https://doi.org/10.1093/sleep/zsad306
Advance access publication 1 December 2023
Original Article

Original Article

Studying sleep: towards the identification of hypnogram features that drive expert interpretation

Caspar van der Woerd[1,2], Hans van Gorp[2,3], Sylvie Dujardin[4], Manuel Sastry[5], Humberto Garcia Caballero[1], Fokke van Meulen[3,4], Stef van den Elzen[1], Sebastiaan Overeem[3,4] and Pedro Fonseca[2,3,*]

[1]Department Mathematics and Computer Science, Eindhoven University of Technology
[2]Remote Patient Management and Chronic Care, Philips Research, Eindhoven, The Netherlands
[3]Department Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[4]Sleep Medicine Center, Kempenhaeghe, Heeze, The Netherlands
[5]Academic Sleep Clinic, CIRO, Horn, The Netherlands

[*]Corresponding author. Pedro Fonseca, Philips Research Eindhoven, High Tech Campus, Eindhoven 34 5656AE, The Netherlands. *Email:* pedro.fonseca@philips.com

Abstract

Study Objectives: Hypnograms contain a wealth of information and play an important role in sleep medicine. However, interpretation of the hypnogram is a difficult task and requires domain knowledge and “clinical intuition.” This study aimed to uncover which features of the hypnogram drive interpretation by physicians. In other words, make explicit which features physicians implicitly look for in hypnograms.

Methods: Three sleep experts evaluated up to 612 hypnograms, indicating normal or abnormal sleep structure and suspicion of disorders. ElasticNet and convolutional neural network classification models were trained to predict the collected expert evaluations using hypnogram features and stages as input. The models were evaluated using several measures, including accuracy, Cohen’s kappa, Matthew’s correlation coefficient, and confusion matrices. Finally, model coefficients and visual analytics techniques were used to interpret the models to associate hypnogram features with expert evaluation.

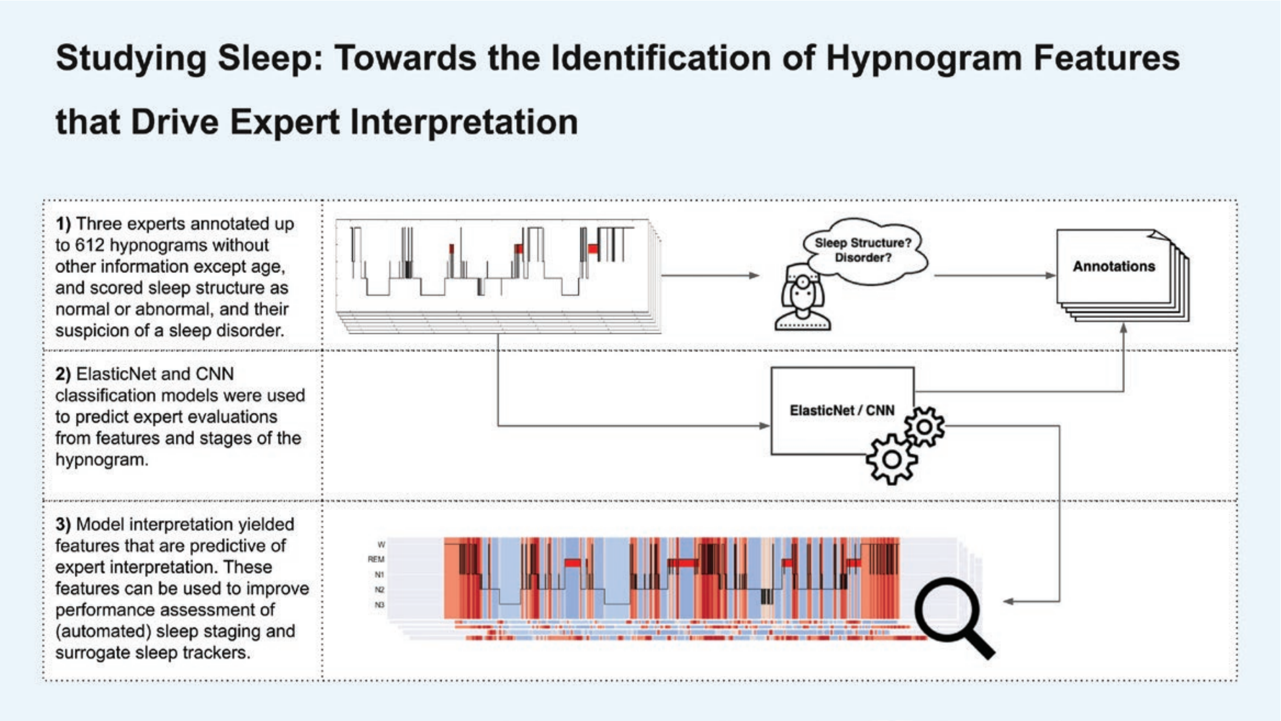
Results: Agreement between models and experts (Kappa between 0.47 and 0.52) is similar to agreement between experts (Kappa between 0.38 and 0.50). Sleep fragmentation, measured by transitions between sleep stages per hour, and sleep stage distribution were identified as important predictors for expert interpretation.

