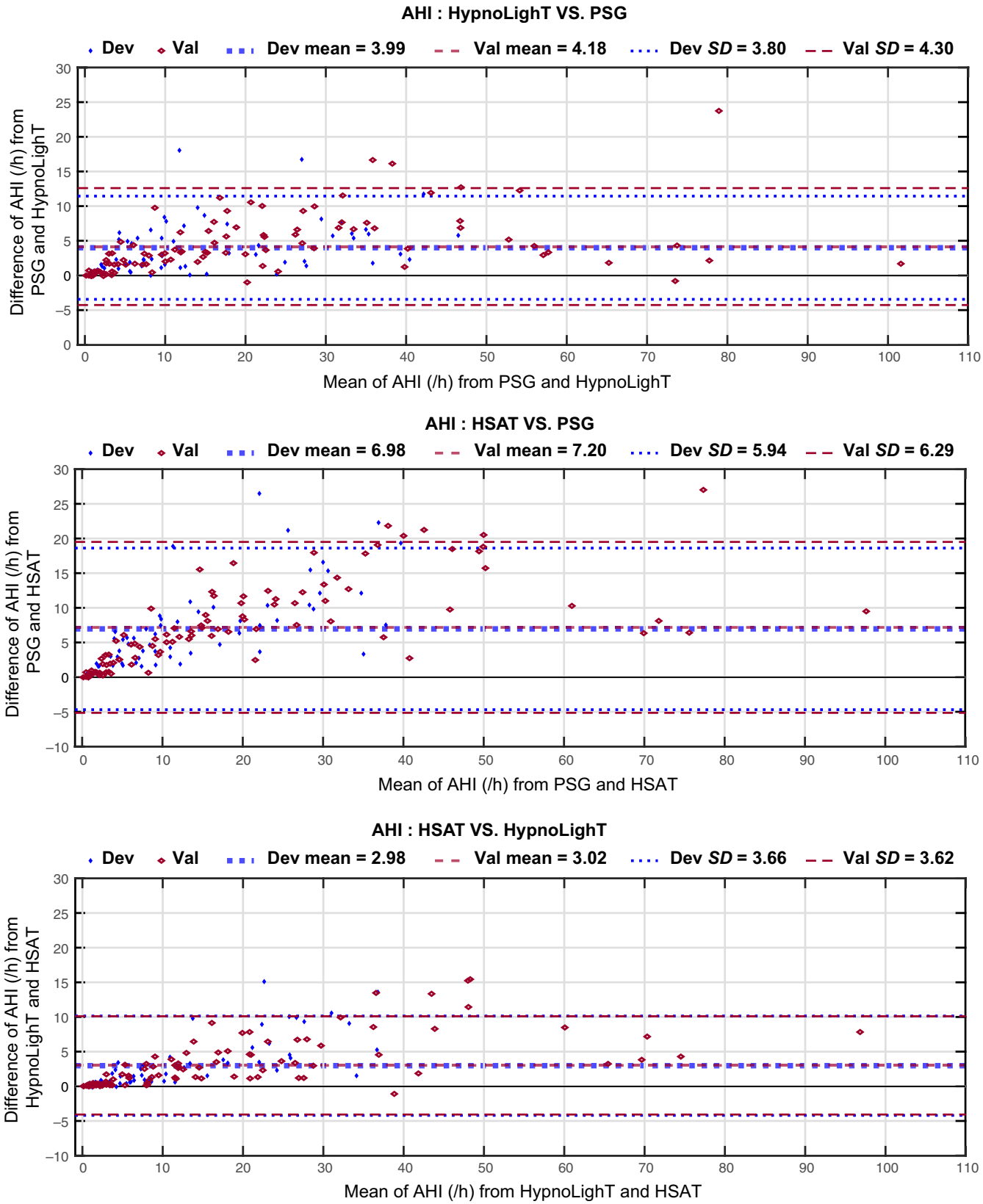


FIGURE 6 Bland–Altman plots comparing the apnea hypopnea index for PSG, HSAT and HypnoLighT (AHI_{PSG}, AHI_{HSAT} and $AHI_{HypnoLighT}$). Results for the development group are given in blue and the validation group in red. Compared with PSG scoring, the mean difference value and the standard deviation were smaller for $AHI_{HypnoLighT}$ than for AHI_{HSAT}. The dashed lines in the figure indicate the mean ± 2 *SD* limits. PSG, polysomnography; AHI, apnea–hypopnea index; Dev, development; Val, validation; *SD*, standard deviation; HSAT, home sleep apnea testing

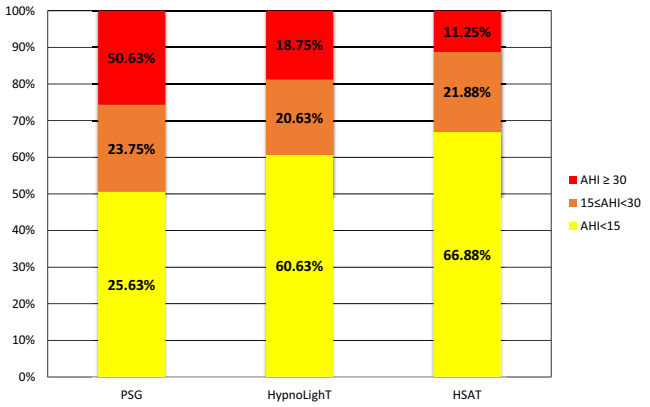


FIGURE 7 Distribution of AHI severity categories for PSG, HypnoLight and HSAT. Comparing AHI_{HSAT} with AHI_{PSG}, 48 patients changed categories (25 from no or mild [AHI < 15] OSA to moderate [15 ≤ AHI < 30] OSA and 23 from mild or moderate to severe [AHI > 30] OSA). Twenty-seven of these 48 patients were successfully reclassified (16 from no or mild OSA to moderate OSA and 11 from mild or moderate to severe OSA) by adding the HypnoLighT algorithm to a simple HSAT. PSG, polysomnography; AHI, apnea–hypopnea index; OSA, obstructive sleep apnea; HSAT, home sleep apnea testing

5 | CONCLUSIONS

A fully automated analysis of a single-lead EEG channel (FP2-A1) combined with HSAT signals was reliable for wake/sleep identification and improved AHI calculation compared with HSAT. This could decrease the number of PSG recordings needed for a therapeutic decision in patients with borderline AHI and/or OSA with comorbid insomnia.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the help of the sleep technicians at the University Hospital of Angers.

CONFLICT OF INTEREST

AbdelKebir Sabil and Guillaume Baffet are fully employed by CIDELEC. All other authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

ORCID

AbdelKebir Sabil http://orcid.org/0000-0002-8966-5441

REFERENCES

Al-Jawder, S. E., & Bahammam, A. S. (2012). Comorbid insomnia in sleep-related breathing disorders: An under-recognized association. *Sleep and Breathing*, *16*(2), 295–304.

Amaddeo, A., Fernandez-Bolanos, M., Olmo Arroyo, J., Khirani, S., Baffet, G., & Fauroux, B. (2016). Validation of a suprasternal pressure sensor for sleep apnea classification in children. *Journal of Clinical Sleep Medicine*, *12*(12), 1641–1647.

Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., … American Academy of Sleep Medicine. (2012). Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events. Deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *Journal of Clinical Sleep Medicine*, *8*(5), 597–619.

Berthomier, C., Drouot, X., Herman-Stoica, M., Berthomier, P., Prado, J., Bokar-Thire, D., … d'Ortho, M. P. (2007). Automatic analysis of single-channel sleep EEG: Validation in healthy individuals. *Sleep*, *30*(11), 1587–1595.

Brown, P. (2000). Cortical drives to human muscle: The Piper and related rhythms. *Progress in Neurobiology*, *60*(1), 97–108.

Corral, J., Sanchez-Quiroga, M. A., Carmona-Bernal, C., Sanchez-Armengol, A., de la Torre, A. S., Duran-Cantolla, J., … Spanish Sleep, N. (2017). Conventional polysomnography is not necessary for the management of most patients with suspected obstructive sleep apnea. Noninferiority, randomized controlled trial. *American Journal of Respiratory and Critical Care Medicine*, *196*(9), 1181–1190.

Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., … Dorffner, G. (2009). Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, *18*(1), 74–84.

Ehlert, I., Danker-Hopfe, H., Höller, L., Von Rickenbach, P., Baumgart-Schmitt, R., & Herrmann, W. M. (1998). A comparison between EEG-recording and scoring by QUISI version 1.0 and standard PSG with visual scoring. *Somnologie-Schlafforschung und Schlafmedizin*, *2*, 104–116.

Escourrou, P., Grote, L., Penzel, T., McNicholas, W. T., Verbraecken, J., Tkacova, R., … ESADA Study Group. (2015). The diagnostic method has a strong influence on classification of obstructive sleep apnea. *Journal of Sleep Research*, *24*(6), 730–738.

Fietze, I., Penzel, T., Partinen, M., Sauter, J., Kuchler, G., Suvoro, A., & Hein, H. (2015). Actigraphy combined with EEG compared to polysomnography in sleep apnea patients. *Physiological Measurement*, *36*(3), 385–396.

Finan, P. H., Richards, J. M., Gamaldo, C. E., Han, D., Leoutsakos, J. M., Salas, R., … Smith, M. T. (2016). Validation of a wireless, self-application, ambulatory electroencephalographic sleep monitoring device in healthy volunteers. *Journal of Clinical Sleep Medicine*, *12*(11), 1443–1451.

Fonseca, P., Long, X., Radha, M., Haakma, R., Aarts, R. M., & Rolink, J. (2015). Sleep stage classification with ECG and respiratory effort. *Physiological Measurement*, *36*(10), 2027–2040.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., & Dickhaus, H. (2012). Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, *108*(1), 10–19.

Gagnadoux, F., Le Vaillant, M., Goupil, F., Pigeanne, T., Chollet, S., Masson, P., … IRSR sleep cohort group. (2011). Influence of marital status and employment status on long-term adherence with continuous positive airway pressure in sleep apnea patients. *PLoS One*, *6*(8), e22503.

Garcia-Molina, G., Bellesi, M., Riedner, B., Pastoor, S., Pfundtner, S., & Tononi, G. (2015). Automatic characterization of sleep need dissipation dynamics using a single EEG signal. *Conference Proceedings IEEE Engineering in Medicine and Biology Society*, *2015*, 5993–5997.

Glos, M., Sabil, A., Jelavic, K. S., Schobel, C., Fietze, I., & Penzel, T. (2018). Characterization of respiratory events in obstructive sleep apnea using suprasternal pressure monitoring. *Journal of Clinical Sleep Medicine*, *14*(3), 359–369.

Gwin, J. T., Gramann, K., Makeig, S., & Ferris, D. P. (2010). Removal of movement artifact from high-density EEG recorded during walking and running. *Journal of Neurophysiology*, *103*(6), 3526–3534.

Hauri, P., & Hawkins, D. R. (1973). Alpha-delta sleep. *Electroencephalography and Clinical Neurophysiology*, *34*(3), 233–237.

Heinzer, R., Vat, S., Marques-Vidal, P., Marti-Soler, H., Andries, D., Tobback, N., … Haba-Rubio, J. (2015). Prevalence of sleep-disordered breathing in the general population: The HypnoLaus study. *The Lancet. Respiratory Medicine*, *3*(4), 310–318.

Johns, M. W. (1991). A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep*, *14*(6), 540–545.

Kaplan, R. F., Wang, Y., Loparo, K. A., Kelly, M. R., & Bootzin, R. R. (2014). Performance evaluation of an automated single-channel sleep-wake detection algorithm. *Nature and Science of Sleep*, *6*, 113–122.

Kapur, V. K., Auckley, D. H., Chowdhuri, S., Kuhlmann, D. C., Mehra, R., Ramar, K., & Harrod, C. G. (2017). Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American Academy of Sleep Medicine Clinical Practice Guideline. *Journal of Clinical Sleep Medicine*, *13*(3), 479–504.

Koley, B., & Dey, D. (2012). An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine*, *42*(12), 1186–1195.

Long, X., Foussier, J., Fonseca, P., Haakma, R., & Aarts, R. M. (2013). Respiration amplitude analysis for REM and NREM sleep classification. *Conference Proceedings IEEE Engineering in Medicine and Biology Society*, *2013*, 5017–5020.

Lucey, B. P., McLeland, J. S., Toedebusch, C. D., Boyd, J., Morris, J. C., Landsness, E. C., … Holtzman, D. M. (2016). Comparison of a single-channel EEG sleep study to polysomnography. *Journal of Sleep Research*, *25*(6), 625–635.

Malhotra, A., Younes, M., Kuna, S. T., Benca, R., Kushida, C. A., Walsh, J., … Pien, G. W. (2013). Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep*, *36*(4), 573–582.

Marino, M., Li, Y., Rueschman, M. N., Winkelman, J. W., Ellenbogen, J. M., Solet, J. M., … Buxton, O. M. (2013). Measuring sleep: Accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*, *36*(11), 1747–1755.

Meslier, N., Simon, I., Kouatchet, A., Ouksel, H., Person, C., & Racineux, J. L. (2002). Validation of a suprasternal pressure transducer for apnea classification during sleep. *Sleep*, *25*(7), 753–757.

Norman, R. G., Pal, I., Stewart, C., Walsleben, J. A., & Rapoport, D. M. (2000). Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*, *23*(7), 901–908.

Peppard, P. E., Young, T., Barnet, J. H., Palta, M., Hagen, E. W., & Hla, K. M. (2013). Increased prevalence of sleep-disordered breathing in adults. *American Journal of Epidemiology*, *177*(9), 1006–1014.

Redmond, S. J., & Heneghan, C. (2006). Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea. *IEEE Transactions on Biomedical Engineering*, *53*(3), 485–496.

Rosenberg, R. S., & Van Hout, S. (2013). The American Academy of Sleep Medicine inter-scorer reliability program: Sleep stage scoring. *Journal of Clinical Sleep Medicine*, *9*(1), 81–87.

Sabil, A., Baffet, G., Rakotonanahary, D., Chaufton, C., Launois, S., & Nguyen, X. (2018). Apport des signaux des sons trachéaux et de la pression susternale pour améliorer l'analyse de la polysomnographie ambulatoire chez l'enfant. *Médecine du Sommeil*, *15*(1), 1.

Sadeh, A. (2011). The role and validity of actigraphy in sleep medicine: An update. *Sleep Medicine Reviews*, *15*(4), 259–267.

Stepnowsky, C., Levendowski, D., Popovic, D., Ayappa, I., & Rapoport, D. M. (2013). Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters. *Sleep Medicine*, *14*(11), 1199–1207.

Su, B. L., Luo, Y., Hong, C. Y., Nagurka, M. L., & Yen, C. W. (2015). Detecting slow wave sleep using a single EEG signal channel. *Journal of Neuroscience Methods*, *243*, 47–52.

Wang, Y., Loparo, K. A., Kelly, M. R., & Kaplan, R. F. (2015). Evaluation of an automated single-channel sleep staging algorithm. *Nature and Science of Sleep*, *7*, 101–111.

Yamashiro, Y., Suganuma, Y., Hosaka, K., & Uchida, K. (1998). Usefulness of arousal for the diagnosis of sleep breathing disorder. *Psychiatry and Clinical Neurosciences*, *52*(2), 211–212.

Yamazaki, M., Asakura, H., Fujimura, M., Yoshida, T., & Matsuda, T. (1989). A case of central sleep apnea syndromes and tachypnea at awake time due to arteriosclerotic changes in right vertebral artery. *Nihon Naika Gakkai Zasshi*, *78*(10), 1480–1481.

Zhang, J., & Wu, Y. (2018). Automatic sleep stage classification of single-channel EEG by using complex-valued convolutional neural network. *Biomedizinische Technik. Biomedical Engineering*, *63*(2), 177–190.

Zhu, G., Li, Y., & Wen, P. P. (2014). Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. *IEEE Journal of Biomedical and Health Informatics*, *18*(6), 1813–1821.

How to cite this article: Sabil A, Vanbuis J, Baffet G, et al. Automatic identification of sleep and wakefulness using single-channel EEG and respiratory polygraphy signals for the diagnosis of obstructive sleep apnea. *J Sleep Res*. 2019;28: e12795. https://doi.org/10.1111/jsr.12795

Informations supplémentaires : les différentes étapes de la classification éveil/sommeil

Nous allons ici développer un peu plus les différentes étapes de la classification éveil/sommeil mises en place. Ces différentes étapes ont été construites principalement à partir des connaissances médicales et en raisonnant sur les signaux. La classification a consisté à considérer, dans un premier temps, tout l'enregistrement comme étant du sommeil, puis à détecter les époques d'éveil. La Figure C3 présente les différentes blocs de détection implémentés.

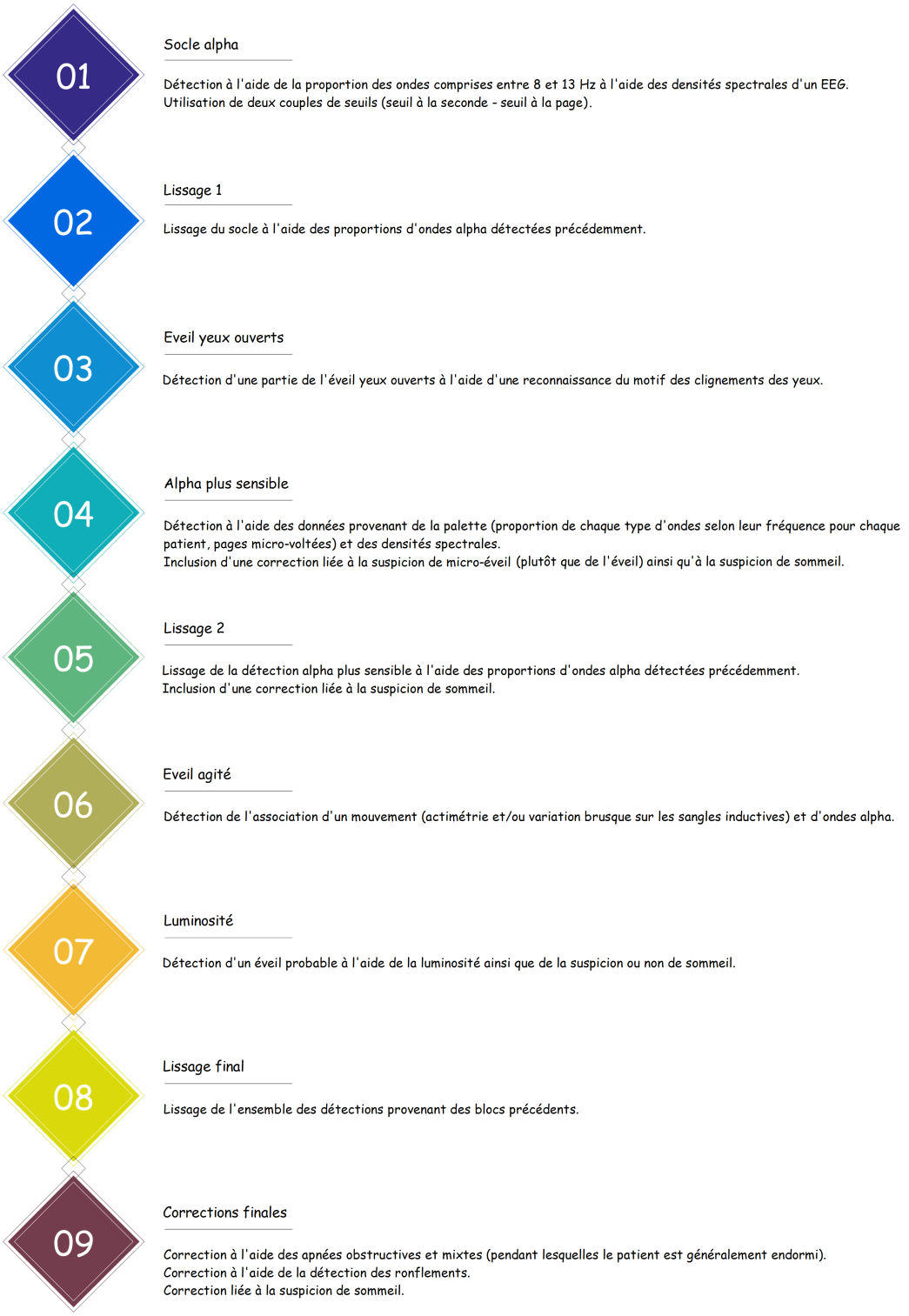


Figure C3 – Diagramme récapitulatif des blocs de détection mis en place pour HypnoLighT.

La détection de l'éveil calme yeux fermés est réalisée en combinant les blocs socle alpha, lissage 1, alpha plus sensible et lissage 2.
En bref, le bloc socle alpha consiste à évaluer la quantité d'ondes alpha à l'aide de deux densités spectrales de puissances (la différence résidant en un décalage d'une demi-seconde), évaluées à partir de la voie EEG. On définit ensuite un couple de seuils (association d'un seuil à la seconde et d'un seuil à l'époque) pour chacune des densités spectrales de puissance, afin d'obtenir deux détections d'éveil calme par époque. Les secondes et époques micro-voltées (dont l'amplitude est extrêmement faible et complique l'estimation des fréquences présentes) sont détectées et les seuils sont adaptés en fonction. La détection finale est un simple « ou logique » entre les deux détections.
Le bloc alpha plus sensible fonctionne sur le même principe (utilisation de couples de seuils appliqués sur les densités spectrales de puissances), mais a également recourt à la « palette », une représentation graphique du contenu fréquentiel de l'EEG calculée grâce au logiciel CIDELEC. Elle est l'équivalent d'une transformée de Fourier condensée et adaptée de manière à être plus facilement visualisée. En effet, elle est constituée de 4 couleurs indiquant le niveau de présence de chaque fréquence, pour chaque époque (rouge = très fréquente, jaune = plutôt fréquente, vert = moyennement fréquente et bleu = peu fréquente). Un exemple est présenté Figure C4.

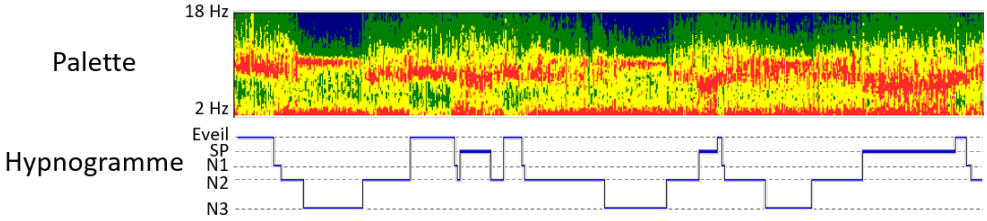


Figure C4 – Exemple de palette et hypnogramme lu manuellement correspondant.

La fréquence relative aux ondes alpha du patient et spécifique à l'éveil est détectée à partir de cette palette. Tout comme pour le socle alpha, des couples de seuils sont ensuite utilisés pour obtenir une détection cette fois-ci très sensible, mais peu spécifique, de l'éveil calme. Grâce à la palette, les zones de suspicion de sommeil ou de micro-éveils sont identifiées et utilisées pour corriger cette détection et permettre l'obtention d'une spécificité de nouveau satisfaisante.
Les blocs lissage 1 et lissage 2 consistent quant à eux à forcer à l'éveil les époques dont la proportion d'ondes alpha reste tout de même relativement importante (simples seuillages) et qui sont adjacentes à une époque mise à l'éveil. Ce lissage est répété itérativement trois fois.

La détection de l'éveil yeux ouverts est réalisée à partir du bloc du même nom. Elle repose principalement sur la détection des clignements des yeux qui sont visibles sur la voie EEG. Dans un premier temps, l'EEG est filtré une première fois afin de retirer l'offset du signal. Un filtre passe-bas de fréquence de coupure 4 Hz est ensuite utilisé afin de nettoyer le signal (les clignements des yeux ont des fréquences généralement comprises entre 0,5 et 2 Hz selon les recommandations de l'AASM). Une détection de pics est ensuite réalisée, et un seuillage sur l'écart-type relatif du signal permet de ne considérer que les pics qui sont suffisamment différents de l'activité cérébrale « de fond ». Combinées à ce seuillage, les informations temporelles sont ensuite utilisées afin de reconnaître les formes semblables à un clignement (détection d'un minima suivi d'un maxima, importance de la proximité du minima et du maxima ainsi que de leur différence d'amplitude, et absence de proximité avec d'autres pics). La Figure C5 présente une exemple de détection des clignements des yeux dans une époque. Le ratio des pics relatifs aux clignements de yeux et des pics initialement détectés est ensuite utilisé pour supprimer d'éventuels faux positifs. La détection de l'éveil yeux ouverts est alors réalisée en seuillant la proportion de clignements dans chaque époque.

La détection de l'éveil agité est réalisée à partir du bloc du même nom. Après une reconstruction des signaux provenant des sangles inductives lorsqu'ils sont saturés, la détection de l'éveil agité se fait grâce à la combinaison de la détection de mouvements ventilatoires brusques (sur les sangles), de l'actimétrie et des ondes alpha détectées à la demi-seconde. L'ordre d'apparition des différents éléments révélateurs d'éveil (mouvements, ondes alpha) est aussi pris en compte

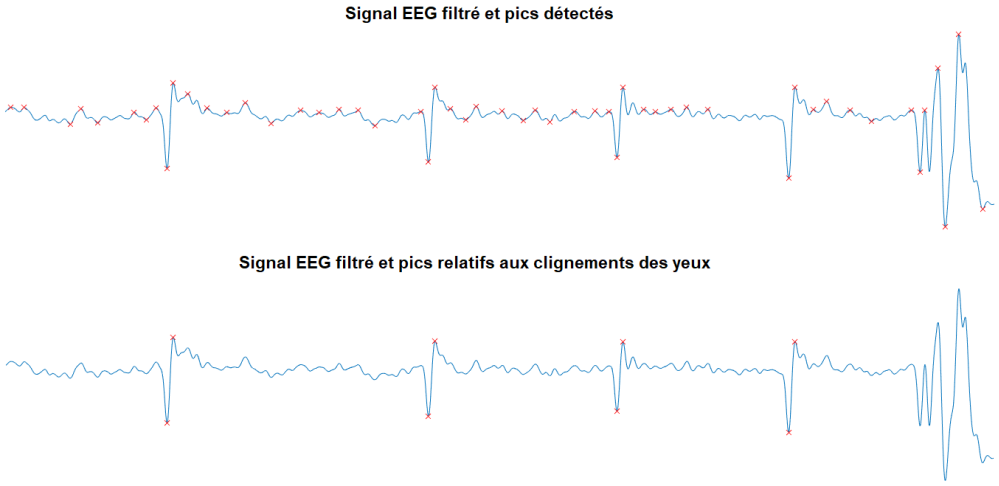


Figure C5 – Exemple de détection des clignements des yeux dans une époque d'éveil yeux ouverts.

afin d'éviter la détection d'époques qui contiendraient un micro-éveil.

La luminosité de la pièce est également enregistrée. Un seuillage nous permet de détecter si la lumière de la pièce est allumée ou non. Il y a en effet de grandes chances que le patient soit éveillé lorsqu'elle l'est. Cependant, cette information n'étant pas directement liée au patient, nous ne détectons de l'éveil que lorsque, en plus de la lumière allumée, nous ne sommes pas dans une zone de suspicion de sommeil (zones précédemment identifées grâce à la palette).

Pour finir, l'ensemble de ces détections sont combinées (« ou logique »), puis un lissage et différentes corrections sont appliquées afin d'obtenir la classification éveil/sommeil finale.

Valorisations supplémentaires

Ce travail a également été présenté lors de congrès :

— [par un autre auteur] Le Congrès du Sommeil : présentation orale de type symposium, Marseille, novembre 2017 (Vanbuis *et al.*, 2018b) ;
— [par un autre auteur] World Sleep Congress : poster, Prague (République tchèque), décembre 2017 (Sabil *et al.*, 2017) ;
— [par un autre auteur] Congrès de Pneumologie de Langue Française : poster, Lyon, janvier 2018 (Sabil *et al.*, 2018) ;
— European Respiratory Society : courte présentation orale et poster, Paris, septembre 2018 (Vanbuis *et al.*, 2018a).
— Dafka Medical Events : présentation orale de type symposium, Tel-Aviv (Israël), octobre 2019 ;

À RETENIR

Un outil de détection d'éveil en PV améliorée par l'ajout d'une voie EEG a été développé. La dérivation EEG choisie permet la détection de certains mouvements oculaires et est facile à poser puisqu'elle se situe sur le haut du front, à la base des cheveux. Différents types d'éveil ont été distingués pour faciliter la détection : l'éveil agité, l'éveil yeux-ouverts et l'éveil calme yeux fermés. Pour ce dernier type d'éveil, il est nécessaire d'évaluer la quantité d'ondes alpha dans chaque époque. La fréquence des ondes alpha relatives à l'éveil est évaluée pour chaque patient et permet l'obtention d'une détection patient-dépendante.
Les résultats suggèrent qu'il est possible d'obtenir un IAH proche de celui de la PSG avec ce système de PV améliorée. Le diagnostic est donc facilité.

C.3.2 Estimation des hypopnées micro-éveillantes en PV améliorée par une voie EEG

Le travail ci-avant présenté permet la détection éveil/sommeil et donc l'estimation du TST (dénominateur du calcul de l'IAH). Cependant, avec le système présenté, les hypopnées micro-éveillantes (H_{AR} dans l'article précédent) ne sont pas identifiées. Cela impacte le numérateur du calcul de l'IAH, qui est donc souvent sous-estimé par rapport à l'IAH de PSG.
Dans la suite, on présentera une méthode d'estimation des hypopnées micro-éveillantes, valable dans le cas du système de PV améliorée par une voie EEG.

Ce travail d'estimation automatique des hypopnées micro-éveillantes nécessite à la fois l'estimation des diminutions de la ventilation (les hypopnées et apnées) et des micro-éveils.

Détection des micro-éveils

La société CIDELEC possède depuis plusieurs années un algorithme de détection des micro-éveils valable en PSG. Cet algorithme a été repris et modifié de manière à pouvoir fonctionner dans notre nouveau système.

En bref, trois indicateurs sont générés pour la détection automatique des micro-éveils :

— l'indicateur d'activité autonomique est estimé à partir de l'actimétrie, du rythme cardiaque (extrait de l'oxymètre de pouls) et du photopléthysmogramme (également extrait de l'oxymètre de pouls) ;

— l'indicateur d'activité ventilatoire est estimé à partir de la pression sus-sternale (extraite du PneaVoX®), du rapport des énergies inspiratoires et expiratoires (également extrait du PneaVoX®), des sangles inductives et du flux nasal ;

— l'indicateur d'activité électrophysiologique est estimé à partir de l'unique voie EEG disponible pour ce type d'enregistrement.

Des signes de variations courtes et soudaines sont recherchés sur ces différents indicateurs, à l'aide de différents seuillages.
Les performances, estimées sur 160 enregistrements polysomnographiques, obtiennent une forte sensibilité (74 %) et une faible VPP (33 %). Il n'est pas possible d'estimer le nombre de FP puisque les micro-éveils sont des événements succincts et de durée variable. De ce fait, la spécificité, la VPN, le taux d'accord et le κ ne peuvent pas être estimés. On utilise donc généralement les indices appelés *precision* et *recall* en anglais, qui correspondent respectivement à la VPP et à la sensibilité.

La détection automatique des micro-éveils n'est pas fiable pour une utilisation seule. Elle a été cependant mise en place pour une estimation des hypopnées micro-éveillantes et c'est pourquoi la sensibilité a été favorisée par rapport à la VPP.

Détection des diminutions de la ventilation

Ce travail a été réalisé par le biais d'un stage encadré durant l'année 2019.
Plusieurs détections ont été effectuées puis combinées. Ces détections ont été effectuées à partir de différents signaux :

À partir des sangles inductives Le débit des sangles inductives (Equation C1) permet une estimation du volume d'air entrant et sortant. Il est donc particulièrement intéressant pour l'identification des diminutions de la ventilation.

$$debitSangle = (sangleAbdominale + sangleThoracique)' \tag{C1}$$

D'après les recommandations AASM, une diminution de la ventilation se caractérise en comparant le niveau d'un signal de débit par rapport au niveau de base. Le niveau de base correspond à l'amplitude moyenne de la respiration stable dans les 2 minutes précédant le début de l'évènement, ou à l'amplitude moyenne des 3 cycles respiratoires les plus amples au cours des 2 minutes précédant le début de l'évènement (chez les sujets n'ayant pas une respiration stable).

L'identification des diminutions de la ventilation à partir du débit des sangles a donc été effectué à l'aide d'un seuil variable, calculé à partir de l'amplitude crête-à-crête lors de la minute précédente (la présence d'artéfacts est à l'origine du choix d'une minute au lieu de deux). Des limitations de durée et d'espacement entre les diminutions ont également été implémentées.

À partir de la lunette nasale Le flux nasal est parfois plus robuste que les sangles inductives (lors des mouvements par exemple), ou parfois moins (notamment lorsque le patient respire par la bouche ou que la lunette nasale est déplacée). Le signal a été intégré et filtré (passe-haut avec fréquence de coupure de 0,2 Hz), permettant l'obtention d'un signal très proche du débit des sangles.
L'algorithme de détection mis en place est ainsi relativement similaire à celui des sangles. Cependant, le signal du flux nasal étant plus fluctuant que celui des sangles, le seuil variable est cette fois-ci estimé à l'aide de l'amplitude moyenne lors des 30 secondes précédentes. Des limitations de durée et d'espacement entre le diminutions ont également été implémentées.

À partir du PneaVoX® Plusieurs signaux ont été testés, mais c'est l'énergie acoustique respiratoire signée (positif pour l'inspiration et négatif pour l'expiration) qui a été retenue [3].
Comme précédemment, un seuil variable a été mis en place et utilisé pour la détection des diminutions. Cette fois-ci, le seuil a été estimé à l'aide des maximas lors des 30 secondes précédentes. Des limitations de durée et d'espacement entre le diminutions ont également été implémentées.

La détection finale a été réalisée en combinant ces trois détections (Equation C2, avec $\wedge \Leftrightarrow$ ET logique et $\vee \Leftrightarrow$ OU logique). Le tableau C1 présente les scores de sensibilité et de VPP obtenus pour chaque détection et pour leur combinaison.

$$diminutions = (dimSangles \wedge dimFluxNasal) \vee dimPneaVoX \tag{C2}$$

Table C1 – Sensibilité et VPP obtenues pour les différentes détections des diminutions de la respiration et leur combinaison.

Détection	Se	VPP
Sangles inductives	62 %	68 %
Flux nasal	76 %	59 %
PneaVoX®	44 %	71 %
Combinaison	69 %	69 %

La détection des diminutions de la ventilation est ici évaluée sans prise en compte des stades de sommeil (il a donc fallu lire les événements à l'éveil pour éviter tout biais de ce côté). Les apnées sont plus faciles à détecter, mais ne nous intéressent pas puisqu'elles sont identifiées par le spécialiste du sommeil en PV. Dans la suite, l'estimation des micro-éveils et des diminutions de la ventilation sont combinées afin d'estimer les hypopnées micro-éveillantes.

Détection des hypopnées micro-éveillantes

On s'intéresse ici uniquement aux événements qui ne peuvent pas être identifiés en PV, c'est-à-dire aux hypopnées associées à un micro-éveil mais non suivies d'une désaturation.
L'association des diminutions de la ventilation et des micro-éveils a été réalisée en s'assurant qu'une diminution ne peut être associée qu'à un seul micro-éveil, et inversement, et que le micro-éveil intervient moins de 15 secondes après la fin de la diminution de la ventilation.

Les résultats sont faibles, avec une sensibilité de 37 % et une VPP de 38 %.
L'observation des hypopnées micro-éveillantes identifiées montre tout de même que de nombreux FP sont situés à proximité d'hypopnées micro-éveillantes non reconnues. Dans la majorité des cas, les FP ne sont pas aberrants. Les FN sont généralement liés à un manque de sensibilité de l'une ou de l'autre détection, ou à un délai trop long entre l'hypopnée et le micro-éveil identifié. Il semblerait également que les zones comportant un grand nombre d'événements ventilatoires obtiennent plus d'erreurs (FP ou FN). Cela peut être dû à l'instabilité des signaux, qui peut être

3. Un descriptif des différents signaux enregistrés par le PneaVoX® est disponible Section E.3.1.

à l'origine de problèmes de seuillage.

La détection des hypopnées micro-éveillantes développée obtient des résultats mitigés. Il est nécessaire d'évaluer son impact sur le diagnostic du patient.

Impact sur le diagnostic du SAHS

Les IAHs résultants sont comparés avec ceux obtenus en PSG.
La Figure C6 présente les graphes de Bland-Altman et les courbes de corrélation résultantes.

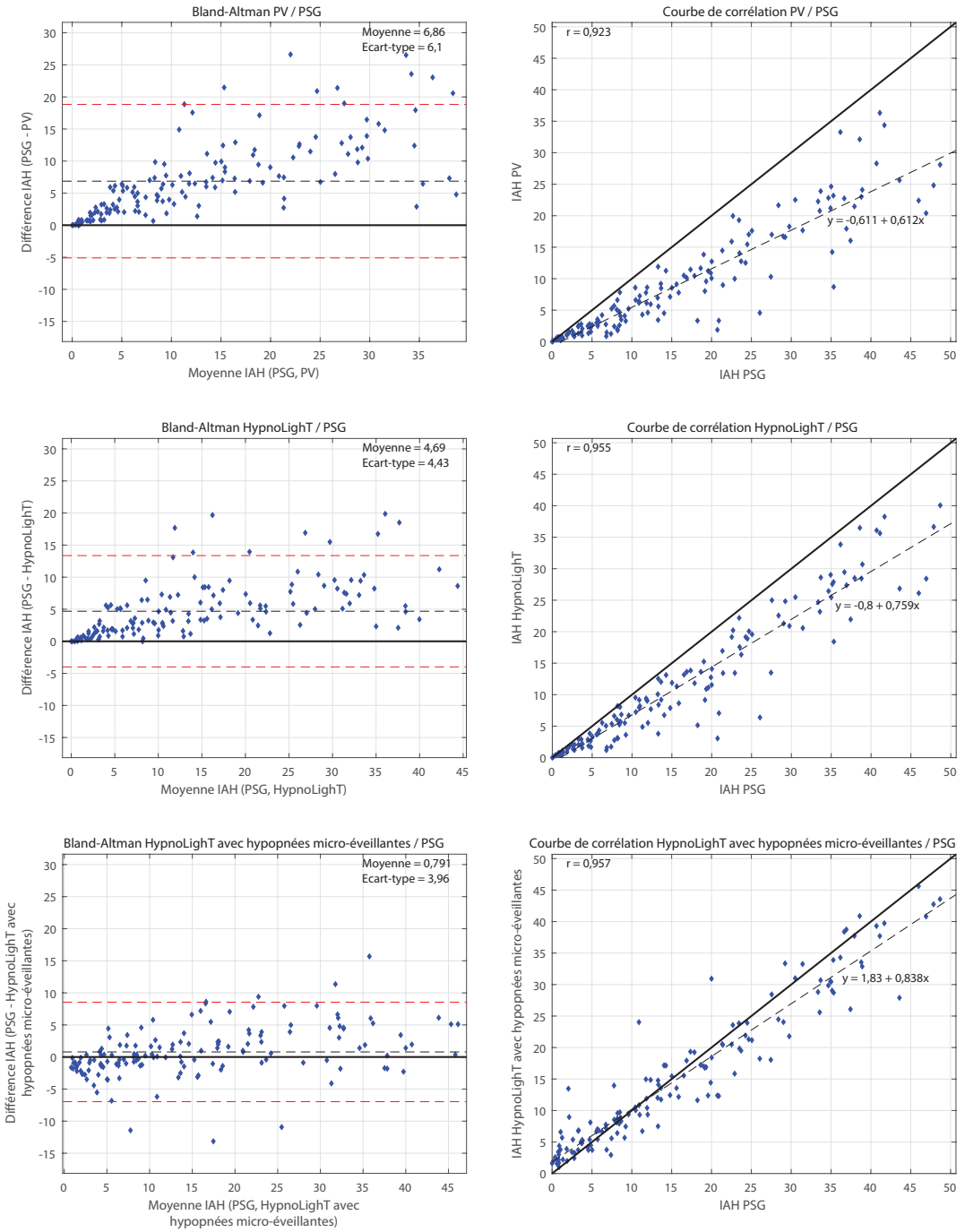


Figure C6 – Graphes de Bland-Altman et courbes de corrélation dans le cas de la comparaison de la PSG avec la PV, avec HypnoLighT et avec HypnoLighT avec détection des hypopnées micro-éveillantes.

Sur ces graphiques, on peut voir que les IAHs résultants de l'analyse HypnoLighT et de la détec-

tion des hypopnées micro-éveillantes est plus proche des IAHs de référence (de PSG), que ceux résultants de l'analyse HypnoLighT seule ou encore que ceux résultant de la PV. Cependant, au contraire de la PV ou de l'analyse HypnoLighT seule, les IAHs résultants de l'analyse HypnoLighT et de la détection des hypopnées micro-éveillantes sont parfois sur-estimés par rapport à la référence.

Pour aller plus loin, on évalue la sévérité du SAHS qui découle de ces différentes méthodes [4]. La Table C2 reporte le nombre de patients, selon la sévérité de SAHS qui aurait été obtenue en PV, comparée à celle obtenue en réalité lors de la PSG. Sur les 160 patients, 73 patients auraient été sous-diagnostiqués (23+4+22+2+22). En particulier, 26 auraient été diagnostiqués comme ayant un SAHS absent ou léger au lieu de modéré, et 24 comme ayant un SAHS léger ou modéré au lieu de sévère.
L'*Acc* et le κ associés à l'estimation de la sévérité du SAHS en PV sont de 54 % et 0,39, respectivement.

Table C2 – Matrice de confusion associée à la sévérité du SAHS, dans le cas de la comparaison PV/PSG.

		PV				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
PSG	Pas de SAHS	37	0	0	0	37
	SAHS léger	23	22	0	0	45
	SAHS modéré	4	22	11	0	37
	SAHS sévère	0	2	22	17	41
	Total	64	46	33	17	160

La Table C3 reporte le nombre de patients, selon la sévérité de SAHS qui aurait été obtenue avec HypnoLighT, comparée à celle obtenue en réalité lors de la PSG. Sur les 160 patients, 54 patients auraient été sous-diagnostiqués (14+1+20+19). En particulier, 21 auraient été diagnostiqués comme ayant un SAHS absent ou léger au lieu de modéré, et 19 comme ayant un SAHS modéré au lieu de sévère.
L'*Acc* et le κ associés à l'estimation de la sévérité du SAHS en PV avec l'utilisation d'HypnoLighT sont de 66 % et 0,55, respectivement.
On estime donc que 19 % ($\frac{26-21}{26}$) des patients qui auraient été sous-diagnostiqués en PV comme ayant un SAHS absent ou léger au lieu de modéré auraient été correctement diagnostiqués grâce à HypnoLighT. De la même manière, 21 % ($\frac{24-19}{24}$) des patients qui auraient été sous-diagnostiqués en PV comme ayant un SAHS absent, léger ou modéré au lieu de sévère auraient été correctement diagnostiqués grâce à HypnoLighT.

Table C3 – Matrice de confusion associée à la sévérité du SAHS, dans le cas de la comparaison HypnoLighT/PSG.

		Analyse HypnoLighT				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
PSG	Pas de SAHS	37	0	0	0	37
	SAHS léger	14	31	0	0	45
	SAHS modéré	1	20	16	0	37
	SAHS sévère	0	0	19	22	41
	Total	52	51	35	22	160

La Table C4 reporte le nombre de patients, selon la sévérité de SAHS qui aurait été obtenue avec HypnoLighT plus l'estimation des hypopnées micro-éveillantes, comparée à celle obtenue en réalité lors de la PSG. Sur les 160 patients, 17 patients auraient été sous-diagnostiqués (3+7+7). En particulier, 7 auraient été diagnostiqués comme ayant un SAHS léger au lieu de modéré, et 7 comme ayant un SAHS modéré au lieu de sévère. 16 patients auraient également été sur-diagnostiqués (11+3+2), dont 3 comme ayant un SAHS modéré au lieu de léger, et 2 comme

4. Les résultats présentés dans cette section ne peuvent être comparés avec ceux introduits dans l'article précédent. En effet, certains biais ont été corrigés après l'écriture de l'article. Par exemple, les événements respiratoires à l'éveil ont été lus, afin que la PV reconstruite à partir de la PSG soit la plus réaliste possible.

ayant un SAHS sévère au lieu de modéré.
L'*Acc* et le κ associés à l'estimation de la sévérité du SAHS en PV avec l'utilisation d'HypnoLighT et l'estimation des hypopnées micro-éveillantes sont de 79 % et 0,72, respectivement.
On estime donc que 73 % ($\frac{26-7}{26}$) des patients qui auraient été sous-diagnostiqués en PV comme ayant un SAHS absent ou léger au lieu de modéré auraient été correctement diagnostiqués grâce à HypnoLighT et à l'estimation des hypopnées micro-éveillantes. De la même manière, 71 % ($\frac{24-7}{24}$) des patients qui auraient été sous-diagnostiqués en PV comme ayant un SAHS absent, léger ou modéré au lieu de sévère auraient été correctement diagnostiqués grâce à HypnoLighT et à l'estimation des hypopnées micro-éveillantes.

Table C4 – Matrice de confusion associée à la sévérité du SAHS, dans le cas de la comparaison HypnoLighT avec hypopnées micro-éveillantes/PSG.

		Analyse HypnoLighT + hypopnées micro-éveillantes				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
PSG	Pas de SAHS	26	11	0	0	37
	SAHS léger	3	39	3	0	45
	SAHS modéré	0	7	28	2	37
	SAHS sévère	0	0	7	34	41
	Total	29	57	38	36	160

Ajoutée à l'analyse HypnoLighT, la détection automatique des hypopnées micro-éveillantes ci-avant présentée permet l'obtention d'un diagnostic de SAHS plus précis, se rapprochant fortement de celui obtenu lors d'un examen de PSG.
Cela suggère qu'avec l'ajout d'une unique voie EEG et l'utilisation des algorithmes développés, il est possible de diminuer grandement le nombre d'examens polysomnographiques de seconde intention.

À RETENIR

Un algorithme de détection d'hypopnées micro-éveillantes en PV améliorée par l'ajout d'une voie EEG a été développé. Cette détection repose sur l'association des estimations automatiques des diminutions de la ventilation et des micro-éveils. L'estimation des micro-éveils repose sur le seuillage de trois indicateurs : un indicateur d'activité autonomique (signaux de l'actimètre et de l'oxymètre de pouls), un indicateur d'activité ventilatoire (signaux du PneaVoX®, des sangles inductives et de la lunette nasale) et un indicateur d'activité électrophysiologique (voie EEG unique). L'estimation des diminutions de la ventilation repose sur les signaux des sangles inductives, de la lunette nasale et du PneaVoX®. Malgré de faibles scores pour la détection des hypopnées micro-éveillantes, les erreurs réalisées ne sont pas aberrantes. De plus, l'IAH obtenu se rapproche de celui de PSG. L'estimation de la sévérité du SAHS est grandement améliorée, mais quelques enregistrements sont sur-estimés.
Les résultats montrent que les estimations de l'IAH et de la sévérité du SAHS sont améliorées, mais que quelques patients sont sur-diagnostiqués. Il est nécessaire d'étudier précisément ces enregistrements afin d'évaluer quel pourrait être l'impact sur la santé du patient.

C.4 Conclusion du chapitre

Dans ce chapitre, un nouveau système a été présenté. Constitué d'une PV améliorée par une unique voie EEG, ce système est plus pratique à mettre en place et à étudier qu'une PSG. La voie EEG est positionnée à la limite des cheveux et est facile à mettre en place.

