Étude de la généralisation d'une approche multimodale par apprentissage pour l'estimation du sommeil calme chez des nouveau-nés prématurés

Houda JEBBARI Sandie CABON Céline CITTÉ Patrick PLADYS Guy CARRAULT Fabienne PORÉE
Univ Rennes, CHU Rennes, INSERM, LTSI - UMR 1099, F-35000, France

Résumé – Cette étude présente une méthode d'estimation du sommeil calme pour le nouveau-né prématuré à partir de signaux cardio-respiratoires et du mouvement issu de vidéos. Celle-ci repose sur le calcul d'un nombre élevé (120) de paramètres et d'une étape de classification intégrant une étape de sélection de caractéristiques pour aboutir à un modèle compact. L'originalité de ces travaux repose sur le fait que deux stratégies sont comparées : l'application d'un modèle entraîné sur un jeu de données antérieur et le calcul d'un modèle nouveau adapté aux données à traiter. Les résultats obtenus montrent que les variations entre le jeux de données, et entre les annotations, impactent les performances.

Abstract – This study presents a method for estimating quiet sleep in premature newborns using cardiorespiratory signals and motion from videos. The method is based on the calculation of a large number (120) of parameters and a classification step including a feature selection step to produce a compact model. The originality of this work lies in the fact that two strategies are compared: the application of a model trained on a previous dataset and the calculation of a new model adapted to the data processed. The results show that variations between datasets and between annotations have an impact on performance.

1 Introduction

La prématurité est une naissance survenue avant 37 semaines de grossesse, ou Âge Gestionnel (AG), et concerne à peu près 55 000 naissances en France par an, soit 7% des naissances. Les prématurés ont plusieurs fonctions immatures et font l'objet d'une surveillance spécifique dans les unités de soins intensifs néonatals. Le sommeil fait partie des données mesurées car il est directement lié au développement du cerveau. Il est donc primordial de préserver au mieux le sommeil des bébés prématurés. Mais il est aussi nécessaire d'évaluer la qualité de leur sommeil, car un sommeil altéré peut être révélateur d'une pathologie sous-jacente.

Chez le nouveau-né, les stades de sommeil sont dit comportementaux car ils dépendent de son activité corporelle, de l'état de ses yeux, de son rythme cardio-respiratoire.... La classification la plus souvent retenue est celle de Brazelton, incluant les stades suivants : le Sommeil Calme (SC), le sommeil agité, la somnolence et la veille [1]. Ils sont "annotés" par des cliniciens, par observation directe du bébé. Ces analyses, subjectives et chronophages, ne peuvent donc être que ponctuelles. L'estimation automatique des stades de sommeil a donc un intérêt clinique évident qui a fait l'objet de plusieurs études récentes.

La plupart des méthodes de la littérature ont été mises en œuvre à partir de signaux électrophysiologiques cardiaques et/ou respiratoires [7, 8]. Quelques travaux sont basés sur l'analyse de l'expression faciale [4]. Pour être aussi proches que possible de la pratique clinique appliquée lors des annotations manuelles, dans nos travaux, nous exploitons les deux types de données, à savoir les signaux électrophysiologiques (électrocardiogramme et respiration) et la vidéo, dont on extrait le mouvement [2]. Par ailleurs, nous nous concentrons actuellement sur l'estimation du SC, car c'est le stade le plus important pour suivre la maturation [3]. Il est caractérisé par une absence d'activité motrice et un rythme cardio-respiratoire régulier.

Cette étude s'inscrit dans la continuité de nos travaux précédents. Dans [5], une méthode supervisée, basée sur l'extraction de 120 paramètres, a été mise en oeuvre sur un petit jeu de données annoté (200 heures), extrait de la base de données du projet Digi-NewB (2016-20) [6]. Elle intègre une étape de sélection de paramètres qui permet d'aboutir à un modèle compact (10 paramètres). Plusieurs algorithmes ont été optimisés et comparés et les meilleures performances ont été obtenues avec l'algorithme des Random Forest.