Conclusions: By comparing hypnograms not solely on an epoch-by-epoch basis, but also on these more specific features that are relevant for the evaluation of experts, performance assessment of (automatic) sleep-staging and surrogate sleep trackers may be improved. In particular, sleep fragmentation is a feature that deserves more attention as it is often not included in the PSG report, and existing (wearable) sleep trackers have shown relatively poor performance in this aspect.

Key words: Sleep staging; hypnogram; sleep structure; inter-rater agreement; sleep fragmentation; sleep diagnosis

Submitted for publication: July 6, 2023; **Revised:** October 18, 2023

Graphical Abstract



Studying Sleep: Towards the Identification of Hypnogram Features that Drive Expert Interpretation

1) Three experts annotated up to 612 hypnograms without other information except age, and scored sleep structure as normal or abnormal, and their suspicion of a sleep disorder.

2) ElasticNet and CNN classification models were used to predict expert evaluations from features and stages of the hypnogram.

3) Model interpretation yielded features that are predictive of expert interpretation. These features can be used to improve performance assessment of (automated) sleep staging and surrogate sleep trackers.

Statement of Significance

The hypnogram contains a wealth of information, and analysis of the hypnogram is a core part of the sleep diagnostic process. However, there is surprisingly little research on how hypnograms are interpreted and which aspects are relevant to experts. Nevertheless, assessment of hypnograms requires expertise, obtained through training and clinical experience. We believe that there must be certain features of the hypnogram that implicitly drive experts in their assessment. This motivated us to conduct the present study, where we aim to identify and quantify these features that are implicitly used by experts. Identifying and understanding these features is meaningful for several reasons and has important implications for automatic scoring algorithms and inter-rater (dis)agreement programs.

Introduction

Polysomnography (PSG) remains the current gold standard for assessment of sleep and sleep quality. Using standardized scoring rules, provided by the American Academy of Sleep Medicine, sleep stages are assigned to each 30-second epoch in the recording [1]. The sequence of annotated sleep epochs throughout the night is then visualized in a hypnogram. Additionally, a limited number of quantitative measures are extracted from the hypnogram, such as sleep onset latency, sleep efficiency, and percentages spent in each sleep stage [1].

From a clinical perspective, the hypnogram contains a wealth of information. Therefore, it is one of the most important instruments in sleep medicine. Together with other information from the PSG report, a patient's medical history, and experienced symptoms, the hypnogram is used to determine if and which sleep disorders are present and how sleep quality is affected. Despite playing such an important role, there are no formal guidelines on how to visually interpret a hypnogram. Instead, interpretation of the hypnogram is subjective and requires domain knowledge and "clinical intuition." Nevertheless, there must be certain patterns or features of the hypnogram that are important to experts and that drive their interpretation. These patterns are not explicitly described in sleep medicine manuals or guidelines. Instead, the recognition of these features implicitly develops with clinical experience, after visualizing large numbers of hypnograms during training and in practice, parsed together with clinical presentation and medical history.

Our aim is to uncover these features of the hypnogram and make them explicit and quantifiable. Furthermore, we want to find out how strongly and in which manner they contribute to expert interpretation of the hypnogram.