Ce système permet l'identification automatique et patient-dépendante de l'éveil, en faisant la distinction entre l'éveil agité, l'éveil yeux-ouverts et l'éveil calme yeux fermés (κ moyen de 0,74). Grâce à cette identification, le TST peut être évalué.
Afin d'améliorer encore l'estimation de l'IAH pour ce système, les hypopnées micro-éveillantes (uniquement disponibles en PSG) ont été estimées automatiquement. Grâce à cela, le système de PV améliorée par une unique voie EEG obtient des IAHs plus proches de ceux obtenus en PSG (coefficient de corrélation r passant de 0,923 entre la PV et la PSG à 0,957 entre ce nouveau système et la PSG).
La sévérité du SAHS, évaluée à partir des IAHs, est mieux estimée également (κ de 0,39 en PV contre 0,72 avec le nouveau système). Quelques enregistrements sont cependant sur-diagnostiqués. Il est nécessaire d'évaluer la gravité de l'erreur qui en découle.

Que ce soit avec ou sans l'estimation automatique des hypopnées micro-éveillantes, ce nouveau système permet, grâce à l'identification automatique de l'éveil, d'obtenir un diagnostic de SAHS plus précis qu'en PV et se rapprochant de la PSG.

Par le biais de ce premier travail, la mise en place d'algorithmes pour l'étude de différents signaux (électrophysiologiques mais aussi cardio-respiratoires) a pu être effectuée. La détection de l'un des stades de sommeil les plus importants pour le diagnostic du SAHS, à savoir l'éveil, a été réalisée en s'inspirant des recommandations AASM et des pratiques des spécialistes. Ce travail est en cours d'industrialisation. La seconde partie de ce travail, à savoir l'estimation des hypopnées micro-éveillantes, sera potentiellement améliorée en vue d'être industrialisée également, après ce travail de thèse.

Dans la suite de ce manuscrit, nous irons plus loin et détecterons l'ensemble des stades de sommeil.

Bibliographie

Berthomier, C., Drouot, X., Herman-Stoïca, M., Berthomier, P., Prado, J., Bokar-Thire, D., Benoit, O., Mattout, J. et D'Ortho, M.-P. (2007). Automatic Analysis of Single-Channel Sleep EEG : Validation in Healthy Individuals. *Sleep*, 30(11):1587–1595.

Bresch, E., Grossekathöfer, U. et Garcia-Molina, G. (2018). Recurrent Deep Neural Networks for Real-Time Sleep Stage Classification From Single Channel EEG. *Frontiers in Computational Neuroscience*, 12:85.

Dong, H., Supratak, A., Pan, W., Wu, C., Matthews, P. M. et Guo, Y. (2018). Mixed Neural Network Approach for Temporal Sleep Stage Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):324–333. arXiv : 1610.06421.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H. et Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, 108(1):10–19.

Hsu, Y.-L., Yang, Y.-T., Wang, J.-S. et Hsu, C.-Y. (2013). Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing*, 104:105–114.

Koley, B. et Dey, D. (2012). An ensemble system for automatic sleep stage classification using single channel EEG signal. *Computers in Biology and Medicine*, 42(12):1186–1195.

Popovic, D., Khoo, M. et Westbrook, P. (2014). Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead : validation in healthy adults. *Journal of Sleep Research*, 23(2):211–221.

Sabil, A., Vanbuis, J., Baffet, G., Feuilloy, M., Le Vaillant, M., Meslier, N. et Gagnadoux, F. (2019). Automatic identification of sleep and wakefulness using single-channel EEG and respiratory polygraphy signals for the diagnosis of obstructive sleep apnea. *Journal of Sleep Research*, 28(2).

Sabil, A., Vanbuis, J., Baffet, G., Feuilloy, M., Meslier, N. et Gagnadoux, F. (2017). Automatic estimation of sleep and wakefulness using a single-channel EEG and home polygraphy signals. *Sleep Medicine*, 40:e287.

Sabil, A., Vanbuis, J., Baffet, G., Feuilloy, M., Meslier, N. et Gagnadoux, F. (2018). Détection automatique de l'éveil et du sommeil basée sur l'association d'un signal EEG et de signaux de polygraphie ventilatoire. *Revue des Maladies Respiratoires*, 35:A257.

Sors, A., Bonnet, S., Mirek, S., Vercueil, L. et Payen, J.-F. (2018). A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42:107–114.

Vanbuis, J., Sabil, A., Baffet, G., Feuilloy, M., Meslier, N. et Gagnadoux, F. (2018a). Estimation of AHI using a single-channel EEG and home polygraphy signals. *European Respiratory Journal*, 52(suppl 62). Publisher : European Respiratory Society Section : Sleep and control of breathing.

Vanbuis, J., Sabil, A., Baffet, G., Feuilloy, M., Meslier, N. et Gagnadoux, F. (2018b). Utilisation d'une voie EEG couplée aux signaux de polygraphie ventilatoire pour la détection automatique de l'éveil et du sommeil. *Médecine du Sommeil*, 15(1):17.

Virkkala, J., Velin, R., Himanen, S.-L., Varri, A., Muller, K. et Hasan, J. (2008). Automatic sleep stage classification using two facial electrodes. *In Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 1643–1646. IEEE.

Chapitre D

Examen polysomnographique

D.1 Objectif et contexte

Dans le cas d'un examen polysomnographique, l'intérêt de l'analyse automatique du sommeil consiste principalement à un gain de temps pour le spécialiste du sommeil. En effet, l'ensemble des signaux nécessaires à l'étude du sommeil est disponible pour ce type d'enregistrements. L'objectif est donc de créer un ou plusieurs outils permettant de soutenir le médecin lors de la lecture des stades de sommeil, c'est de l'aide à la décision.

Dans le domaine de la santé en général, de nombreuses avancées reposant sur des algorithmes de ML ou DL ont fait leurs preuves. Les études de Topol (2019) et Fogel et Kvedar (2018), qui présentent de nombreux exemples d'applications médicales utilisant de l'intelligence artificielle, insistent sur les impacts que peuvent avoir ces nouvelles avancées sur la relation patient-médecin. En effet, l'utilisation d'intelligence artificielle, en réduisant le temps alloué par le médecin à l'étude des signaux ou des images, peut lui permettre de passer plus de temps auprès du patient tout en rendant le rendez-vous médical plus humain.
Il est également important de prendre en compte les recommandations des experts concernant l'utilisation clinique des outils développés. Dans le domaine de l'analyse automatique du sommeil, on recommande au médecin de ne pas utiliser les outils à l'aveugle, et d'effectuer une relecture après l'analyse automatique (Penzel et Conradt, 2000). Plusieurs raisons motivent cette méfiance vis-à-vis des analyses automatiques. Elles sont détaillées dans la section ci-dessous.

D.2 Problématiques identifiées

a) La première problématique identifiée qui motive cette méfiance dans l'analyse automatique est l'opacité de l'algorithme. Elle est particulièrement vraie pour les analyses utilisant du DL. Une étude récente de Fiorillo *et al.* (2019) a montré que le comportement en boîte noire était l'une des principales causes limitant l'utilisation clinique des analyses automatiques du sommeil. Cependant, l'utilisation croissante de DL et de ML a également entraîné un regain d'intérêt pour les systèmes optimisés non seulement pour l'exécution des tâches prévues, mais également pour leur interprétation (Doshi-Velez et Kim, 2017).
Dans le domaine de l'analyse automatique du sommeil, une approche hybride a ainsi été présentée par Chen (2016); Chen *et al.* (2019); Ugon (2015). Ces études sont particulièrement intéressantes car elles reposent sur la fusion de données (Castanedo, 2013). La fusion de données consiste à traiter les informations provenant de plusieurs capteurs de manière à améliorer les performances d'un problème. C'est un domaine multidisciplinaire qui regroupe différents concepts allant du traitement de l'information à l'intelligence artificielle. Plusieurs architectures ont ainsi été proposées pour le décrire. Celles-ci peuvent être basées :

— sur les relations entre les données d'entrée, telle-que l'architecture de Durrant-Whyte (Durrant-Whyte, 1990) ;

— sur les relations entre les données d'entrée et de sortie, telle-que l'architecture de Dasarathy (Dasarathy, 1997 - voir détail ci-après) ;

— sur les niveaux d'abstraction des données (Luo *et al.*, 2002) ;

— ou sur les niveaux tels que définis par les directeurs de laboratoires conjoints (architecture *Joint Directors of Laboratories* ou JDL, White, 1991).

Fusion de données - architecture de Dasarathy :

L'architecture de Dasarathy est l'une des plus connues. Elle est composée de 5 catégories comme illustré Figure D1.
Le premier niveau, DAI-DAO, est le plus élémentaire puisque ses entrées et sorties sont des signaux. Ce niveau de fusion de données est réalisé immédiatement après le chargement des signaux bruts tels qu'enregistrés par les différents capteurs. On retrouve ici le traitement du signal et de l'image.
Le deuxième niveau, DAI-FEO, correspond à l'évaluation de caractéristiques.
Le troisième niveau, FEI-FEO, consiste à l'évaluation de nouvelles caractéristiques à partir de celles déjà extraites. On retrouve donc ici la réduction de la dimensionnalité. Ce niveau est également connu sous le nom de fusion de caractéristiques ou fusion symbolique.
Le quatrième niveau, FEI-DEO, permet d'obtenir un ensemble de décisions à partir de caractéristiques en entrée. On retrouve ici tous les systèmes de classification fonctionnant à partir de caractéristiques.
Le cinquième niveau, DEI-DEO, est également connu sous le nom de fusion de décision. À partir des décisions d'entrées, de nouvelles décisions (souvent meilleures) sont ainsi estimées. On retrouvera donc ici toutes les sortes de corrections utilisant en entrée des décisions déjà obtenues.

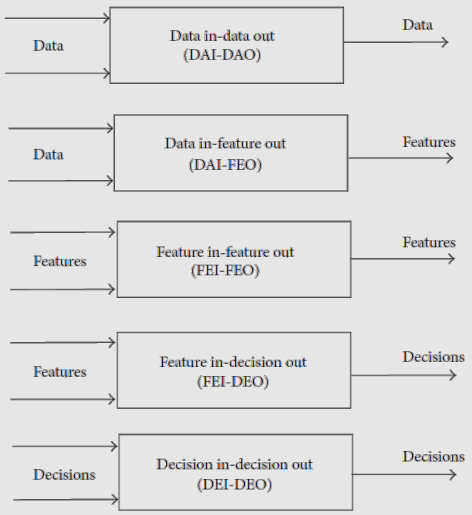


Figure D1 – Architecture de Dasarathy (figure provenant de l'étude de Castanedo, 2013).

Cette architecture propose ainsi, de manière structurée, une classification des différentes étapes formant la structure algorithmique d'une approche de ML.

En reprenant les terminologies proposées par Dasarathy, les travaux de Ugon et Chen portent principalement sur une approche FEI-FEO (fusion symbolique). En effet, les caractéristiques d'entrée correspondent à des caractéristiques quantitatives évaluées à partir des signaux électrophysiologiques (par exemple l'énergie des ondes lentes à partir des EEGs). Les caractéristiques de sortie sont qualitatives et correspondent aux différents niveaux que peuvent prendre ces caractéristiques d'entrée (par exemple pour l'énergie des ondes lentes : haute, moyenne ou basse). La classification des stades de sommeil est ensuite réalisée à l'aide de règles logiques, à partir de ces caractéristiques de sortie. L'objectif visé est ainsi de réaliser une classification automatique par le biais d'une traduction des recommandations AASM.

b) La deuxième problématique mise en avant dans l'article de revue de Fiorillo *et al.* (2019) est la disparité des jeux des données utilisés pour l'apprentissage et le test des méthodes implémentées. De nombreuses études sont en effet développées à partir d'enregistrements provenant de patients sains. Une fois utilisées sur des patients atteints de troubles du sommeil, les méthodes développées obtiennent des résultats bien plus limités.
L'étude de Roy *et al.* (2019) a comparé un nombre conséquent d'études portées sur l'analyse de l'EEG par le biais du DL. Ces études n'étaient pas consacrées uniquement au sommeil. Il a été rapporté que parmi 154 études, la moitié utilisent moins de 13 patients, et seules 6 études dépassent les 250 patients (dont deux avec plus de 10000 patients).
L'utilisation d'un jeu de données le plus grand possible et comportant des enregistrements provenant de patients sains mais aussi touchés par des troubles du sommeil est donc recommandé.

c) Pour finir, la troisième problématique pointée du doigt par Fiorillo *et al.* est l'ergonomie des approches développées. Ces approches doivent être faciles à utiliser par l'expert du sommeil.

Nous avons donc choisi de développer une approche ne nécessitant pas de travail préalable par le spécialiste. Ainsi, on évitera d'utiliser une lecture partielle des stades de sommeil, d'utiliser des éléments détectés par le spécialiste (les événements ventilatoires, les micro-éveils ou encore les artéfacts), et on ne limitera pas la lecture automatique à la période entre le coucher et lever du patient (qui ne sont pas systématiquement annotées).

D.3 Stratégie

À la lumière des trois précédents points, la solution que nous préconisons est une approche hybride, combinant connaissances des experts et méthodes de ML.

Pour répondre au point a) de la Section D.2, la méthode a été pensée pour être facilement compréhensible par la communauté médicale. Nous avons ainsi implémenté un algorithme qui reproduit, étape par étape, le processus de lecture du sommeil tel que réalisé par les spécialistes du sommeil. On distingue plusieurs étapes :

1. adaptation à chaque enregistrement (généralement, le spécialiste visualise très rapidement l'ensemble de l'enregistrement afin de se familiariser avec les signaux du patient et ses spécificités). Dans notre approche, ceci se traduit par la mise en place d'un algorithme de seuillage auto-adaptatif nommé SATUD ;
2. étude des composantes temporelles et spectrales des époques (telles que décrites dans les recommandations AASM). Dans notre approche, ceci se traduit par l'évaluation de caractéristiques très proches des composantes définies dans les recommandations AASM ;
3. étude des grapho-éléments liés aux différents stades de sommeil. Dans notre approche, ceci se traduit par la mise en place de différents outils de traitement du signal pour la reconnaissance des grapho-éléments ;
4. classification des stades de sommeil en prenant en compte l'ensemble de ces éléments, mais également le contenu des époques à proximité et les règles de transition. Dans notre approche, ceci se traduit par l'utilisation en cascade de plusieurs classifieurs, par l'emploi d'un MMC de Viterbi et par l'implémentation des règles de transition.

Pour répondre au point b) de la Section D.2, la méthode a été entraînée et testée sur un grand nombre d'enregistrements polysomnographiques (400 enregistrements) provenant de patients sains ou avec troubles du sommeil. Le jeu de données est donc représentatif des enregistrements sur lesquels la méthode devrait être utilisée par la suite. L'apprentissage étant supervisé, les enregistrements ont préalablement été annotés par des spécialistes du sommeil du Laboratoire du Sommeil du CHU d'Angers, en suivant les recommandations AASM. Notez qu'une fois notre approche entraînée et intégrée dans le système CIDELEC, les spécialistes du sommeil n'auront **en théorie** plus à annoter les enregistrements.

Pour répondre au point c) de la Section D.2, nous nous sommes fixés l'objectif de fournir aux médecins, en plus de l'hypnogramme automatique, d'autres outils qui permettent notamment d'identifier les époques qui nécessitent une relecture en priorité.

Dans la suite de ce chapitre, l'approche mise en place est décrite en détail. L'algorithme SATUD, qui constitue un élément clé de cette méthode, est également présenté et sa robustesse au bruit et aux artéfacts est mise à l'épreuve. L'impact possible de ce travail sur le diagnostic des troubles du sommeil et plus particulièrement du SAHS est ensuite évalué.

D.3.1 Approche pour la classification des stades de sommeil

Ce travail a été présenté en détail par le biais d'un article en deux parties, publié dans le journal Informatics in Medicine Unlocked (références Vanbuis *et al.* (2020b) et Vanbuis *et al.* (2020c)).
La classification en elle-même constitue la première partie de cet article. Une description plus précise des classifieurs utilisés est disponible en Annexe : fonctionnement de la forêt d'arbres décisionnels (page 114) et en Annexe : fonctionnement du modèle de Markov à états cachés de Viterbi (page 116).

Résumé traduit

La lecture manuelle du sommeil est une tâche chronophage nécessitant un haut niveau d'expertise médicale. C'est pourquoi de nombreux algorithmes d'analyse automatique du sommeil ont été récemment développés. Toutefois, leur utilisation clinique reste limitée pour diverses raisons : les approches proposées sont souvent opaques et difficilement interprétables, l'hétérogénéité des patients utilisés pour valider la méthode est rarement suffisante et les approches développées manquent fréquemment de praticabilité.
Cet article présente un système de lecture automatique du sommeil fonctionnant à partir des signaux électrophysiologiques et permettant de surmonter ces limitations. Simple d'utilisation, le système proposé a été entraîné et testé sur un nombre conséquent d'enregistrements (300 et 100 respectivement) provenant de patients avec diverses pathologies du sommeil. La méthode imite le processus manuel d'évaluation du sommeil, tout en suivant les recommandations de l'American Academy of Sleep Medicine (AASM). En plus d'une classification des stades de sommeil patient-dépendante (à l'aide du système SATUD), cet outil d'aide au diagnostic génère un tableau fournissant des indications sur le niveau de confiance de l'algorithme lors de la classification. Contrairement aux récentes approches d'apprentissage profond, les algorithmes implémentés ont été choisis pour leur robustesse et leur compréhensibilité, et les connaissances médicales ont été incluses dans le processus autant que possible.
Les résultats ont montré que le système atteint un accord élevé avec la lecture manuelle (Kappa de Cohen et taux d'accord moyens de 0,69 et 77,8 %, respectivement). Cela démontre qu'une interprétation facilitée du modèle, très importante dans des domaines tels que le diagnostic du sommeil, peut être fournie lors de la mise en oeuvre d'outils automatiques.
Ce nouveau système offre ainsi des outils d'aide au diagnostic du sommeil qui devraient permettre au spécialiste de réduire significativement le temps alloué à la lecture du sommeil.

Aide lexicale

— *Obstructive Sleep Apnea (OSA) syndrome* est le terme utilisé en anglais pour SAHS [a]
— *REM sleep* ou *R stage* : Sommeil Paradoxal (SP)
— *sleep patterns* : grapho-éléments
— *sleep spindles* : fuseaux du sommeil
— *Rapid Eye Movements (REMs)* : Mouvements Oculaires Rapides (MORs)
— *arousal* : micro-éveil

a. Les évènements ventilatoires peuvent être d'origine obstructive (dus à l'obstruction des voies aériennes supérieures), centrale (d'origine neurologique) ou mixte (dont l'origine est d'abord centrale puis devient obstructive). En France, on parle du Syndrome d'Apnées Hypopnées du Sommeil en général, mais il est également possible de distinguer le Syndrome d'Apnées-Hypopnées Obstructives du Sommeil (SAHOS), le Syndrome d'Apnées Obstructives du Sommeil (SAOS) ou le Syndrome d'Apnées Centrales du Sommeil (SACS). Dans les publications internationales, et malgré l'existence de termes équivalents, c'est généralement le terme « Obstructive Sleep Apnea (OSA) syndrome », qui est utilisé (même si l'origine obstructive n'a pas d'importance pour l'étude en question).

Towards a user-friendly sleep staging system for polysomnography Part I: automatic classification based on medical knowledge

Jade Vanbuis[a,b,*], Mathieu Feuilloy[a,b], Guillaume Baffet, Nicole Meslier[c,d], Frédéric Gagnadoux[c,d] and Jean-Marc Girault[a,b]

[a]ESEO, Angers, France
[b]LAUM, UMR CNRS 6613, Le Mans, France
[c]Angers sleep laboratory, University Hospital, Angers, France
[d]INSERM UMR 1063, University of Angers, Angers, France

ARTICLE INFO

Keywords:
Automatic sleep staging for polysomnography
Decision support system
User-friendly and interpretable sleep scoring
Patient-dependent sleep scoring using the SATUD system
Respect of AASM Guidelines

ABSTRACT

Manual sleep scoring is a time-consuming task that requires a high level of medical expertise. For this reason, a number of automatic sleep scoring algorithms have recently been implemented. However, their use by physicians remains limited for various reasons: a lack of transparency of the approach used, insufficient heterogeneity among the patients used for testing, or a lack of practicality.
This paper presents a system for facilitated sleep scoring that will overcome these limitations. The proposed system, a user-friendly tool based on electrophysiological channels, was trained and tested on large datasets of 300 and 100 distinct recordings from patients with various sleep disorders. The method replicates the manual sleep scoring process, in accordance with the American Academy of Sleep Medicine (AASM) guidelines and generates patient-dependent sleep scoring (using the SATUD system). For an improved level of precision and confidence with regard to scoring, our approach also provides a table that gives indications about the confidence level of the algorithm when scoring sleep. In contrast to recent deep learning approaches, the algorithms used were chosen for their resilience and as they are easy to understand. Medical knowledge was included in the process as much as possible. Results showed that the system is consistent with manual scoring (mean Cohen's Kappa of 0.69 and accuracy rate of 77.8 %). It proves that a facilitated interpretation of the model, crucial in such fields as sleep diagnosis, can be provided when using automatic tools.
This new system thereby generates sleep scoring decision support tools, which should easily contribute to significant time-saving and help sleep specialists to perform sleep diagnosis.

1. Introduction

Sleep-disordered breathing (SDB) is a common health issue affecting approximatively a third of the population [1–3]. Symptoms often go unnoticed, since they are not specific to SDB and are quite common [4]. However, bad sleep quality can affect several vital functions, such as learning, memorization and adaptation, resulting in a deterioration of the quality of life.
Over the past few decades, there has been an increasing need for sleep diagnosis [5, 6]. The gold-standard procedure for SDB diagnosis, called polysomnography (PSG), involves the recording of electrophysiological (EP) and cardio-respiratory (CR) signals throughout an entire night [7]. Once recorded, signals are manually studied by a sleep specialist: respiratory events are identified using CR channels, and sleep is scored using EP channels. A diagnosis is reached by cross-checking this information with patient symptoms [7].
Sleep scoring is a time-consuming and complex task. It involves the assessment of each 30-second section's (called an epoch) degree of vigilance [7]. To do so, EP channels such as electroencephalograms (EEG), electrooculograms (EOG) and electromyograms (EMG) are visualized epoch by epoch. Each epoch is identified as belonging to the W stage (wakefulness), the N1 stage (light sleep), the N2 stage (also light sleep), the N3 stage (deep sleep) or the R stage (rapid eye movement sleep). The resulting succession of sleep stages is called a hypnogram. The American Academy of Sleep Medicine (AASM) manual for the scoring of sleep and associated events [7] describes each sleep stage's properties in detail, along with possible transitions between sleep stages. Despite the AASM guidelines, sleep staging remains time-consuming and complex, and the inter-scorer agreement rate hardly exceeds 80-90 % [8].

Lately, artificial intelligence and more specifically learning algorithms have proven their ability to solve complex problems in many healthcare sectors [9, 10]. New algorithms for automatic events or sleep analysis have emerged and been recognized by experts in the field for potentially helping to improve our understanding of sleep [11] and simplifying the scoring procedure [12].
A number of systems for automatic sleep scoring using EP signals have been developed. In such systems, algorithms are trained so they can classify each epoch into a sleep stage, using manual scoring as the reference. The algorithms developed can be broken down into three categories: deep learning [13], machine learning [14] and hybrid approaches [15]. Deep learning is generally applied directly on raw signals. In [16, 17], a combination of a Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN) is used to estimate relevant features and classify sleep, taking tempo-

*Corresponding author
jade.vanbuis@eseo.fr
ORCID(s): 0000-0001-6437-1597 (J. Vanbuis)

rality into consideration.
On the contrary, machine learning usually requires the extraction of descriptors, called features, before classification. Various machine learning classifiers were tested. Among the most widespread, we will mention Support Vector Machines (SVM) in [18–20], Multi-Layer Perceptrons (MLP) in [21, 22] and Random Forests (RF) in [23, 24].
Lastly, hybrid approaches combine deep learning or machine learning with expert knowledge. In [25, 26], symbolic fusion was applied so the features used for classification are qualitative, particularly close to the AASM guidelines.

However, automatic sleep staging faces several challenges. Firstly, the channels used for classification vary greatly from one patient to another. This high variability between subjects is caused by many factors including the subject's age, condition, drug intake, but also the positioning of sensors, the quality of signals (which can be altered by movements or sweating, for example) and the presence of accessories. It particularly complicates the learning process for machine learning approaches, where features are highly impacted by subject specific characteristics. This can be overcome if trained on a large number of subjects with various disorders. In [27], a review of 154 deep learning-based EEG analysis studies showed that half of them included less than 13 subjects, which is reported as being insufficient to illustrate human heterogeneity.
Another challenge regarding machine learning or deep learning approaches is acceptance by the medical community. A recent review paper by Fiorillo et al. [12] presented the barriers for the clinical use of automated scoring on a daily basis. The main limitation was the black box behavior of deep learning algorithms. Nowadays, a certain number of researchers attempt to improve the interpretability of their models [28], using, for example, hybrid approaches incorporating medical knowledge.
The final criteria is for the model to be easy for physicians to use. There is often a preliminary human action needed (for example, identification of artifacted epochs or partial scoring), meaning that methods cannot be applied immediately once the channels have been recorded.
In summary, the clinical use of automatic sleep scoring remains controversial because of three limitations:

a) lack of confidence in the developed approach (algorithms are often considered as a black box);
b) insufficient heterogeneity of the dataset, nonetheless necessary for assessing real-life performances;
c) lack of practicality of the developed approach, which sometimes requires human intervention before use.

The main aim of this study is to implement a user-friendly automatic sleep scoring system to overcome the three previous limitations. Unlike most of the recent studies, we chose to prioritize the understanding of the algorithm's operating mode, addressing issue *a)*. To provide an answer to limitation *b)*, the system was tested on a large dataset of patients with varying levels of SDB severity. Issue *c)* was also taken into consideration, since the system was designed to be used without any preliminary human action (for example partial sleep scoring or invalidation of epochs) and provides a probability table to further assist in scoring.
The work proposed here has been designed on EP channels obtained from polysomnographic recordings. Designed to assist physicians in their diagnosis, the developed system combines artificial intelligence and expert knowledge. Medical practitioners' concerns were considered and the result is a user-friendly tool providing scoring support to avoid sleep scorers spending excessive time on sleep staging.

The present article is the first of a two-part paper. In the following section (Section 2), the recordings used for training and testing are presented first of all. The algorithm architecture is then detailed and shows how the three limitations have been addressed. The SATUD algorithm, also introduced in this section, is detailed in the companion article [REF]. The elements for system evaluation are then presented. In Section 3, the system is compared with manual scoring and its performance is reported. Its results and impact on sleep diagnosis are discussed in Section 4. Finally, the conclusion is presented in Section 5.

2. Methods and materials

First of all, this section presents the database on which the algorithm was trained and tested. The methodology used in the system is then detailed, followed by a presentation of the system performance assessment.

2.1. Data acquisition

A total of 400 anonymous sleep recordings were included in this study thanks to the sleep cohort of Pays de La Loire. This cohort is operated under the aegis of the Institut de Recherche en Santé Respiratoire. Approval was obtained from the University of Angers ethics committee and the "Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé" (CCTIRS; 07.207bis). The recordings used in this study were acquired between 2012 and 2018, and all patients gave their written informed consent. The 400 recordings were divided into two datasets (D1 and D2). A random selection was made, ensuring equal representation of the severity of Obstructive Sleep Apnea (OSA) and the year of recording (see Table 1). D1 was made up of 300 recordings from 182 males and 117 females while D2 was made up of 100 recordings from 66 males and 34 females. Recordings from D1 were used as the training dataset and D2 as the test dataset. D1 dataset was made up of 329,911 epochs (W: 76,897 - R: 52,036 - N1: 24,283 - N2: 122,266 - N3: 54,429). As for D2 dataset, it was made up of 110,978 epochs (W: 24,172 - R: 18,194 - N1: 8,095 - N2: 41,095 - N3: 19,422).
The subjects, who were suspected of having OSA, underwent one-night's PSG in the sleep laboratory of Angers University Hospital (FRANCE). Sleep was recorded following the AASM guidelines [7]. The CID102L8D polysomnograph (CIDELEC St Gemmes-sur-Loire, FRANCE) used pro-

Table 1
OSA severity represented by age, using quartiles, evaluated for D1 and D2.

	D1 dataset				D2 dataset			
	Q1 19-43 y.o.	Q2 44-53 y.o.	Q3 54-62 y.o.	Q4 63-86 y.o.	Q1 19-41 y.o.	Q2 42-53 y.o.	Q3 54-63 y.o.	Q4 64-79 y.o.
No	32 %	18 %	9 %	5 %	44 %	8 %	8 %	8 %
Mild	31 %	23 %	23 %	25 %	32 %	28 %	20 %	24 %
Moderate	17 %	24 %	30 %	36 %	12 %	36 %	36 %	24 %
Severe	20 %	35 %	38 %	34 %	12 %	28 %	36 %	44 %

y.o. = years old

vided the usual electrophysiological (EP) and cardio-respiratory (CR) signals. It also included the PneaVoX® sensor, from which tracheal sounds and respiratory efforts are estimated to facilitate event scoring [29–31]. Once the signals were acquired, each sleep recording was manually scored by a single sleep specialist in accordance with AASM guidelines, although several sleep specialists were involved in this study. The hypnogram established by the sleep specialist was considered as our reference, and was referred to as $hypno_{ref}$ in the rest of this paper. Unlike other studies, epochs with artifacts were not discarded (neither manually nor automatically) to ensure the algorithm's efficiency in real-life conditions. The only epochs rejected were those with extremely bad quality signals preventing the manual scoring of events and/or sleep (for example epochs with missing signals). They were automatically invalidated by the CIDELEC user interface prior to scoring.

2.2. Algorithm structure

The algorithm presented in this section was implemented using Matlab® software to provide sleep scoring support tools. It was designed based on the scoring rules described in the AASM guidelines [7], and behaves similarly to the manual scoring process. Its inputs are EP signals and a priori medical knowledge (AASM guidelines). It classifies all epochs to provide an automatic hypnogram $hypno_{EP}$, with a probability table $probabilities_{EP}$. This table gives information about the confidence level of the algorithm when epochs were classified. Figure 1 illustrates the algorithm structure, described in this section. The architecture is composed of several main functions: **F1**, **F2**, **F3** and **F4**. Each function aims to reproduce one of the tasks realized by sleep specialists when scoring sleep. Firstly, an adaptation to each recording is achieved and provides patient-dependent features (**F1**). Using those features, a rough hypnogram is estimated (**F2**). Similarly to the manual scoring process, the hypnogram is then adjusted with regard to sleep patterns (identified in **F3**), surrounding epochs and transition rules (**F4**).

2.2.1. *F1 - The SATUD system*

Before scoring sleep stages, sleep specialists visualize all epochs to adapt their scoring to the patient's specific characteristics. This process was implemented automatically with the SATUD system, which does not require any training/test steps. The SATUD system was thus applied to all D1 and D2 recordings individually.
The SATUD system is fully presented in the companion paper [REF], in which its functioning and a simplified example of its use for sleep stage classification are detailed.
Briefly, the SATUD system extracted 41 patient-specific qualitative features from EP channels, using a priori medical knowledge obtained from the AASM guidelines. The 41 patient-

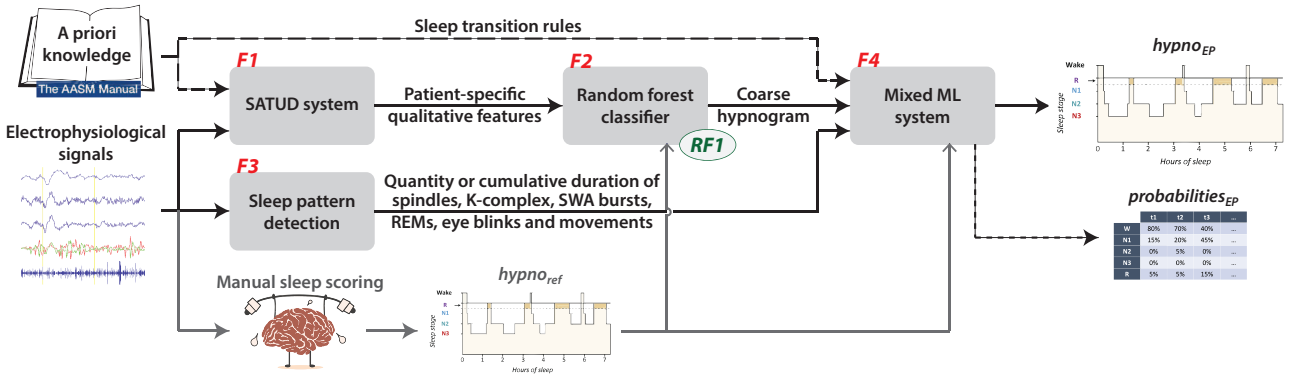


Figure 1: Functional architecture of the user-friendly automatic sleep staging system, composed of four main functions (F1, F2, F3 and F4). Medical knowledge from the AASM manual and electrophysiological signals are used as inputs. The outputs are a hypnogram $hypno_{EP}$ and the associated probability table $probabilities_{EP}$.

specific qualitative features were defined as the levels of various sleep stage descriptors (quantitative features), as presented in Table 2. For example, the AASM guidelines mention the importance of the EEG amplitude (1st line of Table 2) to score sleep stages. EEG amplitude is the highest in N3 and the lowest in N1. It is also generally higher in N2 than in W and R. However, EEG amplitude in R can increase when there are some artifacts (called Rapid Eye Movement artifacts). It can also be really high in W when the patient is agitated (due to movements). With regard to these elements, we decided to use 2 thresholds so we can associate EEG amplitude levels (or qualitative features) with stages as following: N3→High, N2→Mid or High, R→Low or Mid and N1→Low (nothing for W as it can be Low, Mid or High). As there was no need for a Mid EEG amplitude level alone, we did not compute it.

Table 2
List of the 41 sleep stages' qualitative features (right column) obtained using the SATUD system. Each qualitative feature corresponds to a level associated with a descriptor or quantitative feature (left column), identified as relevant for sleep stage scoring using the AASM guidelines.

Quantitative features	Th[a]	Qualitative features used	N[b]
EEG amplitude	2	Low | Low or Mid | Mid or High | High	4
EEG instability	1	No | Yes	2
Slow wave activity quantity	2	Low | Low or Mid | High	3
Alpha waves quantity	2	Low | Low or Mid | Mid | Mid or High | High	5
Beta waves quantity	2	Low | Mid | Mid or High	3
Delta waves quantity	2	Low | Mid or High	2
Theta waves quantity	2	Low | Low or Mid | Mid or High	3
Chin level	2	Low | Low or Mid | Mid or High | High	4
Chin instability	2	Low | Low or Mid | Mid or High | High	4
Summed EOG level	2	Low | Low or Mid | Mid or High	3
Summed EOG instability	2	Low | Low or Mid | Mid or High	3
Substracted EOG level	2	Low | Mid or High	2
Substracted EOG instability	2	Low | Mid or High | High	3
		Total	41

[a] Number of Thresholds used.
[b] Number of qualitative features used for each quantitative feature.

EEG waves (lines 3 to 7 of Table 2) are usually evaluated using frequency ranges with fixed boundaries. In this work, those ranges were redefined for each recording to adapt even further to each patient. In particular, the alpha waves frequency range during wakefulness was adjusted to avoid W overestimation[1].

[1] Alpha waves are representative of the W stage. For some patients, they can also occur in the R stage, or even throughout the entire recording in the case of alpha-delta sleep [32] patients.

2.2.2. F2 - Random forest classifier

Using the 41 patient-specific qualitative features obtained with the SATUD system (F1) as input, F2 aims to generate an initial hypnogram. This hypnogram was referred to as a 'coarse hypnogram' in the rest of this paper, since it will be used to obtain a more precise hypnogram (in F4). Because F2 required training/test steps, all D1 features were concatenated into a single large training matrix. In the same way, all D1 references $hypno_{ref}$ were concatenated.
The implemented classifier was a random forest [33], a machine learning algorithm that combines decision trees. This model was chosen because it is powerful, robust, and not opaque like deep learning methods. The random forest was developed using the TreeBagger function from Matlab®. This function bagged 100 classification trees using bootstrap samples of the data and randomly selecting a subset of 6 features at each node.
Once the model had been trained, it was saved under the name ***RF1*** to be reused for each test recording individually, resulting in one coarse hypnogram per D2 recording.

2.2.3. F3 - Sleep pattern detection

Besides continuous features, so-called 'sleep patterns' are essential for sleep scoring. F3 aims to identify a majority of them within EP signals, using their description as mentioned in the AASM guidelines. Sleep spindles, K-complex, Slow Wave Activity (SWA) bursts, Rapid Eye Movements (REMs), eye blinks and movements were detected using signal processing algorithms previously implemented and not detailed in the present paper (filters, wavelets, empirical reasoning, etc.). For each sleep pattern, depending on its nature, its quantity or cumulative duration was computed for each half-epoch (for better respect of the AASM guidelines). The resulting features were computed for each recording individually (D1 and D2), and will be used to enhance the coarse hypnogram.

2.2.4. F4 - Mixed ML system

The last major element for sleep scoring are sleep transition rules. As described in the AASM guidelines, efficient sleep scoring requires the knowledge of surrounding epochs (especially for N2 and R sleep stages). F4 is a mixed machine learning system that combines the coarse hypnogram (obtained in F2), sleep pattern quantity or duration (obtained in F3) and a priori knowledge of sleep transition rules (obtained from the AASM guidelines). F4 outputs are the final hypnogram $hypno_{EP}$, along with the probability table $probabilities_{EP}$. Figure 2 illustrates F4 structure. Its architecture is composed of several main functions: **F4.1**, which estimates a hypnogram using the coarse hypnogram and sleep patterns by taking temporality into consideration, **F4.2** and **F4.3**, which adjust and correct the obtained hypnogram ensuring that there are no forbidden epoch sequences, and **F4.4**, which is an independent function using the coarse hypnogram and sleep patterns to provide a supplementary scoring support tool.

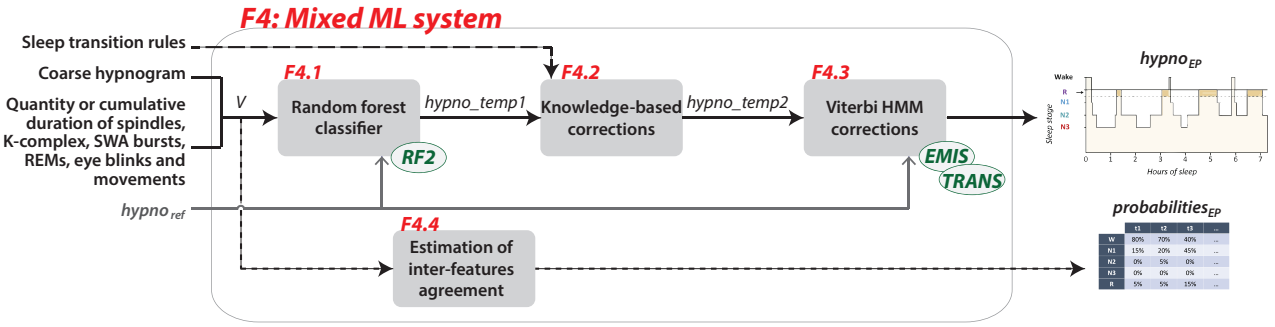


Figure 2: Functional architecture of the mixed ML system, composed of four main functions (F4.1, F4.2, F4.3 and F4.4). Inputs are sleep transition rules (from the AASM manual), the coarse hypnogram (from F2) and sleep patterns (from F3). The manual scoring $hypno_{ref}$ was used for training. The outputs are the hypnogram $hypno_{EP}$ and the associated probability table $probabilities_{EP}$.

F4.1 To address the timeline, nine features were identified within the coarse hypnogram (five features) and sleep patterns (four features). These features were extracted in regard to the current epoch, as shown in Figure 3.

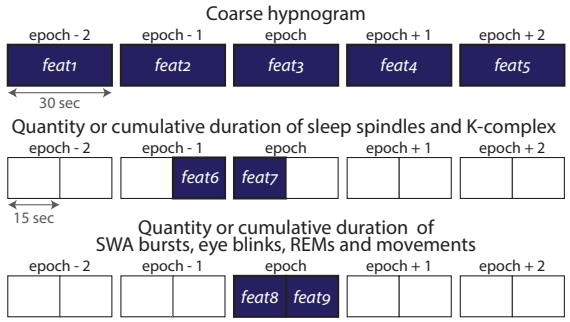


Figure 3: Features extracted from the coarse hypnogram and sleep patterns, in regard to the current epoch.

These features constitute the classifier input vector, noted V. As a consequence, the surrounding epochs are taken into consideration by the classifier. The chosen classifier was a random forest, developed with the same settings as in F2. It was trained concatenating all V vectors from D1 recordings as the input, and all D1 references $hypno_{ref}$ as the reference. The resulting hypnogram will be referred to as *hypno_temp1* in the rest of this paper. The trained model, called ***RF2***, was saved to be reused for each test recording individually, resulting in one *hypno_temp1* per D2 recording.

F4.2 *hypno_temp1* was smoothed using the transition rules described in the AASM guidelines. These rules, that define the possible or forbidden transitions between sleep stages, were implemented using the AASM guidelines but also empirically, by studying the errors that occurred the most. The smoothed hypnogram will be referred to as *hypno_temp2* in the rest of this paper.

F4.3 *hypno_temp2* was further smoothed using a Viterbi hidden Markov model [34], trained to identify and correct sequence mistakes [35, 36]. To do so, we used the hmmviterbi Matlab® function, which required the computation of two matrices:

- the emission probability matrix ***EMIS***, which corresponds to the probability of each sleep stage being emitted depending on the reference sleep stage. It was estimated using the confusion matrix between *hypno_temp2* and $hypno_{ref}$;
- the transition probability matrix ***TRANS***, which corresponds to the probability of transition between each sleep stage. It was estimated from $hypno_{ref}$.

Using these matrices, the Viterbi hidden Markov model smooths hypnogram *hypno_temp2*, resulting in the final one $hypno_{EP}$. ***EMIS*** and ***TRANS*** were both estimated from D1 and then saved to be reused for each test recording individually, resulting in one $hypno_{EP}$ per D2 recording.

F4.4 Independently from F4.1, F4.2 and F4.3, a table called $probabilities_{EP}$ was also computed. This table contains, for each epoch, the estimated probabilities of being in each sleep stage. It can be used to determine which epochs were more or less easily scored by the algorithm. $probabilities_{EP}$ computation is not detailed in this paper. In brief, it was established by measuring the agreement between sets of features which were selected as being representative of each specific sleep stage. $probabilities_{EP}$ thus reflects the algorithm's doubts when scoring sleep. Thanks to this, the medical practitioner knows which epochs the algorithm found difficult or easy to score.

2.3. System evaluation

Results were estimated from the recordings included in the test dataset D2. As a reminder, this dataset is made up of independent recordings not used during training. Each recording was processed using the previously trained models. Then, the resulting $hypno_{EP}$ and $probabilities_{EP}$ were gauged using the associated $hypno_{ref}$ (in 2.3.1 and 2.3.2, respectively). In both sections, reported scores correspond to the averaged individual scores.

2.3.1. Evaluation of $hypno_{EP}$ accuracy

For each recording, $hypno_{EP}$ was compared to $hypno_{ref}$ using a contingency table (see Table 3).

Table 3
Contingency table.

		Automatic analysis					
		W	N1	N2	N3	R	Total
Reference	W	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	$n_{1.}$
	N1	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	$n_{2.}$
	N2	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	$n_{3.}$
	N3	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	$n_{4.}$
	R	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	$n_{5.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	n

The overall accuracy of the automatic scoring is assessed using Cohen's Kappa κ [37] and accuracy rate Acc.
κ is probably the most used index for automatic sleep scoring evaluation. Indeed, it measures the agreement between the reference and the automatic analysis by taking into consideration the random component of this agreement (expected agreement on the assumption that the manual and automatic analyses are totally independent). κ is calculated from the proportion of observed agreement P_o and the proportion of random agreement P_e:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

with $P_o = \frac{1}{n}\sum_{i=1}^{5} n_{ii}$ and $P_e = \frac{1}{n^2}\sum_{i=1}^{5} n_{i.} \times n_{.i}$.

It is usually interpreted using six ranges:

(i) $\kappa < 0.0$: Poor agreement
(ii) $0.0 \leq \kappa < 0.2$: Slight agreement
(iii) $0.2 \leq \kappa < 0.4$: Fair agreement
(iv) $0.4 \leq \kappa < 0.6$: Moderate agreement
(v) $0.6 \leq \kappa < 0.8$: Substantial agreement
(vi) $0.8 \leq \kappa$: Almost perfect agreement

Acc corresponds to the percentage of correctly scored epochs:

$$Acc(\%) = \frac{1}{n}\sum_{i=1}^{5} n_{ii} \times 100$$

Overall scores were also reported while considering subjects by age or OSA severity.

Furthermore, scores were estimated for each sleep stage individually. To do so, we adopted a one-vs.-rest approach where each stage is alternatively considered as the positive class and the others are combined into a single negative class. From the resulting true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), we estimated each stage's Cohen's Kappa, accuracy rate, sensitivity (defined as $\frac{TP}{TP+FN}$) and specificity (defined as $\frac{TN}{TN+FP}$).

2.3.2. Decision support with $probabilities_{EP}$

Since the probability table $probabilities_{EP}$ has no reference with which to be compared, we evaluated the mean probability given by $probabilities_{EP}$ when epochs are correctly versus erroneously scored in a specific sleep stage. The greater the difference between those mean probabilities, the better the probability table. Indeed, we want the mistakes in the algorithm to be limited to epochs where the features designated multiple sleep stages rather than a single one (suggesting manual scoring may also have been complex).

3. Results

3.1. Evaluation of $hypno_{EP}$ accuracy

Among the 100 recordings, 84 % obtained an overall Cohen's Kappa κ above 0.60, showing a substantial or almost perfect agreement with the manual scoring.
Table 4 gives, for D2 recordings, the mean value and the standard deviation of the overall Cohen's Kappa κ and accuracy rate Acc of $hypno_{EP}$. Scores per sleep stage are also reported. The overall κ and Acc were 0.69 and 77.8 %, respectively. If we consider each sleep stage detection individually, stage R obtained the best scores with a κ reaching 0.80 (almost perfect agreement with the reference). κ mean values indicate all other sleep stages got a substantial agreement with the manual scoring, except stage N1. Sensitivities were all above 82 %, except for N2 and N1 sleep stages (74.9 % and 20.8 %, respectively). Specificities were all above 94 %, except for N2 sleep stage (83.3 %).

Table 5 and Table 6 report the performances obtained depending on the subject's age (using quartiles) and OSA severity (based on a physician's diagnosis).
There was a substantial agreement with the manual scoring for all groups. For age groups, the lowest scores were obtained for 53-63 year old patients, and the better ones for patients above 63 years old. For OSA severities, κ were above 0.70 for patients with mild or moderate OSA, and below for patients with no or severe OSA.

Table 4
Overall and individual performances obtained from automatic sleep staging on D2 dataset.

D2 dataset	W	N1	N2	N3	R	All
Cohen's Kappa	0.74 ± 0.14	0.23 ± 0.11	0.64 ± 0.14	0.71 ± 0.20	0.80 ± 0.14	0.69 ± 0.10
Accuracy rate (%)	92.3 ± 4.5	92.1 ± 3.4	83.3 ± 6.2	92.7 ± 4.5	95.1 ± 3.0	77.8 ± 7.0
Sensitivity (%)	82.3 ± 15.5	20.8 ± 9.0	83.9 ± 9.5	74.9 ± 20.3	83.3 ± 16.5	N.A.
Specificity (%)	94.0 ± 5.7	97.8 ± 1.1	83.3 ± 8.0	96.5 ± 4.3	97.4 ± 2.2	N.A.

N.A. = Not Applicable

Table 5
Overall performances depending on the subject's age, using quartiles. Q1 = 19-41 years old, Q2 = 41-53 years old, Q3 = 53-63 years old and Q4 = 63-79 years old.