Le travail que nous présentons ici a été réalisé dans le cadre du projet Neovideo (2020-23), centré sur l'analyse du sommeil des prématurés. Un de ses objectifs était le développement d'un outil d'estimation automatique des stades de sommeil. Une nouvelle base de données a été acquise, avec le même système que dans Digi-NewB. De plus, un travail important d'annotation a été réalisé, conduisant à plus de 2600 heures annotées au total.

Dans cet article, nous présentons l'ensemble des méthodes développées et appliquées dans le projet Neovideo. Une attention particulière a été portée sur la stratégie d'entraînement des modèles. Il s'agit d'étudier les questions de généralisation des méthodes par rapport aux changements de base de données. Ayant à notre disposition un modèle entraîné sur des données Digi-NewB, nous avons cherché à savoir si celui-ci pouvait être utilisé pour traiter les données Neovideo, ou s'il était préférable de réentraîner un modèle nouveau.

L'article est organisé comme suit. La section 2 décrit le protocole, la base de données et les méthodes mises en œuvre. Dans la section 3, les résultats obtenus sont présentés. La discussion et la conclusion sont données dans la section 4.

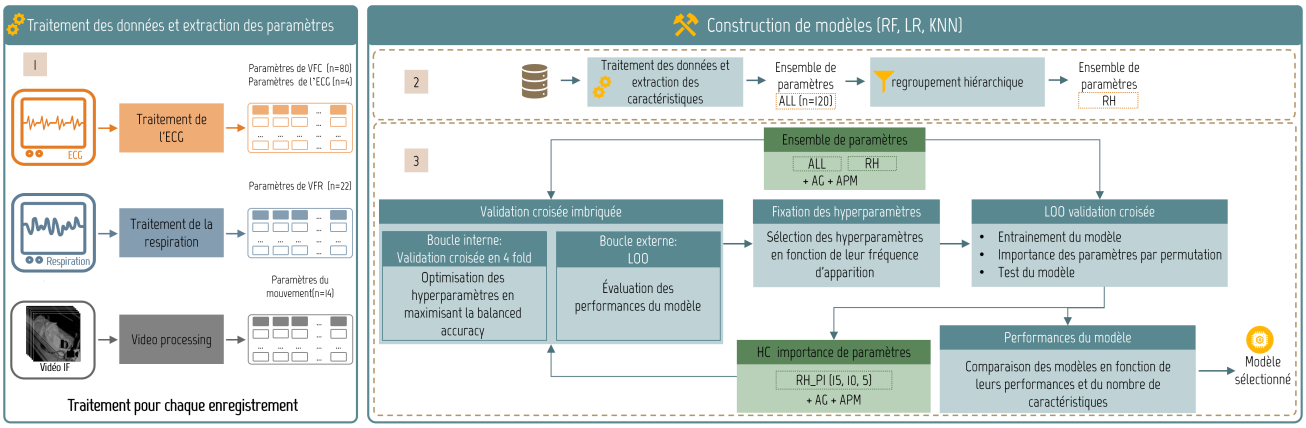

FIGURE 1 : Description de la méthode d'estimation du sommeil calme.

2 Méthode

2.1 Protocole et base de données

La base de données a été acquise dans le cadre du projet Neovideo. Ce projet a été approuvé par le comité de protection des personnes (CPP Paris Ile de France 2020/59). Les données ont été collectées dans deux hôpitaux français (Brest et Rennes). Pour chaque bébé, on a enregistré les signaux physiologiques et deux vidéos, pendant 8 jours consécutifs. L'ECG et la respiration ont été acquis à une fréquence d'échantillonnage de 500 Hz et 62.5 Hz, respectivement. Les vidéos ont été enregistrées avec des caméras infrarouges à la fréquence de 25 images par secondes avec un encodage MPEG-4. Un total de 74 bébés prématurés a été inclus, avec un âge gestationnel compris entre 25+1 et 31+6 semaines.