Understanding how this "clinical intuition" works and which features of the hypnogram drive expert interpretation, is important for several reasons. First, the creation of a hypnogram is a costly and a time-consuming process. It should therefore be studied if and how hypnograms are used in clinical practice. Second, knowing which features of the hypnogram drive interpretation enables us to focus on more meaningful metrics for the evaluation of automatic scoring algorithms, besides the commonly used accuracy and Cohen's Kappa of agreement. These metrics would allow for a clearer distinction between scoring algorithms which might give similar accuracy scores and yet display different sleep behavior, for example for sleep fragmentation (number of sleep stage transitions per hour). Third, it could provide a new lens to look at human inter-rater (dis)agreement in sleep stage scoring. It is well-documented that there is a limited agreement between human scorers on an epoch-by-epoch basis of only around 82.6%. In particular, N1 and N3 suffer from this with reported agreement of 63% and 67.4%, respectively [2]. While this disagreement itself has often been studied, its effect on how hypnograms are clinically used and interpreted has not. If we know which hypnogram features drive clinical interpretation, we know which forms of disagreement are meaningful and which are not. Last but not least, understanding which hypnogram features are important may inspire the derivation of new clinically meaningful quantifiable measures besides the standard measures such as sleep onset latency and sleep efficiency, usually provided in PSG reports.

There are few studies that assess how hypnograms are interpreted and used in clinical settings. It therefore remains largely unknown which features of the hypnogram drive expert interpretation. To the best of our knowledge, there is only one study that aims to relate features of the hypnogram with subjective evaluations by sleep experts. In a small experiment, it was shown that the interpretation of 52 hypnograms in terms of normal or abnormal structure could be predicted accurately from the sleep stage distribution (percentage per sleep stage) [3]. Moreover, it was demonstrated that specific disorders, such as sleep-disordered breathing and insomnia, are associated with differences in sleep structure [4, 5]. However, the diagnostic value of the hypnogram is limited when considered without clinical information. For example, a healthy person may also show an abnormal sleep structure because of the “first-night effect” [6]. Conversely, in some cases, a person with a clinically relevant disorder may have an apparently normal sleep structure.

In this study, we took a quantitative approach to identify which features of the hypnogram contribute to interpretation by experts. To that end, hypnograms were presented to sleep physicians without any additional clinical information. Their assessments regarding whether hypnograms were normal or not, or possibly corresponding to a sleep disorder, were recorded. We then trained a logistic regression (LR) and a neural network (NN) model to predict these expert assessments, using both traditional as well as automatically learned features of the hypnogram. By attempting to “mimic” the expert classification, we expect these models to use similar features as those implicitly used by the physicians. The most important features driving the classification with both models were finally uncovered using correlation analysis, and a NN visual analysis technique.

Methods

Study outline

Hypnograms were obtained from PSG recordings available in the Sleep and Obstructive Sleep Apnea Monitoring with Noninvasive Applications (SOMNIA) database [7]. After exclusion of participants younger than 18 or older than 80 years, 1096 hypnograms were available. These included participants with a wide variety of sleep disorders. The recordings were scored by experienced sleep technicians at the Sleep Medicine Center Kempenhaeghe for each 30-second epoch using the AASM guidelines [8]. Additionally, 97 hypnograms from healthy participants were included in the HealthBed dataset [9]. The HealthBed dataset comprises healthy adults without sleep disorders or other medical or psychiatric comorbidity, who underwent a PSG using the same protocol as the SOMNIA recordings [7]. The SOMNIA and HealthBed studies were reviewed by the medical ethical committee of the Maxima Medical Center (Veldhoven, the Netherlands. File no: N16.074 and W17.128). The protocol for data analysis was approved by the Institutional Review Board of the Kempenhaeghe Hospital and by the Internal Committee of Biomedical Experiments of Philips Research.

The hypnograms were assessed by three clinical sleep experts, all experienced physicians with backgrounds in general medicine, pulmonology, and neurology, referred to as experts A, B, and C, respectively. Since there are no clear guidelines on hypnogram interpretation, and we are interested in subjective assessments, we did not provide instructions on how the assessment should be done. Instead, we left this up to the experience and clinical intuition of the physicians. For each hypnogram, the experts had the choice options of *normal*, *abnormal but healthy,* or *abnormal with a suspected disorder*. In other words, *normal* and *abnormal* specifically referred to the visualized sleep structure. For example, an abnormal sleep structure in a healthy participant could very well occur because of the first night effect [6]. During a pilot study, it was found that, since sleep experts are used to visualize hypnograms when diagnosing disorders, it was helpful to explicitly assess both sleep structure (normal vs. abnormal) and the presence of a disorder as separate choice options.

Importantly, the hypnograms were presented in a restricted setting, with no clinical information except the participants’ age. Other factors were omitted to focus the interpretation on the structure of the hypnogram itself. Age was included as it is a well-known factor that influences sleep structure regardless of the presence of a disorder, and which could affect expert interpretation. Because of the omission of other clinically relevant information, the outcome of the assessments does not represent a formal clinical diagnosis in any way. Finally, the experts were asked to rate how certain they were of their assessment, on a 5-point Likert scale ranging from *very uncertain* (1) to *very certain* (5).