	Q1	Q2	Q3	Q4
κ	0.69 ± 0.10	0.69 ± 0.10	0.67 ± 0.10	0.70 ± 0.09
Acc	78.2 ± 6.6	77.9 ± 6.5	75.8 ± 7.1	78.9 ± 6.7

Table 6
Overall performances depending on the subject's OSA severity (obtained by the sleep expert).

	No	Mild	Moderate	Severe
κ	0.67 ± 0.12	0.70 ± 0.10	0.73 ± 0.06	0.65 ± 0.10
Acc	76.6 ± 7.7	79.1 ± 6.7	80.4 ± 4.7	74.6 ± 7.7

Figure 4 presents the confusion matrix related to Table 4. The quantity of epochs manually scored in each stage (W: 24,172 - R: 18,194 - N1: 8,095 - N2: 41,095 - N3: 19,422) should be borne in mind while interpreting the confusion matrix. Common mistakes appeared to be N1 and N3 epochs being automatically scored as N2 sleep stage.

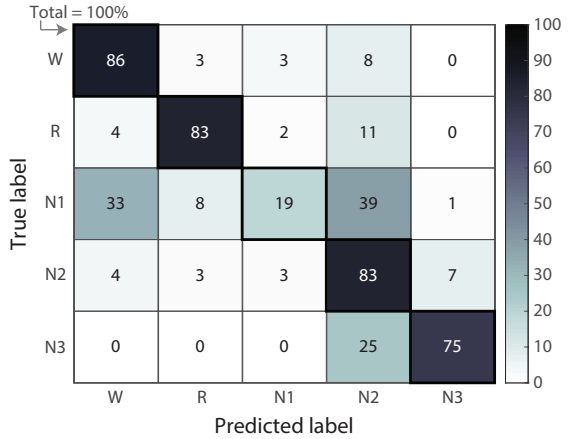


Figure 4: Confusion matrix (percentage over lines) obtained from automatic sleep staging on D2 dataset.

3.2. Decision support with $probabilities_{EP}$

$probabilities_{EP}$ is a supplementary scoring support tool that helps make the system more practical to use and improves physicians' confidence in the algorithm.
Table 7 reports the mean probabilities associated with epochs correctly and erroneously estimated by $hypno_{EP}$, depending on sleep stages. In this table, we can see that epochs correctly identified as N2 have a reported mean probability of 66 % of being N2 stage, whereas the epochs over-detected as N2 have a reported mean probability of only 44 % of being N2. It makes a difference of 22 % between actual N2 epochs and not. Algorithm confidence when scoring an epoch into N2 sleep stage is thus greater when it is an actual N2 epoch. Considering the mean probabilities reported for the other sleep stages, we can see that all differences are above 15 %, except N1 sleep stage.
Erroneous epochs are thus more likely to have small and uniform $probabilities_{EP}$ values than correctly classified ones. Indeed, the latter should show a clear superiority of the probability associated with their sleep stage compared with others. $probabilities_{EP}$ points out epochs which need to be checked as a priority, and manually corrected if necessary.

Table 7
Mean probabilities per sleep stage while the reference agrees or disagrees with the stage detected by the system.

	$hypno_{ref}$ agrees	$hypno_{ref}$ disagrees
$hypno_{EP}$ = W	79 %	60 %
$hypno_{EP}$ = N1	65 %	56 %
$hypno_{EP}$ = N2	66 %	44 %
$hypno_{EP}$ = N3	87 %	71 %
$hypno_{EP}$ = R	69 %	53 %

4. Discussion

The main goal of this study was to implement a user-friendly automatic sleep scoring system.

Several factors have helped to overcome the three limitations presented above:

a) the developed system was designed to be as easy as possible to interpret. To do so, its construction replicates manual scoring and uses medical knowledge extracted from the AASM guidelines. Firstly, the channels chosen were the same ones as when manually scoring sleep. Secondly, we know that medical practitioners quickly visualize the recording before scoring it, to become familiar with its specific characteristics and score accordingly. This step, replicated in the SATUD system (set out in the companion paper [ref]), causes the automatic hypnogram to be patient-dependent. Thirdly, most of the elements described in the AASM guidelines were included in the system: continuous features, sleep patterns, surrounding knowledge and transition rules. Lastly, the methodologies implemented in the system were established using interpretable algorithms or medical knowledge;
b) algorithm performances were evaluated on a hundred independent recordings, from patients with and without sleep-disordered breathing;
c) the system works in real-life conditions and does not require any previous human intervention. There is no need for invalidation of epochs or event scoring. $probabilities_{EP}$ also improves system practicality by pointing out epochs that should be checked as a priority.

Despite these restrictions, the method showed it could get results comparable to those obtained with manual scoring, reaching Cohen's Kappa values around 0.69 and accuracy rates around 78 % (see Table 4). There was no significant impact of age on performance (see Table 5). As for OSA severity (see Table 6), performance was the lowest for patients with severe OSA syndrome. Surprisingly,

Table 8
Performance of 5-stage classification using electrophysiological channels, compared with related works.

	Number of one-night PSG (training and test)	Subject diagnosis	Approach	Overcome limitations*	Acc (%)	κ
Zokaeinikoo et al. 2016 [24]	20 (LOOCV**)	Healthy only	ML	c	74	
Biswal et al. 2018 [16]	10,000 (9,000-1,000)	Healthy and SDB	DL	b and c	**88**	0.81
Zhang et al. 2019 [17]	5,804 (5,213-580)	Healthy and SDB	DL	b and c	87	**0.82**
Chen al. 2019 [26]	16***	Healthy and SDB	Interpretable	a and b	80	0.72
This work	400 (300-100)	Healthy and SDB	Interpretable	a, b and c	78	0.69

ML = Machine Learning, DL = Deep Learning
* Limitations overcome from our point of view.
Limitation a: model opacity, b: insufficient heterogeneity of dataset and c: lack of practicality
** Leave-One-Out Cross-Validation: one by one, each subject's recording was selected as the test dataset, and the others were combined into the training dataset. Final results are provided by the model with the highest scores.
*** Semi-automated method that requires the manual scoring of 5 % of each recording.

the group with the second worst performance was patients with no OSA syndrome. Nonetheless, the results obtained for all groups were acceptable (the lowest κ was 0.65), indicating the algorithm is robust to more or less fragmented sleep recordings. Considering each sleep stage individually (see Figure 4 and Table 4), it seems that the errors were mainly N3 epochs misclassified as N2 sleep stage. Also, sleep stage N1 had very low performance compared to other stages. This is not surprising, since sleep stage N1 is a transitional stage representing approximately 5 % of the night with a very likely overlap with W and N2 stages. In fact, inter-scorer agreement for N1 sleep stage is the lowest [8]. Table 7 showed that misclassified epochs have lower probability values in the returned table $probabilities_{EP}$, compared with correctly scored epochs. Using $probabilities_{EP}$, some of the mistaken epochs could thus be identified and rescored by the manual scorer. A possible strategy would be to highlight epochs with no obvious superiority of one stage probability among the others. The scorer could consider reviewing only those epochs.

Table 8 presents the obtained results, compared with the literature. Only studies using EEG, EOG and EMG, compared with hypnograms manually scored following the AASM guidelines and indicating the number of patients and training/test repartitions were considered for comparison. Deep learning approaches, which reached better scores, do not overcome limitation a). Chen et al. approach [26], which is also interpretable, do not overcome limitation c) since it is semi-automated. Our method showed that despite the inclusion of transparency, hybrid methods can still reach adequate scores.

The proposed system provides resilient tools to facilitate sleep scoring, thus assisting sleep experts in diagnosing sleep disorders. As we are aware of sleep specialists' mistrust in automatic approaches, we identified three limitations to their use and designed the system to overcome them. To go even further, several perspectives are considered. Firstly, we would like to evaluate this methodology using recordings from other sleep laboratories. Indeed, our recordings were all provided by one sleep laboratory and, even if several manual scorers established the references, local scoring practices may have influenced the algorithms. Secondly, micro-arousals[2] (used when manually scoring sleep as they are required by some transition rules), were not detected in this system. Work is currently being done to identify them. Thirdly, the automatic sleep scoring impact on sleep diagnosis should be evaluated. The Apnea Hypopnea Index (AHI) resulting from the current system should be compared with the one resulting from manual scoring. Lastly, since sleep diagnosis is sometimes performed using devices that do not record EP channels, it would be interesting to see how good automatic sleep scoring from cardio-respiratory channels could be.

5. Conclusion

In this paper, a new approach for automatic sleep staging was presented. Its architecture was designed to reproduce step-by-step the manual scoring tasks realized by sleep experts: adaptation to each recording's specific characteristics, study of temporal and spectral content, identification of sleep patterns and classification regarding surrounding epochs and transition rules. As it is easy to understand, this model is well suited to the medical community which lacks confidence in the models generally implemented.
The method was evaluated on 100 patients with and without sleep-disordered breathing, and results showed the automatic hypnogram made a good performance regardless of subjects' age or OSA severity. With a mean Cohen's Kappa and accuracy rate of 0.69 and 77.8 %, respectively, the algorithm obtained a high agreement with the manual scorer.
The proposed method was put together as a scoring support tool for sleep scoring and is thus useable immediately after the polysomnographic recording, without the need for any preliminary human action. Besides the automatic hypnogram, it also points out epochs which should be checked and rescored if necessary.
Given that sleep scoring is a time-consuming and complex task, the presented user-friendly tool should greatly support sleep specialists in their diagnosis.

[2] Short awakenings, markers of sleep disruption

Funding

This study was supported by grants from the Institut de Recherche en Santé Respiratoire des Pays de La Loire.

Acknowledgements

The authors would like to thank Christelle Gosselin and Jean-Louis Racineux, from the Institut de Recherche en Santé Respiratoire des Pays de La Loire, and Margaux Blanchard, from the ESEO. Thanks to Alain Le Duff and Lucile Riaboff, previously from the ESEO. We thank Julien Godey, Laetitia Moreno and Marion Vincent, sleep technicians in the Department of Respiratory and Sleep Medicine of Angers University Hospital.

References

[1] R. Heinzer, S. Vat, P. Marques-Vidal, H. Marti-Soler, D. Andries, N. Tobback, V. Mooser, M. Preisig, A. Malhotra, G. Waeber, P. Vollenweider, M. Tafti, J. Haba-Rubio, Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study, The Lancet Respiratory Medicine 3 (2015) 310–318. doi:10.1016/S2213-2600(15)00043-0.

[2] J. B. Croft, CDC's Public Health Surveillance of Sleep Health, 2017.

[3] C. V. Senaratna, J. L. Perret, C. J. Lodge, A. J. Lowe, B. E. Campbell, M. C. Matheson, G. S. Hamilton, S. C. Dharmage, Prevalence of obstructive sleep apnea in the general population: A systematic review, Sleep Medicine Reviews 34 (2017) 70–81. doi:10.1016/j.smrv.2016.07.002.

[4] AASMTaskForce, Sleep-related Breathing Disorders in Adults: Recommendations for Syndrome Definition and Measurement Techniques in Clinical Research, Sleep 22 (1999) 667–689. doi:10.1093/sleep/22.5.667.

[5] K. A. Franklin, E. Lindberg, Obstructive sleep apnea is a common disorder in the population- a review on the epidemiology of sleep apnea, Journal of Thoracic Disease 7 (2015) 1311–1322. doi:10.3978/j.issn.2072-1439.2015.06.11.

[6] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, K. M. Hla, Increased Prevalence of Sleep-Disordered Breathing in Adults, American Journal of Epidemiology 177 (2013) 1006–1014. doi:10.1093/aje/kws342.

[7] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, R. M. lloyd, S. F. Quan, M. M. Troester, B. V. Vaughn, The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, number 2.4 in American Academy of Sleep Medicine, Darien IL, 2017.

[8] R. S. Rosenberg, S. Van Hout, The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring, Journal of Clinical Sleep Medicine (2013). doi:10.5664/jcsm.2350.

[9] A. L. Fogel, J. C. Kvedar, Artificial intelligence powers digital medicine, npj Digital Medicine 1 (2018). doi:10.1038/s41746-017-0012-2.

[10] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Medicine 25 (2019) 44–56. doi:10.1038/s41591-018-0300-7.

[11] T. Penzel, R. Conradt, Computer based sleep recording and analysis, Sleep Medicine Reviews 4 (2000) 131–148. doi:10.1053/smrv.1999.0087.

[12] L. Fiorillo, A. Puiatti, M. Papandrea, P.-L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, F. D. Faraci, Automated sleep scoring: A review of the latest approaches, Sleep Medicine Reviews 48 (2019). doi:10.1016/j.smrv.2019.07.007.

[13] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. doi:10.1038/nature14539.

[14] C. M. Bishop, Pattern Recognition and Machine Learning, 2006.

[15] S. Wermter, R. Sun, An Overview of Hybrid Neural Systems, in: G. Goos, J. Hartmanis, J. van Leeuwen, S. Wermter, R. Sun (Eds.), Hybrid Neural Systems, volume 1778, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 1–13. doi:10.1007/10719871_1, series Title: Lecture Notes in Computer Science.

[16] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, M. T. Bianchi, Expert-level sleep scoring with deep neural networks, Journal of the American Medical Informatics Association 25 (2018) 1643–1650. doi:10.1093/jamia/ocy131.

[17] L. Zhang, D. Fabbri, R. Upender, D. Kent, Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks, Sleep 42 (2019). doi:10.1093/sleep/zsz159.

[18] S. Enshaeifar, S. Kouchaki, C. C. Took, S. Sanei, Quaternion Singular Spectrum Analysis of Electroencephalogram With Application in Sleep Analysis, IEEE Transactions on Neural Systems and Rehabilitation Engineering 24 (2016) 57–67. doi:10.1109/TNSRE.2015.2465177.

[19] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, K. Jerbi, Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines, Journal of Neuroscience Methods 250 (2015) 94–105. doi:10.1016/j.jneumeth.2015.01.022.

[20] S. Mahvash Mohammadi, S. Kouchaki, M. Ghavami, S. Sanei, Improving time-frequency domain sleep EEG classification via singular spectrum analysis, Journal of Neuroscience Methods 273 (2016) 96–106. doi:10.1016/j.jneumeth.2016.08.008.

[21] S. Charbonnier, L. Zoubek, S. Lesecq, F. Chapotot, Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging, Computers in Biology and Medicine 41 (2011) 380–389. doi:10.1016/j.compbiomed.2011.04.001.

[22] G. Garcia-Molina, F. Abtahi, M. Lagares-Lemos, Automated NREM sleep staging using the Electro-oculogram: A pilot study, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, pp. 2255–2258. URL: http://ieeexplore.ieee.org/abstract/document/6346411/.

[23] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, H. Dickhaus, Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier, Computer Methods and Programs in Biomedicine 108 (2012) 10–19. doi:10.1016/j.cmpb.2011.11.005.

[24] M. Zokaeinikoo, Automatic Sleep Stages Classification, Ph.D. thesis, 2016. URL: http://trace.tennessee.edu/utk_gradthes/4088/.

[25] A. Ugon, Fusion Symbolique et Données Polysomnographiques, Ph.D. thesis, 2015.

[26] C. Chen, A. Ugon, C. Sun, W. Chen, C. Philippe, A. Pinna, Towards a Hybrid Expert System Based on Sleep Event's Threshold Dependencies for Automated Personalized Sleep Staging by Combining Symbolic Fusion and Differential Evolution Algorithm, IEEE Access 7 (2019) 1775–1792. doi:10.1109/ACCESS.2018.2887082.

[27] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, Journal of Neural Engineering 16 (2019). doi:10.1088/1741-2552/ab260c.

[28] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, arXiv:1702.08608 [cs, stat] (2017). ArXiv: 1702.08608.

[29] M. Glos, A. Sabil, K. S. Jelavic, C. Schöbel, I. Fietze, T. Penzel, Characterization of Respiratory Events in Obstructive Sleep Apnea Using Suprasternal Pressure Monitoring, Journal of Clinical Sleep Medicine 14 (2018) 359–369. doi:10.5664/jcsm.6978.

[30] T. Penzel, A. Sabil, The use of tracheal sounds for the diagnosis of sleep apnoea, Breathe 13 (2017) e37–e45. doi:10.1183/20734735.008817.

[31] T. Penzel, A. Sabil, Physics and Applications for Tracheal Sound Recordings in Sleep Disorders, in: K. N. Priftis, L. J. Hadjileontiadis, M. L. Everard (Eds.), Breath Sounds, Springer International Publishing, Cham, 2018, pp. 83–104. doi:10.1007/978-3-319-71824-8_6.

[32] P. Hauri, D. R. Hawkins, Alpha-delta sleep, Electroencephalography and Clinical Neurophysiology 34 (1973) 233–237. doi:10.1016/

0013-4694(73)90250-2.

[33] T. K. Ho, Random Decision Forests (1995). doi:10.1109/ICDAR.1995.598994.

[34] L. E. Baum, T. Pietrie, Statistical Inference for Probabilistic Functions of Finite State Markov Chains, The Annals of Mathematical Statistics (1966) 1554–1563. doi:10.1214/aoms/1177699147.

[35] J. Yang, Toward physical activity diary: motion recognition using simple acceleration features with mobile phones, in: Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics - IMCE '09, ACM Press, Beijing, China, 2009, p. 1. doi:10.1145/1631040.1631042.

[36] L. Riaboff, S. Poggi, A. Madouasse, S. Couvreur, S. Aubin, N. Bédère, E. Goumand, A. Chauvin, G. Plantier, Development of a methodological framework for a robust prediction of the main behaviours of dairy cows using a combination of machine learning algorithms on accelerometer data, Computers and Electronics in Agriculture 169 (2020). doi:10.1016/j.compag.2019.105179.

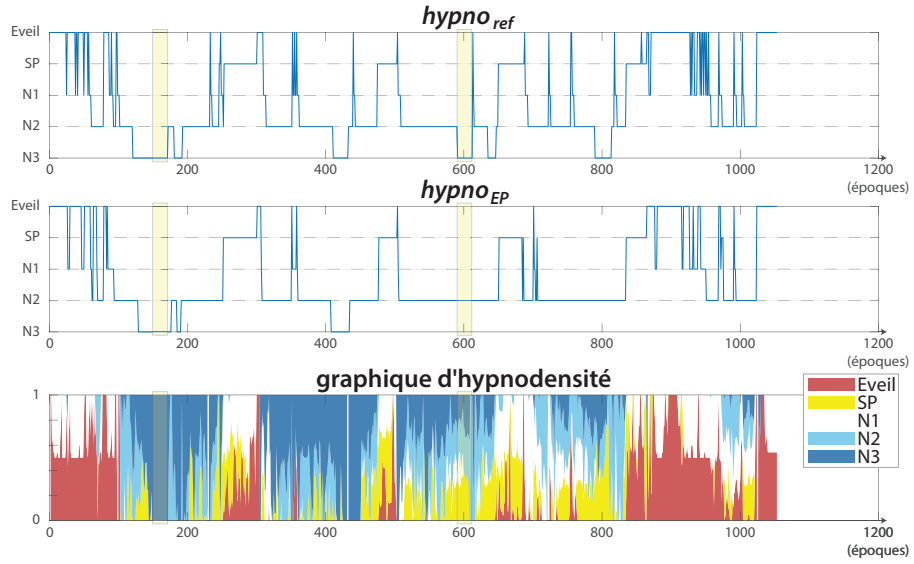
[37] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement 20 (1960) 37–46. doi:10.1177/001316446002000104.

Pour aller plus loin : un nouvel outil d'aide à la décision

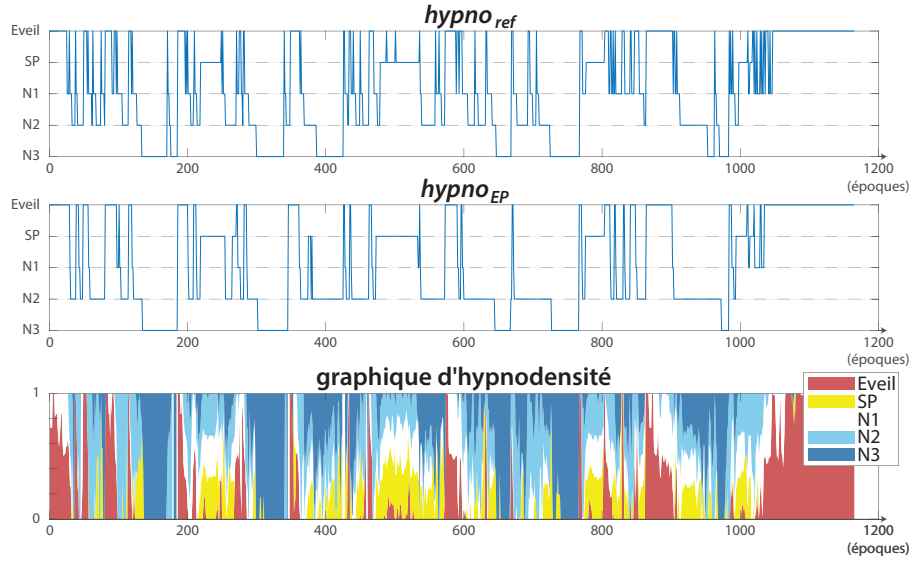
Pour compléter le travail présenté dans l'article, nous proposons d'ajouter un autre outil basé sur la table de probabilité $probabilities_{EP}$: un graphique d'hypnodensité.

> **Graphique d'hypnodensité :**
> Le graphique d'hypnodensité (Stephansen *et al.*, 2018), contrairement à l'hypnogramme, n'impose pas un unique stade de sommeil pour chaque époque. Il permet exprime plutôt un continuum entre chaque stade de sommeil, transmettant ainsi davantage d'informations.

Le graphique d'hypnodensité proposé ici est donc une représentation graphique de la table de probabilité $probabilities_{EP}$ et permet de visualiser, pour chaque époque de l'ensemble de la nuit, les stades de sommeil vers lesquels la classification automatique s'est orientée.
La Figure D2 présente les graphiques d'hypnodensité obtenus pour deux patients, ainsi que la classification manuelle du sommeil $hypno_{ref}$ et l'hypnogramme obtenu automatiquement $hypno_{EP}$.



(a)



(b)

Figure D2 – Exemple des classification manuelles ($hypno_{ref}$), automatiques ($hypno_{EP}$) et des graphiques d'hypnodensité pour : (a) le premier enregistrement du jeu de test et (b) le quatrième enregistrement du jeu de test, choisi pour son sommeil fragmenté par de nombreux courts éveils.

Pour le premier enregistrement, on observera particulièrement la zone autour de la 600ème époque, incorrectement classifiée comme étant du N2 au lieu de N3. On remarquera alors sur le graphique d'hypnodensité une indécision entre le SP, le N2, le N3 et même le N1. Au contraire, pour le N3 correctement classifié avant la 200ème époque, le graphique d'hypnodensité suggère une confiance de l'algorithme très forte dans sa classification. En effet, la couleur bleue occupe la gamme toute entière de 0 à 1, indiquant que le stade N3 est classé sans doute possible.
On remarquera aussi qu'à certains instants d'éveil, la classification automatique hésite avec le N1. En observant les signaux lors de ces époques, on se rend compte que ce sont des époques d'éveil calme. Au contraire, les époques d'éveil pour lesquelles la classification automatique est sûre d'elle sont des époques d'éveil agité.
En ce qui concerne le second enregistrement, le graphique d'hypnodensité nous permet de visualiser en un coup d'oeil que le sommeil est fortement fragmenté.

Le graphique d'hypnodensité est donc un outil supplémentaire d'aide à la décision. Il permettra au spécialiste du sommeil de jauger la complexité de l'enregistrement et d'identifier les époques nécessitant ou non une reclassification manuelle.

Informations supplémentaires : la détection des grapho-éléments du sommeil

La détection automatique des grapho-éléments mentionnés dans l'article ci-avant présenté est détaillée dans cette section. Ce travail de traitement du signal a été réalisé par des étudiants que j'ai supervisés lors de stages ou projets de fin d'études. Actuellement intégré avec peu de modifications par rapport à la version rendue, ce travail pourra faire l'objet d'améliorations par la suite.

Fuseaux : la détection des fuseaux a été réalisée en couplant les informations fréquentielles des EEGs (11-16 Hz ou plus généralement 12-14 Hz) et les amplitudes des signaux filtrés selon les différentes bandes de fréquence électroencéphalographiques. Le classifieur utilisé est un arbre de décision. La sensibilité et la VPP, évaluées sur dix enregistrements, étaient de 83 % et 35 %, respectivement [1]. Pour comparaison, l'étude de Parekh *et al.* (2015) obtient une sensibilité de 71 % pour une VPP de 68 %. Ici, les performances sont indiquées en utilisant uniquement la sensibilité et la VPP. En effet, les grapho-éléments sont des événements succincts et de durée variable. Il n'est donc pas possible d'estimer le nombre de FP. En conséquence, la spécificité, la VPN, le taux d'accord et le κ ne peuvent pas être estimés.

Complexes K et bouffées d'activité EEG en ondes lentes : la détection des complexes K et bouffées d'activité EEG en ondes lentes ont été réalisées en utilisant une décomposition en ondelettes. Dans le cas des complexes K, les zones d'intérêt obtenues ont ensuite été filtrées afin de respecter les recommandations AASM (onde négative nette immédiatement suivie d'un composant positif et dont la durée est supérieure à 0,5 secondes). Pour finir, les EOGs ont été utilisés afin d'éviter les fausses détections lors des Mouvements Oculaires Rapides (MORs). Ce travail a été validé de manière visuelle et les performances n'ont donc pas été estimées.

Mouvements oculaires : la détection des mouvements oculaires (MORs, clignements des yeux et mouvements de lecture) a été réalisée en utilisant des méthodes de filtrage sur les signaux des deux yeux pris séparément et considérés comme un ensemble. En effet, il n'est pas rare d'observer des variations sur les voies EOGs dont l'origine est une activité cérébrale très importante (par exemple en N3). Dans ce cas, les variations sur chaque oeil sont corrélées, au contraire des réels mouvements oculaires. Cela est dû au placement des électrodes EOG, l'une étant positionnée légèrement au-dessus du coin externe de l'oeil et l'autre légèrement au-dessous (voir Figure D3 extraite du manuel de recommandations de l'AASM). Les mouvements oculaires sont donc en opposition de phase sur les signaux EOGs. Ce travail a été validé de manière visuelle et les performances n'ont donc pas été estimées.

1. Les performances ont été évaluées en comparant les fuseaux détectés automatiquement avec les fuseaux identifiés manuellement en amont du travail de développement du stage.

Figure D3 – Placement des électrodes EOG. Figure extraite des recommandation de l'AASM.

Mouvements corporels : la détection des mouvements corporels a été simplement réalisée en utilisant les EMGs enregistrés sur les jambes. Ce travail a été validé de manière visuelle et les performances n'ont donc pas été estimées.

Pour information, le κ de l'hypnogramme obtenu à la sortie de F2 avant utilisation des grapho-éléments (*coarse hypnogram*, voir la Figure 1 page 3 de l'article précédent) était de $0{,}58 \pm 0{,}10$, contre $0{,}67 \pm 0{,}10$ à la sortie de F4.1 (*hypno_temp1*, voir la Figure 2 page 5 de l'article précédent).

Valorisations supplémentaires

Ce travail a également été présenté lors de congrès :

— World Sleep Congress : présentation orale, Vancouver (Canada), septembre 2019 (Vanbuis *et al.*, 2019) ;

— Dafka Medical Events : présentation orale de type symposium, Tel-Aviv (Israël), octobre 2019 ;

— Le Congrès du Sommeil : poster, Lille, novembre 2019 (Vanbuis *et al.*, 2020a).

À RETENIR

Un outil de classification automatique des stades de sommeil à partir des voies électrophysiologiques a été mis en place. Pensé pour la communauté médicale, il reproduit les étapes de classification manuelle. Il est ainsi capable de produire un hypnogramme patient-dépendant en prenant en compte les grapho-éléments du sommeil, les époques à proximité et les règles de transition. Son utilisation est simple et fournit, en plus de l'hypnogramme, un graphe d'hypnodensité permettant d'orienter le médecin vers les époques à relire en priorité. Testé sur un nombre important de patients avec et sans pathologies du sommeil, il a montré qu'il était capable de classifier le sommeil avec des performances similaires à celles obtenues par un expert du sommeil.

D.3.2 SATUD : seuillage auto-adaptatif et non supervisé

Ce travail constitue la seconde et dernière partie de l'article en deux parties.
Dans la première partie, un algorithme de classification des stades de sommeil a été mis en place. Cet algorithme, afin de répondre aux différentes limitations allant à l'encontre de l'utilisation d'une analyse automatique en routine clinique (présentées Section D.2), reproduit le processus de lecture manuelle du sommeil, étape par étape.
La première étape réalisée par le spécialiste du sommeil consiste ainsi à visualiser rapidement l'ensemble des époques constituant l'enregistrement, afin de se familiariser avec les signaux propres au patient et à leurs spécificités. Cette étape est particulièrement importante de par la forte variabilité des signaux entre les patients.
Cette deuxième partie d'article est dédiée à l'algorithme SATUD, implémenté pour reproduire cette étape d'ajustement à chaque enregistrement. Cet algorithme permet en effet l'obtention de caractéristiques propres à chaque patient.
C'est grâce à SATUD que l'hypnogramme final obtenu est patient-dépendant.

Résumé traduit

La lecture manuelle du sommeil est une tâche chronophage, complexe et nécessitant un haut niveau d'expertise médicale. Les chercheurs se penchent donc sur la classification automatique des stades de sommeil, généralement basée sur des méthodes d'apprentissage supervisées. Sa mise en place reste cependant un défi en raison de la grande variabilité entre les patients, qui n'est pas considérée avec de tels algorithmes.
Cet article présente une méthode permettant l'extraction de caractéristiques qualitatives et patient-dépendantes à partir des signaux électrophysiologiques. Les caractéristiques, extraites par segment de 30 secondes (appelé époque), sont destinées à être utilisées en entrée de la classification. À la place de seuils fixes, la méthode développée appelée « Auto-adaptative Thresholding Using Descriptors » (SATUD), propose un seuillage auto-adaptatif non supervisé. Les seuils sont automatiquement ajustés à chaque enregistrement, de manière à maximiser à la fois la similarité au sein des époques d'un même stade de sommeil et la dissimilarité entre les époques de différents stades.
Cette méthode a été évaluée à partir de lectures manuelles du sommeil provenant de 60 patients présentant des pathologies diverses, garantissant une grande variabilité entre les enregistrements. Comparée à deux autres seuillages implémentés dans cette étude, la méthode SATUD montre une meilleure adaptation aux spécificités des patients. En effet, le nombre d'époques respectant toutes les propriétés associées à leur stade de sommeil a augmenté de plus de 80 % avec l'utilisation du SATUD, par rapport aux autres techniques de seuillage. La méthode SATUD s'est également avérée robuste au bruit et aux artefacts de sudation. Elle fournit ainsi des caractéristiques qualitatives et patient-dépendantes pouvant être utilisées pour la classification automatique des stades de sommeil. Une classification utilisant les caractéristiques obtenues a été présentée dans l'article compagnon.

Towards a user-friendly sleep staging system for polysomnography Part II: patient-dependent features extraction using the SATUD system

Jade Vanbuis[a,b,*], Mathieu Feuilloy[a,b], Lucile Riaboff[a,b], Guillaume Baffet, Alain Le Duff[a,b,1], Nicole Meslier[c,d], Frédéric Gagnadoux[c,d] and Jean-Marc Girault[a,b]

[a]ESEO, Angers, France
[b]LAUM, UMR CNRS 6613, Le Mans, France
[c]Angers sleep laboratory, University Hospital, Angers, France
[d]INSERM UMR 1063, University of Angers, Angers, France

ARTICLE INFO

Keywords:
The SATUD system
Data-dependent features
Unsupervised thresholding
Expert knowledge
Interpretable classifier
Sleep scoring

ABSTRACT

Manual sleep stages scoring is time-consuming, complex and requires specific medical knowledge. Automatic sleep stages classification, usually based on supervised methods of machine learning, is the object of researchers interest. However, it remains challenging because of the high variability among patients which is not considered with such algorithms. This paper presents a method to extract patient-dependent qualitative features from electrophysiological signals, preceding a supervised machine learning classifier. Instead of using fixed thresholds, the developed method called "Self-Adaptative Thresholding Using Descriptors" (SATUD), proposes an unsupervised self-adjusting thresholding. Thresholds are automatically adjusted to maximize both the similarity within a same sleep stage and the dissimilarity between different ones. This method is evaluated using manual sleep stages scoring from 60 patients with various pathologies to ensure high variability. The SATUD shows a better adaptation to the patient specificities, compared with two other thresholding methods implemented in this study. Indeed, the number of 30-seconds recording segments respecting all their sleep stage properties increased by more than 80 % with the use of the SATUD, compared to other thresholding techniques. It was also proved robust to noise and sweat artifacts. The SATUD thereby provides patient-dependent qualitative features which can be used for automatic sleep stages scoring using a machine learning method. This last point was presented in the companion paper.

1. Introduction

Sleep-related disorders nearly affect the third of the population and are increasingly recognized as real public health problems [1]. The need of sleep diagnosis has then increased during the last few years [2, 3].
Gold-standard procedure for sleep diagnosis consists of electrophysiological and respiratory signals recording with the use of a polysomnograph. With all those signals, medical staff manually scores both sleep events and sleep stages. Sleep events, as apneas and hypopneas, are detected through respiratory signals. On the other hand, sleep stages are scored based on the electrophysiological signals. Among them, electroencephalograms (EEG), electrooculograms (EOG) and electromyograms (EMG) are used for the measurement of cerebral, ocular and muscular activities, respectively. Sleep scoring consists on the classification of wakefulness, stage N1 and stage N2 (both light sleep), stage N3 (deep sleep) and REM sleep (Rapid Eye Movement, also called R stage or paradoxical sleep: stage with active brain but reduced muscle tone) in 30-second sections, also called epochs. Besides being a time-consuming task, sleep scoring requires specific medical knowledge. A manual of recommendations published by the American Academy of Sleep Medicine (AASM) in 2007 [4] describes the temporal and spectral contents of each sleep stage, as well as sleep patterns that can be recognized by the scorer and the possible transitions from one sleep stage to another.

There was a growing number of automatic sleep scoring algorithms developed those last few years, spread out into three categories: deep learning [5, 6], machine learning [7–13] and hybrid approaches [14–16]. Despite the increasing number of models based on artificial intelligence (AI), only a few are routinely employed by sleep specialists. Several reasons for that have been identified and detailed in the companion paper [REF] and in [17]. One of them is the lack of transparency of the developed approaches. Indeed, deep learning approaches, which often reach the best scores, are opaque and their lack of transparency raises skepticism among physicians. With opaque approaches, medical practitioners must accept to loose their control upon the task that is realized by the algorithm. It could prevent them to properly react when dealing with an unusual pathology (that was not necessarily represented when training the model). For this reason, some researchers attempt to improve the interpretability of their models [18], and provide some concrete elements for the practitioner to relate to. Hybrid approaches [14–16] were then designed to overcome several identified limitations, including the lack of transparency of the implemented method.

The present article is the second part of a two-part paper. In the companion paper [REF], a novel hybrid approach replicating the steps of a manual scoring was developed and

*Corresponding author
jade.vanbuis@eseo.fr
ORCID(s): 0000-0001-6437-1597 (J. Vanbuis)
[1]Present address: Olympus NDT Canada, Québec, Canada

presented. This hybrid system is composed of several functions, including one dedicated to features extraction, necessary for classification. Called SATUD, this features extractor will be fully detailed in the following section, since it is the core of the present paper. Extracted features describe each epoch. They are qualitative and represent concrete elements mentioned in the AASM guidelines. However, the extraction of such features can be problematic for classification if not correctly carried out. Indeed, there is a strong variability within subjects. This variability is one of the major difficulties encountered by sleep specialists while manually scoring sleep. To score sleep stages appropriately, medical practitioners generally need to get familiar with the recording specificities. To do so, they start with a quick visualization of the entire recording, before doing the scoring. However, amongst the many automatic scoring algorithms developed the last few years [5–16], only a few were adapted to the variability between subjects [14–16]. Furthermore, adaptation to each recording specificities implies having a sufficient number of recordings from patients with various pathologies, which is rarely the case.

This study presents the Self-Adaptative Thresholding Using Descriptors (SATUD) method. Developed for the extraction of subject-dependent features, the SATUD reduces the impact of subjects variability. Features values are adjusted to each patient, as sleep specialists naturally do when scoring sleep. The SATUD principle is explained in the following section, and its ability to correctly adjust thresholds is then evaluated and compared with other thresholding methods.

2. Material and method

In this section, the data used for the implementation of the SATUD is first presented (2.1). The chosen approach is then detailed in Section 2.2.

2.1. Data

Patients with various sleep pathologies underwent one-night polysomnography (PSG) in Angers sleep laboratory (University Hospital, INSERM UMR1063, Angers, FRANCE). PSG is a sleep diagnostic device acquiring electrophysiological signals, respiratory signals, body movements and position. Recordings were part of the "Institut de Recherche en Santé Respiratoire des Pays de la Loire" [IRSR] sleep cohort. Approval was obtained from the University of Angers ethics committee and from the "Comité Consultative sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé" [CCTIRS] (07.207bis). All patients included in the IRSR sleep cohort have given their written informed consent. Recordings were anonymous. Sleep stages and events were recorded and scored following the AASM recommendations [4] using CID102L8D polysomnographs. Besides those standard polysomnographic signals, tracheal sounds were recorded using a PneaVoX® device for enhanced ventilatory events recognition [19]. A total of 60 anonymous patients recordings was scored by sleep specialists (three sleep specialists were involved in the study, but each recording was scored by a single scorer). In this study, the automatic algorithm only employed EEG, EOG and EMG signals used by physicians to score sleep stages. The sequence of wake, N1, N2, N3 and REM sleep in epochs of 30 seconds is called hypnogram and constitutes our reference.

2.2. Method

The hypothesis made in this paper is that concrete features can help overcome the model lack of transparency. Qualitative (ordinal) features allow close translation of medical knowledge (AASM recommendations). Their estimation requires the discretization of quantitative features. For example, the quantitative feature *amplitudeEEG*[1] is discretized into qualitative features as *amplitudeEEGHigh* and *amplitudeEEGLow*, employing a certain number of thresholds needing to be estimated. For this example, the problem is to know the threshold for which we can consider having a low or high EEG amplitude. The SATUD algorithm aims to adjust those thresholds automatically, for each recording. The integration of the SATUD in the sleep staging system implemented in this study is schematized in Figure 1. The architecture is composed of several main functions: **F1**, **F2**, **F3** and **F3.A**. A detailed example can be found in Appendix A.

2.2.1. SYSTEM

The proposed system aims to provide a set of patient-specific qualitative features. As described in Figure 1, it was composed of three main functions (*F1*, *F2*, *F3*). For applications concerning sleep staging, the system inputs were:

- **a priori knowledge** giving information about sleep scoring through AASM manual. For example, this manual indicates that N3 sleep stage can be recognized using low-frequency (0.5-2 Hz) high-amplitude ($> 75\,\mu V$) EEG frontal waves. A high proportion of those waves called 'Slow wave activity', indicates N3;
- **electrophysiological signals** obtained from the recorded polysomnographies, and more precisely three electroencephalograms (EEG), two electrooculograms (EOG) and the chin electromyogram (EMG).

2.2.2. F1 - Sleep stages description

F1 aims to build one list of properties for each sleep stage. The a priori knowledge from AASM manual (as detailed in SYSTEM) was the input of the first function F1. In order to build the list, we carefully studied the AASM manual for all sleep stages, and translated subsequently into lists of properties which represent time and frequency information used to differentiate sleep stages. The upper part of the Figure 2 shows an example of two properties that were used to describe sleep stage N3. The lists of properties associated with each sleep stage are essential for the SATUD proper functioning, as they replace labelled data.

[1] Note the signals were recorded with a bit-depth of at least 8 bits.

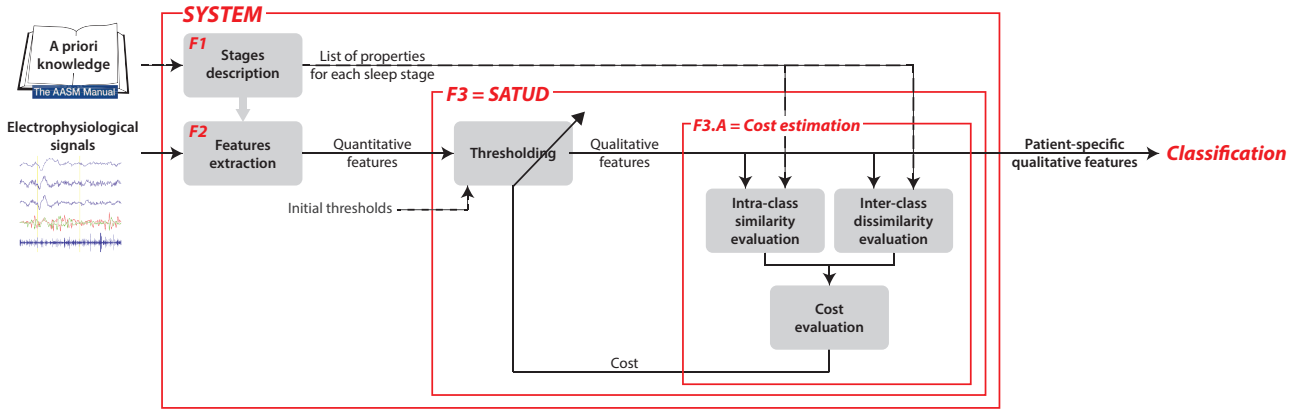
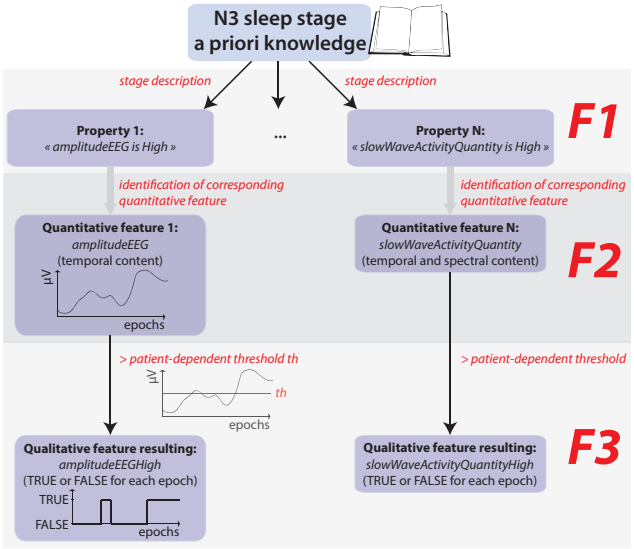


Figure 1: Functional architecture composed of three main functions (F1, F2 and F3). Medical knowledge from the AASM manual and electrophysiological signals are used as inputs. The output is a set of qualitative features with less vulnerability to patients specificities.



Slow wave activity consists of low-frequency (0.5 - 2 Hz) high-amplitude (peak-to-peak >75 µV) EEG waves measured over the frontal region of the brain.

Figure 2: Simplified example illustrating the links between sleep stage N3 properties (defined in F1), quantitative features (defined in F2) and qualitative features (defined in F3).

2.2.3. F2 - Features extraction

F2 aims to extract the quantitative features corresponding to the properties identified in function F1. Using the electrophysiological channels split into epochs, 13 quantitative features were extracted. Those quantitative features are listed in the 1st column of Table 1. Those quantitative features can reflect temporal or spectral content, but also a combination of both temporal and spectral content. They were identified in the AASM guidelines as being required for sleep scoring. In the middle part of Figure 2 are presented two quantitative features identified thanks to the previous properties.

2.2.4. F3 - SATUD

The SATUD aims to deduce patient-specific information from quantitative features using specific knowledge. It decreases subjects variability with the estimation of patient-specific thresholds. Thresholds were applied on the 13 quantitative features obtained in F2 to generate 41 qualitative features. The qualitative features and the associated number of thresholds were chosen in agreement with the properties described in F1. To do so, for each quantitative feature from F2, and depending on the properties from F1 (which translate the AASM guidelines), the appropriate qualitative features were extracted[2]. Those qualitative features are listed in the 3rd column of Table 1.

Different ways of defining initial thresholds were tested. In this application, initial thresholds were chosen using a statistical approach (percentiles). Thresholds values were then adjusted in order to minimize the cost function described in F3.A. To do so, several meta-heuristics were tested: global search algorithms such as simulated annealing [20] and genetic algorithms [21, 22] and also local algorithms such as gradient descent methods [23, 24]. For our application, the use of local search algorithms alone was ineffective due to the number of thresholds to adjust. We thus chose to use a global search algorithm to initialize the search zone, but still combined it with a final local search algorithm for better precision. As indicated in the pseudo-code reported in Appendix B, the method chosen for this study was the combination of simulated annealing followed by a gradient descent.

As there is a high variability between subjects, optimal thresholds values could be very different from one patient to another. Thresholds adaptation to each patient is explained in the following section.

[2]Only qualitative features which correspond to sleep stages properties described in the AASM guidelines were computed, even if the number of thresholds allowed the estimation of more features. For example, 2 thresholds (Low → Mid and Mid → High) can generate 6 qualitative features: low, low or mid, mid, mid or high, high and low or mid or high. However, the study of the AASM guidelines rarely indicates there is a need for all those features. Only few of them can be sufficient to translate the guidelines.

Table 1
List of the 41 qualitative features extracted for sleep scoring, depending on the 13 quantitative features extracted in F2 (ranging from EEG amplitude to Substracted EOG instability) and the number of associated thresholds.

Quantitative features	Th[a]	Qualitative features used	N[b]
EEG amplitude	2	Low | Low or Mid | Mid or High | High	4
EEG instability	1	No | Yes	2
Slow wave activity quantity	2	Low | Low or Mid | High	3
Alpha waves quantity	2	Low | Low or Mid | Mid | Mid or High | High	5
Beta waves quantity	2	Low | Mid | Mid or High	3
Delta waves quantity	2	Low | Mid or High	2
Theta waves quantity	2	Low | Low or Mid | Mid or High	3
Chin level	2	Low | Low or Mid | Mid or High | High	4
Chin instability	2	Low | Low or Mid | Mid or High | High	4
Summed EOG level	2	Low | Low or Mid | Mid or High	3
Summed EOG instability	2	Low | Low or Mid | Mid or High	3
Substracted EOG level	2	Low | Mid or High	2
Substracted EOG instability	2	Low | Mid or High | High	3
		Total	41

[a] Number of Thresholds employed.
[b] Number of qualitative features used for each quantitative feature.

2.2.5. F3.A - Cost estimation

The final cost was defined as a weighted sum of the costs associated to each class (sleep stage in this study):

$$finalCost = \sum_{c=1}^{L} w_c \times cost_c \quad (1)$$

where L is the number of classes, and w_c and $cost_c$ are the weight and cost associated to the c^{th} class respectively. Weights are optional. In our application, they have been chosen to promote sleep stages that are difficult to identify or demote the ones that occur rarely (N1 sleep stage only represents approximately 5 % of the night).
$cost_c$ was defined as:

$$cost_c = \frac{1}{conc(R_{P1}, R_{P2}, ..., R_{PN}) \times std(antiScore_c)} \quad (2)$$

with:

$$antiScore_c = \sum_{i=1}^{N} v_i \times \overline{R_{Pi}} \quad (3)$$

where N is the number of properties of class c, R_{Pi} is a binary variable representing the respect of the i^{th} property of class c (as explained hereafter), and v_i is the weight associated to the i^{th} property of class c. This time, weights have been chosen to better translate the AASM guidelines. Indeed, some properties are clearly indicated as being more important for sleep staging.
The cost function is minimized for each recording individually. Defined as above, it is equivalent to i) maximize the similarity between the epochs of a same sleep stage of the same patient, hereinafter referred to as intra-class similarity and ii) maximize the differences between the epochs belonging to different sleep stages of the same patient, hereinafter referred to as inter-class dissimilarity.

i) **maximize intra-class similarity**: intra-class similarity is optimized by maximizing $conc(R_{P1}, R_{P2}, ..., R_{PN})$. This term represents the concordance between the respect of the properties used to describe class c. The respect of a property is assessed using the corresponding qualitative feature. For our application, the concordance between the respect of the properties was estimated using the Fleiss' Kappa [25]. This is a statistical measure used to assess the reliability of agreement between several categorical vectors. Indeed, we can consider that when a patient is in a specific sleep stage, most of the properties related to this sleep stage are respected. When the sleep stage changes, a certain number of those properties will no longer be respected. If we focus on the properties related to a particular sleep stage, the transitions to and from this sleep stage will be highlighted by a simultaneous change of the respect of its properties. Those simultaneous changes can be evaluated by assessing its properties respects interdependency. Therefore, a high interdependency will indicate a general concordance between thresholds. Thresholds were adjusted unsupervisingly, until they agreed and confirmed each other, leading to a higher concordance measure (Fleiss' Kappa). This is illustrated in Figure 3.