2.2 Annotations

Les annotations des stades de sommeil ont été réalisées, par une experte du sommeil, sur un ensemble de 47 bébés. Pour cela, elle a utilisé un logiciel développé par notre équipe [3], permettant la visualisation des deux vidéos et de trois signaux : l'électrocardiogramme, le rythme cardiaque et la respiration. Les courbes de mouvement (issues des vidéos) étaient aussi représentées. Trois états ont été annotés : le SC, l'éveil et un état 'NOK', correspondant aux périodes non annotables (par exemple quand le bébé n'est pas visible). Les périodes restantes ont été affectées au label 'sommeil non calme', celui-ci regroupant le sommeil agité et la somnolence. Un total de 2600 heures ont été annotées.

Notre étude se focalisant sur l'estimation du SC, l'éveil et le sommeil non calme sont fusionnés dans le stade noté $\overline{SC}$.

2.3 Méthode d'estimation du sommeil calme

La méthode d'estimation du SC repose sur trois étapes : (i) le calcul de paramètres, (ii) la sélection de paramètres et (iii) la classification avec une deuxième sélection de paramètres (Figure 1). Le détail de la méthode est présenté dans [5].

2.3.1 Calcul des paramètres

Un total de 120 paramètres est extrait à partir du signal ECG, de la respiration et du mouvement issu des vidéos, comme illustré dans la Figure 1.1. Concernant l'ECG, 80 paramètres liés à la Variabilité de la Fréquence Cardiaque (VFC) ainsi que 4 paramètres relatifs à l'amplitude du signal sont calculés. Pour la respiration, 22 paramètres décrivant la Variabilité de la Fréquence Respiratoire (VFR) sont extraits [5]. Ces paramètres de variabilité sont répartis en trois catégories :

- Temporels : 25 paramètres VFC et 13 VFR
- Fréquentiels : 8 paramètres VFC et 4 VFR
- Non linéaires : 47 paramètres VFC et 5 VFR

De plus, 14 paramètres liés au nombre et à la durée des intervalles de mouvement et de non-mouvement ont été calculés [2]. L'âge gestationnel et l'âge postmenstruel ont aussi été inclus.

Enfin, les valeurs des paramètres ont été standardisées pour chaque enregistrement en utilisant le z-scoring, afin d'assurer une échelle uniforme et réduire l'impact de la variabilité individuelle.

2.3.2 Classification

Dans [5], trois méthodes de classification ont été testées : le Random Forest, la régression logistique et le KNN. Pour chacune d'elles, les hyperparamètres ont été optimisés, dans le but d'obtenir un modèle performant, compact et interprétable.

La stratégie appliquée est illustrée dans la Figure 1.3. Une boucle imbriquée a été utilisée, comprenant une boucle extérieure pour évaluer la généralisation et une boucle intérieure pour l'optimisation des hyperparamètres, afin de minimiser le biais et l'overfitting. La boucle extérieure repose sur une validation croisée "leave-one-out" (LOO), où le modèle est entraîné sur tous les enregistrements des patients, en réservant les données d'un seul bébé pour le test à chaque itération. Cette boucle se répète autant de fois qu'il y a de bébés dans l'ensemble de données. La boucle intérieure optimise les hyperparamètres des modèles à l'aide d'une grille de recherche, en maximisant la balanced accuracy, qui est la moyenne de la sensibilité et de la spécificité. Elle applique une validation croisée à 4 fold sur les données d'entraînement, en regroupant les données au niveau du patient et en stratifiant par état SC. Le meilleur ensemble d'hyperparamètres a été choisi en fonction de la balanced accuracy moyenne la plus élevée obtenue sur les ensembles de validation. Ensuite, les hyperparamètres optimaux pour chaque modèle ont été sélectionnés en fonction de leur fréquence d'apparition dans la boucle extérieure. Enfin,

une validation croisée LOO standard a été effectuée sur les hyperparamètres fixés pour évaluer les performances du modèle final.

2.3.3 Ensembles des paramètres à l'entrée du modèle

Trois stratégies ont été testées.