To assist in the assessment, we developed a web application where experts could login and rate hypnograms at their own pace. To ensure a diverse range of assessments, the order of the hypnograms was randomized, with the first 200 hypnograms being the same for all experts, and each expert subsequently receiving their own unique sequence. A subset of the HealthBed hypnograms was given priority in the sequence to increase the diversity of the assessed hypnograms, as it was not expected that a single physician would evaluate all available 1193 hypnograms. The distribution of hypnograms and the order of the assessment sequence is also illustrated in Figure 1. The main interface of the web app that was used to collect the assessments is shown in Figure 2. A meeting was held to introduce the study and the scoring application, where experts were informed that there were no correct or incorrect answers and asked to rely on their initial impressions of the hypnograms.

Analysis

During a pilot study with the experts, it was found that the first hypnograms were more difficult to assess since the task of interpreting hypnograms without additional clinical context was novel to them. Therefore, the first 50 responses were discarded for each of the raters. Using simple visualizations and statistics, we inspected the distribution of the collected evaluations and certainty scores across the experts. The distribution of evaluations was also compared across participants with a diagnosis of sleep disorder (SOMNIA) and healthy participants (HealthBed).

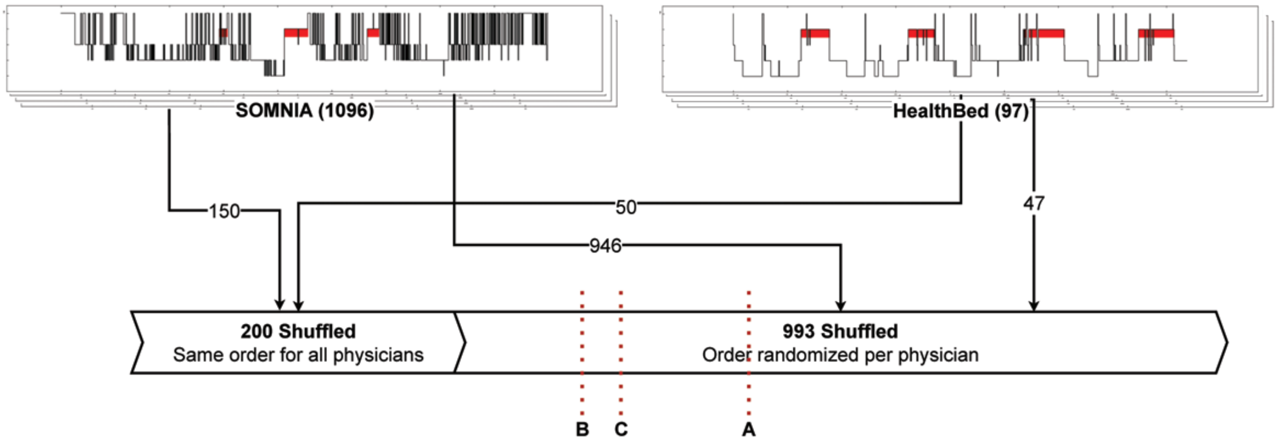


Figure 1. Hypnogram distribution for assessment. Distribution of SOMNIA and HealthBed hypnograms. The order in which they were presented in the web app was randomized with the first 200 hypnograms having a fixed order across experts. HealthBed hypnograms were prioritized in the sequence to ensure a more diverse set of hypnograms early on. Sleep experts assessed a variable amount of hypnograms as indicated by the dotted, red lines.