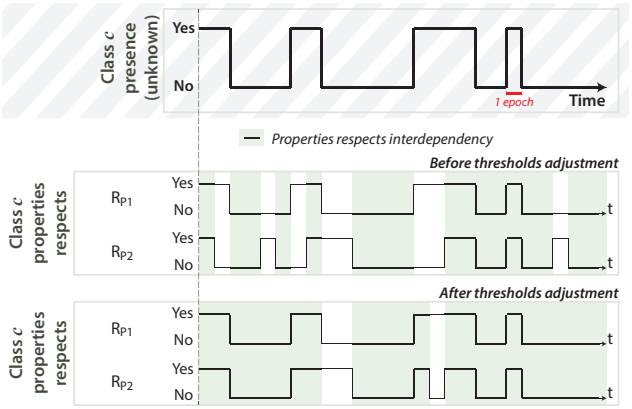


Figure 3: Simplified example of the properties respects behaviour before and after thresholds adjustment. Class c is described by only two properties ($P1$ and $P2$). The interdependency between properties respect R_{P1} and R_{P2} increases when thresholds are being adjusted.

In this example, that has a length of 26 epochs, class c presence is unknown for thresholds adjustment (unsupervised functioning). The better the thresholds are, the more properties are respected (R_{P1} and R_{P2}) when and only when the patient is in class c. Before thresholds adjustment, 18 epochs

among the 26 were in concordance with each other (highlighted areas). This number increased from 18 to 23 after thresholds adjustment, leading to a better Fleiss' Kappa.

ii) **maximize inter-class dissimilarity**: inter-class dissimilarity is optimized by maximizing $std(antiScore_c)$. The anti-score function is defined as a weighted sum of the no-respect of a class properties (Equation 3). Weights are optional. For our application, they were empirically chosen to translate the degree of importance according to the AASM manual.
$std(antiScore_c)$ represents the fluctuation (using standard deviation value) of class c anti-score function. The anti-score function of a particular sleep stage varies from 0, when all the properties associated to this sleep stage are respected, to 1 when none of the properties are respected. When thresholds are being adjusted, the anti-score values are getting closer to its limits (0 when in the sleep stage and 1 when in another sleep stage), with fewer values in between. Consequently, the anti-score standard deviation on the entire recording increases. This is illustrated in Figure 4, which pursues the example presented in Figure 3.

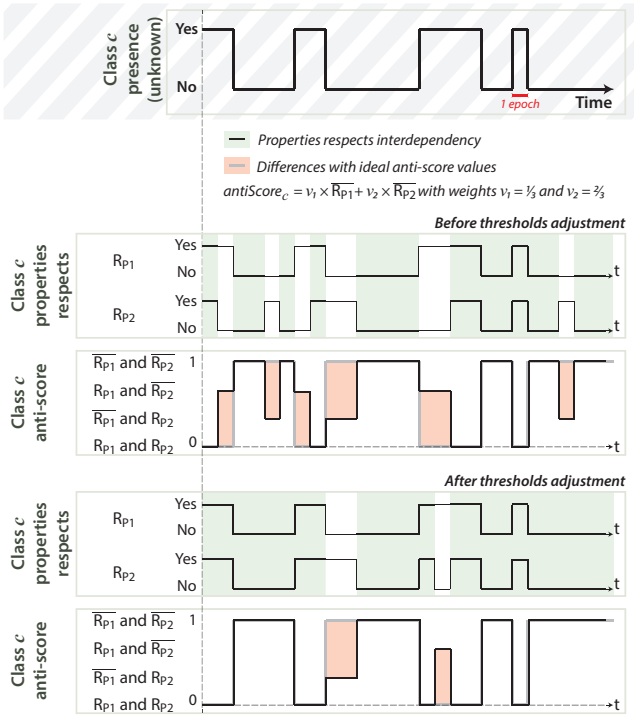


Figure 4: Simplified example of the anti-score behaviour before and after thresholds adjustment. Class c is described by only two properties ($P1$ and $P2$). The anti-score is a weighted sum of the respect of each property (R_{P1} and R_{P2}). Its standard deviation value increases when thresholds are being adjusted.

In this example, the anti-score standard deviation increased from 0.41 to 0.47 with thresholds adjustment.

To summarize, the cost depends on the intra-class similarity and inter-class dissimilarity, both associated with the respect of each class properties. The respect of each class properties are evaluated from the qualitative features values. Those values are directly related to the thresholds. For each recording individually, and without using the manual scoring as the reference, the thresholds are thus adjusted by minimizing the cost. At the end of the process, the patient-specific qualitative features are extracted (using the final thresholds) and can be used for classification.

3. SATUD evaluation methods

This section is dedicated to the assessment of the SATUD performance, which was evaluated from the totality of the 60 recordings (since the method is unsupervised). The SATUD was employed to extract qualitative features that are understandable and represent concrete information from the AASM guidelines. It was thus compared with other thresholding methods employed the same way. The complete system reported in Figure 1 was replaced by GST (General Statistical Thresholding) and IST (Individual Statistical Thresholding) as described thereafter.

- *General Statistical Thresholding (GST)*: with this method, thresholds were adjusted using statistical information of the entire database, as percentiles (thresholds were not patient-dependent).
- *Individual Statistical Thresholding (IST)*: with this method, thresholds were adjusted for each patient depending on statistical information (thresholds were patient-dependent).

Impact on the obtained qualitative features was first estimated (3.1). The SATUD behaviour in presence of noise or artifacts was then tested (3.2) and, eventually, the impact on classification was evaluated (3.3).

3.1. SATUD impact on qualitative features

Using the lists describing each sleep stage, we quantified the agreement between each epoch qualitative features and the properties of the associated sleep stage. For example if there were 10 properties in the list describing sleep stage Wake, then a Wake epoch with qualitative features respecting only 5 of them had a global respect value R of 50 % with its sleep stage properties.
Outcomes will be presented in Section 4.1.

3.2. Robustness test

To evaluate the robustness of the SATUD, we assessed the quantity of epochs highly respecting the properties associated to their sleep stage under several situations:

- *noise amplification*: white Gaussian noise was added to all raw electrophysiological signals. Raw signals presented a native Signal-to-Noise Ratio (SNR) of approximately 46 dB. Several SNR levels were tested, ranging from 30 dB to 0 dB, in 10 dB intervals.
- *artifacts addition*: several artifact types are usually present on electrophysiological signals. All artifacts were kept in our recordings to evaluate our algorithm under real-life

conditions. To further test the robustness of the method, we added artificial artifacts to the signals. In this study, we chose sweat artifacts, considered as the most disruptive ones. Indeed, during such artifacts, both temporal and spectral information are compromised or even lost. Artificial sweat artifacts were created using random slopes and durations within intervals defined after visualization of several natural sweat artifacts. A duration limit was used to prevent an entire 30-seconds segment to be exclusively composed of sweat artifacts. Natural sweat artifacts occur more often in some sleep stages compared to others and their quantity is generally limited [26, 27]. For this reason, we tested the results with the addition of artificial artifacts on 0 % (raw signals) to 100 % of the epochs composing recordings. Figure 5 presents natural (5a) and artificial (5b) sweat artifacts examples.

Outcomes will be presented in Section 4.2.

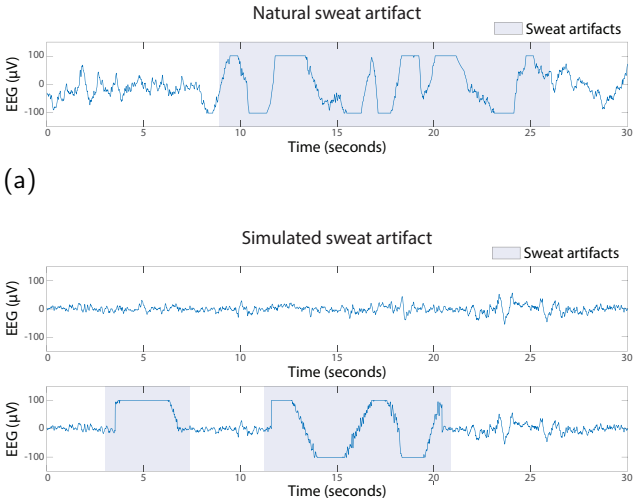


Figure 5: Examples of EEG signals during 30-seconds epochs: (a) an epoch with natural sweat artifacts and (b) an epoch with added simulated sweat artifacts.

3.3. SATUD impact on classification

Patient-dependent qualitative features described in the previous section are used for sleep stages classification. To properly estimate the impact of the SATUD, a first simple classification model was tested in this paper and compared with manual sleep scoring. The implementation of an advanced classifier was presented in the companion paper [REF]. However, it is necessary to assess the SATUD efficiency by avoiding any bias that could be linked to the use of powerful classification models. The classifier described in the current paper was built to be relatively transparent and easy to understand. It did not required training and testing steps. Indeed, using the adjusted thresholds, percentages of agreement with each sleep stage list of properties were evaluated for each epoch. The sleep stage with the higher agreement rate between the properties and its patient-dependent qualitative features was the chosen one. In the current paper, results were simply expressed in term of accuracy rate with the manual scoring, named *Acc*:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

where *TP*, *FN*, *FP* and *TN* represent the number of true positives, false negatives, false positives and true negatives, respectively.
Outcomes will be presented in Section 4.3.

4. Results

The SATUD performance was estimated through the qualitative features generated (see Section 4.1) and compared with the two other thresholding methods: GST and IST. Robustness to noise or artifacts tests results were then reported in Section 4.2. The impact on classification were evaluated in Section 4.3.

4.1. SATUD impact on qualitative features

Figure 6 shows the number of epochs that respect at least 0 % to 100 % of their sleep stage properties, according to the thresholding method used. Of course, all epochs respected at least 0 % of their sleep stage properties whatever the method used. With the SATUD, the number of epochs respecting at least 60 % to 100 % of the properties associated with its sleep stage were higher than with GST and IST. It means that qualitative features obtained with the SATUD better respect the properties expected for their sleep stage. There were indeed relative increases of 81 % ($\frac{8135-4506}{4506}$) and 89 % ($\frac{8135-4306}{4306}$) of the number of epochs that perfectly respect the properties associated with their sleep stage when using the SATUD compared to GST and IST, respectively.

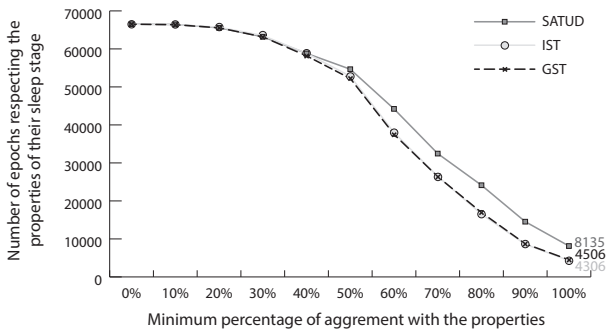


Figure 6: Evolution of the number of epochs respecting at least 0 % (the total number of epochs is obtained) to 100 % of the properties associated to their sleep stage, depending on the method used.

When focusing on epochs that highly ($60\,\% < R \leq 80\,\%$) or almost perfectly ($R > 80\,\%$) respect the properties associated with their sleep stage, we compared the different methods depending on each sleep stage. Figure 7 shows the impact of the SATUD on those epochs compared to GST and IST. When comparing the SATUD with GST and IST, we registered improvements for sleep stages W, N2, N3 and REM sleep but not for N1 sleep stage.

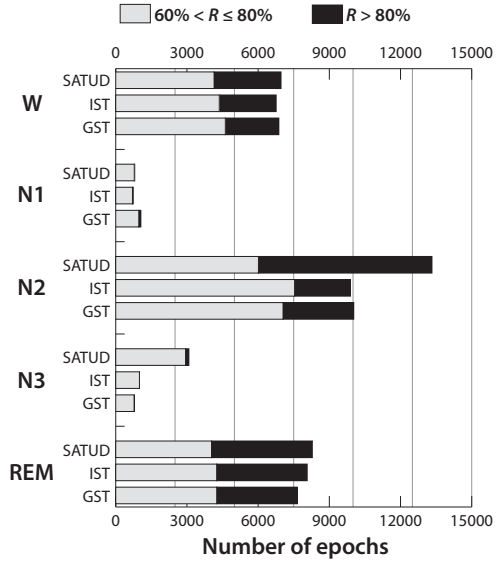


Figure 7: For each sleep stage, quantity of epochs highly ($60\,\% < R \leq 80\,\%$) or almost perfectly ($R > 80\,\%$) respecting the properties associated with their sleep stage, according to the method used.

4.2. Robustness test

Robustness tests described in Section 3.2 were conducted on GST, IST and the SATUD methods.

We first studied noise and sweat artifacts impact on qualitative features. The quantity of epochs that almost perfectly respect the properties associated with their sleep stage ($R > 80\,\%$) were evaluated. Results were compared for different levels of noise and artifacts for GST, IST and the SATUD (Figure 8). The SATUD behaviour to noise was globally similar than GST and IST, but with higher scores. Almost no N1 and N3 epochs respected almost perfectly the N1 and N3 properties. The quantity of epochs that almost perfectly respect the properties associated with their sleep stage dropped from a SNR level of 10 dB for all sleep stages except for Wake, where it remained quite constant with GST and IST, and slowly decreased for the SATUD.

Considering the addition of artificial sweat artifacts, the use of the SATUD once again positively impacts the respect of sleep stages with their properties. However this time, it behaved differently than GST and IST. For those last ones, artifacts had a very small impact, with a slight increase for Wake and N2 epochs and a progressive decrease for REM sleep. Using the SATUD, N2 and N3 epochs underwent an improvement whereas adding sweat artifacts until they are applied on 30 % - 40 % of the recording epochs. Afterwards, it decreased progressively.

The SATUD obtained better results than GST and IST in terms of quantity of epochs respecting almost perfectly properties associated with their sleep stage. While remaining superior, its behaviour in noise and artifact situations was globally similar to GST and IST thresholding methods.

Table 2
Accuracy rate with the manual scoring according to the different thresholding methods while testing robustnes.

	SATUD	IST	GST
	Robustness to noise test results		
Raw signals	55 %	46 %	45 %
30 dB SNR	55 %	46 %	45 %
20 dB SNR	54 %	46 %	45 %
10 dB SNR	44 %	34 %	34 %
00 dB SNR	43 %	32 %	33 %
	Robustness to sweat artifacts test results		
Raw signals	55 %	46 %	45 %
20 %	55 %	45 %	41 %
40 %	55 %	45 %	37 %
60 %	54 %	44 %	36 %
80 %	54 %	44 %	35 %
100 %	54 %	44 %	33 %

4.3. SATUD impact on classification

Classification global accuracy rates were estimated for all three methods. For raw signals, the SATUD obtained the best agreement with the reference with $Acc = 55\,\%$, versus $Acc = 45\,\%$ and $Acc = 46\,\%$ for GST and IST respectively (Table 2). Confusion matrices are shown in Figure 9. The SATUD confusion matrix (Figure 9a) registered significant improvements for sleep stages N2 and N3, if compared to IST and GST (Figure 9b and Figure 9c). Results for different levels of noise and artifacts are reported in Table 2. The SATUD obtained the best results for all levels of signals degradation. For IST and the SATUD, sweat artifacts did not have an important impact on the classification scores.

5. Discussion

The method developed in our study, called SATUD (Self-Adaptative Thresholding Using Descriptors algorithm), aims to facilitate data-dependent features obtainment to be used for classification. It automatically and unsupervisingly adjust thresholds by minimizing a cost function. The cost function was determined based on mathematical reasoning and translation of the expert knowledge. For our application, the SATUD was employed to reduce the impact of subjects variability before automatic sleep staging. It reproduces the first screening done by experts when manually scoring sleep.

This pre-processing step proved its value since the SATUD showed an improvement of epoch agreement with the properties associated with their sleep stage, compared to two other thresholding methods. Compared to GST (General Statistical Thresholding) and IST (Individual Statistical Thresholding), the quantity of epochs almost perfectly respecting their sleep stage properties increased from less than 3000 to more than 7300 epochs with the use of the SATUD for sleep stage N2. This stage had the best classification improvement. We can consider that the SATUD highlighted this sleep stage, probably because it represents almost 50 % of a night. On

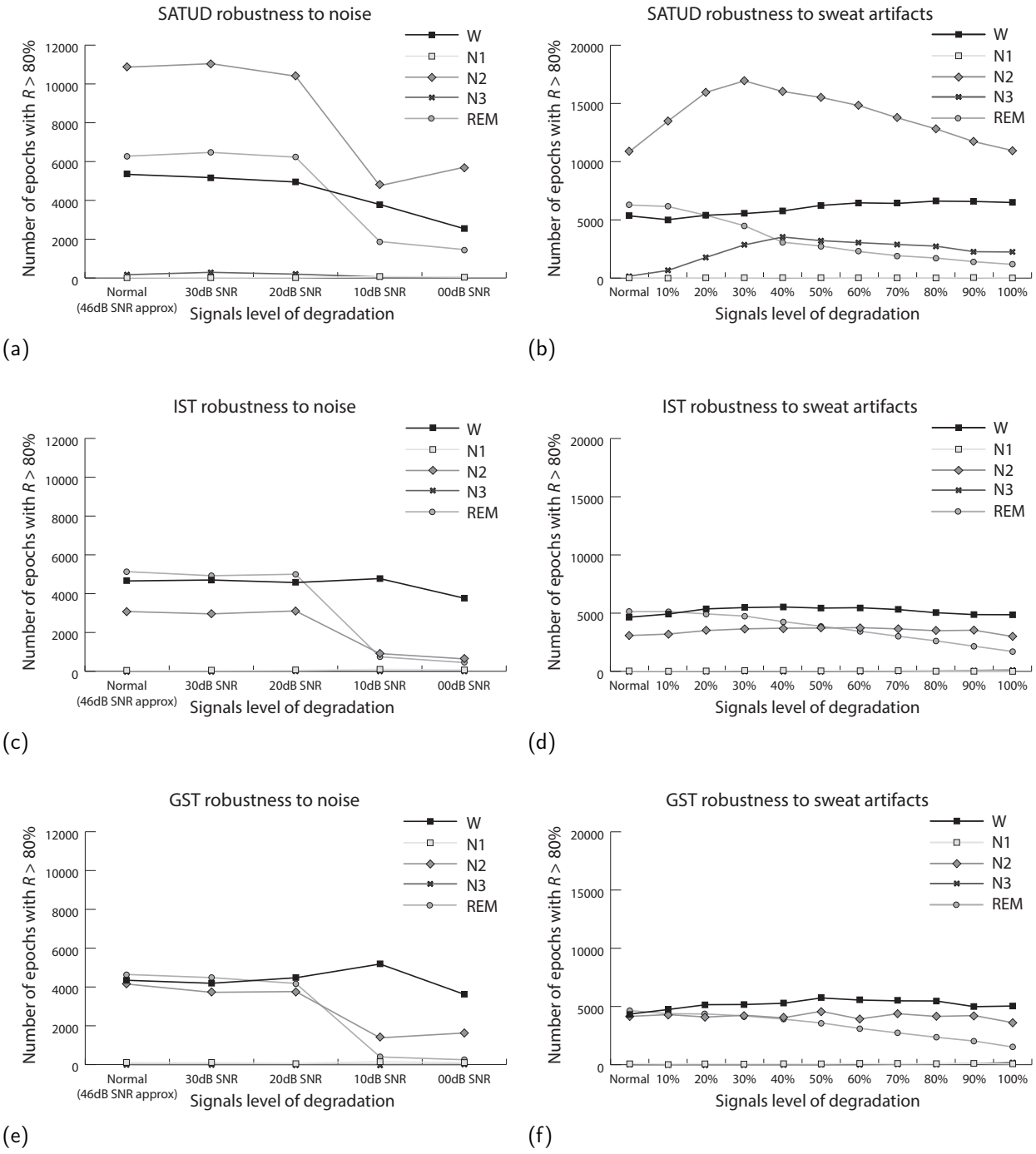


Figure 8: Evolution of the number of epochs with an almost perfect agreement ($R > 80\,\%$) with the properties associated to their sleep stage for: (a) the SATUD thresholding tested with increasing noise, (b) the SATUD thresholding tested with increasing number of artifacts, (c) IST thresholding tested with increasing noise, (d) IST thresholding tested with increasing number of artifacts, (e) GST thresholding tested with increasing noise and (f) GST thresholding tested with increasing number of artifacts.

the opposite, and for all methods, N1 was the sleep stage with the worst results. This could be explained by the fact that N1 is a transitional stage that occurs during only 5 % of the night. This limitation does not have a significant impact on the classification performance because N1 appears to be rare, and N1 errors are common in literature, even within manual scorers [28]. The primary classification tool tested also showed higher results for the SATUD than GST and IST. Improvements were more important for the N3 sleep stage. The proposed method also showed a good performance regarding noise and artifacts. The SATUD was not sensitive to additional white noise until the SNR level of 10 dB was reached. Note that this is an important noise level that would be difficult to reach using sensors as used in polysomnog-

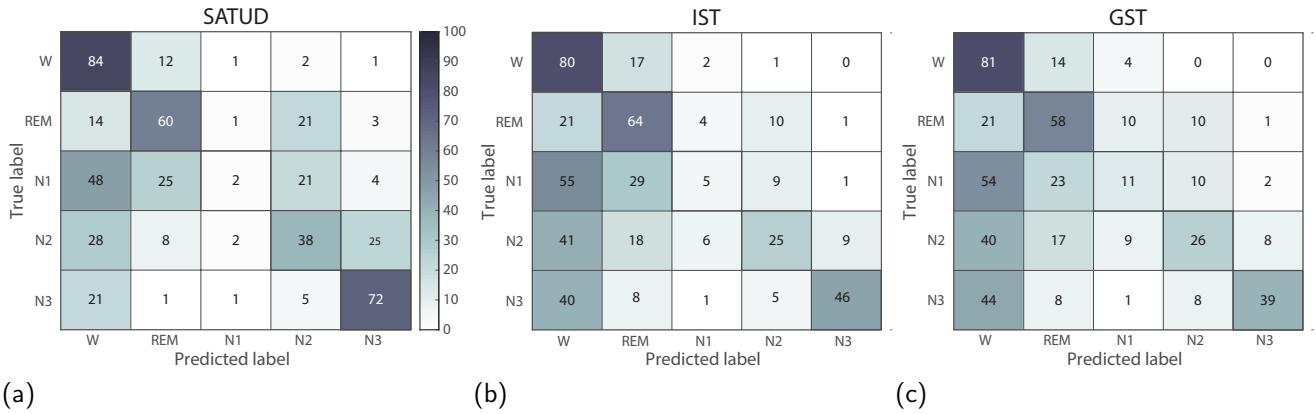


Figure 9: Classification agreement rate (percentage) between the predicted and true labels for: (a) the SATUD thresholding method, (b) IST thresholding method and (c) GST thresholding method.

raphy. Moreover such noise would also make the manual scoring a lot more complicated, leading to a probable invalidation of the recording. Artificial sweat artifacts did not have an important impact on the SATUD and IST classification agreement rate. Regarding the SATUD, they however tended to increase the number of epochs almost perfectly respecting the properties associated with their sleep stage for sleep stages N2 and N3, until a limit around 30 %-40%. It could be explained by the fact that one property of those sleep stages is the high amplitude, more often respected with the addition of sweat artifacts. Also, the number of Wake epochs almost perfectly respecting Wake properties increases with sweat artifacts when using the SATUD. The reason would be that agitated wake can make saturation appear on signals, like in presence of sweat artifacts.
The proposed method showed its efficiency, especially since the database used is composed of many patients with various pathologies.

Based on medical knowledge, this method was built to contribute to the development of a user-friendly sleep staging system. As discussed in the companion paper [REF], very few sleep staging systems are considered by physicians because of several limitations. One of them is the lack of transparency of models, often considered as black boxes. To overcome this limitation, a methodology replicating the manual scoring process was implemented in the companion paper [REF]. To give the medical practitioner concrete elements to relate to, qualitative features were chosen as the input of the classifier. However, such features can be problematic if not adjusted to each patient. The SATUD was designed to extract patient-dependent qualitative features, without the need of previous partial scoring by a sleep specialist. The SATUD method is potentially transposable to other applications. It however requires a good knowledge of the specific field, with the identification of a maximum of rules describing each class. Those numerous rules are key points for the SATUD functioning, enabling the use without the need of knowledge on the classification output. For this reason, this algorithm can run under real time conditions.

6. Conclusion

This paper presents an approach to extract features from a dataset. This approach, called SATUD, is at the core of the companion paper [REF], devoted to an automatic sleep staging for polysomnographic recordings.

The SATUD method was the outcome of an approach combining mathematical reasoning and medical knowledge. Compared to other thresholding method, the SATUD allowed a better generalization of features depending on each sleep stage. The performance proved that resulting features are significantly in agreement with the expected sleep stage properties. A straightforward classification model was also developed (to test the SATUD without being biased by the abilities of a complex classifier). Reported results were better for the SATUD, compared to the other thresholding methods. This performance, obtained on 60 patients with various sleep pathologies, confirmed that this approach is suitable for sleep staging. Tests with noise and artifacts showed this algorithm had a sufficient robustness for the application.

The companion paper [REF] presented an entire and user-friendly classification model based on the extracted features. A detailed analysis of obtained classification impact on each patient diagnostic is also worthy of investigation.

Funding

This work was supported by grants from the Institut de Recherche en Santé Respiratoire des Pays de La Loire.

Acknowledgment

The authors would like to thank Christelle Gosselin and Jean-Louis Racineux, from the Institut de Recherche en Santé Respiratoire des Pays de La Loire. We thank Julien Godey, Laetitia Moreno and Marion Vincent, sleep technicians in the Department of Respiratory and Sleep Medicine of Angers University Hospital. We thank Roberto Longo for its contribution.

A. Simplified example of the SATUD employment

This appendix presents a simplified version of the use of the SATUD for sleep stage classification. We thus consider only two patients, Patient1 and Patient2 and we admit they had a similar night. For a better understanding, we merged sleep stages N1, N2 and N3 into the so-called NREM sleep. We also considered in this appendix that wake was only composed of calm wake with eyes closed. The simplified hypnogram is represented in Figure A.1.

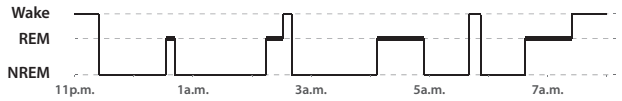


Figure A.1: Simplified hypnogram presenting the nights of patients Patient1 and Patient2.

The goal is then to retrieve the hypnogram using each patient signals, knowing that there is a variability between each patient. A highly simplified version of sleep stages description (**F1**) is presented in Table A.1. For this example, only three properties were used to describe wake, REM sleep and NREM sleep.

The SATUD aims to estimate the thresholds following the properties indicated in Table A.1. Then, a classification tool will be able to rebuild the hypnogram. For both Patient1 and Patient2, thresholds have to be adjusted until *H - L - L* $\leftrightarrow$ *Wake*, *M - MH - M* $\leftrightarrow$ *REM* and *L - L - H* $\leftrightarrow$ *NREM*.

If thresholds are well adjusted, each patient properties will be respected, as in Figure A.2. Because of the rapid eye movements that is Low during wake but also during NREM, we can see that the properties respects interdependency is not at 100 % for Wake and NREM sleep.
If thresholds are not correctly settled, and for example the rapid eye movements threshold is too high, the property *rapid eye movements is Mid or High* associated to REM sleep will never be respected. The result on the properties respects interdependency of REM sleep would then be deteriorated, as shown in Figure A.3a. For each sleep stage, the properties respects interdependency has to be maximized to find the best thresholds. However thresholds that are too far from correct values (highly erroneous) can generate a situation where the properties respects interdependency would be maximum as in Figure A.3b. To prevent this situation, fluctuation has been considered. Indeed, we can assume that all sleep stages will appear during a whole night. To estimated fluctuation, anti-scores were created using a weighted sum of the disrespect of each property.

Table A.1
Simplified sleep stage description.

	Alpha waves	Rapid eye movements	EEG amplitude
Wake[a]	High	Low	Low
REM	Mid	Mid or High	Mid
NREM	Low	Low	High

[a] Only wake with eyes closed was considered for this example.

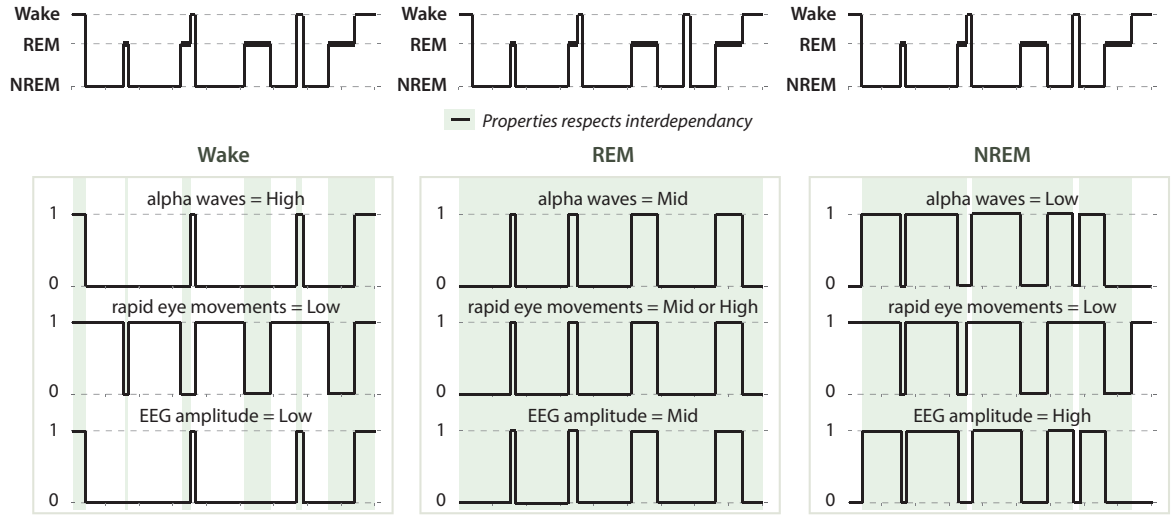


Figure A.2: Example of properties respects for each sleep stage, when thresholds are perfectly adjusted. Using Table A.1, we understand that even with perfectly adjusted thresholds, there is not always 100 % interdependency between properties respects. This is due to the possibility of some properties to be respected in different sleep stages.

Informations supplémentaires : les listes de propriété

Les listes de propriétés associées à chaque stade et utilisées par SATUD sont décrites dans cette partie. Pour traduire au mieux les recommandations AASM, des poids ont été attribués à chaque propriété (1 étant le moins important et 10 le plus important).

De la même manière que pour le projet HypnoLighT, plusieurs types d'éveil ont été différenciés : l'Éveil Yeux Fermés (EYF), l'Éveil Yeux Ouverts (EYO) et l'Éveil Agité (EA). La distinction entre ces types d'éveil permet une meilleure traduction des recommandations AASM. Ainsi, une époque d'éveil n'est pas nécessairement caractérisée par des ondes alpha.
Ces différents types d'éveil sont uniquement considérés de manière indirecte lors de la classification manuelle, et regroupés dans le stade unique Éveil. De façon identique, l'algorithme proposé dans ce travail utilise 7 listes de propriétés (il différencie les différents types d'éveil) mais ne fournit en sortie qu'une classification en 5 stades.

Table D1 – Liste des différentes propriétés (caractéristique et état attendu) de chaque stade et poids associés.

	Caractéristique	État attendu	Poids
Éveil Agité (EA)	Amplitude EEG	Haute	10
	Instabilité EEG	Oui	10
	Quantité d'activité en ondes lentes	-	-
	Quantité d'ondes alpha	-	-
	Quantité d'ondes beta	Moyenne ou haute	5
	Quantité d'ondes delta	-	-
	Quantité d'ondes theta	-	-
	Amplitude du menton	Haute	8
	Instabilité du menton	Haute	8
	Amplitude oculaire (somme)	Moyenne ou haute	4
	Instabilité oculaire (somme)	Moyenne ou haute	4
	Amplitude oculaire (soustraction)	Moyenne ou haute	4
	Instabilité oculaire (soustraction)	Moyenne ou haute	4
Éveil Yeux Ouverts (EYO)	Amplitude EEG	Moyenne ou haute	6
	Instabilité EEG	-	-
	Quantité d'activité en ondes lentes	-	-
	Quantité d'ondes alpha	-	-
	Quantité d'ondes beta	Moyenne ou haute	5
	Quantité d'ondes delta	-	-
	Quantité d'ondes theta	-	-
	Amplitude du menton	Haute	8
	Instabilité du menton	Moyenne ou haute	7
	Amplitude oculaire (somme)	Moyenne ou haute	8
	Instabilité oculaire (somme)	Moyenne ou haute	8
	Amplitude oculaire (soustraction)	Moyenne ou haute	8
	Instabilité oculaire (soustraction)	Haute	10
Éveil Yeux Fermés (EYF)	Amplitude EEG	Basse ou moyenne	3
	Instabilité EEG	Non	6
	Quantité d'activité en ondes lentes	Basse	8
	Quantité d'ondes alpha	Haute	10
	Quantité d'ondes beta	Moyenne ou haute	5
	Quantité d'ondes delta	Basse	5
	Quantité d'ondes theta	Basse	5
	Amplitude du menton	Haute	8
	Instabilité du menton	-	-
	Amplitude oculaire (somme)	Basse	8
	Instabilité oculaire (somme)	Basse	7
	Amplitude oculaire (soustraction)	Basse	8
	Instabilité oculaire (soustraction)	Basse	7

poursuivi page suivante

suite de la page précédente

	Caractéristique	État attendu	Poids
N1	Amplitude EEG	Basse	3
	Instabilité EEG	Non	6
	Quantité d'activité en ondes lentes	Basse	8
	Quantité d'ondes alpha	Moyenne	10
	Quantité d'ondes beta	Moyenne ou haute	8
	Quantité d'ondes delta	Basse	5
	Quantité d'ondes theta	Moyenne ou haute	5
	Amplitude du menton	Moyenne ou haute	8
	Instabilité du menton	Basse ou moyenne	6
	Amplitude oculaire (somme)	Moyenne ou haute	6
	Instabilité oculaire (somme)	-	-
	Amplitude oculaire (soustraction)	Basse	6
	Instabilité oculaire (soustraction)	Basse	7
N2	Amplitude EEG	-	-
	Instabilité EEG	-	-
	Quantité d'activité en ondes lentes	Basse ou moyenne	6
	Quantité d'ondes alpha	Basse ou moyenne	4
	Quantité d'ondes beta	Moyenne	4
	Quantité d'ondes delta	Moyenne ou haute	8
	Quantité d'ondes theta	Moyenne ou haute	5
	Amplitude du menton	Basse ou moyenne	8
	Instabilité du menton	-	-
	Amplitude oculaire (somme)	Basse	6
	Instabilité oculaire (somme)	Basse	6
	Amplitude oculaire (soustraction)	Basse	6
	Instabilité oculaire (soustraction)	Basse	6
N3	Amplitude EEG	Haute	8
	Instabilité EEG	-	-
	Quantité d'activité en ondes lentes	Haute	8
	Quantité d'ondes alpha	Basse	4
	Quantité d'ondes beta	Basse	4
	Quantité d'ondes delta	Moyenne ou haute	8
	Quantité d'ondes theta	Basse ou moyenne	5
	Amplitude du menton	Basse ou moyenne	8
	Instabilité du menton	Basse	8
	Amplitude oculaire (somme)	Basse	8
	Instabilité oculaire (somme)	Basse	8
	Amplitude oculaire (soustraction)	Basse	8
	Instabilité oculaire (soustraction)	Basse	8
SP	Amplitude EEG	Basse ou moyenne	3
	Instabilité EEG	-	-
	Quantité d'activité en ondes lentes	Basse ou moyenne	4
	Quantité d'ondes alpha	Moyenne ou haute	8
	Quantité d'ondes beta	Moyenne ou haute	5
	Quantité d'ondes delta	Moyenne ou haute	5
	Quantité d'ondes theta	Moyenne ou haute	7
	Amplitude du menton	Basse	10
	Instabilité du menton	Basse	7
	Amplitude oculaire (somme)	Basse ou moyenne	7
	Instabilité oculaire (somme)	Basse ou moyenne	7
	Amplitude oculaire (soustraction)	Moyenne ou haute	5
	Instabilité oculaire (soustraction)	Moyenne ou haute	5

Une représentation des propriétés organisée non pas par stade mais par caractéristique est présentée Figure D4. Dans cette figure, malgré l'absence des poids, il est plus aisé de prendre connaissance des évolutions attendues pour chaque caractéristique selon les différents stades.

Par exemple, pour le cas de l'amplitude EEG, il est attendu qu'elle soit maximale pour l'EA et le N3, plutôt haute pour l'EYO, plutôt basse pour l'EYF et le SP et minimale pour le N1. Aucun état particulier n'est attendu pour le stade N2, puisqu'il est possible que l'amplitude EEG y soit très faible (au début du N2, pour les époques sans complexe K) comme très élevée (à la fin du N2, juste avant le N3).

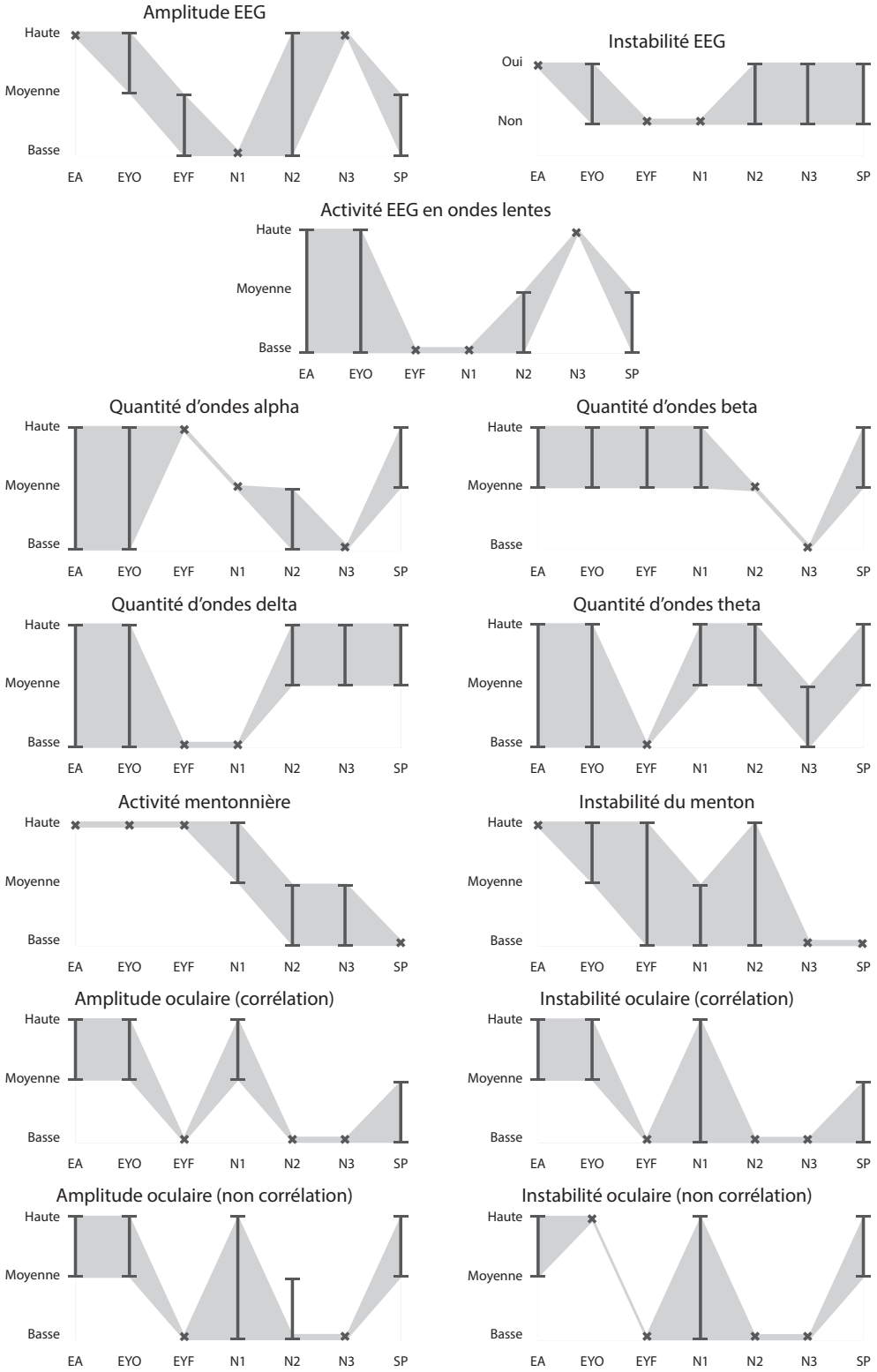


Figure D4 – États attendus pour les différentes caractéristiques selon chaque stade.

Pour aller plus loin : un exemple concret

Dans cet exemple, on considère les quantités d'ondes alpha obtenues pour deux patients distincts. Ces quantités représentent la proportion d'ondes alpha évaluée dans chaque époque [2]. La Figure D5 illustre ces quantités ainsi que les hypnogrammes de référence correspondants.

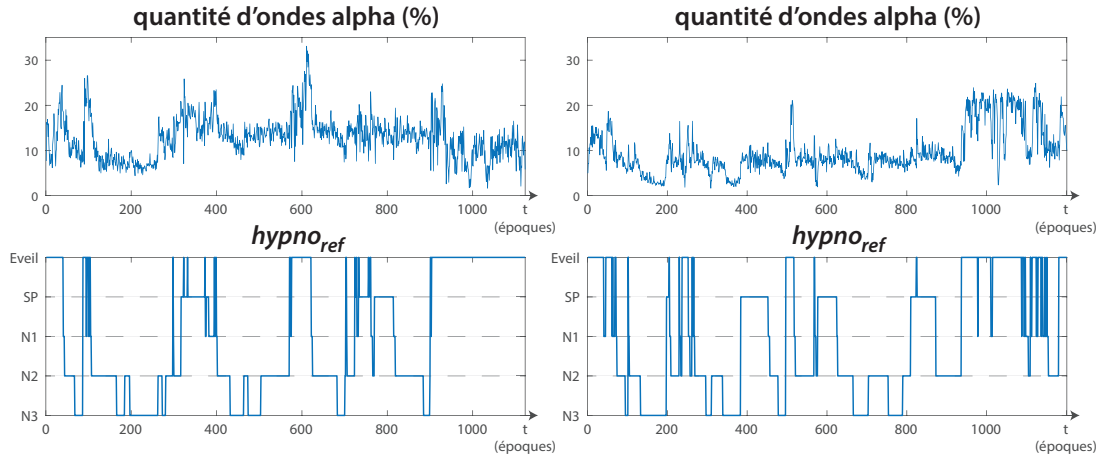


Figure D5 – Quantité d'ondes alpha et hypnogrammes de référence (lus manuellement par un spécialiste du sommeil) chez deux patients.

À l'aide des diagrammes en boîte à moustaches (Figure D6), on remarque que la quantité d'ondes alpha est généralement plus faible chez le deuxième patient que chez le premier, quelque soit le stade de sommeil [3].

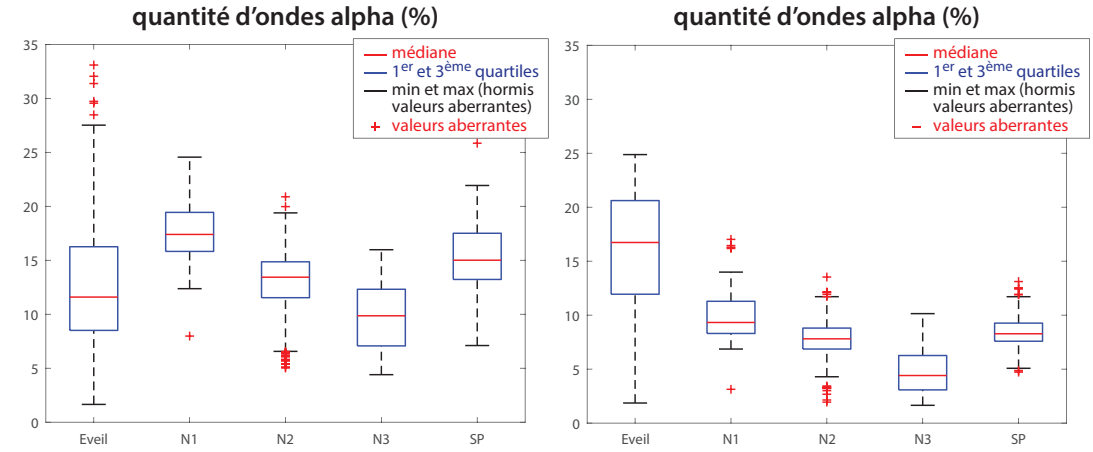


Figure D6 – Diagrammes en boîte à moustaches correspondants aux enregistrements de la Figure D5.

Avec SATUD, deux seuils ont été évalués pour chaque enregistrement. Pour le premier, le seuil 1 (basse -> moyenne) était de 14, et le seuil 2 (moyenne -> haute) de 18. Pour le second, le seuil 1 et le seuil 2 étaient de 8 et 12, respectivement (voir Figure D7).
En utilisant les règles de propriété correspondant aux ondes alpha (présentées Tableau D1 et

2. Attention, il ne faut pas comparer les proportions d'ondes alpha reportées dans ce travail avec les valeurs indiquées dans les recommandations AASM. En effet, il est inscrit dans celles-ci que la présence de plus de « 50 % d'ondes alpha » dans une époque indique que cette dernière est à l'éveil. Cependant, ces 50 % sont estimés visuellement. A l'inverse, les valeurs mentionnées dans le présent travail sont des valeurs estimées à l'aide d'une transformée de Fourier, et prennent donc en compte le contenu fréquentiel difficilement visible à l'oeil nu.

3. On observe également que la quantité d'ondes alpha à l'éveil est plutôt faible à modérée (comparée aux autres stades) dans le cas du premier patient, contrairement au second pour lequel la quantité d'ondes alpha à l'éveil est bien supérieure. Cette différence est notamment visible sur la Figure D5, lors des longues périodes d'éveil situées à la fin des enregistrements. Une explication possible serait qu'il s'agisse d'éveil agité pour le premier patient (qui a probablement été actif en attendant la fin de l'examen), et d'éveil calme yeux fermés pour le second (qui a même réussi à se rendormir).