- L'ensemble des paramètres (ALL) : Cet ensemble de paramètres regroupe les 120 paramètres (106 issus des signaux ECG et respiratoires, et 14 liés au mouvement) et sert de référence pour la comparaison des modèles entrainés sur les autres ensembles.

- Le regroupement hiérarchique (RH) : Compte tenu du grand nombre de paramètres extraits, en particulier ceux issus de l'ECG, une sélection préliminaire a été effectuée afin de réduire la redondance due à la multicolinéarité entre les paramètres. Pour ce faire, un regroupement hiérarchique non-supervisé des paramètres a été appliqué à chaque modalité, en s'appuyant sur leur degré de corrélation. La méthode Silhouette a permis de déterminer un seuil optimal pour fixer le nombre de groupes, et un unique paramètre, sélectionné aléatoirement, a été conservé au sein de chaque groupe.

- La permutation d'importance (PI) : Cet ensemble de caractéristiques étend RH en intégrant l'importance des variables. L'importance par permutation a été appliquée aux modèles entraînés sur RH afin d'identifier les caractéristiques les plus influentes (configuration RH_PI). Lors de la validation croisée LOO, la diminution de la balanced accuracy après permutation de chaque variable a été enregistrée. Trois sous-ensembles de 15, 10 et 5 caractéristiques ayant le plus grand impact ont été retenus.

2.4 Comparaison des estimateurs

La Figure 2 illustre la problématique centrale abordée dans cet article.

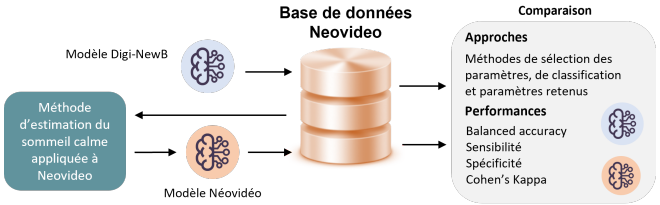

FIGURE 2 : Stratégie de comparaison des estimateurs.

Comme détaillé dans [5], la méthode d'estimation du SC a été appliquée sur la base de données de Digi-NewB. Après une évaluation exhaustive des performances sur l'ensemble des configurations possibles, le modèle le plus compact et performant a été entraîné sur l'ensemble du jeu de données, donnant naissance au Modèle Digi-NewB. Parallèlement, la méthode d'estimation du SC a été appliquée à la base de données Neovideo, où elle a été évaluée, aboutissant ainsi au Modèle Neovideo. Afin d'analyser la capacité de généralisation du Modèle Digi-NewB, celui-ci a été appliqué en inférence sur la base de données Neovideo sans réentrainement. Les estimateurs peuvent ainsi être comparés au regard :

- des méthodes de sélection et de classification retenues et des paramètres sélectionnés ;
- de leurs performances en termes de discrimination : balanced accuracy, sensibilité, spécificité et Cohen Kappa.

Cette expérimentation permet d'évaluer la généralisation à la fois de notre approche mais aussi des modèles à une nouvelle base de données.

3 Résultats

Nous proposons d'observer les résultats en deux temps. En premier, nous nous concentrons sur la comparaison entre les méthodes de sélection de paramètres et de classification retenues et les paramètres intégrés lors de l'apprentissage du Modèle Neovideo, nouvellement produit et celles et ceux retenus lors de notre précédente étude. Ensuite, nous comparons les performances de ce modèle avec celles obtenues lorsque le Modèle Digi-NewB est appliqué en inférence.

3.1 Comparaison des méthodes et paramètres retenus

Les méthodes de mise en oeuvre de l'estimateur ont été appliquées pour identifier la meilleure approche sur la base Neovideo (Modèle Neovideo). Parmi les 9 configurations étudiées (3 sets de paramètres sélectionnés x 3 classifieurs), la meilleure en terme de balanced accuracy et minimisant le nombre de paramètres s'est avérée être la même que lors de notre expérimentation menant au Modèle Digi-NewB. Il s'agit de la configuration RH_PI avec 10 paramètres mis en entrée d'un Random Forest. Cette stabilité entre les expériences rassure quant à la modélisation du problème par cette approche. Les paramètres inclus dans les modèles Digi-NewB et Neovideo sont rapportés dans la Figure 3. Leur description exacte est donnée dans [5].