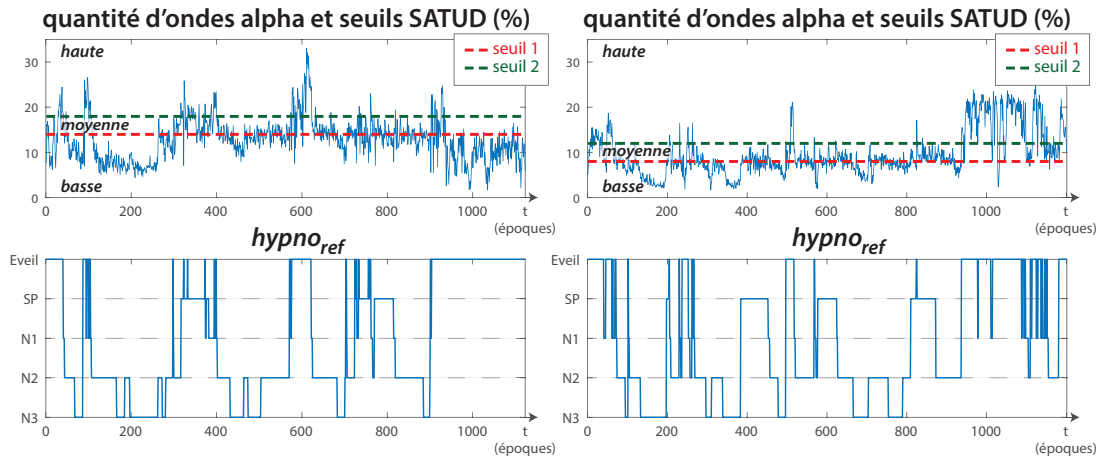


Figure D7 – Quantité d'ondes alpha, hypnogrammes de référence et seuils estimés par SATUD dans le cas des deux enregistrements des figures D5 et D6.

Figure D4), on peut en déduire le nombre d'époques de chaque stade qui respectent ou non la propriété associée à leur stade. Le Tableau D2 compare le pourcentage d'époques de chaque stade qui respectent la propriété relative à la quantité d'ondes alpha de leur stade.

Table D2 – Pour chaque stade, comparaison des proportions d'époques respectant la propriété de leur stade relative à la quantité d'ondes alpha, dans le cas de seuils moyens et des seuils obtenus grâce à SATUD.

	Enregistrement 1		Enregistrement 2	
Stade	Seuils moyens*	Seuils SATUD	Seuils moyens*	Seuils SATUD
E**	-	-	-	-
N1	17.0 %	61.7 %	26.7 %	65.3 %
N2	77.5 %	98.2 %	100.0 %	99.3 %
N3	60.9 %	92.3 %	100.0 %	96.0 %
SP	96.2 %	63.2 %	8.6 %	60.0 %

* seuil 1 = 11, seuil 2 = 15

** Le stade E n'étant pas limité en terme de quantité d'ondes alpha (puisqu'il regroupe l'EA, l'EYO et l'EYF), il n'est pas possible d'estimer ses proportions. On relève tout de même que 22.0 % et 74.6 % des époques d'éveil ont une quantité d'ondes alpha dite haute (en utilisant les seuils SATUD) pour les enregistrement 1 et 2, respectivement. Cela suggère bien un éveil en grande partie agité pour le premier enregistrement et un éveil plutôt calme pour le second.

Dans cet exemple, l'utilisation de SATUD a permis une estimation des caractéristiques qualitatives relatives à la quantité d'ondes alpha mieux ajustée à chaque enregistrement, permettant une meilleure analyse de tous les stades.

Informations supplémentaires : le choix des artéfacts utilisés

Parmi les artéfacts présentés Section B.4, les artéfacts de sudation, ou artéfacts électrodermographiques, ont été choisis pour tester la robustesse de SATUD. Plusieurs raisons expliquent ce choix :

— les artéfacts extrinsèques n'ont pas été retenus, de par leur nature ;

— les artéfacts intrinsèques liés aux mouvements oculaires sont pris en compte pour la classification, en accord avec les recommandations AASM (fonction détection des grapho-éléments) ;

— les artéfacts intrinsèques liés à l'activité musculaire et de type électromyographique sont pris en compte pour la classification, en accord avec les recommandations AASM (fonction détection des grapho-éléments). Les artéfacts de type photomyogénique ou glossokinétique sont trop peu présents et trop peu disruptifs ;

— les artéfacts intrinsèques liés à l'activité cardiaque sont trop peu disruptifs ;

— les artéfacts intrinsèques liés à l'activité ventilatoire sont également trop peu disruptifs.

L'étude de Arnardottir *et al.* (2010) a également montré que la sudation était plus forte chez les patients atteints du SAHS (voir Figure D8 extraite de leur étude). Dans le cas de notre étude, l'artéfact de sudation semblait donc être le plus pertinent.

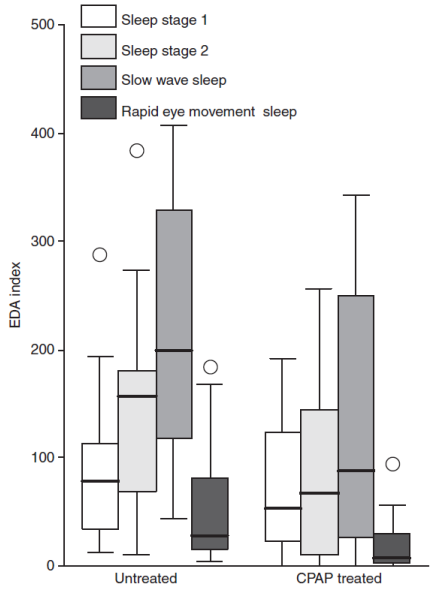


Figure D8 – Indice d'activité électrodermique (EDA index) pour différents types de sommeil (sleep stage 1 = N1, sleep stage 2 = N2, slow wave sleep = N3 et rapid eye movement sleep = SP) pour des patients souffrant de SAHS et traités ou non traités par Pression Positive Continue (PPC) ou *Continuous Positive Airway Pressure* en anglais (CPAP). Figure extraite de Arnardottir *et al.* (2010).

Valorisation supplémentaire

Ce travail a également été valorisé dans le congrès francophone GRETSI (Lille, août 2019), par le biais d'une présentation orale.

À RETENIR

Une méthode de seuillage auto-adaptatif fonctionnant de manière non supervisée a été mise en place. Utilisée dans le cadre de la classification automatique des stades de sommeil, elle permet d'extraire des caractéristiques qualitatives ajustées à chaque patient. Ces caractéristiques, en permettant la traduction des propriétés décrites dans les recommandations AASM, améliorent la confiance du spécialiste du sommeil dans l'outil de classification automatique. Robuste au bruit et aux artéfacts de sudation, on considère que cette méthode peut être utilisée sur des enregistrements réels, avec tous les artéfacts qui peuvent en découler.

D.3.3 Mesure de l'impact clinique

Pour aller plus loin, l'impact clinique de cette classification sur différents enregistrements a été évalué. Ce travail fait l'objet d'un article clinique actuellement en cours de rédaction. Cette section reporte une partie des éléments qui seront présentés dans l'article.
L'objectif est ainsi de mesurer l'impact de la classification automatique en PSG sur le diagnostic du patient, et de le comparer au diagnostic résultant de la lecture manuelle de référence.

Méthode

Données

Nous avons cette fois-ci utilisé un jeu de 1291 enregistrements, provenant une fois encore de la cohorte du sommeil des Pays de La Loire. Ces enregistrements n'ont fait partie ni du jeu d'apprentissage, ni du jeu de test précédent. De la même manière que pour les études précédentes, ils ont été enregistrés et lus durant les années 2012-2018 au laboratoire du sommeil du CHU d'Angers, en utilisant un appareil polysomnographique de type CID102L8D, et en respectant les recommandations de l'AASM. Tous les patients, investigués pour suspicion de SAHS, ont donné leur consentement écrit et informé. Le tableau D3 présente les caractéristiques des enregistrements utilisés.

Table D3 – Description du jeu de données utilisé pour la mesure de l'impact clinique de l'analyse automatique en PSG.

Variable	Valeur
Sexe féminin	$\approx 37\,\%$
Âge (moyenne ± écart-type)	52 ± 14 ans
IAH* (moyenne ± écart-type)	23 ± 21 par heure
Nombre d'époques total	1414095
Nombre d'époques d'éveil*	314292 ($\approx 22\,\%$)
Nombre d'époques de N1*	100346 ($\approx 7\,\%$)
Nombre d'époques de N2*	539798 ($\approx 38\,\%$)
Nombre d'époques de N3*	236268 ($\approx 17\,\%$)
Nombre d'époques de SP*	223391 ($\approx 16\,\%$)

* D'après la lecture manuelle qui constitue notre référence

Approche

Nous considérerons deux hypnogrammes. Le premier est l'hypnogramme de référence (toujours obtenu par lecture manuelle), annoté $hypno_{ref}$ dans la suite de cette section. Le second est l'hypnogramme obtenu à partir de l'analyse automatique en PSG, auquel il sera fait référence sous le nom $hypno_{EP}$. Nous n'étudierons ici que l'impact de l'hypnogramme automatique et ne prendrons pas en considération la table de probabilité.

Comme présenté dans la Section B.3.2, la mesure de l'impact clinique est généralement effectuée en considérant plusieurs éléments : les indicateurs de proportions et latences, l'IAH et la sévérité du SAHS. Les proportions et latences sont utilisées pour détecter une possible anomalie. Par exemple, les patients souffrants de narcolepsie ont souvent une latence de SP particulièrement faible. L'IAH est utilisé pour déterminer la sévérité des syndromes liés aux diminutions des la ventilation durant le sommeil (syndrome d'apnées du sommeil, syndrome d'apnées-hypopnées du sommeil, syndrome d'apnées obstructives du sommeil, syndrome d'apnées-hypopnées obstructives du sommeil, syndrome d'apnées centrales du sommeil, etc.). Pour finir, la sévérité du SAHS est l'indicateur de plus haut niveau. Il permet ainsi de mesurer l'impact concret sur le diagnostic du patient.

1) Proportions et latences

Le Temps de Sommeil Total (TST) obtenu avec l'analyse automatique à partir des signaux électrophysiologiques en PSG (TST_{EP}) a été comparé avec le TST découlant de l'analyse manuelle en PSG (TST de référence TST_{ref}). Ces durées sont exprimées en minutes. La comparaison a été effectuée en utilisant des graphiques de corrélation et de Bland-Altman.

De la même manière, les proportions relatives aux stades de sommeil et obtenues avec l'analyse automatique ont été comparées avec celles résultant de la lecture manuelle. Ces proportions sont usuellement exprimées par rapport au TST, et non à la durée totale d'enregistrement. Le stade N1 étant très peu présent comparé aux autres stades et, qui plus est, le stade avec le taux d'accord inter-scorer le plus faible, on regroupera le N1 et le N2. Les différentes proportions étudiées seront donc notés *propN1N2*, *propN3* et *propSP*. La proportion d'éveil n'a pas été estimée, puisqu'elle est directement reliée au TST. Cette fois-ci, on étudiera les erreurs relatives ($\frac{\text{valeur estimée - valeur de référence}}{\text{valeur de référence}}$) associées à chaque proportion.
La latence d'endormissement *SOL*, de *Sleep Onset Latency* en anglais, et les latences de chaque stade (*latN2*, *latN3* et *latSP*), ont également été évaluées et comparées avec la référence. L'endormissement a été ici défini comme étant la première époque de sommeil, à condition que celle-ci soit suivie d'au moins deux autres époques de sommeil. La latence d'endormissement correspond donc au délai (en minutes) entre le début d'enregistrement et l'endormissement comme précédemment défini. Étant donné que l'endormissement se fait nécessairement par le stade N1, la latence du N1 est équivalente à la *SOL* et n'a donc pas été estimée. Les latences *latN2*, *latN3* et *latSP* correspondent au délai, en minutes, entre l'endormissement (comme précédemment défini) et la première époque du stade en question. Ici, on étudiera les valeurs de la différence entre les latences obtenues grâce à la lecture automatique et manuelle.

2) IAH

L'IAH découlant de l'analyse automatique, IAH_{EP}, est comparé à l'IAH de référence, IAH_{ref}. Cependant, en dehors de l'hypnogramme, les événements pris en compte lors de l'analyse manuelle ou automatique peuvent différer. En effet, en dépit du fait qu'ils soient utilisés pour déterminer les hypopnées, les micro-éveils n'ont pas été détectés automatiquement. Nous considérerons donc deux IAHs pour l'analyse automatique. Le premier, $IAH_{EPsansMEV}$, est l'IAH qui serait obtenu sans lecture manuelle des micro-éveils. Dans cette situation, seule la lecture de la ventilation est réalisée manuellement. Le second, $IAH_{EPavecMEV}$, est l'IAH qui serait obtenu avec lecture manuelle des micro-éveils. Dans cette situation, une lecture manuelle des micro-éveils est réalisée en plus de la lecture de la ventilation. Le tableau D4 récapitule les éléments considérés lors de l'estimation des différents IAHs.

Table D4 – Estimation des différents IAHs comparés dans cette partie. *A* définit les apnées, et *H* définit les hypopnées.

	Examen de PV	Examen de PSG		
	pas d'hypnogramme	$hypno_{EP}$		$hypno_{ref}$
	sans micro-éveils	sans micro-éveils	avec micro-éveils	avec micro-éveils
Événements respiratoires	A*(toutes)* H_{desat}*(toutes)*	A *en sommeil*$_{EP}$ H_{desat} *en sommeil*$_{EP}$ H_{Ev} *en sommeil*$_{EP}$	A *en sommeil*$_{EP}$ H_{desat} *en sommeil*$_{EP}$ H_{Ev} *en sommeil*$_{EP}$ H_{MEV} *en sommeil*$_{EP}$	A *en sommeil*$_{ref}$ H_{desat} *en sommeil*$_{ref}$ H_{Ev} *en sommeil*$_{ref}$ H_{MEV} *en sommeil*$_{ref}$
Durée	TE	TST_{EP}	TST_{EP}	TST_{ref}
IAH	$\frac{A+H_{desat}}{TE}$	$\frac{A+H_{desat}+H_{Ev}}{TST_{EP}}$	$\frac{A+H_{desat}+H_{Ev}+H_{MEV}}{TST_{EP}}$	$\frac{A+H_{desat}+H_{Ev}+H_{MEV}}{TST_{ref}}$

En pratique, le calcul est légèrement plus complexe puisqu'une hypopnée peut être à la fois désaturante et micro-éveillante ou éveillante. Dans ce cas, on ne comptabilise qu'une seule fois l'hypopnée en question. Les événements non comptabilisés dans le cas de l'$IAH_{EPsansMEV}$, comparé à l'$IAH_{EPavecMEV}$, sont donc les hypopnées qui ne sont qu'uniquement micro-éveillantes. On comparera les différents IAHs en utilisant une fois encore les graphes de corrélation et de Bland-Altman.

3) Sévérité du SAHS

La sévérité du SAHS est estimée selon la valeur de l'IAH obtenu. On peut ainsi comparer les sévérités dans le cas de la lecture manuelle, de la lecture automatique sans micro-éveils et dans le cas de l'analyse automatique avec micro-éveils.
On définira un SAHS absent lorsque l'IAH est inférieur à 5/h, un SAHS léger lorsque l'IAH est compris entre 5/h et 15/h, modéré lorsqu'il est compris entre 15/h et 30/h, et sévère lorsqu'il est supérieur à 30/h. Il est important de savoir qu'il est possible de traiter un patient

lorsque son IAH excède 30/h, ou 15/h selon ses symptômes et caractéristiques. De ce fait, les patients qui seraient diagnostiqués comme ayant un SAHS absent au lieu de léger (ou inversement), sont des patients pour lesquels l'erreur réalisée n'aurait eu que pas ou peu d'impact. Au contraire, les erreurs pour lesquelles l'impact possible sur le traitement du patient n'est pas négligeable, sont les patients diagnostiqués avec un IAH inférieur à 15/h au lieu de supérieur (ou inversement), ou inférieur à 30/h au lieu de supérieur (ou inversement).
On mesurera l'exactitude de l'estimation de la sévérité du SAHS à l'aide de matrices de confusion.

Résultats

Les résultats sont organisés en considérant les différents éléments présentés précédemment. L'ordre de ces éléments allant du moins au plus proche du diagnostic du patient, nous pouvons considérer les résultats comme étant présentés des moins au plus importants.

1) Proportions et latences

La Figure D9 présente les graphes de corrélation et de Bland-Altman liés à l'estimation du TST.

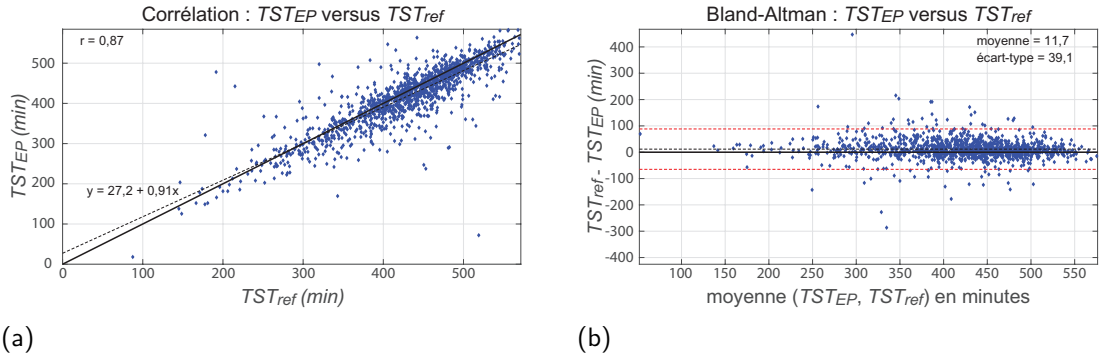


Figure D9 – a) graphe de corrélation du TST estimé grâce à l'analyse automatique du sommeil en PSG, versus le TST obtenu avec lecture manuelle du sommeil, et b) graphe de Bland-Altman correspondant.

L'estimation du TST obtient une forte corrélation. De plus, le TST_{EP} est en moyenne sous-diagnostiqué d'uniquement 11,7 minutes comparé à TST_{ref}. Sur des valeurs de TST avoisinant les 400 minutes, cette différence est convenable.

La Figure D10 présente la distribution des proportions des différents stades de sommeil.
Les erreurs relatives de ces proportions sont estimées en retirant les patients dont la proportion de référence est inférieure à 5 % (valeurs peu représentatives et divisions par zéro). Pour *propN1N2*, l'erreur relative (moyenne ± écart-type) est de $3{,}6\,\% \pm 18{,}6\,\%$. Pour *propN3*, elle est de $-1{,}0\,\% \pm 41{,}5\,\%$. Enfin, *propSP* obtient une erreur relative de $4{,}0\,\% \pm 32{,}3\,\%$.

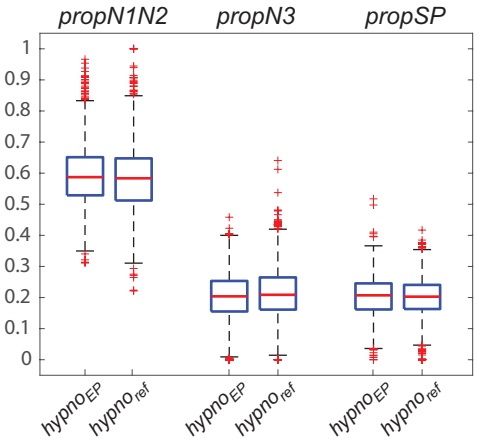


Figure D10 – Diagrammes en boîte à moustaches des proportions associées au N1 et N2, au N3, et au SP, dans le cas des analyses automatiques et manuelles.

La Figure D11 représente un histogramme empilé reportant le nombre d'enregistrements, en ordonné, selon la durée de la latence de référence, en abscisse, en fonction de plusieurs catégories. Ces catégories per-

mettent de différencier les erreurs d'estimation de la latence inférieures à 2 minutes, comprises entre 2 et 5 minutes, entre 5 et 15 minutes, entre 15 et 30 minutes, et supérieures à 30 minutes.

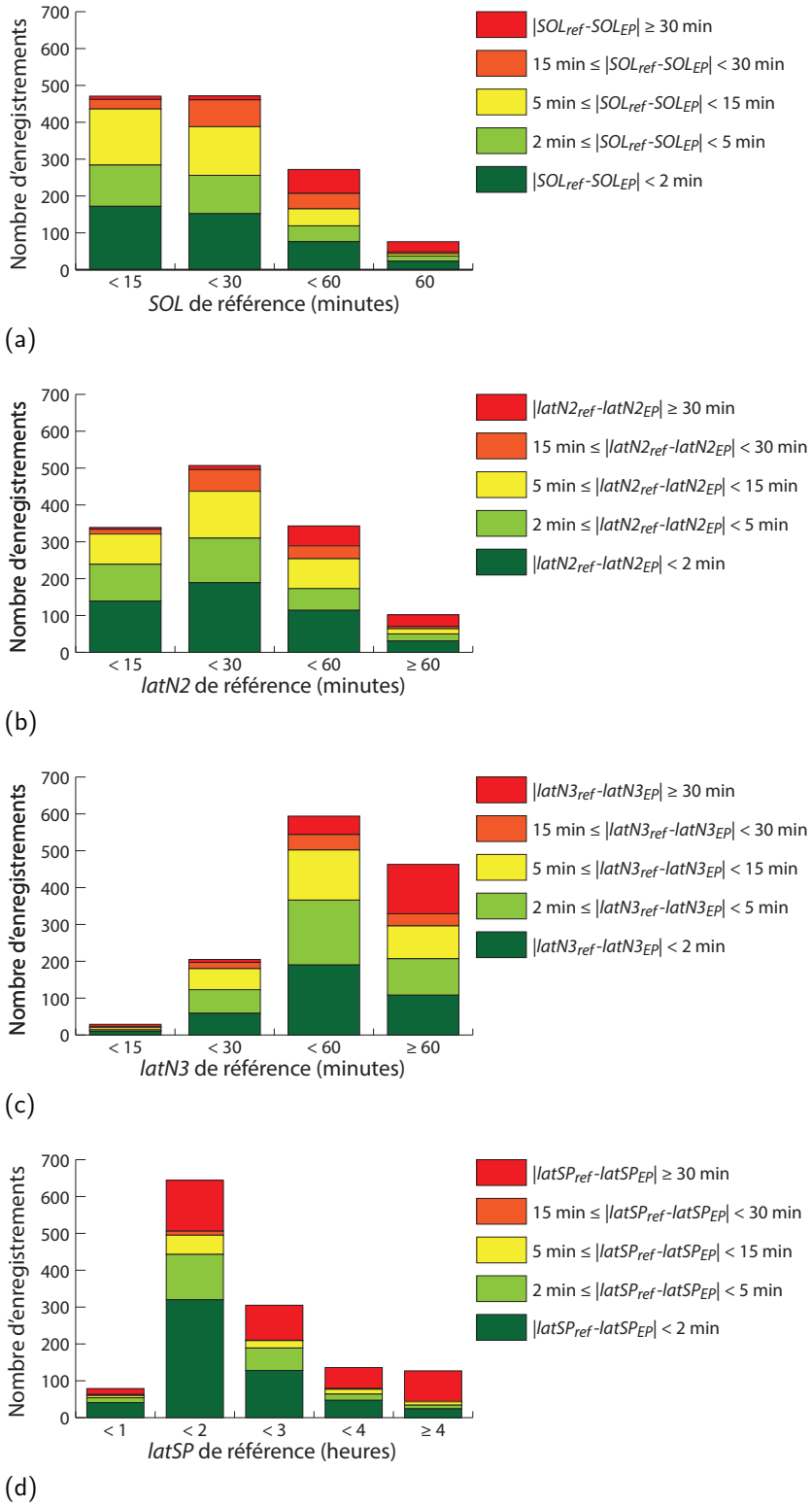


Figure D11 – Distribution des enregistrements selon l'erreur d'estimation de latence et la valeur de la latence, dans le cas : a) de la latence d'endormissement *SOL*, b) de la latence de N2 *latN2*, c) de la latence de N3 *latN3* et d) de la latence de SP *latSP*.

De manière générale, le nombre d'enregistrements pour lesquels les estimations de latences sont correctes à 2 minutes près sont nombreux. On retrouve tout de même certains enregistrements pour lesquels les latences sont estimées avec plus de 15 minutes (voire plus de 30 minutes) d'écart avec la référence. On observe également que plus la latence de référence est élevée, plus la proportion des enregistrements avec une erreur d'estimation importante augmente, à l'inverse de la proportion des enregistrements avec une faible erreur d'estimation. Cela apparaît à la fois en considérant l'évolution au sein d'une même latence, qu'en comparant les latences entre elles.

2) IAH

Les graphes de corrélation et de Bland-Altman dans le cas des IAHs avec ou sans micro-éveils sont présentés Figure D12. L'IAH en PV a été ajouté pour mieux évaluer la progression.

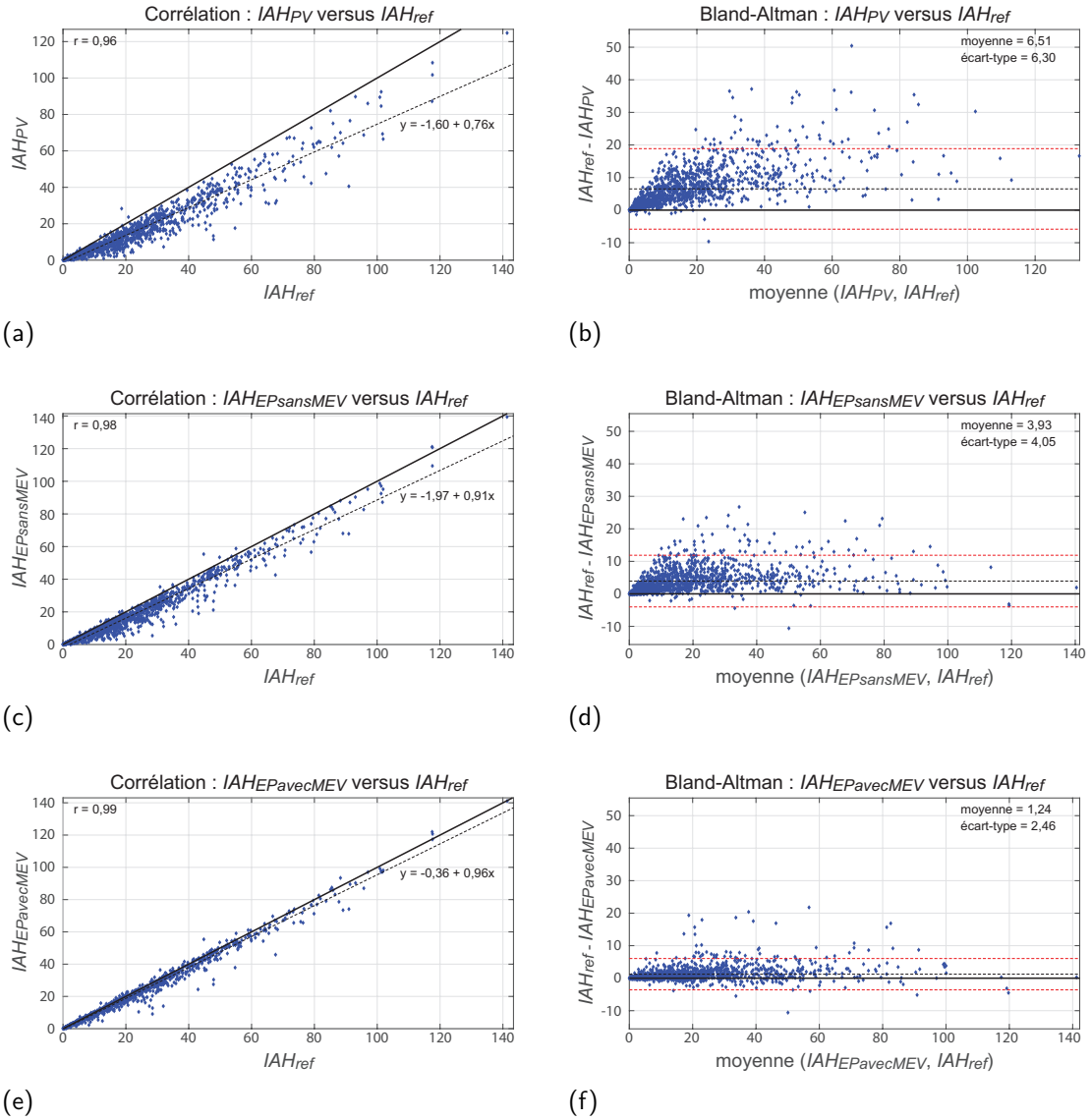


Figure D12 – a) et b) graphes de corrélation et de Bland-Altman des IAHs obtenus sans analyse automatique du sommeil et des micro-éveils (donc équivalent aux IAHs en PV), versus les IAHs obtenus avec analyse manuelle du sommeil et des micro-éveils en PSG, c) et d) graphes de corrélation et de Bland-Altman des IAHs obtenus avec analyse automatique du sommeil en PSG et sans lecture manuelle des micro-éveils, versus les IAHs obtenus avec analyse manuelle du sommeil et des micro-éveils, et e) et f) graphes de corrélation et de Bland-Altman des IAHs obtenus avec analyse automatique du sommeil en PSG et avec lecture manuelle des micro-éveils, versus les IAHs obtenus avec analyse manuelle du sommeil et des micro-éveils.

L'analyse automatique du sommeil permet d'améliorer grandement l'évaluation de l'IAH, qui se rapproche de celui obtenu suite à l'analyse du sommeil manuel. L'$IAH_{EPsansMEV}$ est ainsi sous-estimé de 3,93/h ± 6,30/h par rapport à l'IAH_{ref}, contre 6,51/h ± 6,30/h sans utiliser d'analyse automatique du sommeil.
Enfin, l'ajout d'une lecture manuelle des micro-éveils permet d'obtenir un $IAH_{EPavecMEV}$ surestimé de seulement 1,24/h ± 2,46/h par rapport à l'IAH_{ref}. On observe également que très peu d"enregistrements ont un $IAH_{EPsansMEV}$ surestimé par rapport à l'IAH_{ref}. Il arrive plus souvent que l'$IAH_{EPavecMEV}$ soit surestimé, mais dans ce cas la surestimation reste faible. C'est pour cette raison qu'il est nécessaire d'évaluer l'impact des IAHs estimés sur la sévérité du SAHS.

3) Sévérité du SAHS

La matrice de confusion associée à l'estimation du SAHS sans analyse automatique versus avec lecture manuelle est présentée Table D5.

Table D5 – Matrice de confusion associée à la sévérité du SAHS sans lecture du sommeil et des micro-éveils. Les valeurs en rouge indiquent une sous-estimation qui peut potentiellement impacter le traitement du patient.

		Sans lecture du sommeil ni des micro-éveils				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
Référence	Pas de SAHS	303	1	0	0	304
	SAHS léger	131	198	0	0	329
	SAHS modéré	10	199	149	0	358
	SAHS sévère	0	8	122	170	300
	Total	444	406	271	170	1291

Sans analyse du sommeil ni des micro-éveils (en PV), le Kappa de Cohen et le taux d'accord associés à la matrice de confusion sont de 0,51 et 64 %, respectivement. Cela signifie un accord modéré avec la référence.
On observe également qu'environ 36 % (N=131+10+0+199+8+122=470) des enregistrements sont sous-diagnostiqués. Plus particulièrement, approximativement 26 % (N=10+199+8+122=339, en rouge) des enregistrements pourraient engendrer, en l'état (sans relecture), le non traitement de patients qui l'auraient peut-être étés suite à la lecture manuelle. Seul 1 enregistrement est au contraire sur-diagnostiqué.

Sont ensuite indiquées les matrices de confusion associées à l'estimation du SAHS avec analyse automatique du sommeil (tables D6 et D7).

Table D6 – Matrice de confusion associée à la sévérité du SAHS lors de la lecture automatique du sommeil sans lecture des micro-éveils. Les valeurs en rouge indiquent une sous-estimation qui peut potentiellement impacter le traitement du patient, et la valeur en bleu indique une sur-estimation qui peut potentiellement impacter le traitement du patient.

		Sommeil automatique sans lecture des micro-éveils				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
Référence	Pas de SAHS	304	0	0	0	304
	SAHS léger	94	234	1	0	329
	SAHS modéré	3	130	225	0	358
	SAHS sévère	0	3	83	214	300
	Total	401	367	309	214	1291

Dans le cas de l'analyse automatique du sommeil mais sans lecture manuelle des micro-éveils, le Kappa de Cohen et le taux d'accord sont cette fois-ci de 0,68 et 76 %, respectivement. Cela signifie un accord fort avec la référence.
On observe également qu'environ 24 % (N=94+3+0+130+3+83=313) des enregistrements sont sous-diagnostiqués. Plus particulièrement, approximativement 17 % (N=3+130+3+83=219, en rouge) des enregistrements pourraient engendrer, en l'état (sans relecture), le non traitement de patients qui l'auraient peut-être été suite à la lecture manuelle. Encore une fois,

1 enregistrement est au contraire sur-diagnostiqué. Sans relecture de cet enregistrement (en bleu), le spécialiste du sommeil pourrait décider de traiter le patient, alors qu'il ne l'aurait pas nécessairement fait en lisant manuellement le sommeil.

Table D7 – Matrice de confusion associée à la sévérité du SAHS lors de la lecture automatique du sommeil avec lecture manuelle des micro-éveils. Les valeurs en rouge indiquent une sous-estimation qui peut potentiellement impacter le traitement du patient, et les valeurs en bleu indiquent une sur-estimation qui peut potentiellement impacter le traitement du patient.

		Sommeil automatique avec lecture des micro-éveils				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
Référence	Pas de SAHS	298	6	0	0	304
	SAHS léger	21	303	5	0	329
	SAHS modéré	0	31	325	2	358
	SAHS sévère	0	1	39	260	300
	Total	319	341	369	262	1291

Dans le cas de l'analyse automatique du sommeil avec lecture manuelle des micro-éveils, le Kappa de Cohen et le taux d'accord atteignent 0,89 et 92 %, respectivement. Cela signifie un accord excellent avec la référence.
On observe également qu'environ 7 % (N=21+0+0+31+1+39=92) des enregistrements sont sous-diagnostiqués. Plus particulièrement, approximativement 5 % (N=31+1+39=71, en rouge) des enregistrements pourraient engendrer, en l'état (sans relecture), le non traitement de patients qui l'auraient peut-être été suite à la lecture manuelle. Environ 1 % (N=6+0+5+0+0+2=13) des enregistrements sont au contraire sur-diagnostiqués, dont 7 (5+2, en bleu) pour lesquels l'analyse automatique utilisée sans relecture pourrait aboutir au traitement de patients qui n'auraient pas nécessairement été traités à la suite d'une lecture manuelle.

Discussion

Bien qu'il soit fortement conseillé d'utiliser l'analyse automatique du sommeil tel un outil d'aide au diagnostic, avec relecture manuelle, nous avons étudié l'impact que pourrait avoir l'utilisation de cette analyse dans le cas où il n'est pas effectué de relecture par le spécialiste du sommeil.

Dans un premier temps, nous avons évalué l'habileté de l'analyse automatique à estimer différents indicateurs utilisés par le spécialiste du sommeil lors du diagnostic. Le Temps de Sommeil Total (TST) et les proportions de N2, N3 et SP ont obtenu des résultats satisfaisants avec des différences avec la référence faibles par rapport à leur valeur de référence (voir Figures D9 et D10).
Les latences d'endormissement ainsi que les latences du N2, N3 et SP ont aussi été comparées avec celles obtenues par lecture manuelle du sommeil (Figure D11). Encore une fois, la différence entre les valeurs de latences estimées grâce à l'analyse automatique et manuelle sont peu importantes comparées aux valeurs de référence, obtenues par lecture manuelle du sommeil. Ainsi, la proportion d'enregistrements pour lesquels les latences sont estimées avec une erreur de plus de 30 minutes est plus importante pour le SP, qui apparaît généralement plus tard dans la nuit, que pour le N3 ou, plus flagrant encore, le N2. Le N1 étant un stade de sommeil très bref, nous pouvons également observer des résultats sensiblement similaires pour la latence d'endormissement et la latence du N2.

En ce qui concerne l'estimation de l'IAH (Figure D12), les résultats varient de manière importante selon que les micro-éveils soient utilisés ou non. Cela n'est pas surprenant mais confirme la nécessité d'une analyse automatique qui détecterait également ces micro-éveils.
Ainsi, la valeur de la pente de la courbe de corrélation passe de 0,76 sans analyse du sommeil, à 0,91 avec analyse automatique du sommeil mais sans micro-éveils, et enfin à 0,96 avec analyse automatique du sommeil et lecture manuelle des micro-éveils. Sur les graphes de Bland-Altman, on observe également que la moyenne et l'écart-type de l'erreur d'estimation d'IAH diminue avec l'analyse automatique du sommeil puis de l'ajout de la lecture manuelle des micro-éveils.

Enfin, la sévérité du SAHS a été évaluée en fonction des trois situations précédentes :

— sans analyse du sommeil ni des micro-éveils versus avec analyse manuelle du sommeil et des micro-éveils (Tableau D5) : la matrice de confusion obtient un accord modéré avec la référence, selon l'interprétation du Kappa de Cohen. On observe un grand nombre de patients sous-diagnostiqués. Plus d'un quart des patients n'auraient pas pu être traités sans examen de seconde intention ;

— avec analyse du sommeil mais sans lecture des micro-éveils versus avec analyse manuelle du sommeil et des micro-éveils (Tableau D6) : la matrice de confusion obtient un accord fort avec la référence. On observe bien moins de patients sous-diagnostiqués, et un seul patient sur-diagnostiqué. Plus que 17 % des patients auraient été dans une situation où un examen de seconde intention aurait été nécessaire pour se voir proposer un éventuel traitement ;

— avec analyse du sommeil et lecture manuelle des micro-éveils versus avec analyse manuelle du sommeil et des micro-éveils (Tableau D7) : la matrice de confusion obtient un accord excellent avec la référence. On observe encore moins de patients sous-diagnostiqués, et un seul patient sur-diagnostiqué. Seuls 5 % des patients auraient été dans une situation où un examen de seconde intention aurait été nécessaire pour se voir proposer un éventuel traitement. Au contraire, un risque de traitement proposé à des patients qui pourraient ne pas en avoir besoin apparaît pour environ 1 % des patients.

La sévérité du SAHS nous permet ainsi d'avoir une idée très précise de l'impact concret que pourrait avoir l'analyse automatique du sommeil sur le patient. Cependant, il serait nécessaire d'estimer cet impact après la mise en place d'une estimation automatique des micro-éveils, mais également après relecture du sommeil par le spécialiste du sommeil. Comme mentionné dans les sections précédentes, l'utilisation conseillée de l'analyse automatique du sommeil consiste à utiliser l'hypnogramme généré ainsi que le graphique d'hypnodensité pour effectuer une lecture manuelle plus rapide et efficace.

Conclusion

L'analyse du sommeil présentée précédemment a été appliquée sur un jeu de données composé de 1291 nouveaux enregistrements polysomnographiques. Ces enregistrements proviennent de patients qui ont consulté pour suspicion de troubles du sommeil lors des années 2012 à 2018. L'estimation des performances cliniques a été réalisée en comparant l'enregistrement constitué de l'hypnogramme résultant de l'analyse automatique du sommeil, avec l'enregistrement tel que lu manuellement par le spécialiste du sommeil. L'impact clinique de cette analyse sur le diagnostic a été évalué sur plusieurs niveaux : les indicateurs de proportions et latences, l'IAH et la sévérité du SAHS.
Il en ressort que l'analyse automatique obtient des résultats cliniques satisfaisants, mais qui restent néanmoins limités par le manque de détection automatique des micro-éveils, nécessaires pour la prise en compte des hypopnées micro-éveillantes. Couplé à une lecture manuelle des micro-éveils, les résultats deviennent excellents et promettent une estimation de l'IAH et du SAHS très précise.

À RETENIR

L'analyse automatique et patient-dépendante des stades de sommeil à partir des voies électrophysiologiques a montré qu'elle était suffisamment précise pour permettre une estimation des différents paramètres cliniques utilisés pour le diagnostic (latences et proportions), de l'IAH et de la sévérité du SAHS. Pour améliorer encore le diagnostic, il s'agirait non pas d'améliorer cette analyse des stades de sommeil mais de mettre en place une détection automatique des micro-éveils, à ce jour non réalisée.

D.4 Conclusion du chapitre

Dans ce chapitre, une analyse automatique des stades de sommeil en PSG a été présentée. Cette analyse utilise les voies électrophysiologiques habituelles (EEGs, EOGs et EMG du menton).

Plusieurs limitations à l'utilisation des analyses automatiques en pratique clinique ont été identifiées. La méthodologie développée a été pensée pour solutionner ces limitations.
Algorithmiquement, le modèle implémenté reproduit le cheminement des spécialistes du sommeil lors d'une lecture manuelle. Une première fonction sert à ajuster les caractéristiques de l'enregistrement de manière non supervisée. Cette fonction, appelée SATUD (*Self-Adaptative Thresholding Using Descriptors*), est un élément clé qui permet, à terme, d'obtenir des caractéristiques patient-dépendantes. Robuste au bruit et aux artéfacts (notamment ceux de sudation qui sont très disruptifs), cette fonction peut être comparée à la visualisation rapide de l'ensemble des époques par le spécialiste du sommeil, en début d'enregistrement, lui permettant d'ajuster sa lecture aux spécificités du patient. Une seconde fonction permet ensuite la construction d'un hypnogramme à partir des caractéristiques ajustées. Une troisième fonction est ensuite utilisée pour l'identification des grapho-éléments et des événements notables pour la lecture du sommeil. Pour finir, une dernière fonction combine les différents éléments précédemment identifiés et génère un hypnogramme qui, comme lors d'une lecture manuelle, est construit en prenant en considération les époques à proximité et les règles de transition entre les différents stades.
À l'inverse des comportements en boîte noire et du manque de transparence que l'on retrouve parfois dans les analyses utilisant de l'intelligence artificielle, la méthodologie mise en place reste donc très proche du processus de lecture des médecins.

Afin de s'assurer de la robustesse de l'analyse automatique, cette dernière a été entraînée et testée sur un grand nombre d'enregistrements provenant de patients suspectés d'avoir des troubles du sommeil. Les résultats sont convaincants puisqu'ils varient peu selon l'âge et la sévérité du SAHS du patient.

Pour finir, l'analyse présentée ne nécessite aucune action préalable du médecin (par exemple l'invalidation de certaines époques ou la lecture partielle du sommeil).
Cette facilité d'utilisation est complétée par la génération, en plus de l'hypnogramme automatique, d'une matrice de confusion. Cette dernière peut être visualisée sous la forme d'un graphique d'hypnodensité, outil visuel permettant d'identifier les époques qui nécessiteraient une relecture en priorité.

Avec un Kappa de Cohen moyen de 0,69 sur le groupe de test (taux d'accord moyen de 77,8 %), l'analyse automatique obtient des performances très proches des performances inter-scorer. Les différents indices utilisés lors du diagnostic (latences et proportions des stades) ont pu être estimés avec une bonne précision. L'analyse permet d'améliorer grandement l'estimation de l'IAH et ainsi de la sévérité du SAHS. Pour l'améliorer encore plus, il serait nécessaire de mettre en place une analyse automatique des micro-éveils. Ce travail n'a pas encore été réalisé à ce jour, mais a fait l'objet d'un stage.

Nous rappelons que, malgré les résultats présentés, il reste de la responsabilité du spécialiste du sommeil de relire le sommeil, et que les algorithmes développés constituent des outils d'aide au diagnostic et non des outils de diagnostic en tant que tels.

Annexes

Fonctionnement de la forêt d'arbres décisionnels

La forêt d'arbres décisionnels ou forêt aléatoire est un algorithme ensembliste. En effet, son fonctionnement est basé sur la combinaison de plusieurs classifieurs : des arbres de décision.

Un arbre de décision est un classifieur qui divise de manière successive l'espace, de manière à obtenir des sous-parties permettant de distinguer les catégories.
Prenons un exemple simple dans lequel nous n'avons que deux caractéristiques : la quantité d'ondes alpha, et l'amplitude des ondes EEG. Dans cet exemple simplifié, nous ne considérerons également que deux catégories : l'éveil et le sommeil. La Figure D13 présente la répartition, dans cet espace 2D, de différentes époques. L'arbre de décision sépare cet espace de manière répétitive, en identifiant des valeurs de seuils pertinentes pour la distinction des différentes classes. Il réalise cela jusqu'à ce que chaque sous-espace soit pur (constitué d'époques d'une unique classe) ou qu'un critère du classifieur soit atteint. Nous reviendrons sur cet aspect par la suite.
La Figure D14 présente un exemple des différentes séparations en sous-ensembles qui peuvent être réalisées.
L'arbre de décision équivalent est illustré Figure D15.

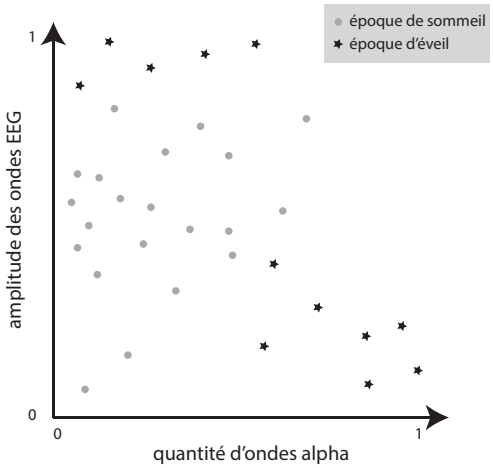


Figure D13 – Répartition de différentes époques dans un espace constitué de deux caractéristiques quantitatives comprises entre 0 et 1.

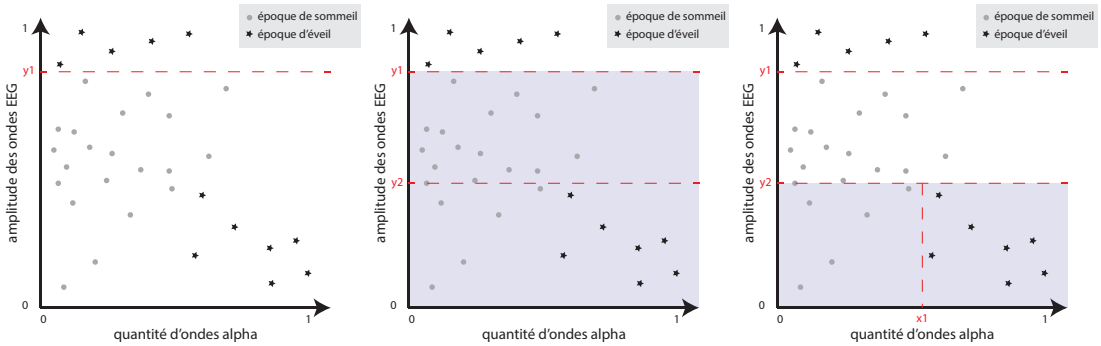


Figure D14 – Différentes étapes de séparation de l'espace avec l'objectif de distinguer les époques à l'éveil et au sommeil.

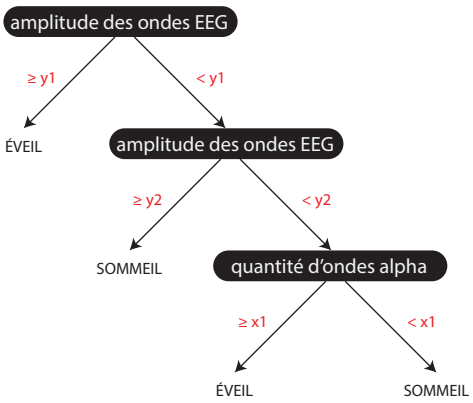


Figure D15 – Arbre de décision équivalent.

À chaque étape, l'arbre de décision doit choisir la caractéristique la plus adaptée pour séparer l'espace en deux sous-ensembles, avec le moins d'impureté possible. Pour cela, il évalue le gain d'information :

$$Gain\,d'information = entropie(espace) - ([moyenne\,pondérée] \times entropie(sous-espaces\,résultants))$$

L'entropie est une mesure statistique permettant d'estimer le caractère aléatoire de l'espace. Elle reflète ainsi son degré d'impureté. Plus le gain d'information est élevé, plus cela indique que la séparation est pertinente pour la distinction des classes.
En réalité, cette opération est réalisée dans un espace souvent constitué de bien plus que deux dimensions. Il n'est pas rare qu'il soit compliqué d'obtenir des sous-catégories pures. On tolère alors un certain niveau d'impureté. Elle peut être liée à un manque de caractéristiques pertinentes, mais peut aussi résulter d'un choix puisqu'il est parfois préférable de limiter la complexité de l'arbre, au profit de l'espace mémoire, du temps d'exécution et de son interprétation. Aussi, plus un arbre est complexe, moins il permet la généralisation et plus le risque de surapprentissage est grand.

La forêt d'arbres décisionnels est composée de plusieurs arbres de décision, entraînés sur différents sous-ensembles des données d'entrée. Ces sous-ensembles sont générés au hasard et peuvent se chevaucher. Une fois appliqués sur une nouvelle donnée, il en résulte plusieurs prédictions, qui ne sont pas nécessairement toutes identiques. Un vote est donc réalisé afin de définir la classification finale.
Il existe différents paramètres permettant d'ajuster, par exemple, le nombre d'arbres, la répartition des données servant d'apprentissage à chaque arbre, ou encore le déroulement du vote (en pondérant les arbres selon leurs performances).
Au contraire d'un seul arbre, qui peut être sensible au bruit, la forêt d'arbres décisionnels est plus robuste et obtient donc, généralement, de meilleures performances.

Un site internet bien vulgarisé (arbres de décision et forêt d'arbres décisionnels) : le blog pour l'apprentissage du Machine Learning « Machine Learning-101 » (2017) consultable ici : `https://medium.com/machine-learning-101`.
Une référence pour aller plus loin : Chapitre 4 de Criminisi et Shotton (2013), sur les forêt d'arbres décisionnels dans le cas d'une problématique de classification.

Fonctionnement du modèle de Markov à états cachés de Viterbi

Le modèle de Markov à états cachés (MMC) est un modèle statistique dans lequel le système modélisé est supposé être un processus markovien de paramètres inconnus.

Dans un premier temps, nous allons nous concentrer sur le processus markovien. On abordera donc ici le principe d'un modèle de Markov à états observables. Pour cela, il nous faut considérer chaque stade de sommeil comme étant un « état », et l'hypnogramme comme étant une séquence d'« observations ». On notera donc les états de notre modèle de Markov à états observables : $états = \{E, N1, N2, N3, SP\}$ et, dans le cas où notre enregistrement du sommeil comporte 1000 époques, nous aurons 1000 observations.
La probabilité de passer d'un état à l'autre dépend des différents états. Par exemple, la probabilité de passer de l'éveil au N3 est nulle, puisque l'endormissement se fait nécessairement par le stade N1. La probabilité de passer de l'éveil au N1 n'est pas nulle, mais est plus faible que celle de rester en éveil. Nous pouvons donc, en utilisant des données d'apprentissage, évaluer une matrice de transition indiquant les probabilités de passer d'un état à un autre (voir Table D8).

Table D8 – Exemple de matrice de transition.

Époque actuelle \ Époque suivante	E	N1	N2	N3	SP
E	0,9	0,1	0,0	0,0	0,0
N1	0,2	0,5	0,3	0,0	0,0
N2	0,1	0,1	0,7	0,1	0,0
N3	0,1	0,0	0,1	0,6	0,2
SP	0,2	0,0	0,1	0,1	0,6

Le modèle de Markov à états cachés, au contraire du modèle de Markov à états observables, considère qu'en plus des probabilités liées aux transitions entre les états, il existe des probabilités d'émission des observations. Les états, cachés, ne sont désormais plus équivalents aux stades de sommeil. On les notera $états = \{C1, C2, C3, C4, C5\}$. Ces états n'indiquent donc plus nécessairement un unique stade de sommeil. Cependant, c'est à partir de chacun d'entre eux que seront désignés, selon des probabilités d'émission, les observations. Par exemple, à partir de l'état caché C1, il serait possible que l'observation émise soit de l'éveil, mais également tout autre stade selon différentes probabilités.

L'algorithme de Viterbi se sert de cela pour corriger une séquence d'entrée. En effet, dans notre situation, les probabilités d'émission peuvent être estimées à partir des probabilités de classification d'une époque dans un certain stade, au lieu d'un autre. La matrice d'émission indique donc les probabilités de confondre un stade de sommeil avec un autre. Un exemple est disponible Table D9. Dans cet exemple, l'état caché C1 peut émettre une observation de type éveil (p = 0,7), N1 (p = 0,1), N3 (p = 0,1) ou SP (p = 0,1). Cela voudrait dire qu'en entrée du modèle, on considère qu'il est possible qu'une époque d'éveil soit confondue avec une époque de N1, N3 ou SP. En pratique, la matrice d'émission est donc équivalente à la matrice de confusion de l'hypnogramme à corriger (provenant bien sûr toujours d'un jeu d'apprentissage).

Table D9 – Exemple de matrice d'émission.

	E	N1	N2	N3	SP
C1	0,7	0,1	0,0	0,1	0,1
C2	0,3	0,3	0,1	0,0	0,3
C3	0,1	0,1	0,5	0,2	0,1
C4	0,0	0,0	0,3	0,7	0,0
C5	0,2	0,2	0,0	0,0	0,6

La Figure D16 illustre les automates de Markov à états observables et cachés correspondants. Pour faciliter leur lecture, nous ne considérons que l'éveil E et le sommeil S (qui regroupe tous les autres stades). Les probabilités de transition et d'émission ont été évaluées à partir des Tables D8 et D9.

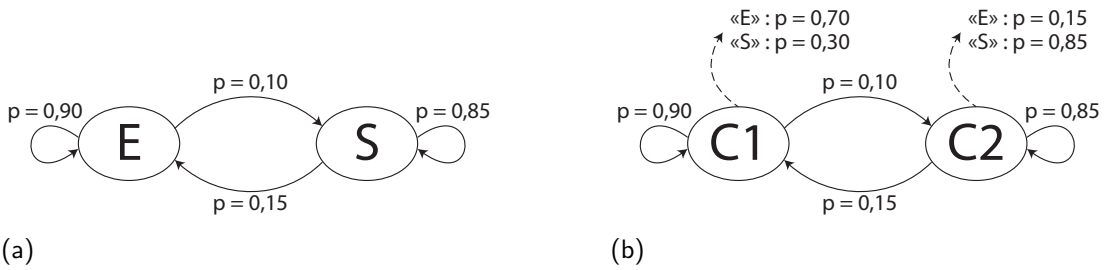


Figure D16 – Exemple d'automates de Markov : a) à états observables et b) à états cachés.

L'algorithme de Viterbi est donc utilisé pour déterminer la meilleure suite d'états à partir d'un modèle de Markov à états cachés dont les matrices de transition et d'émission sont connues. Pour rappel, c'est en effet notre cas puisque la matrice de transition est obtenue à partir des hypnogrammes de référence de notre jeu d'apprentissage, et que la matrice d'émission équivaut à la matrice de confusion (entre les hypnogrammes à corriger et les hypnogrammes de référence) de notre jeu d'apprentissage.
Pour cela, à partir d'une suite d'observations O, il cherche à trouver une séquence d'états C pour laquelle la probabilité *a posteriori* $\mathbf{P}(C|O)$ est maximale. En d'autres termes, l'algorithme de Viterbi détermine la séquence la plus probable d'états ayant conduit à la séquence d'observations. L'hypnogramme corrigé est ensuite déduit de cette séquence d'états.

En pratique, la première étape de l'algorithme de Viterbi consiste à déterminer les probabilités associées à la première observation (la première époque). Pour des raisons pédagogiques, nous considérerons dans notre exemple que la probabilité que l'état initial soit C1 vaut $0,6$, et $0,4$ pour C2. Ainsi, dans le cas où la première observation est de l'éveil, la probabilité que l'état caché menant à cette observation soit C1 vaut $0,6 \times 0,7 = 0,42$, et $0,4 \times 0,15 = 0,06$ pour C2. La deuxième étape de l'algorithme consiste à passer en revue tous les états possibles de la deuxième observation, et d'estimer le meilleur état de l'observation précédente. La Figure D17 présente un exemple dans lequel la deuxième observation serait du sommeil. Dans cet exemple, on voit dans le (a) que la meilleure probabilité pour le cas où l'état de la deuxième observation est C1 est $0,1134$. Cette probabilité est obtenue si l'état de la première observation est C1. On voit dans le (b) que la meilleure probabilité pour le cas où l'état de la deuxième observation est C2 est environ $0,0434$. Cette probabilité est obtenue si l'état de la première observation est C2. Seules ces deux meilleures transitions seront donc gardées pour la suite.

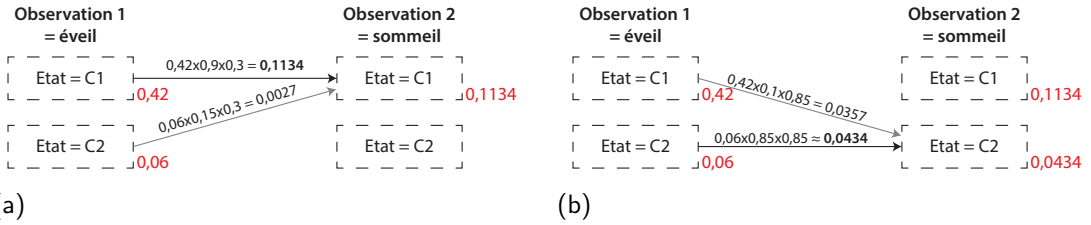


Figure D17 – Algorithme de Viterbi : exemple de la deuxième étape.

Cette deuxième étape est ensuite répétée pour chaque nouvelle observation. Nous poursuivrons notre exemple en imaginant que la troisième observation est de l'éveil. La Figure D18 présente un exemple dans lequel la troisième observation serait de l'éveil. Dans cet exemple, on voit dans le (a) que la meilleure probabilité pour le cas où l'état de la troisième observation est C1 est environ $0,0714$. Cette probabilité est obtenue si l'état de la première observation est C1. On voit dans le (b) que la meilleure probabilité pour le cas où l'état de la troisième observation est C2 est environ $0,0055$. Cette probabilité est obtenue si l'état de la première observation est C2. Encore une fois, seules ces deux meilleures transitions seront donc gardées pour la suite.

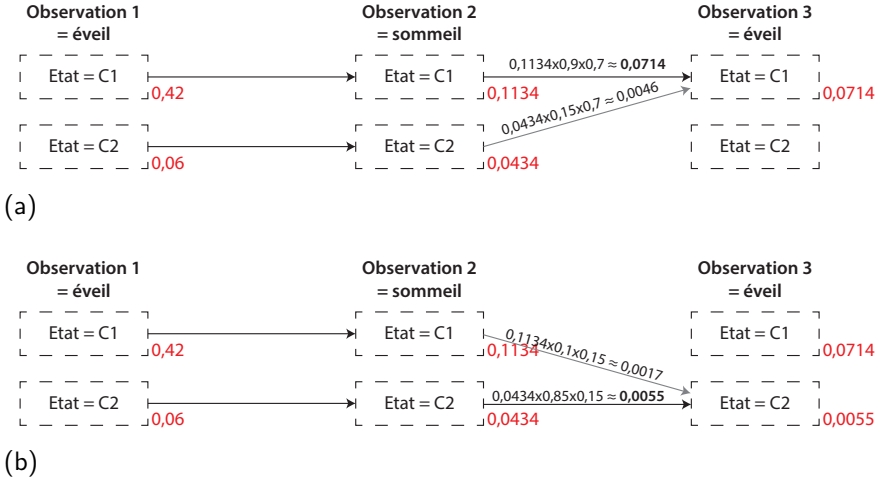


Figure D18 – Algorithme de Viterbi : exemple de la troisième étape.

Si l'on devait s'arrêter ici, nous nous retrouverions avec les probabilités indiquées dans la Figure D19.

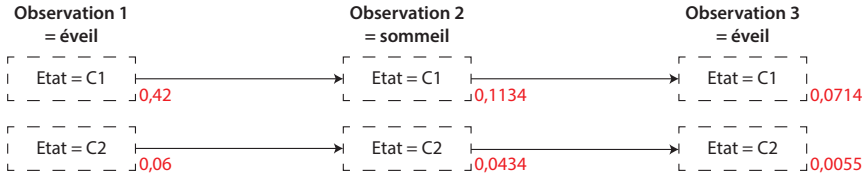


Figure D19 – Algorithme de Viterbi : exemple de la finalisation.

Dans ce cas, la séquence d'observation éveil→sommeil→éveil aurait une plus grande probabilité de découler des états C1→C1→C1 $(0,0714)$. En sortie, l'hypnogramme corrigé serait donc éveil→éveil→éveil (puisque pour rappel, la matrice d'émission est équivalente à la matrice de confusion).

Une référence pour aller plus loin : Chapitre 18 de Cornuéjols *et al.* (2018).

Bibliographie

Arnardottir, E. S., Thorleifsdottir, B., Svanborg, E., Olafsson, I. et Gislason, T. (2010). Sleep-related sweating in obstructive sleep apnoea : association with sleep stages and blood pressure. *Journal of Sleep Research*, 19(1p2):122–130.

Castanedo, F. (2013). A Review of Data Fusion Techniques. *The Scientific World Journal*, 2013:1–19.

Chen, C. (2016). *An e-health system for personalized automatic sleep stages classification.* Thèse de doctorat, Université Pierre et Marie Curie - Paris VI.

Chen, C., Ugon, A., Sun, C., Chen, W., Philippe, C. et Pinna, A. (2019). Towards a Hybrid Expert System Based on Sleep Event's Threshold Dependencies for Automated Personalized Sleep Staging by Combining Symbolic Fusion and Differential Evolution Algorithm. *IEEE Access*, 7:1775–1792.

Cornuéjols, A., Miclet, L. et Barra, V. (2018). *Apprentissage artificiel : Deep learning, concepts et algorithmes.* Algorithmes. Eyrolles, 3eme édition.

Criminisi, A. et Shotton, J. (2013). Classification Forests. *In* Criminisi, A. et Shotton, J., éditeurs : *Decision Forests for Computer Vision and Medical Image Analysis*, Advances in Computer Vision and Pattern Recognition, pages 25–45. Springer, London.

Dasarathy, B. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, 85(1):24–38.

Doshi-Velez, F. et Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv :1702.08608 [cs, stat].* arXiv : 1702.08608.

Durrant-Whyte, H. F. (1990). Sensor Models and Multisensor Integration. *In* Cox, I. J. et Wilfong, G. T., éditeurs : *Autonomous Robot Vehicles*, pages 73–89. Springer, New York, NY.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L. et Faraci, F. D. (2019). Automated sleep scoring : A review of the latest approaches. *Sleep Medicine Reviews*, 48.

Fogel, A. L. et Kvedar, J. C. (2018). Artificial intelligence powers digital medicine. *npj Digital Medicine*, 1(1).

Luo, R. C., Chih-Chen Yih et Kuo Lan Su (2002). Multisensor fusion and integration : approaches, applications, and future research directions. *IEEE Sensors Journal*, 2(2):107–119.

Parekh, A., Selesnick, I. W., Rapoport, D. M. et Ayappa, I. (2015). Detection of K-complexes and sleep spindles (DETOKS) using sparse optimization. *Journal of Neuroscience Methods*, 251:37–46.

Penzel, T. et Conradt, R. (2000). Computer based sleep recording and analysis. *Sleep Medicine Reviews*, 4(2):131–148.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H. et Faubert, J. (2019). Deep learning-based electroencephalography analysis : a systematic review. *Journal of Neural Engineering*, 16(5).

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y. L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S. C., Kim, T. W., Pizza, F., Plazzi, G., Vandi, S., Antelmi, E., Perrin, D., Kuna, S. T., Schweitzer, P. K., Kushida, C., Peppard, P. E., Sorensen, H. B. D., Jennum, P. et Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, 9(1):5229.

Topol, E. J. (2019). High-performance medicine : the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56.

Ugon, A. (2015). *Fusion Symbolique et Données Polysomnographiques*. Thèse de doctorat.

Vanbuis, J., Baffet, G., Feuilloy, M., Le Duff, A., Girault, J.-M., Meslier, N. et Gagnadoux, F. (2020a). Classification automatique et patient-dépendante des stades de sommeil, basée sur un processus de machine learning s'inspirant de la lecture manuelle. *Médecine du Sommeil*, 17(1):59–60.

Vanbuis, J., Feuilloy, M., Baffet, G., Meslier, N., Gagnadoux, F. et Girault, J.-M. (2020b). Towards a user-friendly sleep staging system for polysomnography part I : Automatic classification based on medical knowledge. *Informatics in Medicine Unlocked*, 21:100454.

Vanbuis, J., Feuilloy, M., Le Duff, A., Baffet, G., Girault, J.-M., Meslier, N. et Gagnadoux, F. (2019). Automated sleep stage classification using an adaptive patient-dependent algorithm based on physician-mimicking process. *Sleep Medicine*, 64:S400.

Vanbuis, J., Feuilloy, M., Riaboff, L., Baffet, G., Le Duff, A., Meslier, N., Gagnadoux, F. et Girault, J.-M. (2020c). Towards a user-friendly sleep staging system for polysomnography part II : Patient-dependent features extraction using the SATUD system. *Informatics in Medicine Unlocked*, 21:100453.

White, F. E. (1991). Data Fusion Lexicon :. Rapport technique, Defense Technical Information Center, Fort Belvoir, VA.

Chapitre E

Examen polygraphique

E.1 Objectif et contexte

Pour aller plus loin, nous avons tenté de réaliser une classification sur des examens polygraphiques. Dans le cas d'une PV, les informations relatives au sommeil ne sont pas disponibles. En effet, malgré de nombreux liens entre les stades de sommeil et les systèmes cardiaque et vasculaire, il est à l'heure actuelle impossible de lire le sommeil à partir des voies cardio-respiratoires.

L'objectif est donc ici d'estimer si certaines des informations relatives au sommeil peuvent être obtenues notamment grâce à des méthodes de ML ou DL. Avec ces informations, le spécialiste du sommeil serait en mesure d'affiner son diagnostic en PV. Dans certains cas, cela permettra au patient d'éviter un examen polysomnographique de seconde intention. En effet, lors d'une suspicion forte de SAHS, une PV est souvent réalisée. Si le résultat est négatif, non concluant ou techniquement inadéquat, il est alors recommandé de réaliser une PSG (Kapur *et al.*, 2017).

E.2 Problématiques identifiées

On retrouve certaines problématiques déjà mentionnées pour les cas d'un examen polysomnographique, telles que l'insuffisante hétérogénéité des données d'apprentissage et de test, et le manque d'ergonomie des approches développées.

Cependant, puisqu'il n'est pas possible de classifier les stades de sommeil manuellement à partir de ce type d'enregistrement, l'opacité du modèle n'est cette fois-ci généralement pas problématique. Le modèle doit d'ailleurs apprendre à réaliser une tâche que l'Homme ne sait pas faire. Pour cette raison, il est possible d'adopter des méthodes de DL sur les signaux bruts, ou de faire de la fouille de données.
La fouille de données, également appelée exploration de données ou data mining en anglais, est un domaine multidisciplinaire à l'intersection des statistiques, de l'intelligence artificielle et de l'informatique (bases de données). Elle consiste à extraire des informations d'un jeu de données de manière intelligente. Ces informations, préalablement inconnues, sont ensuite utilisées pour améliorer les connaissances ou, dans notre cas, classifier le sommeil.

Dans la littérature, certaines études adoptent une approche de ML constituée d'une évaluation d'un certain nombre de caractéristiques très précises puis directement d'une classification (Devot *et al.*, 2010; Hayet et Slim, 2012; Long *et al.*, 2015). D'autres études utilisent en plus une véritable étape de réduction de la dimensionnalité (Domingues *et al.*, 2014; Fonseca *et al.*, 2015; Willemen *et al.*, 2015). Cependant, toutes ces études sont réalisées à partir de signaux choisis en amont. La conséquence est que la fouille de données est alors limitée aux caractéristiques extraites desdits signaux. Ces approches ne permettent donc pas d'identifier les signaux de PV les plus judicieux pour la classification automatique du sommeil.

E.3 Stratégie

Nous avons ainsi choisi d'utiliser l'ensemble des signaux disponibles en PV, afin d'évaluer de nombreuses caractéristiques existantes (et déjà présentées dans la littérature). Une étape de réduction de la dimensionnalité est ensuite réalisée et nous permet de tirer des conclusions quant-aux signaux les plus utiles pour la classification automatique des stades de sommeil.

La méthode a été entraînée et testée sur un grand nombre d'enregistrements polysomnographiques (les 400 mêmes enregistrements que ceux utilisés pour la classification en PSG) provenant de patients sains ou avec troubles du sommeil. Le jeu de données est donc représentatif des enregistrements sur lesquels la méthode devrait être utilisée par la suite. L'apprentissage étant supervisé, les enregistrements ont préalablement été annotés par des spécialistes du sommeil du Laboratoire du Sommeil du CHU d'Angers, en suivant les recommandations AASM.

En plus de proposer un hypnogramme automatique, nous fournissons une fois encore aux médecins une table de probabilité. Étant donné que, cette fois-ci, le médecin ne sera pas en mesure de corriger les résultats de l'analyse automatique, cette table a été pensée pour rassurer le médecin sur la qualité de la classification automatique ou, au contraire, pour l'alerter dans le cas d'une classification trop complexe.

Dans la suite de ce chapitre, l'approche mise en place est décrite en détail. L'impact possible de ce travail sur le diagnostic des troubles du sommeil et plus particulièrement du SAHS est ensuite évalué.

E.3.1 Approche pour la classification des stades de sommeil

Ce travail a été présenté en détail par le biais d'un article soumis dans le journal IEEE Transactions on Biomedical Engineering.
Une description plus précise des classifieurs utilisés est disponible en Annexe : fonctionnement du perceptron multicouche (page 145) et en Annexe : fonctionnement du modèle de Markov à états cachés de Viterbi (annexe du chapitre précédent, page 116).

Résumé traduit

Les examens du sommeil de type III (examens de PV) ne permettent l'acquisition que des signaux cardio-respiratoires. Comparé à PSG, qui permet l'enregistrement supplémentaire des signaux électrophysiologiques, la PV présente cependant de nombreux avantages : elle est plus rapide à mettre en place et à lire, moins coûteuse et plus facilement réalisable en ambulatoire. Toutefois, sa précision est limitée par le manque d'informations sur le sommeil. Pour cette raison, de nombreux travaux ont présenté des caractéristiques cardio-respiratoires permettant l'étude de l'influence des stades du sommeil sur les activités cardiaques ou respiratoires.
Dans ce papier, un total 1111 caractéristiques déjà présentées dans la littérature ont été évaluées. L'oxymètre de pouls, le capteur PneaVoX® (enregistrant les sons trachéaux), les sangles thoracico-abdominales, la lunette nasale et l'actimètre ont fourni un ensemble de 112 caractéristiques estimé comme étant le plus pertinent pour l'analyse automatique des stades de sommeil. Ensuite, un modèle de classification en 3 étapes a été mis en place : classification avec un réseau de neurones artificiel, corrections à l'aide des règles de transition du sommeil (issues des recommandations de l'*American Academy of Sleep Medicine* ou AASM), et correction des séquences à l'aide d'un Modèle de Markov à états Cachés (MMC) de Viterbi. L'ensemble du processus a été entraîné et testé en utilisant 300 et 100 enregistrements indépendants, respectivement, et provenant de patients susceptibles d'avoir des troubles respiratoires du sommeil.
Les résultats ont indiqué que le système obtient un accord fort avec la lecture manuelle lors des classifications en 2 stades (éveil vs. sommeil : Kappa de Cohen κ moyen de 0,63 et taux d'accord *Acc* moyen de 87.8 %) et 3 stades (éveil vs. SP vs. SL : κ moyen de 0,60 et *Acc* moyen de 78.5 %). Cela indique que la méthode pourrait fournir des informations permettant d'aider les spécialistes lors du diagnostic du sommeil.
Le modèle présenté obtient des résultats prometteurs, et son impact clinique devra être étudié plus en détail.

Aide lexicale

- *Respiratory Inductance Plethysmography (RIP) belts* : sangles inductives
- *actigraphy* ou *actimetry* : actimétrie
- *REM sleep* ou *R stage* : Sommeil Paradoxal (SP)
- *NREM* : Sommeil Lent ou regroupement du N1, N2 et N3 (SL)
- *Obstructive Sleep Apnea (OSA) syndrome* est le terme utilisé en anglais pour SAHS [a]
- *HSAT* est le terme utilisé en anglais pour PV [b]

a. Les évènements ventilatoires peuvent être d'origine obstructive (dus à l'obstruction des voies aériennes supérieures), centrale (d'origine neurologique) ou mixte (dont l'origine est d'abord centrale puis devient obstructive). En France, on parle du Syndrome d'Apnées Hypopnées du Sommeil en général, mais il est également possible de distinguer le Syndrome d'Apnées-Hypopnées Obstructives du Sommeil (SAHOS), le Syndrome d'Apnées Obstructives du Sommeil (SAOS) ou le Syndrome d'Apnées Centrales du Sommeil (SACS). Dans les publications internationales, et malgré l'existence de termes équivalents, c'est généralement le terme « *Obstructive Sleep Apnea (OSA) syndrome* », qui est utilisé (même si l'origine obstructive n'a pas d'importance pour l'étude en question).

b. Dans les publications internationales, le terme « *Home Sleep Apnea Testing* (HSAT) » est communément utilisé pour mentionner la PV car cette dernière est fréquemment réalisée en ambulatoire.

A New Sleep Staging System for Type III Sleep Studies Equipped with a Tracheal Sound Sensor

Jade Vanbuis, Mathieu Feuilloy, Guillaume Baffet, Nicole Meslier, Frédéric Gagnadoux, and Jean-Marc Girault, *Senior Member, IEEE*

***Abstract*—Type III sleep studies record cardio-respiratory channels only. Compared with polysomnography, which also records electrophysiological channels, they possess many advantages: they are less expensive, less time-consuming, and more likely to be performed at home. However, their accuracy is limited by missing sleep information. This is why many studies presented specific cardio-respiratory parameters to assess sleep stages causal effects upon cardiac or respiratory activities.**
In this paper, many parameters proposed in the literature were gathered, leading to 1,111 features. The pulse oximeter, the PneaVoX® sensor (recording tracheal sounds), respiratory inductance plethysmography belts, the nasal cannula and the actimeter provided the 112 worthiest ones for automatic sleep scoring. Then, a 3-steps model was implemented: classification with a multi-layer perceptron, sleep transition rules corrections (from the AASM guidelines), and sequences corrections using a Viterbi hidden Markov model. The whole process was trained and tested using 300 and 100 independent recordings provided from patients suspected of having sleep breathing disorders.
Results indicated the system achieves a substantial agreement with manual scoring for classifications into 2 stages (wake vs. sleep: mean Cohen's Kappa κ of 0.63 and accuracy rate Acc of 87.8 %) and 3 stages (wake vs. R stage vs. NREM stage: mean κ of 0.60 and Acc of 78.5 %). It indicates the method could provide information to help specialists while diagnosing sleep.
The presented model had promising results, and should be further studied through the estimation of its clinical impact.

***Index Terms*—Automatic sleep staging, Type III sleep study, Cardio-respiratory channels, Tracheal sound sensor, Multilayer perceptron, Viterbi HMM.**

I. Introduction

SLEEP-related disorders are common in the population [1]–[3]. They can be diagnosed using several types of sleep diagnosis devices, classified according to the number of sensors and the conditions of recording:

- Type I = in-laboratory polysomnography (PSG) supervised by trained staff and composed of both electrophysiological (EP) and cardio-respiratory (CR) channels;
- Type II = out of center PSG composed of both EP and CR channels;
- Type III = in-laboratory or at home ventilatory polygraphy, composed of CR channels only;
- Type IV = in-laboratory or at home recording of one or two CR channels only.

Among these, Type I, II and III devices have satisfying precision, and Type I is considered as the gold-standard.
Compared with PSG, Type III is less expensive and the time allocated to the sensors placement is lower (due to the reduced number of recorded channels). Furthermore, a Type III device is easier to perform at home.
When testing for Obstructive Sleep Apnea (OSA), at home Type III diagnosis, also called Home Sleep Apnea Testing (HSAT), is commonly performed. The counterpart of using HSAT instead of PSG is that the diagnosis is less accurate and the obtention of false negatives may be possible. Indeed, no sleep scoring is available with Type III sleep studies, since EP channels are not recorded. Sleep scoring consists of the identification of all 30 seconds segments (called epochs) as belonging to wakefulness (W stage), light sleep (N1 and N2 stages), deep sleep (N3 stage) or rapid eye movement sleep (R stage). When diagnosing OSA, it is therefore recommended to perform a PSG when the first intention HSAT was negative, inconclusive, or technically inadequate [4]. This second intention testing should be avoided as it is a waste of time for the patient and the physician.

As previously mentioned, Type III devices possess many advantages but also drawbacks whose most important is a low accuracy limited by missing sleep information. This is why many studies investigated the links between CR channels and sleep stages. A recent review paper [5] synthesized the knowledge of cardiac activity modulations in PSG. The regulation of the autonomous nervous system across the different sleep stages was studied using several signals as heart rate, electrocardiogram (ECG) and cardio-respiratory coupling. Even if manual sleep scoring from CR channels remains impossible, it has been shown that there was causal relationships between sleep stages and cardiac or respiratory activities. During those last few years, many studies sought to exploit those relationships to propose automatic sleep scoring from CR channels only [6]–[15]. Most of those studies focused on a few channels and aimed to evaluate the impact of new features on an automatic sleep scoring, without dwelling much on classification. Results from those studies proved sleep staging could be performed from CR channels, at least partly. Among all studies, some are focused on the identification of a single sleep stage [6]–[10], [12], [14]. Others combine N1 and N2 sleep stages (result is a

Manuscript received March DAY, 2021; revised MONTH DAY, YEAR.

J. Vanbuis (e-mail: jade.vanbuis@eseo.fr), M. Feuilloy and J-M. Girault are with the ESEO, Angers, France and LAUM, UMR CNRS 6613, Le Mans, France.

G. Baffet was with CIDELEC, Sainte-Gemmes-sur-Loire, France.

N. Meslier and F. Gagnadoux are with the Angers sleep laboratory, University Hospital, Angers, France and INSERM UMR 1063, University of Angers, Angers, France.

4-stage classification) [12], [13], [15], or even N1, N2 and N3 sleep stages in the so-called NREM sleep stage (leading to a 3-stage classification) [8], [11]–[13], [15]. As a matter of fact, even 2-stage classifications greatly improve sleep diagnosis. If CR channels have proven not to be irrelevant for sleep staging, we can question their ability to score sleep while in Type III sleep studies. Indeed, almost all studies employed an electrocardiogram (ECG) sensor, that is not recorded in Type III sleep studies.

The work presented here aims to estimate how good automatic sleep staging can be while in Type III sleep studies. Cardiac information was thus only extracted from the pulse oximeter sensor. The targeted system was the CID-LX polygraph (CIDELEC, France), which has the particularity of being equipped with a tracheal sound sensor. This sensor is employed more and more frequently in recent Type III systems. Indeed, its benefits have been reported for the study of ventilatory activity [16]–[22] and, recently, of cardiac activity [23]. To the best of our knowledge, its use for sleep stages study was not yet investigated. The method proposed in this paper focuses on the classification process rather than the research of new features or their evaluation. It thus involves the evaluation of many well-known features from all recorded channels. The features kept for classification were then automatically selected regardless of their origin. Some available channels could thus be unemployed despite their recording. A complete 5-stage sleep scoring was also realized so the results could be compared with an automatic system for PSG (in this case the one presented in [24] which was implemented using the same recordings) and not only with the manual scoring.

In the following section (Section II), the recordings used for training and testing are first presented. The various features competing for classification are then listed and a reduced set of the best ones is composed. Lastly, the classifier architecture and evaluation process are detailed. In Section III, the system is compared with the manual scoring and its performance is reported. Its efficiency is discussed in Section IV. At last, the conclusion is presented in Section V.

II. Methods and materials

This section is split into three parts: the description of the data used, the description of the system implemented, and the description of the methodology established for the performance assessment.

A. Data acquisition

Four hundred anonymous sleep recordings were included in this study thanks to the sleep cohort of Pays de La Loire, operated under the aegis of the Institut de Recherche en Santé Respiratoire. Approval was obtained from the University of Angers ethics committee and the "Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé" (CCTIRS; 07.207bis).

The datasets used were the exact same ones as in [24]: the training dataset, D1, consisted of 300 PSG recordings, distinct from the 100 PSG recordings of the test dataset, D2. D1 and D2 were randomly built and we ensured there was equal representation of the severity of OSA and the year of recording. D1 dataset was made up of 329,911 epochs (W: 76,897 - R: 52,036 - N1: 24,283 - N2: 122,266 - N3: 54,429). As for D2 dataset, it was made up of 110,978 epochs (W: 24,172 - R: 18,194 - N1: 8,095 - N2: 41,095 - N3: 19,422). Recordings were acquired from patients suspected of having sleep breathing disorders (SBD), using the CID102L8D polysomnograph (CIDELEC St Gemmes-sur-Loire, FRANCE). EP channels were acquired and manually scored following the American Academy of Sleep Medicine (AASM) recommendations of good practice [25]. The resulting hypnogram will be referred to as $hypno_{ref}$ in the rest of this paper. CR channels were also acquired with the CID102L8D polysomnograph. Besides the usual ones, this polysomnograph is equipped with a tracheal sound sensor (the PneaVoX® sensor). Placed over the suprasternal notch, this sensor simultaneously records tracheal sounds, snoring and estimates suprasternal pressure.

B. Algorithm structure

The algorithm was implemented using MATLAB® software. Its inputs were CR signals available in Type III sleep study (with the PneaVoX® sensor), and a priori medical knowledge from the AASM guidelines [25]. Its outputs were the automatically estimated hypnogram $hypno_{CR}$, and a probability table $probabilities_{CR}$ which gives information about the confidence level of the algorithm when it classified the epochs. The proposed approach is the systematic one when doing machine learning, involving several steps: features generation, dimensionality reduction and classification.

1) Features generation: This first step aimed to generate multiple descriptors from each CR signal available in Type III sleep studies. Those descriptors are quantitative features with one value per epoch. The sensors and associated signals used in this study are listed in Table I.
For each sensor, we identified the signals that should be normalized (using MATLAB® zscore function) or not. For example, the signals obtained from respiratory inductance plethysmography (RIP) belts depend on the belts adjustment on the patient. It could thus be interesting to interpret their values with regard to their evolutions over time, within the same recording. In this case, a normalization should be considered. However, since the recordings come from patients with various pathologies, some may have greater cardio-respiratory variations compared with others. This is why features from raw signals were always computed. Only some signals were duplicated to get a normalized version.
Five features categories were identified. The 1st one regroups the usual statistical features. They were estimated for all signals. The 2nd, 3rd and 4th categories were the usual cardiac features. In the literature, heart rate variability (HRV) parameters [26] are often estimated from the RR intervals

TABLE I
LIST OF THE SENSORS AND RECORDED SIGNALS THAT WERE USED FOR FEATURES GENERATION.

Sensor	Signals
Pulse oximeter	Photoplethysmogram
	Saturation in oxygen
PneaVoX®	Breathing sounds intensity
	Breathing sounds energy
	Snoring sounds intensity
	Snoring sounds energy
	Snoring interval
	Snoring regularity
	Energy ratio
	Inspiratory duration
	Breathing cycle duration
	Breathing durations ratio
	Suprasternal pressure
RIP belts	Thoracic belt
	Abdominal belt
	RIP belts ratio
	RIP belts phase shift
Nasal cannula	Nasal airflow
Actimeter	Actigraphy
Light sensor	Ambient light

(RRI)[1] of the ECG signal. In this paper, cardiac activity was available through the photoplethysmogram (PPG) signal (from the pulse oximeter sensor) and not from the ECG signal. In [27], it has been shown that pulse rate variability (PRV) parameters, estimated from the pulse-to-pulse intervals (PPI) of the PPG signal, can globally be used as a replacement for HRV parameters. The 2nd, 3rd and 4th features categories thus regrouped temporal, spectral and non-linear PRV features, respectively. The 5th and last features category was composed of shape-related PPG features. Table II presents all features and references of papers with descriptions of them. Indeed, since our objective was to gather a substantial amount of features from multiple papers, features equations will not be detailed in the present article.
In the end, a total of 1,111 features were estimated.

2) Dimensionality reduction: A dimensionality reduction process was set to reduce the number of features used for classification by selecting the most relevant ones. Indeed, the model can be confused when having too many input features. Plus, the computation time would be problematic for a daily basis use. The dimensionality reduction process is illustrated in Figure 1.
It was composed of three steps. First, the correlation within features was computed to identify similar features (when features were similar, only one of them was kept). Then, a Kruskal-Wallis test was performed to identify features whose p-value suggests its distribution is highly independent from the output ($hypno_{ref}$) distribution. Those features were disqualified for the next selection step. Lastly, ten random forests were trained on each tenth of the input matrix length. The TreeBagger MATLAB® function used gave us measures of relevance for each feature, consequently considered relevant or not for the classification. Only the features which were always considered being relevant for the classification were kept.

[1]Interval between consecutive heart beats, measured using the R waves.

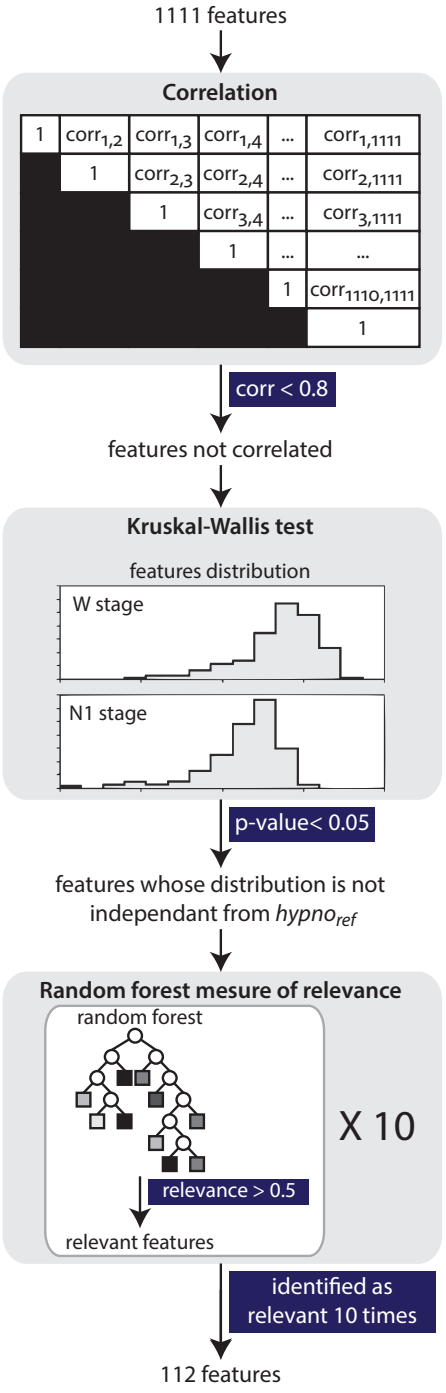
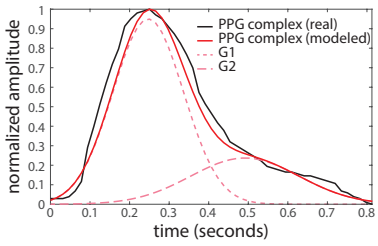


Fig. 1. Dimensionality reduction process implemented to reduce the number of features before classification.

TABLE II
LIST OF ALL FEATURES GENERATED FROM THE SIGNALS, WITH REFERENCES TO DESCRIPTIVE ARTICLES, SPLIT IN FIVE CATEGORIES.

Category	Names	References
Statistical features	1st to 4th order statistical moments: mean, variance, skewness and kurtosis	[28]–[31]
	Means ratio (e/e-1)	
	Standard deviation	
	Root mean square	
	Hjorth parameters: complexity and mobility	[28], [29]
	Percentiles: 10th, 25th, 50th, 75th and 90th	
	Inter-quartile range: 90th-10th and 75th-25th	[32]
	Zero-crossing analysis	[11], [28]
Temporal PRV features*	meanHR: mean heart rate	[11], [15], [32]
	meanNN: mean of NN intervals	[11], [14], [15], [33]
	SDNN: standard deviation of NN intervals	[11], [14], [26], [33]
	CVNN: coefficient of variation (or relative standard deviation) of NN intervals	[11], [33]
	RMSSD: root mean square of the successive NN differences	[11], [14], [26], [33]
	SDSD: standard deviation of the successive NN differences	[26]
	pNN50: proportion of adjacent NN intervals differing by more than 50 ms	[11], [26], [33]
	pNN625: proportion of adjacent NN intervals differing by more than 6.25 ms	[33]
Spectral PRV features†	Frequency ranges: very low frequency (VLF), low frequency (LF), medium frequency (MF), high frequency (HF) and total power (TP)	[11], [26], [32]–[34]
	Frequency ranges ratios: VLF/TP, LF/TP, MF/TP, HF/TP, VLF/HF, LF/HF and MF/HF	[14], [26], [32]–[34]
Non-linear PRV features	Approximate entropy ApEn TRIG‡, centered and not centered	[32], [35], [36]
	Sample entropy SampEnt TRIG‡, centered and not centered	[14], [32], [35], [36]
	Fuzzy SampEnt TRIG‡, centered and not centered	[36], [37]
	Lempel-Ziv complexity	[28]
Shape-related PPG features§	G1 mean and G2 mean	
	difference between G1 and G2 means	
	G1 standard deviation and G2 standard deviation	
	G1 weighting coefficient and G2 weighting coefficient	
	G1 amplitude and G2 amplitude	
	Difference between G1 and G2 amplitudes	
	Difference between PPG complex and G1 barycentres (time and amplitude)	
	PPG complex foot	
	G1 foot and G2 foot	
	Difference between PPG complex barycentre and foot (time)	



* Normal-to-normal (NN) beat intervals were estimated from removed artifacts PPI signal.
† Estimated using *Long et al.* [34] approach.
‡ Estimated for the four types of symmetry: translation (T), vertical reflection (R), inversion (I) and glide reflection (G).
§ Each PPG complex was modeled using a Gaussian mixture of two Gaussian components G1 and G2.

The 112 selected features were estimated from: the pulse oximeter (23 features), the PneaVoX® (23 features), RIP belts (56 features), the nasal cannula (7 features) and the actimeter (3 features). A detailed list is available in the Appendix.

3) Classification: Because no proper rules for sleep staging using only cardio-respiratory signals are available, several classifiers were considered (with some more opaque than others) in order to obtain empirical rules via a training phase. On top of that, we were still able to use some of the medical knowledge described in the AASM guidelines. Indeed, sleep transition rules, which give information about the possible transitions between sleep stages, are not related to specific channels. The classification was thus achieved using the 112 selected features, combined with a priori knowledge of sleep transition rules. Its outputs were $hypno_{CR}$ and the probability table $probabilities_{CR}$. A combination of three main functions (**II-B3**, **II-B3** and **II-B3**) was implemented, as illustrated in Figure 2. The first hypnogram is computed in **II-B3**. It is then corrected in **II-B3** and **II-B3**, which ensure there is no forbidden epochs sequences.

C1 A large matrix of the 112 selected features from all D1 recordings was first created. In the same way, all D1 references $hypno_{ref}$ were combined together. Using those matrices as input and reference, respectively, several classifiers were trained and compared: k-nearest neighbor, random forest, support vector machine, multinomial logistic regression, multilayer perceptron neural network and also long short-term memory recurrent neural network. The best model was the multilayer perceptron model [38]. It is a feedforward neural network model which last layer neurons correspond to the outputs of the system. The implemented multilayer perceptron thus had 5 neurons on its last layer (one per sleep stage). It also had one hidden layer of 10 neurons, and was constructed using the Levenberg-Marquardt optimization. The trained multilayer perceptron, called ***MLP***, was saved

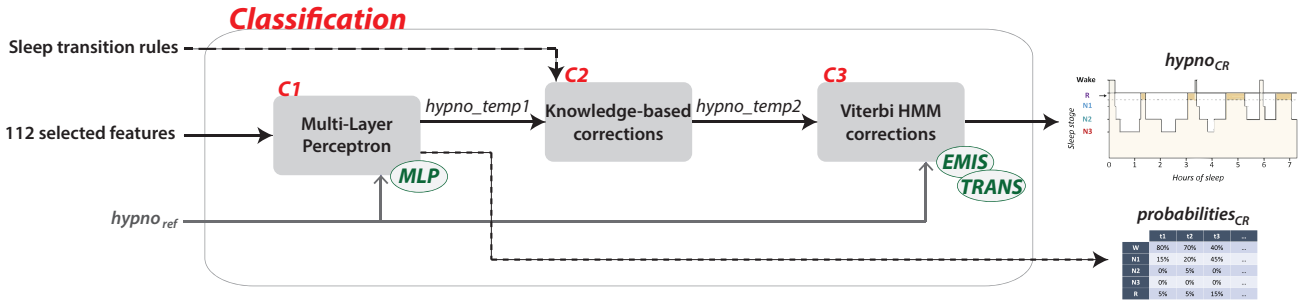


Fig. 2. Functional architecture of the classification step, composed of three main functions (C1, C2 and C3). Inputs are sleep transition rules (from the AASM manual) and the 112 selected features (from dimensionality reduction). The manual scoring $hypno_{ref}$ was used for training. The output are the hypnogram $hypno_{CR}$ and the associated probability table $probabilities_{CR}$.

to be further used for each test recording (D2) individually. The output was a matrix composed of 5 values (one for each sleep stage) per epoch, with the highest values designating the estimated sleep stages. The resulting hypnogram will be referred to as *hypno_temp1* in the rest of this paper, and the entire output matrix was returned under the name $probabilities_{CR}$. This matrix gives indications about the algorithm's doubts when scoring sleep. Thanks to it, the medical practitioner knows if the epochs were difficult or easy to score by the algorithm.

C2 *hypno_temp1* was smoothed using transition rules corrections implemented using the AASM guidelines but also empirically, by studying the most recurring errors. The resulting hypnogram will be referred to as *hypno_temp2* in the rest of this paper.

C3 Lastly, *hypno_temp2* is smoothed by a Viterbi hidden Markov model. This model also aims to correct common sequences mistakes, using an emission probability matrix ***EMIS*** and a transition probability matrix ***TRANS***. Those matrices, which correspond respectively to the confusion matrix and $hypno_{ref}$ transition probability, were estimated from training recordings (D1). ***EMIS*** and ***TRANS*** were saved to be further used for each test recording (D2) individually. The resulting hypnogram is the final one. It will be referred to as $hypno_{CR}$ in the rest of this paper.

C. System evaluation

Results were estimated from the recordings included in the test dataset D2. As a reminder, this dataset is composed of independent recordings not used during training. Each recording was treated using the previously trained models. Then, the resulting $hypno_{CR}$ was compared with its associated $hypno_{ref}$ (in II-C1). Last, $probabilities_{CR}$ was gauged using $hypno_{ref}$ (in II-C2). In both sections, individual scores were averaged for enhanced clarity.

1) Evaluation of $hypno_{CR}$ accuracy: For each recording, the overall accuracy of $hypno_{CR}$ compared to $hypno_{ref}$ was estimated using the usual Cohen's Kappa κ [39] and accuracy rate Acc. Cohen's Kappa index is usually interpreted using six ranges:

1) $\kappa < 0.0$: Poor agreement
2) $0.0 \leq \kappa < 0.2$: Slight agreement
3) $0.2 \leq \kappa < 0.4$: Fair agreement
4) $0.4 \leq \kappa < 0.6$: Moderate agreement
5) $0.6 \leq \kappa < 0.8$: Substantial agreement
6) $0.8 \leq \kappa$: Almost perfect agreement

Furthermore, scores were estimated for each sleep stage individually. To do so, we adopted a one-vs.-rest approach where each stage is alternatively considered as the positive class and the others are combined into a single negative class. From the resulting true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), we estimated Cohen's Kappas (detailed in [24]), accuracy rates (defined as $\frac{TP+TN}{TP+TN+FP+FN}$), sensibilities (defined as $\frac{TP}{TP+FN}$) and specificities (defined as $\frac{TN}{TN+FP}$).
Some sleep parameters missing in Type III sleep studies can be estimated from a simplified sleep stage scoring. For example, the estimation of the total sleep duration, called Total Sleep Time (TST), only requires a 2-stages classification W / sleep (wake versus the combination of N1, N2, N3 and R sleep stages). Another common 2-stages scoring is R / other (R sleep stage versus the combination of W, N1, N2 and N3 sleep stages). The 3-stages classification W / R / NREM, which corresponds to a wake versus R sleep stage versus the so-called NREM stage (combination of N1, N2 and N3 sleep stages), is also frequent in the literature. For those three simplified scorings, reported results were assumed from the confusion matrix obtained while doing a 5-stage classification. No new training was realized even though it could deliver a better performance for 2 or 3-stage classification.