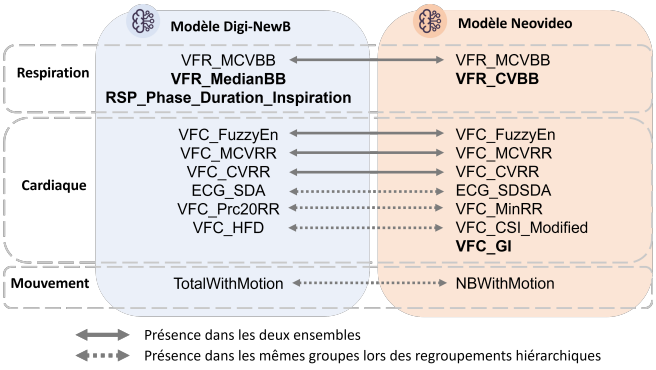

FIGURE 3 : Comparaison des ensembles de paramètres retenus pour les modèles Digi-NewB et Neovideo. Les paramètres pour lesquels un lien fort n'a pas été établi sont marqués en gras.

Quatre paramètres ont été retenus pour les deux modèles. Pour les 12 autres, nous avons analysé les groupes de paramètres générés lors des regroupements hiérarchiques lors des développements du modèle Digi-NewB et du modèle Neovideo. Huit d'entre eux se retrouvaient au sein des mêmes regroupements, indiquant qu'ils partagent des similarités fortes en lien avec la physiologie, et que c'est seulement par hasard que les mêmes n'ont pas été sélectionnés (un seul paramètre par groupe est aléatoirement retenu). Pour les quatre restants,

les liens ne sont pas établis. Il s'agirait alors d'une différence intrinsèque aux jeux de données étudiés, en faveur d'un ré-apprentissage des modèles en fonction de la base étudiée.

3.2 Comparaison des performances

Les performances obtenues avec chacun des modèles sont rapportées dans la Figure 4. Pour le modèle Neovideo, nous rapportons les moyennes et écart-types obtenus sur chacun des jeux de validation lors de la validation croisée (LOO) et pour le modèle Digi-NewB, celles obtenues en appliquant le modèle en inférence sur ces mêmes jeux (sans ré-entraînement). Pour comparaison, les performances obtenues lors de notre précédente expérimentation ont aussi été intégrées.

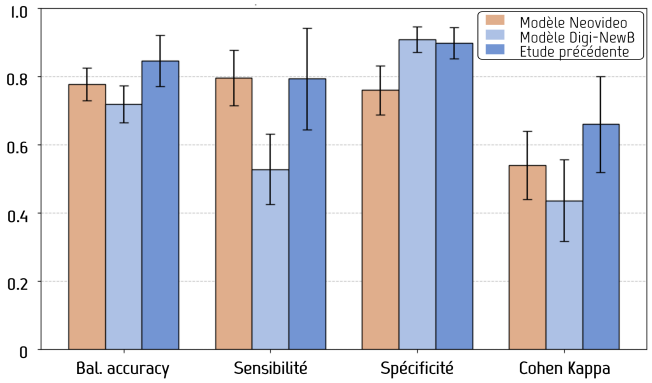

FIGURE 4 : Comparaison des performances obtenues avec les différents modèles.

Le modèle Neovideo est plus performant en termes de balanced accuracy (77.8±4.8%), sensibilité (79.6±8.1%) et Cohen's Kappa (0.54±0.10). La variabilité des métriques est comparable entre les deux modèles. Le modèle Digi-NewB présente une meilleure spécificité (90.9±3.4% vs 76.0±7.1).