2) Decision support with $probabilities_{CR}$: Since $probabilities_{CR}$ has no reference to be compared with, we evaluated the mean probability given by $probabilities_{CR}$ when epochs are correctly versus erroneously scored in a specific sleep stage. The greater the difference between those mean probabilities will be, the better. Indeed, we want the algorithm mistakes to be limited to epochs where the features designated multiple sleep stages rather than a single one

TABLE III
OVERALL AND INDIVIDUAL PERFORMANCE OBTAINED FROM AUTOMATIC SLEEP STAGING ON D2 DATASET.

D2 dataset	W	N1	N2	N3	R	All
Cohen's Kappa	0.63 ± 0.15	0.09 ± 0.06	0.36 ± 0.17	0.48 ± 0.23	0.53 ± 0.25	0.48 ± 0.13
Accuracy rate (%)	87.8 ± 5.6	91.7 ± 3.3	70.5 ± 9.0	86.3 ± 7.1	88.5 ± 8.1	62.4 ± 10.1
Sensibility (%)	80.5 ± 14.2	8.0 ± 4.5	64.1 ± 18.0	59.7 ± 30.4	58.3 ± 29.7	N.A.
Specificity (%)	89.5 ± 7.5	98.3 ± 0.8	73.4 ± 14.9	92.1 ± 8.1	94.3 ± 9.3	N.A.

N.A. = Not Applicable

(suggesting manual scoring may has been complex as well).

III. RESULTS

A. Evaluation of $hypno_{CR}$ accuracy

Among the 100 recordings, 17 % obtain an overall Cohen's Kappa κ above 0.60, and 73 % above 0.40.
Table III gives, for D2 recordings, the mean value and the standard deviation of the overall Cohen's Kappa κ and accuracy rate Acc of $hypno_{CR}$. Scores per sleep stage are also reported. The overall κ and Acc are 0.48 and 62.4 %, respectively. If we consider each sleep stage detection individually, stage W obtains the best scores with a κ reaching 0.63. It is the only stage with a substantial agreement with the manual scoring. Cohen's Kappa mean values indicate a moderate agreement with the manual scoring for sleep stages N3 and R, a fair agreement for stage N2, and a slight agreement for stage N1. κ mean values indicate all other sleep stages get a substantial agreement with the manual scoring, except stage N1. Sensibilities are all above 82 %, except for N2 and N1 sleep stages (74.9 % and 20.8 %, respectively). Specificities are all above 94 %, except for N2 sleep stage (83.3 %).
Figure 3 presents the confusion matrix related to Table III. N2 sleep stage is clearly over-scored, and some N2 sleep epochs are mistaken in N3 sleep stage.

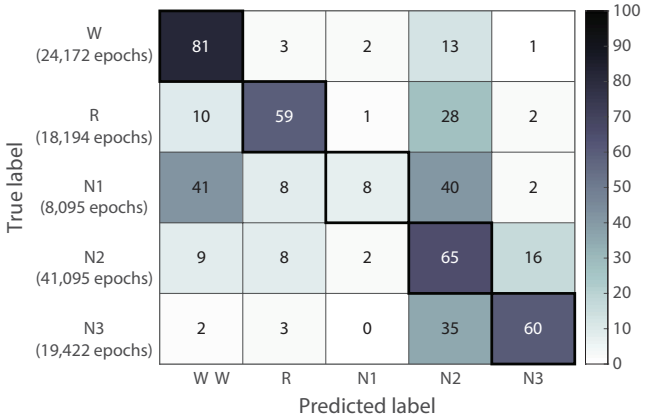


Fig. 3. Confusion matrix (percentage) obtained from automatic sleep staging on D2 dataset.

The κ and Acc of the simplified classifications are also reported in Table IV.
Because no specific training was realized, results for the 2-stages classifications are the same as the individual results presented in Table III. For the 3-stages scoring, mean κ and Acc are 0.60 and 78.5 %, respectively (showing a substantial agreement with the manual scoring).

TABLE IV
PERFORMANCE OBTAINED FOR SIMPLIFIED SCORINGS (2 OR 3 STAGES CLASSIFICATIONS) ON D2 DATASET.

D2 dataset	W / sleep	R / other	W / R / NREM
κ	0.63 ± 0.15	0.53 ± 0.25	0.60 ± 0.14
Acc (%)	87.8 ± 5.6	88.5 ± 8.1	78.5 ± 9.0

B. Decision support with $probabilities_{CR}$

$probabilities_{CR}$ is an additional tool to $hypno_{CR}$. Table V reports the mean probabilities associated with epochs correctly and erroneously estimated by $hypno_{CR}$, depending on the sleep stages.

TABLE V
MEAN PROBABILITIES PER SLEEP STAGE WHILE THE REFERENCE AGREES OR DISAGREES WITH THE STAGE DETECTED BY THE SYSTEM.

	$hypno_{ref}$ agrees	$hypno_{ref}$ disagrees
$hypno_{CR}$ = W	66 %	45 %
$hypno_{CR}$ = N1	16 %	14 %
$hypno_{CR}$ = N2	56 %	51 %
$hypno_{CR}$ = N3	57 %	48 %
$hypno_{CR}$ = R	58 %	44 %

In this table, we can see that epochs correctly identified as W have a reported mean probability of 66 % of being W stage, whereas the epochs over-detected as W have a reported mean probability of only 45 % of being W. It makes a difference of 21 % between actual W epochs and not. The algorithm confidence when scoring an epoch into W sleep stage is thus greater when it is an actual W epoch. Considering the mean probabilities reported for the other sleep stages, we can see that only R stage had a difference higher than 10 %.
Erroneous epochs are thus more likely to have small and uniform $probabilities_{CR}$ values than correctly classified ones. Consequently, $probabilities_{CR}$ gives indications about the level of confidence of the algorithm when scoring sleep.

IV. DISCUSSION

The main goal of this study was to implement an automatic sleep scoring system for Type III sleep studies, in which electrophysiological channels are not recorded. Tracheal sensor, which is increasingly recognized for sleep diagnosis,

was also recorded and employed. Main difficulties were to implement an approach using only cardio-respiratory channels, and suitable for patients with various sleep breathing disorders (SBD). Indeed, by definition, cardiac and respiratory signals from patients with SBD are more disruptive than healthy patients signals. In the literature, causal relationships between new specific features and sleep staging are often studied on patients with normal sleep. The aim of this paper was to gather extensive knowledge about specific features and channels seen in the literature, and develop a method to automatically sort those information and classify sleep stages.

To do so, a total of 1,111 features of different kind (statistical, temporal, spectral non-linear or shape-related) were first generated. Those features were identified as being promising in the field literature. Then, a dimensionality reduction was realized so only best features were kept as inputs of the following classification. Lastly, a classification consisting in several functions was implemented. The estimated hypnogram, $hypno_{CR}$, was thus built taking into consideration the transition rules as described in the AASM guidelines. Along with $hypno_{CR}$, the probability table $probabilities_{CR}$ was returned.

The method showed it could get moderate agreement with manual scoring, reaching Cohen's Kappa values around 0.48 and accuracy rates around 62 %. Considering each sleep stage individually (see Figure 3 and Table III), it seems that the errors were mainly overestimations of N2 sleep stage or N2 epochs erroneously scored in N3 sleep stage. Also, sleep stage N1 had a very low performance compared to other stages. This is not surprising, since sleep stage N1 is a transitional stage representing approximately 5 % of the night with a very likely overlap with W and N2 stages. In fact, inter-scorer agreement for N1 sleep stage is the lowest [40]. Table V showed that W and R misclassified epochs have lower probability values in the returned table $probabilities_{CR}$, compared with correctly scored epochs. A strategy would be to provide a confidence level for each epoch to ponder results.

Because of the moderate agreement with manual scoring, it is common in the literature to express the results using simplified hypnograms (2 or 3 stages classification). Indeed, a reduced number of stages is sufficient to obtain some of the necessary information for sleep diagnosis. For example, total sleep time is an important indicator for OSA screening which can be estimated from W/sleep classifications. On the other hand, distinction of W, R and NREM can help for the diagnosis of some syndromes, like narcolepsy. It should be noted that comparison within studies is very challenging due to the distinct datasets and methods of training/validation/test. However, for information purposes, we indicated in Table VI the results from the literature and from the present paper (simplified sleep scorings results were estimated from the 5-stage classification matrix, without new specific training). Studies validated on healthy patients only were dissociated and the number of patients were indicated for each study, with the training/test repartitions employed. All studies reported in Table VI using cardiac signals included an ECG (which is not normally recorded when using Type III sleep diagnosis devices). On the other hand, only our study included tracheal sound information. Those reinforce the fact that comparison has to be considered with caution. Therefore, we can only state that we are satisfied by our results regarding the literature, in both scenarios (subjects with or without SBD or healthy subjects only). We also note that among the studies using both

TABLE VI
PERFORMANCE OF W/SLEEP, R/OTHER AND W/R/NREM CLASSIFICATION USING CARDIO-RESPIRATORY CHANNELS, OBTAINED IN OTHER STUDIES AND IN THE PRESENT PAPER (COMPARE WITH CAUTION).

	Number of one-night recordings (training and test)	Channels	W / sleep		R / other		W / R / NREM	
			Acc (%)	κ	Acc (%)	κ	Acc (%)	κ
Healthy and SBD subjects								
Devot et al. 2010 [6]	35 (LOOCV*)	C, Re, M	87	0.62				
Hayet et Slim 2012 [7]	16 (2/3-1/3)	C	65					
Domingues et al. 2014 [8]	20 (LOOCV*)	C, Re, M	80				66	
Willemen et al. 2015 [9]	85 (57-28)	C, Re	78	0.44				
Yoon et al. 2017 [10]	51 (26-25)	C			87	**0.61**		
This work	400 (300-100)	C, Re, M	**88**	**0.63**	**88**	0.53	**79**	**0.60**
Healthy subjects only								
Xiao et al. 2013 [11]	45 (LOOCV*)	C	56	0.48	60	0.50	73	0.46
Fonseca et al. 2015 [12]	48 (10-fold CV**)	C, Re	**91**	0.51	**87**	**0.58**	**80**	0.56
Long et al. 2015 [13]	48 (10-fold CV**)	Re					77	0.48
Aktaruzzaman et al. 2017 [14]	18 (LOOCV*)	C, M	76	0.31				
Gaiduk et al. 2018 [15]***	35 (5-30)	C, Re, M	83	0.40	86	0.37	72	0.42
This work	****	C, Re, M	88	**0.59**	**87**	0.51	78	**0.58**

C = cardiac signal(s), Re = respiratory signal(s), M = movement signal(s)

* Leave-One-Out Cross-Validation: one by one, each subject recordings were selected as the test dataset, and the others were combined into the training dataset. Final results are provided by the model with the highest scores.

** 10-fold Cross-Validation: recordings are divided into 10 groups. Nine are used for training and the remaining one for testing. As LOOCV, final results are provided by the model with the highest scores.

*** Scores estimated from the given confusion matrix.

**** **No new training**, results expressed on the 21 healthy patients extracted from the test dataset. For information, there was a total of 74 healthy patients among the 400 in total.

healthy and SBD subjects, the majority are limited to the scoring of a unique sleep stage. In the end, considering the training/test strategy, the fact that all stages were estimated, and the important number of patients (suggesting a high adaptation to sleep diversity), we are fully satisfied with the outcomes of the proposed method.

To go further, we compared the automatic scoring with the one presented in [24] (which was designed for polysomnographic sleep studies), since they were trained and tested on the exact same recordings. We tried to figure out what were the weaknesses of the Type III automatic sleep scoring compared to the polysomnographic one. Table VII reports the Cohen's Kappa scores for both automatic tools. The overall

TABLE VII
MEAN COHEN'S KAPPA OBTAINED FROM BOTH AUTOMATIC SLEEP STAGING TOOLS, ON D2 DATASET.

D2 dataset	W	N1	N2	N3	R	All
Type I or II	0.74	0.23	0.64	0.71	0.80	0.69
Type III	0.63	0.09	0.36	0.48	0.53	0.48
Difference	0.11	0.14	0.28	0.23	0.27	0.21

mean Cohen's Kappa value decreases of 0.21 when using the tool using cardio-respiratory (CR) channels instead of the one based on electrophysiological (EP) channels. Considering each stage individually, W stage seems to be the less impacted. On the contrary, the reported differences and the confusion matrices show that N2, N3 and R sleep stages are the most affected. Interestingly, N2 and R sleep stages are the ones requiring the identification of sleep patterns (only possible on EP channels). This could be a first reason explaining why the algorithm is facing difficulties when scoring N2 and R sleep stages. We also know that N2, N3 and R sleep stages are the ones with the lower muscle tones. However, patients suffering from obstructive sleep apnea (OSA) syndrome are more likely to have apneas or hypopneas when the muscular tone is low. Respiratory and cardiac signals are thus more disrupted by respiratory events during those stages, making sleep-related variations less noticeable. Consequently, sleep staging complexity may be exacerbated for those stages.

The proposed system provides a system for automatic sleep scoring in Type III sleep diagnosis. Results were reasonably acceptable, but showed a 5-stage classification is out of range when using cardio-respiratory channels only. Several perspectives are being considered. First, we would like to evaluate this methodology on recordings from other sleep laboratories to make sure there is no local scoring practices bias. Secondly, a future work focused on a 3-stage classification (W/R/NREM) could be considered. Maybe it would help reaching a better R sleep stage identification as some papers from the literature. Last, the combination of both electrophysiological and cardio-respiratory based tools (presented in the previous and current papers, respectively) is currently ongoing. Even if results are not expected to greatly increase compared with the electrophysiological method, it could provide enhancement with low effort. As it currently stands, our Type III system is of great interest for the estimation of sleep parameters (total sleep time and enhanced apnea hypopnea index for example). The assessment of its impact on sleep diagnosis is being considered.

V. CONCLUSION

Sleep scoring provides information for precise sleep diagnosis. Despite being unfeasible in Type III sleep studies, in which only cardio-respiratory channels are recorded, many researchers have been working on it. Indeed, there are causal relationships between cardiac and respiratory activities and sleep stages.

In this study, we presented an automated sleep staging system for Type III sleep studies equipped with a tracheal sound sensor. After gathering a high number of features described in the literature, the system chooses the relevant features for sleep scoring, an then classifies sleep stages. It was tested on a large dataset of patients with and without sleep breathing disorders, and included sleep transition rules as described in the AASM guidelines.

Results showed the automatic hypnogram obtained reaches a moderate agreement with a manual scorer (mean Cohen's Kappa and accuracy rate of 0.48 and 62.4 %, respectively), due to a general overestimation of N2 sleep stage. When doing a 3-stage classification (W / R / NREM), a substantial agreement with the reference was obtained (mean Cohen's Kappa and accuracy rate of 0.60 and 78.5 %, respectively). Those results are promising when it comes to sleep-related disorders diagnosis.

The presented tool should help having more precise sleep diagnosis in HSAT with a tracheal sound sensor.

ACKNOWLEDGEMENTS

The authors would like to thank Christelle Gosselin and Jean-Louis Racineux, from the Institut de Recherche en Santé Respiratoire des Pays de La Loire, and Margaux Blanchard, from the ESEO. Thanks to Alain Le Duff and Lucile Riaboff, previously from the ESEO. We thank Julien Godey, Laetitia Moreno and Marion Vincent, sleep technicians in the Department of Respiratory and Sleep Medicine of Angers University Hospital.

APPENDIX
LIST OF THE SELECTED FEATURES

Table A.1 details the 112 selected features used for classification in this work.

TABLE A.1
LIST OF THE 112 SELECTED FEATURES.

Sensor	Signals	Features category	Features names
Pulse oximeter	Photoplethysmogram	Statistical features	Skewness Kurtosis Complexity Inter-quartile range 90th-10th Inter-quartile range 75th-25th Mean zero-crossing interval (ZCI) Normalized standard deviation of ZCI
		Spectral PRV features	Standard deviation of VLF Quadratic mean of TP Quadratic mean of VLF/TP Quadratic mean of LF/TP Quadratic mean of HF/TP Quadratic mean of VLF/HF Standard deviation of LF/HF
		Non-linear PRV features	Quadratic mean of ApEn for R symmetry Mean of SampEnt for R symmetry Standard deviation of SampEnt for R symmetry, centered Quadratic mean of FuzzySampEnt for T symmetry Quadratic mean of FuzzySampEnt for G symmetry Mean of LZC
	Saturation in oxygen	Statistical features	Variance Standard deviation 50th percentile
PneaVoX®	Snoring sounds energy	Statistical features	Mean ZCI Normalized standard deviation of ZCI
	Snoring intervals	Statistical features	Mean ZCI Standard deviation of ZCI Mean ZCI when regular
	Energy ratio	Statistical features	Variance Root mean square 25th percentile
	Inspiratory duration	Statistical features	Root mean square Inter-quartile range 75th-25th Normalized: mean Normalized: root mean square Normalized: 25th percentile
	Breathing cycle duration	Statistical features	Variance Root mean square 25th percentile Normalized: 50th percentile Normalized: 90th percentile
	Breathing duration ratio	Statistical features	10th percentile 50th percentile 75th percentile Inter-quartile range 75th-25th Normalized: variance
Nasal cannula	Nasal airflow	Statistical features	Mobility 10th percentile Normalized: 50th percentile Normalized: 75th percentile Normalized: inter-quartile range 90th-10th Normalized: mean ZCI Normalized: normalized standard deviation of ZCI
Actimeter	Actigraphy	Statistical features	Variance Mean ZCI Normalized standard deviation of ZCI

TABLE A.1
LIST OF THE 112 SELECTED FEATURES (CONTINUED).

Sensor	Signals	Features category	Features names
RIP belts	Thoracic belt	Statistical features	Mobility 25th percentile 90th percentile Normalized standard deviation of ZCI Absolute value: standard deviation Absolute value: means ratio Absolute value: 50th percentile Absolute value: mean ZCI Absolute value: normalized standard deviation of ZCI Normalized: standard deviation Normalized: mobility Normalized: 25th percentile Normalized: Mean ZCI Normalized + absolute value: root mean square Normalized + absolute value: means ratio Normalized + absolute value: complexity Normalized + absolute value: mean ZCI Normalized + absolute value: normalized standard deviation of ZCI
	Abdominal belt	Statistical features	Variance 75th percentile Mean ZCI Absolute value: mean Absolute value: means ratio Absolute value: mobility Absolute value: inter-quartile range 75th-25th Absolute value: mean ZCI Absolute value: normalized standard deviation of ZCI Normalized: standard deviation Normalized: 10th percentile Normalized: 75th percentile Normalized: normalized standard deviation of ZCI Normalized + absolute value: mean Normalized + absolute value: means ratio Normalized + absolute value: mobility Normalized + absolute value: mean ZCI Normalized + absolute value: normalized standard deviation of ZCI
	RIP belts ratio	Statistical features	Variance Root mean square Mobility 75th percentile Mean ZCI Normalized: root mean square Normalized: complexity Normalized: mobility Normalized: 25th percentile Normalized: inter-quartile range 75th-25t
	RIP belts phase shift	Statistical features	Root mean square Mobility 25th percentile 75th percentile 90th percentile Normalized: mean Normalized: standard deviation Normalized: complexity Normalized: mobility Normalized: 10th percentile

REFERENCES

[1] R. Heinzer, S. Vat, P. Marques-Vidal, H. Marti-Soler, D. Andries, N. Tobback, V. Mooser, M. Preisig, A. Malhotra, G. Waeber, P. Vollenweider, M. Tafti, and J. Haba-Rubio, "Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study," *The Lancet Respiratory Medicine*, vol. 3, no. 4, pp. 310–318, Apr. 2015.

[2] J. B. Croft, "CDC's Public Health Surveillance of Sleep Health," 2017.

[3] C. V. Senaratna, J. L. Perret, C. J. Lodge, A. J. Lowe, B. E. Campbell, M. C. Matheson, G. S. Hamilton, and S. C. Dharmage, "Prevalence of obstructive sleep apnea in the general population: A systematic review," *Sleep Medicine Reviews*, vol. 34, pp. 70–81, Aug. 2017.

[4] V. K. Kapur, D. H. Auckley, S. Chowdhuri, D. C. Kuhlmann, R. Mehra, K. Ramar, and C. G. Harrod, "Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline," *Journal of Clinical Sleep Medicine*, vol. 13, no. 03, pp. 479–504, Mar. 2017.

[5] T. Penzel, J. W. Kantelhardt, R. P. Bartsch, M. Riedl, J. F. Kraemer, N. Wessel, C. Garcia, M. Glos, I. Fietze, and C. Schöbel, "Modulations of Heart Rate, ECG, and Cardio-Respiratory Coupling Observed in Polysomnography," *Frontiers in Physiology*, vol. 7, Oct. 2016.

[6] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/wake detection based on cardiorespiratory signals and actigraphy," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 5089–5092.

[7] W. Hayet and Y. Slim, "Sleep-wake stages classification based on heart rate variability," in *2012 5th International Conference on BioMedical Engineering and Informatics*. Chongqing, China: IEEE, Oct. 2012, pp. 996–999.

[8] A. Domingues, T. Paiva, and J. M. Sanches, "Hypnogram and Sleep Parameter Computation From Activity and Cardiovascular Data," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1711–1719, Jun. 2014.

[9] T. Willemen, C. Varon, A. Caicedo Dorado, B. Haex, J. Vander Sloten, and S. Van Huffel, "Probabilistic cardiac and respiratory based classification of sleep and apneic events in subjects with sleep apnea," *Physiological Measurement*, vol. 36, no. 10, pp. 2103–2118, Oct. 2015.

[10] H. Yoon, S. H. Hwang, J.-W. Choi, Y. J. Lee, D.-U. Jeong, and K. S. Park, "REM sleep estimation based on autonomic dynamics using R-R intervals," *Physiological Measurement*, vol. 38, no. 4, pp. 631–651, Apr. 2017.

[11] M. Xiao, H. Yan, J. Song, Y. Yang, and X. Yang, "Sleep stages classification based on heart rate variability and random forest," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 624–633, Nov. 2013.

[12] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *Physiological Measurement*, vol. 36, no. 10, pp. 2027–2040, Oct. 2015.

[13] X. Long, J. Yang, T. Weysen, R. Haakma, J. Foussier, P. Fonseca, and R. M. Aarts, "Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging," *Physiological Measurement*, vol. 36, no. 3, pp. 625–625, Mar. 2015.

[14] M. Aktaruzzaman, M. W. Rivolta, R. Karmacharya, N. Scarabottolo, L. Pugnetti, M. Garegnani, G. Bovi, G. Scalera, M. Ferrarin, and R. Sassi, "Performance comparison between wrist and chest actigraphy in combination with heart rate variability for sleep classification," *Computers in Biology and Medicine*, vol. 89, pp. 212–221, Oct. 2017.

[15] M. Gaiduk, T. Penzel, J. A. Ortega, and R. Seepold, "Automatic sleep stages classification using respiratory, heart rate and movement signals," *Physiological Measurement*, Dec. 2018.

[16] M. Glos, A. Sabil, K. S. Jelavic, C. Schöbel, I. Fietze, and T. Penzel, "Characterization of Respiratory Events in Obstructive Sleep Apnea Using Suprasternal Pressure Monitoring," *Journal of Clinical Sleep Medicine*, vol. 14, no. 03, pp. 359–369, Mar. 2018.

[17] T. Penzel and A. Sabil, "The use of tracheal sounds for the diagnosis of sleep apnoea," *Breathe*, vol. 13, no. 2, pp. e37–e45, Jun. 2017.

[18] ——, "Physics and Applications for Tracheal Sound Recordings in Sleep Disorders," *Breath Sounds*, pp. 83–104, 2018.

[19] S. Huq and Z. Moussavi, "Automatic breath phase detection using only tracheal sounds," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. Buenos Aires: IEEE, Aug. 2010, pp. 272–275.

[20] X. Lu, D. Guiraud, S. Renaux, T. Similowski, and C. Azevedo, "Breathing detection from tracheal sounds in both temporal and frequency domains in the context of phrenic nerve stimulation," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Berlin, Germany: IEEE, Jul. 2019, pp. 5473–5476.

[21] N. M. Ghahjaverestan, S. Saha, B. Gavrilovic, and A. Yadollahi, "Removing of Snoring Segments from Tracheal Breathing Sounds using a Wavelet-based Algorithm," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Montreal, QC, Canada: IEEE, Jul. 2020, pp. 764–767.

[22] N. Montazeri Ghahjaverestan, M. Kabir, S. Saha, K. Zhu, B. Gavrilovic, H. Alshaer, B. Taati, and A. Yadollahi, "Automatic Respiratory Phase Identification Using Tracheal Sounds and Movements During Sleep," *Annals of Biomedical Engineering*, Jan. 2021.

[23] N. Freycenon, R. Longo, and L. Simon, "Estimation of heart rate from tracheal sounds recorded for the sleep apnea syndrome diagnosis," *IEEE Transactions on Biomedical Engineering*.

[24] J. Vanbuis, M. Feuilloy, G. Baffet, N. Meslier, F. Gagnadoux, and J.-M. Girault, "Towards a user-friendly sleep staging system for polysomnography part I: Automatic classification based on medical knowledge," *Informatics in Medicine Unlocked*, vol. 21, p. 100454, 2020.

[25] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, R. M. lloyd, S. F. Quan, M. M. Troester, and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, ser. American Academy of Sleep Medicine. Darien IL, 2017, no. 2.4.

[26] "Heart rate variability," *Eur Heart J*, vol. 17, p. 28, 1996.

[27] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?" *International Journal of Cardiology*, vol. 166, no. 1, pp. 15–29, Jun. 2013.

[28] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1186–1195, Dec. 2012.

[29] S. Charbonnier, L. Zoubek, S. Lesecq, and F. Chapotot, "Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging," *Computers in Biology and Medicine*, vol. 41, no. 6, pp. 380–389, Jun. 2011.

[30] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, and K. Jerbi, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," *Journal of Neuroscience Methods*, vol. 250, pp. 94–105, Jul. 2015.

[31] M. Zokaeinikoo, "Automatic Sleep Stages Classification," Ph.D. dissertation, 2016.

[32] F. Ebrahimi, S.-K. Setarehdan, J. Ayala-Moyeda, and H. Nazeran, "Automatic sleep staging using empirical mode decomposition, discrete wavelet transform, time-domain, and nonlinear dynamics features of heart rate variability signals," *Computer Methods and Programs in Biomedicine*, vol. 112, no. 1, pp. 47–57, Oct. 2013.

[33] P. K. Stein and Y. Pu, "Heart rate variability, sleep and sleep disorders," *Sleep Medicine Reviews*, vol. 16, no. 1, pp. 47–66, Feb. 2012.

[34] X. Long, P. Fonseca, R. Haakma, R. M. Aarts, and J. Foussier, "Spectral Boundary Adaptation on Heart Rate Variability for Sleep and Wake Classification," *International Journal on Artificial Intelligence Tools*, vol. 23, no. 03, p. 1460002, Jun. 2014.

[35] A. Delgado-Bonal and A. Marshak, "Approximate Entropy and Sample Entropy: A Comprehensive Tutorial," *Entropy*, vol. 21, no. 6, p. 541, May 2019.

[36] J.-M. Girault and A. Humeau-Heurtier, "Centered and Averaged Fuzzy Entropy to Improve Fuzzy Entropy Precision," *Entropy*, vol. 20, no. 4, p. 287, Apr. 2018.

[37] A. Zaylaa, S. Oudjemia, J. Charara, and J.-M. Girault, "n-Order and maximum fuzzy similarity entropy for discrimination of signals of different complexity: Application to fetal heart rate signals," *Computers in Biology and Medicine*, vol. 64, pp. 323–333, Sep. 2015.

[38] D. E. Rumelhart, G. E. Hintont, and R. J. Williams, "Learning representations by back-propagating errors," p. 4, 1986.

[39] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[40] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring," *Journal of Clinical Sleep Medicine*, no. 9, Jan. 2013.

Pour aller plus loin : score de confiance de l'algorithme

La table de probabilité $probabilities_{CR}$ a été utilisée pour fournir également un score de confiance aux spécialistes du sommeil.
Calculé à partir de la table de probabilité ($probabilities_{CR}$) et de l'hypnogramme estimé automatiquement ($hypno_{CR}$), ce score peut être considéré comme une estimation de la facilité de l'algorithme à classifier le sommeil.

Ce score a été évalué dans le cas d'une classification en trois stades : E/SP/SL (Sommeil Lent ou regroupement du N1, N2 et N3). Pour chaque époque, le stade pour lequel $probabilities_{CR}$ est maximum est identifié. Les époques classifiées (par $hypno_{CR}$) dans le même stade que celui précédemment identifié sont comptabilisées. Le score de confiance correspond à la proportion d'époques dont la classification automatique est identique au stade privilégié par $probabilities_{CR}$. La Figure E1 illustre le dénombrement de ces époques.

époque	$probabilities_{CR}$ E	SP	SL	$hypno_{CR}$	max($probabilities_{CR}$) = $hypno_{CR}$
#1	90%	10%	0%	E	Oui
#2	35%	20%	45%	E	Non
#3	70%	20%	1%	E	Oui
#4	40%	50%	10%	E	Non
...	...	...	...	...	...
#960	70%	30%	0%	E	Oui
				Total	Oui : 60%, Non : 40%

Score de confiance : 60%

Figure E1 – Exemple illustrant l'estimation de score de confiance, en fonction de $probabilities_{CR}$ et $hypno_{CR}$.

Les époques classifiées dans un autre stade sont en fait des époques corrigées par C2 ou C3. Une importante proportion d'époques corrigées par C2 ou C3 indique que l'algorithme a eu des difficultés lors de la classification. Le tableau E1 présente les scores de confiance moyens obtenus pour différentes tranches de taux d'accord.

Table E1 – Score de confiance moyen (et écart-type) obtenu en fonction des valeurs du taux d'accord.

Taux d'accord	Score de confiance (moyenne ± écart-type)
$Acc < 60\,\%$	73.5 ± 10.0
$60\,\% \leq Acc < 70\,\%$	79.0 ± 7.3
$70\,\% \leq Acc < 80\,\%$	85.0 ± 4.0
$Acc \geq 80\,\%$	88.0 ± 3.2

Les enregistrements obtenant un taux d'accord inférieur à 70 % ont été identifiés comme étant ceux pour lesquels l'utilisation de la classification automatique est déconseillée. Sur les 100 enregistrements du groupe de test (D2), ils sont au nombre de 13.
En fixant un seuil d'alerte à 80 % sur le score de confiance, on obtient la matrice de confusion présentée Tableau E2. Les résultats de l'identification des enregistrements « à risque » ($Acc = 83\,\%$, $Se = 54\,\%$ et VPP $= 64\,\%$) montrent que le score de confiance permet, dans une certaine mesure, de jauger la validité de l'analyse automatique.

Table E2 – Matrice de confusion pour l'identification des enregistrements « à risque ».

	$Conf < 80\,\%$	$Conf \geq 80\,\%$
$Acc < 70\,\%$	VP = 7	FN = 6
$Acc \geq 70\,\%$	FP = 4	VN = 83

Le score de confiance est donc un outil supplémentaire permettant au médecin de jauger la complexité de la classification. Ce score peut ainsi permettre d'alerter le médecin lors d'une classification automatique jugée peu sûre.

Informations supplémentaires : les signaux du PneaVoX®

Le capteur PneaVoX® est comparable à un stéthoscope. Composé d'un capteur acoustique et d'un capteur de pression, il est positionné au-dessus de la fourchette sternale.
En sortie du capteur, on retrouve un signal sonore brut donnant des informations à la fois sur le débit et les ronflements, ainsi qu'un signal de pression permettant d'étudier la résistance des voies aériennes supérieures et les efforts respiratoires.

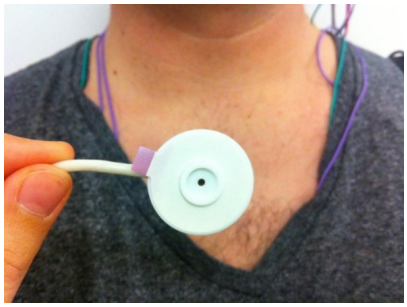


Figure E2 – Le capteur PneaVoX®.

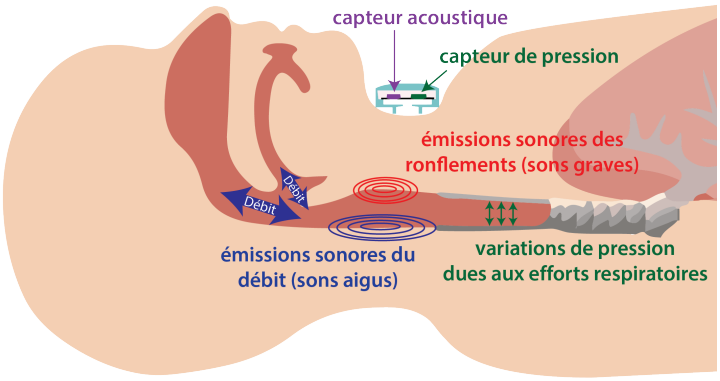


Figure E3 – Informations récupérées par le capteur PneaVoX®.

Les émissions sonores liées au débit ventilatoire (en bleu sur la Figure E3) et aux ronflements (en rouge sur la Figure E3) correspondent aux fréquences aiguës (200-2000 Hz) et graves (20-200 Hz) du signal sonore brut, respectivement.
La pression sus-sternale (en vert sur la Figure E3) est quant à elle évaluée à partir des faibles fréquences (0,2-20 Hz) du signal de pression.

À partir des émissions sonores liées au débit, sont extraits l'intensité et l'énergie des sons respiratoires. En découlent le ratio des énergies liées à l'inspiration et à l'expiration, la durée de l'inspiration, la durée du cycle respiratoire, et le ratio entre ces durées. Dans l'article précédent, ces signaux sont nommés *breathing sounds intensity*, *breathing sounds energy*, *energy ratio*, *inspiratory duration*, *breathing cycle duration* et *breathing durations ratio*.
Les émissions sonores liées aux ronflements permettent l'extraction de leur intensité et de leur énergie, dont il est fait référence sous les noms *snoring sounds intensity* et *snoring sounds energy* dans l'article précédent. L'intervalle entre les ronflements (*snoring interval*) et la régularité des ronflements *snoring regularity*) sont estimés à partir de ces signaux.
Pour terminer, le signal *suprasternal pressure* est estimé à partir du signal de pression.
Ces signaux sont retranscrits visuellement, afin que le spécialiste puisse les exploiter au mieux.

Valorisation supplémentaire

Ce travail a également été valorisé dans le congrès francophone Le Congrès du Sommeil (en virtuel, novembre 2020), par le biais d'un poster.

À RETENIR

Un outil de classification automatique des stades de sommeil à partir des voies disponibles en PV a été mis en place. Cet outil classifie le sommeil à partir d'un ensemble de caractéristiques sélectionnées spécifiquement pour cette tâche. Ces caractéristiques sont extraites de l'oxymètre de pouls, du PneaVoX®, de la lunette nasale, de l'actimètre et des sangles inductives. Les époques à proximité et les règles de transition sont également prises en compte. Cet outil est simple d'utilisation et a été testé sur un nombre important de patients avec et sans pathologies du sommeil. L'hypnogramme obtenu est satisfaisant comparé à la littérature, mais insuffisant comparé à la classification manuelle en PSG. Son utilisation est simple et fournit, en plus de l'hypnogramme, un score de confiance permettant d'alerter le médecin dans le cas où la classification automatique est jugée peu sûre.

E.3.2 Mesure de l'impact clinique

Pour aller plus loin, l'impact clinique de cette classification sur différents enregistrements a été évalué. L'objectif est donc d'estimer l'apport de la classification automatique en PV sur le diagnostic du patient.

Méthode

DONNÉES

Nous avons utilisé les mêmes 1291 enregistrements polysomnographiques que pour l'évaluation de l'impact clinique de l'analyse automatique en PSG. Pour rappel, ces enregistrements font partie de la cohorte du sommeil des Pays de La Loire et proviennent de patients investigués pour suspicion de SAHS. Les enregistrements ont été réalisés et lus durant les années 2012-2018 au laboratoire du sommeil du CHU d'Angers, en utilisant un appareil polysomnographique de type CID102L8D, et en respectant les recommandations de l'AASM. Tous les patients ont donné leur consentement écrit et informé.
L'étude portant sur la PV, nous avons décidé d'évaluer la lecture du sommeil en 3 stades au lieu de 5 (voir article précédent). Nous ne considérerons donc que l'éveil, le Sommeil Lent ou regroupement du N1, N2 et N3 (SL) et le sommeil paradoxal SP. C'est en effet la distinction de ces trois stades qui permettra d'obtenir les éléments principaux pour le diagnostic des pathologies les plus fréquentes. Le tableau E3 présente les caractéristiques des enregistrements utilisés.

Table E3 – Description du jeu de données utilisé pour la mesure de l'impact clinique de l'analyse automatique en PV.

Variable	Valeur
Sexe féminin	$\approx 37\,\%$
Âge (moyenne $\pm$ écart-type)	52 ± 14 ans
IAH* (moyenne $\pm$ écart-type)	23 ± 21 par heure
Nombre d'époques total	1414095
Nombre d'époques d'éveil*	314292 ($\approx 22\,\%$)
Nombre d'époques de SL*	876412 ($\approx 62\,\%$)
Nombre d'époques de SP*	223391 ($\approx 16\,\%$)

* D'après la lecture manuelle qui constitue notre référence

APPROCHE

La démarche d'évaluation de l'impact clinique en PV est la même que celle mise en place dans le cas de l'analyse en PSG (voir section D.3.3). Nous évaluerons donc encore une fois les latences et proportions liées aux stades identifiés, l'IAH et la sévérité associée au SAHS.

Pour cela, nous considérerons également deux hypnogrammes. Ces hypnogrammes seront cependant simplifiés en regroupant le N1, N2 et N3 de façon à n'avoir que 3 stades au lieu de 5. Il est important de noter qu'à l'heure actuelle, il n'y a pas eu de nouvel entraînement, spécifique à cette classification en 3 stades.
Dans la suite de cette section, nous annoterons $hypno_{ref}$ l'hypnogramme de référence simplifié en 3 stades, et $hypno_{CR}$ l'hypnogramme obtenu à partir de l'analyse automatique en PV, simplifié en 3 stades.
Nous n'étudierons ici que l'impact de l'hypnogramme automatique et ne prendrons pas en considération la table de probabilité.

1) Proportions et latences

Le Temps de Sommeil Total (TST) obtenu avec l'analyse automatique à partir des signaux cardio-respiratoires en PV (TST_{CR}) a été comparé avec le TST découlant de l'analyse manuelle en PSG (TST de référence TST_{ref}). Pour mieux interpréter les résultats obtenus, nous avons également comparé le Temps d'Enregistrement (TE) avec TST_{ref}. En effet, c'est le TE qui, en PV, est utilisé en remplacement du TST dans le calcul de l'IAH. Toutes ces durées, exprimées en minutes, sont comparées par le biais de l'erreur relative ($\frac{\text{valeur estimée - valeur de référence}}{\text{valeur de référence}}$).
De la même manière, les proportions de SP et de SL (*propSP* et *propSL*) obtenues avec l'analyse

automatique ont été comparées avec celles résultant de la lecture manuelle. Ces proportions sont usuellement exprimées par rapport au TST, et non à la durée totale d'enregistrement. La proportion d'éveil n'a pas été estimée, puisqu'elle est directement reliée au TST. Ici encore, le graphe de corrélation et de Bland-Altman ont été privilégiés.
La latence d'endormissement *SOL* et la latence de SP *latSP*, ont également été évaluées et comparées avec la référence (étant donné que l'endormissement se fait nécessairement par le stade N1, et donc le SL, la latence du SL est équivalente à la *SOL* et n'a donc pas été estimée). Tout comme pour l'étude de l'impact clinique de l'analyse en PSG, l'endormissement a été défini comme étant la première époque de sommeil, à condition que celle-ci soit suivie d'au moins deux autres époques de sommeil. La latence *latSP* correspond quant-à-elle au délai, en minutes, entre l'endormissement (comme précédemment défini) et la première époque de SP. Ici, on étudiera les valeurs de la différence entre les latences obtenues grâce aux lectures automatiques et manuelles.

2) IAH

L'IAH découlant de l'analyse automatique, IAH_{CR}, est comparé à l'IAH de référence, IAH_{ref}, mais aussi à l'IAH qui aurait été obtenu en PV sans lecture du sommeil, IAH_{PV}. Nous voulons ainsi vérifier que l'IAH_{CR} est plus précis que l'IAH_{PV}, et se rapproche de l'IAH_{ref}. Il est nécessaire de rappeler qu'en dehors de l'hypnogramme, les événements pris en compte lors de l'analyse manuelle ou automatique peuvent différer. Contrairement à la Section D.3.3, nous ne considérerons pas les micro-éveils ici, puisqu'il n'est pas possible de les identifier manuellement dans le cas d'une PV. L'IAH_{CR} est donc évalué sans prendre en compte les micro-éveils. Le tableau E4 récapitule les éléments considérés lors de l'estimation des différents IAHs.

Table E4 – Estimation des différents IAHs comparés dans cette partie. *A* définit les apnées, et *H* définit les hypopnées.

	Examen de PV		Examen de PSG
	pas d'hypnogramme sans micro-éveils	$hypno_{CR}$ sans micro-éveils	$hypno_{ref}$ avec micro-éveils
Événements respiratoires	*A(toutes)* H_{desat}*(toutes)*	*A en sommeil*$_{CR}$ H_{desat} *en sommeil*$_{CR}$ H_{Ev} *en sommeil*$_{CR}$	*A en sommeil*$_{ref}$ H_{desat} *en sommeil*$_{ref}$ H_{Ev} *en sommeil*$_{ref}$ H_{MEV} *en sommeil*$_{ref}$
Temps de référence	*TE*	TST_{CR}	TST_{ref}
IAH	$\frac{A+H_{desat}}{TE}$	$\frac{A+H_{desat}+H_{Ev}}{TST_{CR}}$	$\frac{A+H_{desat}+H_{Ev}+H_{MEV}}{TST_{ref}}$

On comparera les différents IAHs en utilisant une fois encore les graphes de corrélation et de Bland-Altman.

3) Sévérité du SAHS

La sévérité du SAHS est estimée selon la valeur de l'IAH obtenu. On peut ainsi comparer les sévérités dans le cas de la PV sans ou avec lecture automatique du sommeil, et dans le cas de la lecture manuelle en PSG.
Pour rappel, on définit un SAHS absent lorsque l'IAH est inférieur à 5/h, un SAHS léger lorsque l'IAH est compris entre 5/h et 15/h, modéré lorsqu'il est compris entre 15/h et 30/h, et sévère lorsqu'il est supérieur à 30/h. On mesurera l'exactitude de l'estimation de la sévérité du SAHS à l'aide de matrices de confusion.

Résultats

Les résultats sont organisés en considérant les différents éléments présentés précédemment. L'ordre de ces éléments allant du moins au plus proche du diagnostic du patient, nous pouvons considérer les résultats comme étant présentés des moins au plus importants.

1) Proportions et latences

La Figure E4 présente les graphes de corrélation et de Bland-Altman liés à l'estimation du TST. Les mêmes graphiques pour le TE ont été insérés. Ils permettent de mieux se rendre

compte des durées prises en compte pour l'évaluation du SAHS lorsque le sommeil n'est pas estimé.

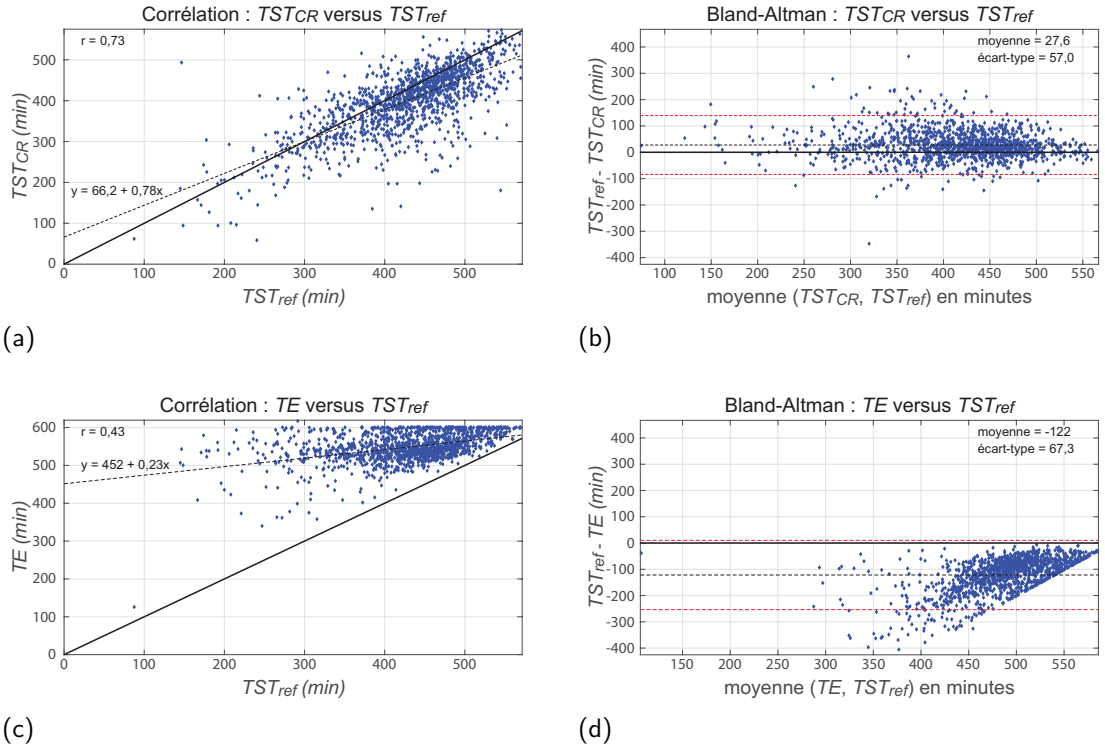


Figure E4 – a) et b) graphes de corrélation et de Bland-Altman du TST estimé grâce à l'analyse automatique du sommeil en PV, versus le TST obtenu avec lecture manuelle du sommeil, et c) et d) graphes de corrélation et de Bland-Altman du TE, versus le TST obtenu avec lecture manuelle du sommeil.

Sans surprises, l'estimation du TST obtient une corrélation satisfaisante lorsqu'elle est comparée à la corrélation obtenue par le TE. Cependant, le graphique de Bland Altman confirme que le TST_{CR} reste moyennement précis, puisqu'il est sous-diagnostiqué d'une demi-heure approximativement en moyenne, avec un écart-type avoisinant 1 heure. Sur des valeurs de TST avoisinant les 400 minutes, cette différence n'est pas négligeable.

La Figure E5 présente la distribution des proportions SL et de SP.
Les erreurs relatives sont estimées en retirant les patients dont la proportion de référence est inférieure à 5 % (valeurs peu représentatives et divisions par zéro). Pour *propSL*, l'erreur relative (moyenne ± écart-type) est de 16,7 % ± 15,1 %. *propSP* obtient quant à elle une erreur relative de 6,0 % ± 75,4 %.

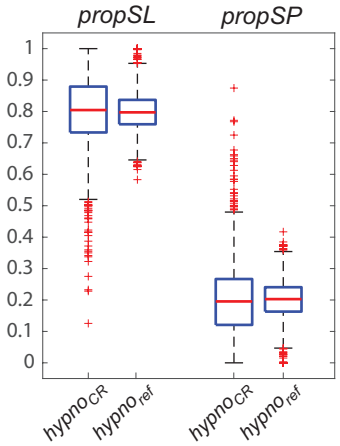


Figure E5 – Diagrammes en boîte à moustaches des proportions associées au SL et au SP, dans le cas des analyses automatiques et manuelles.

La Figure E6 est un histogramme empilé reportant le nombre d'enregistrements, en ordonné, selon la durée de la latence de référence, en abscisse, en fonction de plusieurs catégories. Ces catégories permettent de différencier les erreurs d'estimation de la latence inférieures à 2 minutes, comprises entre 2 et 5 minutes, entre 5 et 15 minutes, entre 15 et 30 minutes, et supérieures à 30 minutes.

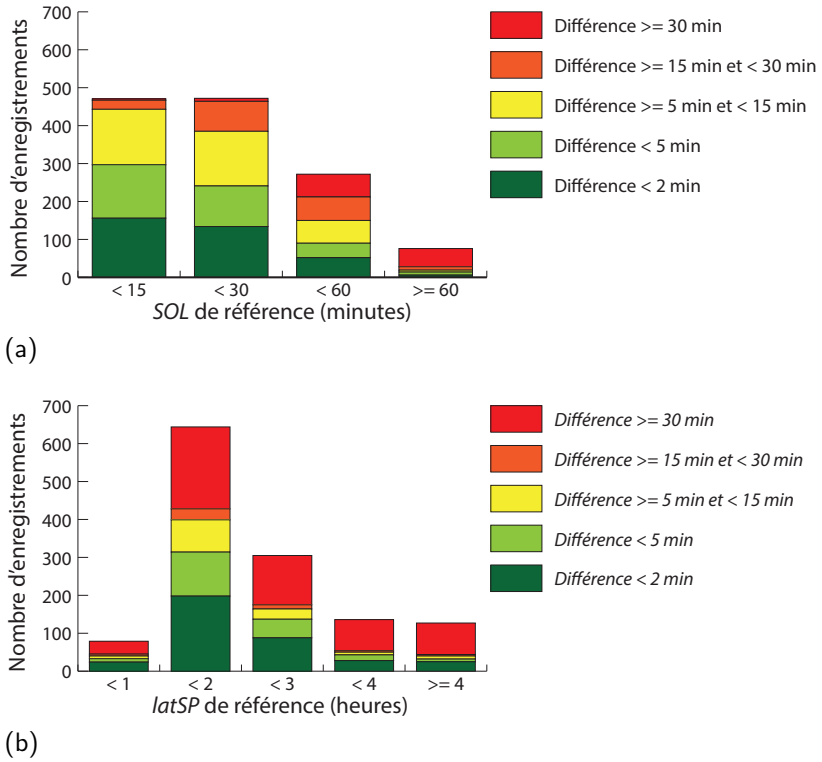


Figure E6 – Distribution des enregistrements selon l'erreur d'estimation de latence et la valeur de la latence, dans le cas : a) de la latence d'endormissement *SOL* et b) de la latence de SP *latSP*.

Pour le *SOL*, le nombre d'enregistrements pour lesquels les estimations de latences sont correctes à moins de 5 minutes près sont relativement nombreux. Cependant, plus la latence d'endormissement de référence est grande, plus les chances que l'erreur d'estimation soit importante augmente. Pour *latSP*, qui est généralement bien plus grande que le *SOL* puisque le SP apparaît plus tard dans la nuit (elle est d'ailleurs exprimée en heures), l'estimation est moins bonne. On observe, quelle que soit la valeur de *latSP*, un grand nombre d'enregistrements pour lesquels la différence d'estimation dépasse les 30 minutes. De plus, la proportion d'enregistrements pour lesquels *latSP* est estimée correctement à moins de 5 minutes près diminue avec l'augmentation de la valeur de la latence de SP de référence.

2) IAH

Les graphes de corrélation et de Bland-Altman permettant l'étude des IAHs résultants sont présentés Figure E7.

L'analyse automatique du sommeil est prometteuse quant-à l'évaluation de l'IAH, qui se rapproche de celui obtenu suite à l'analyse du sommeil manuelle.

L'IAH_{CR} est ainsi sous-estimé de $4{,}31/\text{h} \pm 4{,}59/\text{h}$ par rapport à l'IAH_{ref}, contre $6{,}51/\text{h} \pm 6{,}30/\text{h}$ sans utiliser d'analyse automatique du sommeil. On observe également que très peu d'IAHs sont surestimés par rapport à l'IAH de référence. Dans le cas de la PV sans analyse du sommeil, certains IAHs étaient également surestimés. C'est pour cette raison qu'il est nécessaire d'évaluer l'impact des IAHs estimés sur la sévérité du SAHS.

3) Sévérité du SAHS

Une première matrice de confusion, associée à l'estimation du SAHS sans analyse automatique versus avec lecture manuelle, est rappelée Table E5 (cette table et les résultats associés ont déjà présentés dans la Section D.3.3).

Sans analyse du sommeil ni des micro-éveils, le Kappa de Cohen et le taux d'accord sont de 0,51 et 64 %, respectivement. Cela signifie un accord modéré avec la référence.

On observe également qu'environ 36 % (N=131+10+0+199+8+122=470) des enregistrements sont sous-diagnostiqués. Plus particulièrement, approximativement 26 % (N=10+199+8+122=339) des enregistrements pourraient engendrer, en l'état, le non traitement de patients qui l'auraient

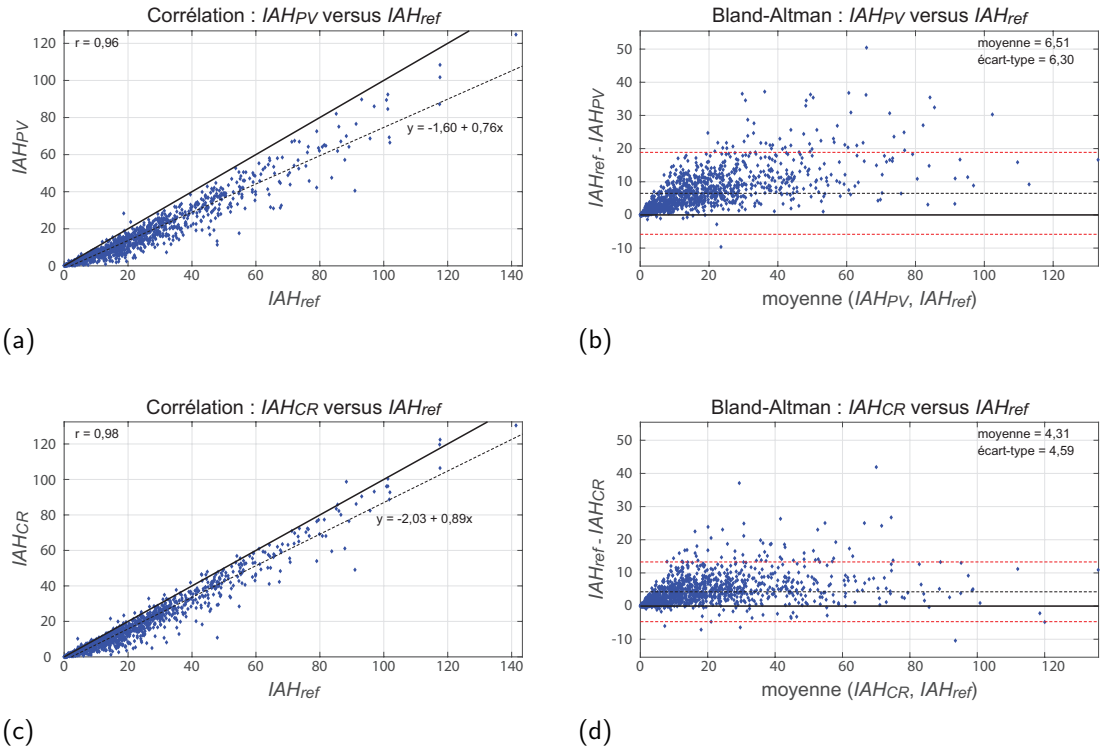


Figure E7 – a) et b) graphes de corrélation et de Bland-Altman des IAHs obtenus sans analyse du sommeil et des micro-éveils (donc équivalent aux IAHs en PV), versus les IAHs obtenus avec analyse manuelle du sommeil et des micro-éveils en PSG et c) et d) graphes de corrélation et de Bland-Altman des IAHs obtenus avec analyse automatique du sommeil en PV et sans lecture des micro-éveils, versus les IAHs obtenus avec analyse manuelle du sommeil et des micro-éveils.

Table E5 – Matrice de confusion associée à la sévérité du SAHS sans lecture du sommeil. Les valeurs en rouge indiquent une sous-estimation qui peut potentiellement impacter le traitement du patient.

		Sans lecture du sommeil ni des micro-éveils				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
Référence	Pas de SAHS	303	1	0	0	304
	SAHS léger	131	198	0	0	329
	SAHS modéré	10	199	149	0	358
	SAHS sévère	0	8	122	170	300
	Total	444	406	271	170	1291

été suite à la lecture manuelle en PSG. Seul 1 enregistrement est au contraire sur-diagnostiqué.

Est ensuite présentée la matrice de confusion associée à l'estimation du SAHS avec analyse automatique en PV versus avec lecture manuelle en PSG (table E6).
On obtient alors un Kappa de Cohen et un taux d'accord de 0,63 et 73 %, respectivement. Cela signifie un accord fort avec la référence.
On observe également qu'environ 27 % (N=101+6+0+151+4+83=349) des enregistrements sont sous-diagnostiqués. Plus particulièrement, approximativement 19 % (N=6+151+4+83=248) des enregistrements pourraient engendrer, en l'état, le non traitement de patients qui l'auraient été suite à la lecture manuelle. Cette fois-ci, 5 (3+1+1) enregistrements sont au contraire sur-diagnostiqués. Parmi ces enregistrements, 2 (1+1, en bleu) pourraient éventuellement engendrer, sans relecture, le traitement du patient alors qu'il n'aurait pas eu lieu avec la lecture manuelle en PSG.

Table E6 – Matrice de confusion associée à la sévérité du SAHS lors de la lecture automatique du sommeil en PV. Les valeurs en rouge indiquent une sous-estimation qui peut potentiellement impacter le traitement du patient, et les valeurs en bleu indiquent une sur-estimation qui peut potentiellement impacter le traitement du patient.

		Sommeil automatique				
		Pas de SAHS	SAHS léger	SAHS modéré	SAHS sévère	Total
Référence	Pas de SAHS	301	3	0	0	304
	SAHS léger	101	227	1	0	329
	SAHS modéré	6	151	200	1	358
	SAHS sévère	0	4	83	209	300
	Total	408	385	288	210	1291

Discussion

Nous avons étudié l'impact que pourrait avoir l'utilisation de l'analyse automatique à partir des signaux cardio-respiratoires en PV. Pour cela, nous avons comparé les résultats obtenus avec ceux estimés par lecture du sommeil manuelle en PSG et, lorsque c'est possible, avec ceux découlant d'une PV classique, sans analyse du sommeil.

Tout d'abord, le TST, les proportions de SL et de SP et les latences d'endormissement et de SP ont été évalués.
Sans surprise, le TST obtenu est bien plus précis que le TE utilisé normalement en PV. Il permet de passer d'une sur-estimation du TE moyenne d'environ 2 heures, à une sous-estimation du TST moyenne d'environ 30 min (voir Figure E4).
Concernant les proportions de SL, les résultats sont une fois encore mitigés avec une différence avec la référence moyenne par rapport à leur valeur de référence (erreur relative moyenne de 16,7 % tout de même). Pour le SP, les résultats sont plus convaincants avec une erreur relative moyenne de 6,0 % (voir Tableau E5).
Les latences d'endormissement et de SP ont aussi été comparées avec les latences obtenues par lecture manuelle du sommeil (Figure E6). Encore une fois, les résultats sont moins bons que pour l'analyse automatique en PSG. La proportion d'enregistrements pour lesquels les latences sont estimées avec une erreur de plus de 30 minutes est plus importante pour le SP, qui apparaît généralement plus tard dans la nuit. Le SOL est mieux estimé, et particulièrement pour les enregistrements pour lesquels le patient s'est endormi rapidement.

En ce qui concerne l'estimation de l'IAH (Figure E7), les résultats montrent très clairement que l'analyse automatique en PV, malgré qu'elle soit moins bonne que l'analyse automatique en PSG, et que la précision de l'estimation des indices précédents soit réduite, constitue un avantage certain pour le diagnostic. Ainsi, la valeur de la pente de la courbe de corrélation passe de 0,76 sans analyse du sommeil, à 0,89 avec analyse automatique du sommeil en PV (pour rappel, nous obtenions 0,91 en analyse automatique du sommeil en PSG sans prendre en compte les micro-éveils). Sur les graphes de Bland-Altman, on observe également que la moyenne et l'écart-type de l'erreur d'estimation d'IAH diminue avec l'analyse automatique du sommeil.
Il est très intéressant de constater que, malgré des erreurs non négligeables quant aux valeurs des indices précédemment présentés, l'estimation de l'IAH est tout à fait correcte. Cela s'explique par le fait que ce dernier ne prend en compte que l'éveil versus le sommeil, et qu'il est également très dépendant des événements ventilatoires (qui sont sensiblement les mêmes que l'on soit en analyse automatique du sommeil en PV, ou en analyse automatique du sommeil en PSG sans prise en compte des micro-éveils).
Cela est d'autant plus visible avec l'estimation de la sévérité du SAHS, puisqu'il n'est alors plus fait de différence entre les IAHs d'une même catégorie.
Dans le cas de la PV sans estimation du sommeil, on obtenait un accord modéré avec la référence, selon l'interprétation du Kappa de Cohen. Un grand nombre de patients étaient également sous-diagnostiqués, et plus d'un quart des patients n'auraient pas pu être traités sans examen de seconde intention.
Avec l'ajout d'une analyse du sommeil en PV, l'accord avec la référence est fort (toujours selon l'interprétation du Kappa de Cohen). Le nombre de patients sous-diagnostiqués diminue grande-

ment, et plus qu'un cinquième des patients n'auraient pas pu être traités sans examen de seconde intention (19 % pour l'analyse automatique en PV, contre 17 % pour l'analyse automatique en PSG sans prise en compte des micro-éveils). À l'inverse, un risque de traitement proposé à des patients qui pourraient ne pas en avoir besoin apparaît pour 2 patients seulement.
La sévérité du SAHS nous permet ainsi d'avoir une idée très précise de l'impact concret que pourrait avoir l'analyse automatique du sommeil sur le patient.

Il est important de noter que, contrairement à la PSG, il n'est pas possible de lire les micro-éveils en PV. Pour la PSG, nous avions conclu qu'une amélioration de l'analyse du sommeil n'aurait pas ou très peu d'impact sur le diagnostic, et que seule une détection des micro-éveils pourrait améliorer les estimations de l'IAH et de la sévérité du SAHS. Dans le cas de la PV, nous obtenons des résultats (en terme d'IAH et de sévérité du SAHS) qui sont très proches de ceux obtenus par analyse automatique en PSG sans prise en compte des micro-éveils. Nous pouvons donc considérer que, dans le cas de la PV, l'impact sur le diagnostic ci-avant présenté ne pourra pas être amélioré significativement.

Conclusion

L'analyse du sommeil en PV présentée précédemment a été appliquée sur un jeu de données composé de 1291 nouveaux enregistrements polysomnographiques. Ces enregistrements proviennent de patients qui ont consulté pour suspicion de troubles du sommeil lors des années 2012 à 2018.
L'estimation des performances cliniques a été réalisée en comparant l'enregistrement constitué de l'hypnogramme résultant de l'analyse automatique du sommeil, avec l'enregistrement tel que lu manuellement par le spécialiste du sommeil. L'impact clinique de cette analyse sur le diagnostic a été évalué sur plusieurs niveaux : les indicateurs de proportions et latences, l'IAH et la sévérité du SAHS.
Il en ressort que l'analyse automatique obtient des résultats cliniques décevants quand il s'agit des indicateurs de proportions et latences, mais satisfaisants lorsqu'il s'agit de l'IAH et de la sévérité du SAHS. Pour ces derniers, les résultats sont même comparables avec ceux précédemment obtenus lors de l'analyse automatique du sommeil en PSG (sans prise en compte des micro-éveils). On en conclura que l'impact sur l'estimation de l'IAH et du SAHS ne pourront pas être significativement améliorés par des modifications sur l'analyse automatique en PV actuellement mise en place.

Valorisation supplémentaire

Ce travail a également été valorisé dans le congrès francophone Le Congrès du Sommeil (en virtuel, novembre 2020), par le biais d'un poster.

À RETENIR

L'analyse automatique des stades de sommeil à partir des voies cardio-respiratoires a montré qu'elle était suffisamment précise pour permettre une estimation de l'IAH et de la sévérité du SAHS.

E.4 Conclusion du chapitre

Dans ce chapitre, une analyse automatique des stades de sommeil en PV a été présentée. Cette analyse utilise un ensemble de voies cardio-respiratoires disponibles en PV et sélectionnées automatiquement après avoir été jugées intéressantes pour la classification automatique des stades de sommeil (oxymètre de pouls, PneaVoX®, lunette nasale, actimètre et sangles inductives).

Parmi 1111 caractéristiques évaluées (de type statistique, temporel, spectral, non-linéaire ou lié à la forme), un ensemble de 112 caractéristiques a été estimé comme étant le plus pertinent pour la classification. Un modèle composé de trois étapes (réseau neuronal multicouche, corrections à l'aide des règles de transition et corrections à l'aide d'un MMC de Viterbi) a ensuite été entraîné puis testé sur un grand nombre d'enregistrements.
Afin de s'assurer du bon fonctionnement de l'analyse en pratique clinique, les enregistrements proviennent de patients avec et sans troubles du sommeil.
Encore une fois, l'analyse présentée ne nécessite aucune action préalable du médecin (par exemple l'invalidation de certaines époques ou la lecture partielle du sommeil). Cette facilité d'utilisation est complétée par la génération, en plus de l'hypnogramme automatique, d'une matrice de confusion. À partir de cette dernière, il est possible de générer un score de confiance. Cet outil permet d'alerter le médecin en cas de faible confiance de l'algorithme dans la classification réalisée.

Les résultats sont sans surprise inférieurs à ceux obtenus pour l'analyse automatique en PSG. Le Kappa de Cohen moyen obtenu est de 0,48 sur le groupe de test (taux d'accord moyen de 62,4 %), contre 0,69 pour l'analyse automatique en PSG (taux d'accord moyen de 77,8 %). Afin de comparer les résultats avec la littérature, nous avons également estimé les résultats pour des classifications réduites en deux ou trois stades. En particulier, la classification en trois stades éveil/sommeil lent (combinaison du N1, N2 et N3)/sommeil paradoxal obtient un Kappa de Cohen moyen de 0,60 et un taux d'accord moyen de 78,5 %.
Les différents indices utilisés lors du diagnostic (latences et proportions des stades) ont pu être estimés avec une précision moyenne. Cependant, l'analyse permet d'améliorer grandement l'estimation de l'IAH et ainsi de la sévérité du SAHS, et obtient des performances extrêmement proches de celles obtenues pour l'analyse automatique en PSG.

Annexe

Fonctionnement du perceptron multicouche

Le perceptron multicouche ou MLP, aussi appelé « *feed forwark network* », est un réseau de neurones artificiel.
Pour rappel, il est composé de plusieurs perceptrons (ou neurones) agencés entre eux. Le perceptron et son fonctionnement ont été illustrés page 29. On y avait vu qu'un perceptron est une unité de calcul élémentaire, qui réalise une somme pondérée de ses n entrées. En sortie, le neurone fournit une image de cette somme, selon sa fonction d'activation f (fonction heaviside, linéaire, sigmoïde ou autre). On définit ainsi la sortie du perceptron $Y = f(\sum_{a=1}^{n} i_a \times w_a)$, avec i_a l'entrée numérotée a et w_a le poids associé à cette entrée.

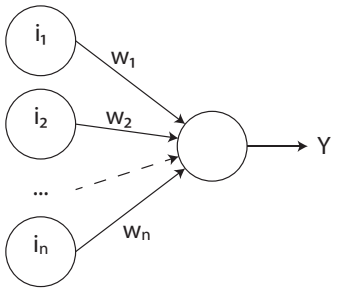


Figure E8 – Schéma d'un perceptron.

Dans le cas d'un MLP, la couche d'entrée, qui reçoit les données, est ainsi suivie d'une ou plusieurs couches cachées composées de plusieurs neurones. Ces couches cachées précèdent la couche de sortie.
Prenons un exemple pour lequel l'entrée n'est constituée que de deux éléments (par exemple, deux caractéristiques). Dans cet exemple, nous considérons par souci de lecture que nous n'avons qu'une couche cachée de 3 neurones. Ce MLP servant à classifier les stades de sommeil, notre sortie est composée de 5 neurones. La Figure E9 illustre cet exemple.
La sortie des neurones de la couche cachée $Y_v(k)$ (k étant 1, 2 ou 3) est donc :

$$Y_v(k) = f_v(\sum_{a=1}^{2} i_a \times v_{ak})$$

De la même façon, la sortie des neurones de la couche de sortie $Y_w(l)$ (l étant 1, 2, 3, 4 ou 5) est donc :

$$Y_w(l) = f_w(\sum_{k=1}^{3} Y_v(k) \times w_{kl})$$

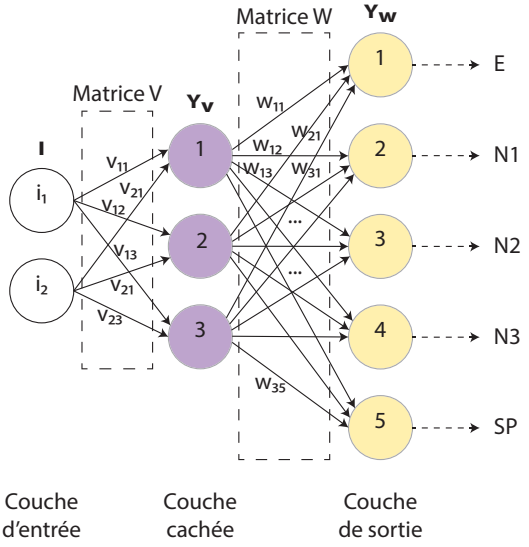


Figure E9 – Exemple de perceptron multicouche.

En considérant que les matrices V et W reportent les différents poids v_{ak} et w_{kl}, on peut aussi écrire $Y_v = f_v(I \times V)$ et $Y_w = f_w(Y_v \times W)$.

C'est lors de l'entraînement que les poids des connexions inter-neuronales sont définis. Cela s'effectue au travers d'un algorithme de rétropropagation du gradient de l'erreur.
Après avoir présenté un premier vecteur d'entrée au MLP juste initialisé, la sortie obtenue est comparée avec la sortie attendue, et l'erreur de sortie est évaluée. Cette erreur est ensuite propagée vers les couches précédentes, en parcourant le MLP en sens inverse. L'erreur des neurones cachés est alors évaluée comme étant l'erreur de sortie pondérée par des poids, et le poids de chaque connexion est alors ajusté. Ces étapes, répétées itérativement durant l'entraînement, permettent d'ajuster les poids de telle sorte à minimiser l'erreur finale.

Une référence pour aller plus loin : Chapitre 4 (« *Multilayer perceptrons : architecture and error backpropagation* ») de Du et Swamy (2014).

Bibliographie

Devot, S., Dratwa, R. et Naujokat, E. (2010). Sleep/wake detection based on cardiorespiratory signals and actigraphy. *In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 5089–5092. IEEE.

Domingues, A., Paiva, T. et Sanches, J. M. (2014). Hypnogram and Sleep Parameter Computation From Activity and Cardiovascular Data. *IEEE Transactions on Biomedical Engineering*, 61(6):1711–1719.

Du, K.-L. et Swamy, M. N. S. (2014). Multilayer Perceptrons : Architecture and Error Backpropagation. *In Neural Networks and Statistical Learning*, pages 83–126. Springer London, London.

Fonseca, P., Long, X., Radha, M., Haakma, R., Aarts, R. M. et Rolink, J. (2015). Sleep stage classification with ECG and respiratory effort. *Physiological Measurement*, 36(10):2027–2040.

Hayet, W. et Slim, Y. (2012). Sleep-wake stages classification based on heart rate variability. *In 2012 5th International Conference on BioMedical Engineering and Informatics*, pages 996–999, Chongqing, China. IEEE.

Kapur, V. K., Auckley, D. H., Chowdhuri, S., Kuhlmann, D. C., Mehra, R., Ramar, K. et Harrod, C. G. (2017). Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea : An American Academy of Sleep Medicine Clinical Practice Guideline. *Journal of Clinical Sleep Medicine*, 13(03):479–504.

Long, X., Yang, J., Weysen, T., Haakma, R., Foussier, J., Fonseca, P. et Aarts, R. M. (2015). Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging. *Physiological Measurement*, 36(3):625–625.

Willemen, T., Varon, C., Caicedo Dorado, A., Haex, B., Vander Sloten, J. et Van Huffel, S. (2015). Probabilistic cardiac and respiratory based classification of sleep and apneic events in subjects with sleep apnea. *Physiological Measurement*, 36(10):2103–2118.

Chapitre F

Discussion générale et perspectives

F.1 Discussion

Environ un tiers de la population est touchée par des troubles du sommeil. Ces troubles, aux symptômes souvent non spécifiques et fréquents dans la population générale, ne suffisent pas pour émettre un diagnostic. Il est nécessaire d'effectuer un enregistrement du sommeil à l'aide d'outils spécifiques et suffisamment précis. C'est grâce à ces outils que le praticien est en mesure d'émettre un diagnostic et de proposer le traitement le plus adapté.
Parmi les outils utilisés pour enregistrer le sommeil, la polysomnographie (PSG) et la Polygraphie Ventilatoire (PV) permettent l'acquisition, respectivement, des signaux électrophysiologiques et cardio-respiratoires, et des signaux cardio-respiratoires uniquement.
L'ensemble de ces signaux est visualisé par le spécialiste du sommeil qui peut alors y déceler de possibles particularités ou anomalies. En effet, les événements cardiaques ou respiratoires, et particulièrement les apnées du sommeil, sont identifiables sur les signaux cardiorespiratoires. Les stades de sommeil (éveil, N1, N2, N3 et sommeil paradoxal) sont quant à eux identifiés en utilisant les signaux électrophysiologiques et permettent de déceler d'autres anomalies. Cette analyse des stades de sommeil, uniquement réalisée lors d'un examen polysomnographique, est particulièrement chronophage et nécessite un haut niveau d'expertise. Elle a de plus une part de subjectivité liée à l'évaluation visuelle du spécialiste du sommeil et à son jugement.

Le Syndrome d'Apnées Hypopnées du Sommeil (SAHS) est l'une des pathologies du sommeil les plus reconnues. Sa détection et sa caractérisation peuvent être effectuées grâce à une PSG ou à une PV. Cependant, son traitement n'est possible que lorsque le syndrome est jugé suffisamment impactant et dangereux pour le patient. L'évaluation de la sévérité du SAHS repose sur les symptômes du patient mais aussi sur un index, l'Index d'Apnées Hypopnées (IAH). Ce dernier exprime le nombre d'événements ventilatoires (dont font partie les apnées par exemple) par heure de sommeil. Dans le cas d'une PV, la précision de l'IAH est moindre puisque la durée du sommeil n'est pas connue.

Dans cette thèse, nous avons étudié la faisabilité d'une analyse automatique des stades de sommeil dans le cas :

— d'une PSG, avec l'objectif de faciliter le travail du spécialiste du sommeil ;

— d'une PV, avec l'objectif de fournir, dans une certaine mesure, des informations essentielles au spécialiste du sommeil pour affiner le diagnostic ;

— d'un nouveau type d'examen, au croisement de la PV et de la PSG, avec encore l'objectif d'affiner le diagnostic.

La première partie de cette thèse est consacrée à ce système intermédiaire, qui est composé d'une PV améliorée par l'ajout d'une unique voie électroencéphalographique (EEG). L'objectif dans cette première partie est d'effectuer une analyse du sommeil simplifiée (de fournir un hypnogramme "*light*" en anglais, d'où le nom du projet HypnoLighT), en ne réalisant qu'une classification éveil versus sommeil. Cette information sur le sommeil, même incomplète, permet déjà d'obtenir un IAH plus précis et d'affiner le diagnostic du SAHS.

La voie EEG retenue a été la voie pré-frontale, située en haut du front, à la limite du cuir chevelu, et dont l'avantage est la facilité de la pose sur le patient. C'est également le meilleur placement pour observer à la fois les ondes alpha, très présentes lors de l'éveil calme, yeux fermés, et les mouvements oculaires observables lors de l'éveil yeux ouverts. Malgré le manque d'électromyogramme (EMG), l'éveil agité peut être identifié grâce aux mouvements visibles sur l'actimètre et les sangles inductives. Afin d'éviter les fausses détections d'éveil calme lors du sommeil paradoxal, la fréquence des ondes alpha spécifiques à l'éveil est identifiée automatiquement pour chaque enregistrement. Ensuite, une algorithme reposant sur des seuillages a été mis en place. Les résultats sont convaincants et montrent qu'il est possible, grâce au simple ajout d'une voie EEG en PV, d'identifier l'éveil de manière simple et robuste.

Pour aller plus loin, un algorithme de détection automatique des hypopnées micro-éveillantes a été implémenté. En effet, les hypopnées (uniquement) micro-éveillantes sont les seuls événements ventilatoires non identifiables en PV, contrairement à la PSG. Les résultats de cette détection sont limités. On estime qu'il serait nécessaire de revenir sur cette partie, et particulièrement sur l'estimation des micro-éveils, qui font l'objet de nombreuses recherches. Leur origine (autonomique ou corticale), leur rôle et leur influence ne sont pas encore tout à fait appréhendés par la communauté médicale. Malgré tout, une fois combiné à la détection automatique de l'éveil, l'impact d'un point de vue clinique est tout à fait satisfaisant.

Ce système de PV améliorée par une unique voie EEG permet d'estimer avec une très bonne précision l'IAH. La sévérité du SAHS est donc évaluée bien mieux qu'avec une simple PV.

En PSG, où l'ensemble des voies électrophysiologiques sont disponibles, il est cette fois-ci possible de réaliser une classification complète, en 5 stades (éveil, N1, N2, N3 et sommeil paradoxal). Ce travail a été présenté dans la deuxième partie de cette thèse.

Pour remédier aux faiblesses qui limitent actuellement l'utilisation clinique d'un algorithme permettant l'analyse automatique des stades de sommeil en PSG, l'approche choisie a consisté à reproduire, une à une, les différentes étapes réalisées par le spécialiste du sommeil lors de son analyse manuelle. Ainsi, une première étape permet l'évaluation de caractéristiques patient-dépendantes. Pour cela, un algorithme de seuillage auto-adaptatif et non-supervisé a été implémenté. Appelé SATUD (de Self-Adaptative Thresholding Using Descriptors), il repose sur la transposition en langage algorithmique des propriétés associées à chaque stade de sommeil et décrites dans le manuel de recommandations de l'*American Academy of Sleep Medicine* (AASM). Grâce à ces propriétés, les seuillages sont automatiquement ajustés à chaque patient, et les caractéristiques évaluées sont donc préservées de l'influence des spécificités du patient. Cette première étape est l'équivalent de la visualisation rapide de l'ensemble de l'enregistrement qui est réalisée par le spécialiste du sommeil avant sa classification, et qui lui permet de se familiariser avec les signaux propres audit patient. Une deuxième étape permet ensuite d'obtenir un premier hypnogramme à partir des composantes temporelles et spectrales (caractéristiques extraites). Le modèle mis en place est une forêt aléatoire, entraînée et testée sur un nombre conséquent d'enregistrements provenant de patients avec différentes pathologies du sommeil. La troisième étape consiste en la détection automatique des nombreux grapho-éléments qui, lorsqu'ils sont présents, permettent également l'identification des stades de sommeil. La quatrième et dernière étape consiste à utiliser ces grapho-éléments, ainsi que les règles de transition et la connaissance des époques à proximité, pour corriger l'hypnogramme obtenu lors de la deuxième étape. Pour une traduction optimale des recommandations AASM, il manquera tout de même, encore une fois, la détection automatique des micro-éveils. En effet, de nombreuses règles de transition reposent sur leur présence. La version actuelle ne prend donc pas en considération ces règles, qui devront être ajoutées dans un second temps, lorsqu'une détection automatique des micro-éveils efficiente sera disponible. Malgré tout, cette étape peut être décomposée en différentes fonctions, dont une classification par forêt aléatoire et un Modèle de Markov à états Cachés de Viterbi. Ce dernier a pour objectif d'apprendre à corriger les erreurs récurrentes. En sortie, en plus de l'hypnogramme obtenu par l'analyse automatique, est fournie une table de probabilités fournissant des indications sur le niveau de confiance de l'algorithme lors de la classification. Cette table est utilisée pour la création d'un graphique d'hypnodensité. Grâce à ce graphique, le spécialiste du sommeil peut rapidement identifier les époques pouvant nécessiter une reclassification manuelle ou non, et jauger la complexité de l'enregistrement.

Les résultats obtenus sont bons et permettent une estimation tout à fait correcte des proportions des différents stades et des différentes latences. L'IAH et la sévérité du SAHS sont

estimées avec une très bonne précision. Pour améliorer encore plus les performances de leur estimation, il faudrait non pas améliorer la classification existante, mais mettre en place une détection automatique des micro-éveils.

Ce système, tout en restant accessible à la communauté médicale, permet d'évaluer cliniquement les principaux indices nécessaires au diagnostic de nombreux troubles associés au sommeil. Facile d'utilisation, il ne nécessite pas l'intervention préalable d'un expert du sommeil (pas d'invalidation ni de lecture partielle requise) et est utilisable sur tout patient suspecté d'avoir des troubles du sommeil. Il procure au spécialiste du sommeil différents outils qui lui permettent, en modulant le niveau de l'aide à la lecture selon ses préférences et selon l'enregistrement, de réduire le temps alloué à l'étude post-enregistrement.

Dans la troisième partie de la thèse, la faisabilité de l'analyse automatique de l'ensemble des stades de sommeil en PV a été évaluée. N'étant pas réalisable manuellement par les experts du sommeil, cette tâche est plus challengeante que les précédentes.

La stratégie a donc été bien différente des deux premières parties pour lesquelles les connaissances médicales étaient suffisamment avancées. Les connaissances quant au lien entre les activités cardiaques et respiratoires, qui peuvent être étudiées en PV, et l'état de vigilance sont assez restreintes. En effet, si l'on sait que l'activité autonomique, qui varie en fonction des stades de sommeil, impacte le système cardiaque, l'évaluation des stades de sommeil grâce aux variations de l'activité autonomique reste extrêmement complexe. De même pour l'activité ventilatoire, qui varie en fonction des stades de sommeil (par exemple, le rythme respiratoire diminue lors de l'endormissement). De nombreuses études ont tout de même développé de nouveaux indicateurs qui permettent, dans une certaine mesure, d'étudier le sommeil à partir d'une voie cardiaque ou respiratoire. Cependant, les résultats sont limités et, surtout, souvent estimés sur des bases de données constituées uniquement de patients sains. Néanmoins, les événements ventilatoires, qui sont liés à certaines pathologies du sommeil (par exemple les apnées pour le SAHS), impactent également l'activité autonomique, et bien entendu l'activité ventilatoire. Notre objectif étant de fournir une classification automatique chez les patients investigués pour suspicion de troubles du sommeil, il n'est pas certain que les indicateurs présentés dans la littérature soient réellement adaptés. Pour cette raison, nous avons choisi une approche de fouille de données. Nous avons ainsi évalué 1111 caractéristiques, toutes déjà présentées dans la littérature, sur l'ensemble des signaux à disposition. Une étape de réduction de la dimensionnalité a ensuite permis d'obtenir un ensemble de 112 caractéristiques jugé comme étant le plus pertinent pour la classification du sommeil. Les caractéristiques sélectionnées proviennent uniquement de l'oxymètre de pouls (nous n'avons pas utilisé d'ECG puisque ce signal n'est pas enregistré en PV), du PneaVoX®, des sangles inductives, de la lunette nasale et de l'actimètre. Une classification constituée de la succession d'un réseau neuronal de type perceptron multicouche, de corrections basées sur les règles de transition et d'un Modèle de Markov à états Cachés de Viterbi est ensuite réalisé. En sortie, en plus de l'hypnogramme obtenu par l'analyse automatique, est encore une fois fournie une table de probabilités fournissant des indications sur le niveau de confiance de l'algorithme lors de la classification. Cette table est cette fois-ci utilisée pour la mise à disposition d'un score de confiance sur l'ensemble de la nuit. Grâce à ce score, le spécialiste du sommeil est en mesure d'estimer si la classification automatique est peu sûre, et de choisir de la prendre en compte ou non.

Les résultats pour une classification en 5 stades confirment la difficulté de la tâche. Cette classification ne doit donc être utilisée que lorsqu'elle est jugée fiable (notamment grâce à l'indice de confiance). La classification en 3 stades (éveil, sommeil paradoxal et sommeil lent) obtient des résultats bien plus convaincants. L'analyse automatique permet une estimation plutôt décevante des proportions des différents stades et des différentes latences. Cependant, l'IAH et la sévérité du SAHS sont estimées avec une très bonne précision.

Ce système a l'avantage de fournir des éléments de diagnostic normalement non disponibles au clinicien lors d'une PV. Avec le complément du score de confiance, on peut tout à fait imaginer l'utiliser dans le cadre de la détection des troubles respiratoires du sommeil comme le SAHS.

F.2 Perspectives

Différents axes de recherche peuvent être envisagés pour la poursuite du travail présenté dans cette thèse.

Premièrement, il est possible d'améliorer les performances du travail déjà réalisé en revenant sur certains aspects.
Par exemple, il y a une marge d'amélioration de la détection des grapho-éléments. À l'heure actuelle, leur détection est suffisante pour qu'ils soient utilisés dans la classification. Cependant, elle est insuffisante pour que l'on puisse imaginer que ces détections soient utilisées par les spécialistes du sommeil afin de faciliter encore plus leur lecture.
Les artéfacts n'ont également pas été identifiés ni supprimés dans le cadre de ce travail de thèse. Le développement d'outils supplémentaires pour leur détection et/ou suppression pourra très certainement améliorer les résultats, puisque les signaux d'entrée seront plus propres.
Comme nous avons pu l'indiquer dans de multiples sections de ce mémoire, la détection des micro-éveils permettra l'obtention d'un diagnostic bien plus précis.
Une autre amélioration facilement réalisable est la combinaison de la détection des stades de sommeil en PV avec le projet HypnoLighT. En effet, tous les signaux utilisés pour l'analyse du sommeil en PV sont disponibles dans le cadre d'un examen HypnoLighT. Il est alors possible d'améliorer les résultats d'HypnoLighT simplement en y ajoutant l'algorithme mis en place pour l'analyse en PV.

Deuxièmement, il est possible d'ouvrir le travail réalisé vers d'autres problématiques.
Les systèmes de diagnostic de Type IV ou de dépistage, composés d'un nombre réduit de voies cardio-respiratoires, n'ont en effet pas été considérés dans ce travail de thèse. Leur développement est pourtant de plus en plus fréquent, et il ne serait pas aberrant d'imaginer qu'ils deviennent reconnus par les spécialistes du sommeil dans les prochaines années. L'algorithme mis en place pour l'analyse automatique en PV est facilement adaptable à ce type de systèmes. En effet, il faut simplement dans ce cas réaliser une nouvelle fois la réduction de la dimensionnalité et l'entraînement du classifieur, mais cette fois-ci à partir des caractéristiques estimées à partir des seuls signaux disponibles pour le système réduit.
Il serait également intéressant de considérer l'adaptation des algorithmes présentés à d'autres populations cibles. En effet, les règles de lecture du sommeil utilisées dans le travail de cette thèse sont spécifiques à la population adulte. Cependant, les troubles du sommeil chez l'enfant ont des conséquences souvent plus graves que celles chez l'adulte. La possibilité d'utiliser les algorithmes en pédiatrie peut donc être pertinente.

Pour terminer, il est possible d'imaginer que l'algorithme SATUD puisse être généralisé et utilisé dans d'autres domaines, pour lesquels un seuillage auto-adaptatif serait avantageux et où il existe une large base de connaissances d'experts. C'est généralement le cas pour les systèmes biomédicaux.

F.3 Bénéfices pour le système de santé

Le travail réalisé dans cette thèse permet une amélioration globale de la qualité du diagnostic. Il sera valorisé par l'entreprise CIDELEC et, de ce fait, pourra être utilisé par les spécialistes du sommeil qui travaillent avec des outils de diagnostic CIDELEC.
À l'heure actuelle, un travail d'industrialisation est en finalisation sur le projet HypnoLighT (PV améliorée par une unique voie EEG). Le nouveau système devrait être disponible durant le premier trimestre 2021.
D'un point de vue medical, les bénéfices des différentes analyses automatiques présentées dans cette thèse peuvent être classifiés selon différents types de finalités : l'amélioration du système de santé, l'amélioration des soins et de la santé publique, et, pour finir, la recherche et l'augmentation des connaissances.

F.3.1 Amélioration du système de santé

Les trois analyses automatiques présentées dans cette thèse permettent de soutenir le spécialiste du sommeil lors de son diagnostic.
Dans le cas d'HypnoLighT, le nouveau système procure un IAH plus précis qu'en PV classique, se rapprochant de celui obtenu en PSG. Cela implique différents avantages :

— une **économie financière** : pour les patients qui effectuent une PV et qui obtiennent un IAH insuffisant pour qu'un traitement puisse être proposé, il est recommandé d'effectuer une PSG de seconde intention. Dans ce cas, l'utilisation en premier lieu d'une PV améliorée peut permettre de réaliser l'économie de ce second examen ;

— une **réduction du temps alloué au diagnostic** : pour les mêmes raisons.

Dans le cas de l'analyse automatique en PSG, le travail de lecture des stades de sommeil est grandement facilité par l'hypnogramme automatique et le graphique d'hypnodensité. L'analyse engendre donc une **diminution de la pénibilité du travail des spécialistes du sommeil**, pour qui la lecture du sommeil est plus rapide et facile, sans perte sur le contrôle de la lecture réalisée.
Dans le cas de l'analyse automatique en PV, des informations supplémentaires sont fournies au spécialiste du sommeil. Dans certains cas, cela peut aboutir, comme pour HypnoLighT, à l'économie d'un examen polysomnographique de seconde intention (économie financière et réduction du temps alloué au diagnostic).

F.3.2 Amélioration des soins et de la santé publique

Ces améliorations du système de santé impactent les soins et la santé publique.
Premièrement, il est probable que cela ait pour conséquence une **facilitation de la prise en charge des patients**. En effet, un impact probable serait la diminution des délais, qu'ils soient liés à l'attente de disponibilités pour effectuer un enregistrement du sommeil, ou à l'attente du verdict une fois l'enregistrement réalisé.
Une **amélioration de l'expérience du patient** devrait également en résulter puisque le spécialiste du sommeil, en passant moins de temps à lire le sommeil, sera en mesure de passer plus de temps au contact de ses patients.
Dans le cas d'HypnoLighT ou de l'analyse en PV, cela implique également une **amélioration de la santé de la population** puisque la qualité du diagnostic est améliorée et permet la prise en charge de plus de patients, d'autant plus que ces examens peuvent être réalisés en ambulatoire.

F.3.3 Recherche et augmentation des connaissances

Ce point concerne particulièrement l'analyse du sommeil en PV, pour laquelle l'approche a permis d'identifier les capteurs et les caractéristiques les plus pertinents pour l'analyse automatique des stades de sommeil. Elle donne d'une certaine façon quelques informations quant à la contribution du système nerveux autonome dans les composantes du sommeil.

F.4 Conclusion

Dans cette thèse, nous avons traité de l'analyse automatique des stades de sommeil, dans le cas de différents systèmes de diagnostic du sommeil, nouveaux ou déjà existants.
Pensées pour être utilisées comme des outils d'aide au diagnostic pour les spécialistes du sommeil, les méthodologies ont été choisies en fonction. Les différentes limites à l'utilisation clinique des systèmes actuels ont également été étudiées, et les stratégies mises en place ont été construites afin de surmonter ces limitations.
Les algorithmes présentés ont prouvé leur intérêt pour l'analyse des troubles du sommeil, et particulièrement du Syndrome d'Apnées Hypopnées du Sommeil. Leur apport, qu'il soit en terme d'amélioration de la prise en charge et de l'expérience du patient, ou de diminution du coût et de la pénibilité du travail du praticien, est indéniable. Il laisse à penser que le système pourra être accepté et utilisé par les spécialistes du sommeil, à l'heure où la place de l'intelligence artificielle dans le domaine est encore limitée.

Titre : Analyse automatique des stades du sommeil à partir des voies électrophysiologiques et cardio-respiratoires

Mots clés : lecture du sommeil, classification, aide au diagnostic, troubles du sommeil, polysomnographie, polygraphie ventilatoire

Résumé : Le diagnostic des troubles du sommeil repose sur l'analyse de différents signaux enregistrés lors d'un examen du sommeil. Cette analyse est réalisée par un spécialiste du sommeil qui étudie la ventilation et, selon l'outil de diagnostic, la succession des stades de sommeil. Cette dernière tâche est particulièrement chronophage et complexe. Trois algorithmes d'aide au diagnostic et dédiés à cette tâche sont présentés.
Le premier permet la classification éveil/sommeil lors de l'utilisation d'un nouvel outil de diagnostic. Il en découle la possibilité pour le médecin de diagnostiquer précisément le syndrome d'apnées du sommeil et à moindre coût.
Le deuxième, fondé sur les voies électrophysiologiques, permet d'obtenir une classification de tous les stades de sommeil à partir de l'outil de diagnostic le plus complet. Il a été implémenté en considérant les limitations à l'utilisation d'un tel algorithme en routine clinique. L'architecture de cet algorithme reproduit ainsi le processus de classification réalisé manuellement par les médecins. Une fonction de seuillage auto-adaptatif a aussi été mise en place afin de fournir une classification patient-dépendante. Les résultats obtenus sont comparables avec ceux des médecins.
Le troisième algorithme, fondé sur les voies cardio-respiratoires, permet de classifier les stades de sommeil à partir d'un outil de diagnostic très utilisé mais pour lequel il n'est normalement pas possible d'étudier les stades de sommeil. La tâche est complexe, mais les résultats obtenus sont satisfaisants vis-à-vis de la littérature.
Les trois algorithmes, destinés aux différents outils de diagnostic, permettront d'aider les spécialistes à analyser le sommeil.

Title: Automatic sleep stage analysis from electrophysiological and cardio-respiratory channels

Keywords: sleep scoring, classification, diagnosis support tools, sleep-disordered breathing, polysomnography, home sleep apnea testing

Abstract: The diagnostic of sleep-disordered breathing requires the analysis of various signals obtained while recording sleep. The analysis is carried by a sleep specialist, which studies the patient's ventilation and, depending on the diagnostic tool used for the record, sleep stages. Sleep stage scoring is a complex and time-consuming task. Three diagnosis support algorithms dedicated to this task are presented in this thesis.
The first one provides a wakefulness versus sleep classification, designed for a new diagnostic tool. It results in the ability to make a precise diagnosis of sleep apnea syndrome, at low cost.
The second algorithm, based on electrophysiological channels, provides a full sleep stage classification while using the most complete diagnosis tool. It was implemented considering the known limitations for the use of algorithms in clinical practice. Its architecture thus reproduces the manual scoring process. A self-adaptative thresholding function was also implemented to provide a patient-dependent classification. The obtained results are comparable with the ones from sleep experts.
The third algorithm, based on cardio-respiratory channels, provides a sleep stage classification while using a diagnostic tool that is insufficient for a manual sleep scoring, yet still highly used. The task is challenging but the obtained results are satisfying compared to literature.
All three algorithms, which were designed for various diagnostic tools, will help sleep experts analyzing sleep.