Lors de notre précédente étude sur la base de données Digi-NewB [5], des performances légèrement supérieures en moyenne mais globalement moins stables avaient été obtenues : balanced accuracy (84.6±7.5%), sensibilité (79.4±14.9%), spécificité (89.8±4.6) et Cohen Kappa (0.66±0.14).

Ces résultats nous indiquent que i) l'approche globale est adaptée pour générer des estimateurs performants au sein d'une même base ii) une faiblesse du modèle Digi-NewB en terme de généralisation, et en particulier critique au niveau de sa sensibilité (-25% lorsqu'il est appliqué sur la base Neovideo).

4 Discussion et Conclusion

Dans cet article, nous nous sommes intéressés à la problématique de généralisation des modèles. Nous avions à notre disposition un modèle précédemment entraîné sur les données Digi-NewB (200h), et un nouveau jeu de données à traiter, celles de Neovideo (2600h).

Les données Digi-NewB et Neovideo ont été acquises avec le même système d'acquisition et dans les mêmes hôpitaux. Pourtant les résultats que nous avons obtenus montrent des performances meilleures quand le modèle est réappris sur le nouveau jeu de données. Quelques éléments d'explication se trouvent sans doute dans des différences liées aux protocoles d'acquisition et d'annotation des données.

D'une part, dans la base Neovideo, le nombre de bébés est plus grand, les enregistrements étaient de plus longue durée, et surtout les bébés étaient en moyenne plus prématurés.

D'autre part, dans Neovideo, les annotations ont été réalisées sur des périodes de journée, alors que dans Digi-NewB, des périodes de nuit avaient été choisies. Les annotations ont été réalisées par deux experts différents. Alors que dans Digi-NewB l'expert ne devait annoter que le sommeil calme, dans Neovideo, l'annotation portait sur trois stades (sommeil calme, éveil et sommeil non calme). Enfin, pour des raisons liées à l'étude clinique, dans Neovideo l'expert avait accès aux courbes de mouvement ; pas dans Digi-NewB.

Ces résultats soulignent l'importance de vérifier la capacité de généralisation des modèles par apprentissage. Les variations entre jeux de données, entre annotations sont difficiles à quantifier et à contrôler automatiquement. Il est aussi complexe d'identifier celles qui auront un impact sur les prédictions d'un modèle et donc d'assurer un maintien des performances.

Remerciements

Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet SLEEPINESS ANR-23-CE19-0020-01.

Références

[1] TB BRAZELTON et JK NUGENT : Neonatal behavioral assessment scale. Cambridge University Press, 1995.

[2] S. CABON, R. WEBER, S. SIMON et al. : Functional age estimation through neonatal motion characterization using continuous video recordings. IEEE J Biomed Health Inform, 27:1500–11, 2023.

[3] L. CAILLEAU, R. WEBER, S. CABON et al. : Quiet sleep organization of very preterm infants is correlated with postnatal maturation. Front Pediatr, 8, 2020.

[4] D. HUANG, D. YU, Y. ZENG et al. : Generalized camera-based infant sleep-wake monitoring in NICUs : A multi-center clinical trial. IEEE JBHI, 28(5):3015–3028, 2024.

[5] H. JEBBARI, S. CABON, P. PLADYS, G. CARRAULT et F. PORÉE : A Compact Quiet Sleep Estimator Based on Cardiorespiratory and Video Motion Features for Maturation Analysis in NICU. IEEE JBHI, pages 1–11, 2025.

[6] Digi-NewB GCS HUGO CHU monitoring SYSTEM : http ://www.digi-newb.eu. 14 April 2020.

[7] T. SENTNER, X. WANG, E. de GROOT et al. : The sleep well baby project : an automated real-time sleep–wake state prediction algorithm in preterm infants. Sleep, 45(10):zsac143, 2022.

[8] J. WERTH, M. RADHA, P. ANDRIESSEN et al. : Deep learning approach for ECG-based automatic sleep state classification in preterm infants. Biomed Signal Process Control, 56:101663, 2020.