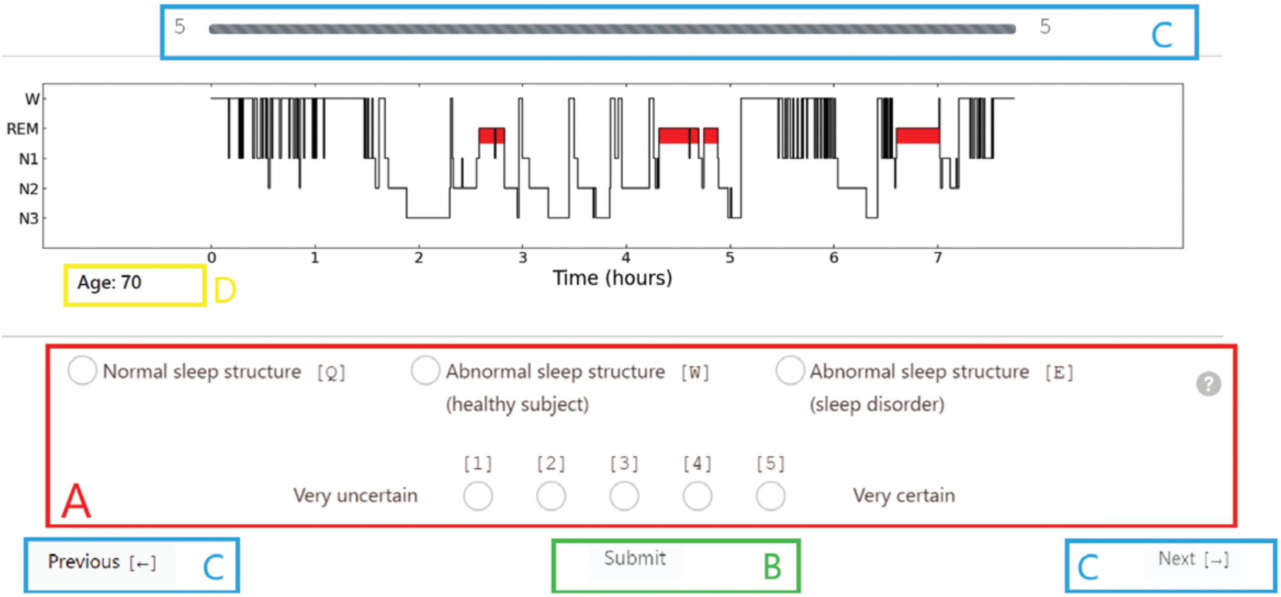


Figure 2. User interface for web-based hypnogram assessment. The hypnogram is shown in the top center of the screen. (A) The evaluation and certainty options can be selected below the hypnogram, keyboard shortcuts can be used to select the buttons (e.g. [Q] for normal sleep structure). (B) When evaluation and certainty are selected, the result can be submitted to continue to the next hypnogram. (C) Using the previous and next buttons the user can scroll back and forth between previously evaluated hypnograms. (D) Age is the only clinical information that is presented.

Agreement between experts was measured using Cohen's Kappa and Matthew's Correlation Coefficient (MCC).

To identify features of the hypnogram that drive interpretation by experts, a LR model and a convolutional NN (CNN) were developed to predict expert evaluation [10]. We selected LR because of its familiarity and interpretability. A more complex CNN model was selected to discover additional, more complex, interrelated, features without relying on manually engineered features.

Logistic regression.

The quantitative features shown in Table 1 were computed for each hypnogram and used as input for the model. The model was trained to predict 1 if the hypnogram was labeled by one of the experts as *normal* or *abnormal-healthy*, and 0 if the expert labeled the hypnogram as *abnormal-disordered*. Grouping of *normal* and *abnormal-healthy* was applied since the *normal* option was hardly used.

All unique hypnograms that were evaluated were split into a training (70%) and test set (30%). For hypnograms that were evaluated by multiple experts, we included all evaluations in the corresponding set (either training or test, but not both). It has been shown that training with one-hot encoded labels in this way results in learning the underlying label distribution [11]. Input features were standardized (zero mean, unit variance) to enable comparison of model coefficients across features. Standardization was applied based on the mean and variance of the unique hypnograms in the training set. To deal with the strong correlation between many of the features, we used Elastic Net, a type of regression that combines L1 and L2 regularization to provide variable selection, deal with multicollinearity, and prevent overfitting [12]. Optimal values for the alpha, which controls the trade-off between L1 and L2 regularization, and lambda hyperparameters were estimated using grid search in combination with 10-fold cross-validation over the set of training evaluations. Sample

Table 1. Hypnogram Features as Input for the Elastic Net Logistic Model

Feature	Abbreviation	Description
% Wake	%W	Percentage of W in hypnogram
% N1	%N1	Percentage of N1 in hypnogram
% N2	%N2	Percentage of N2 in hypnogram
% N3	%N3	Percentage of N3 in hypnogram
% REM	%REM	Percentage REM in hypnogram
Maximum wake	MaxW	Maximum consecutive duration of Wake.
Maximum N1	MaxN1	Maximum consecutive duration of N1 sleep
Maximum N2	MaxN2	Maximum consecutive duration of N2 sleep
Maximum N3	MaxN3	Maximum consecutive duration of N3 sleep
Maximum REM	MaxREM	Maximum consecutive duration of REM sleep
Sleep cycles	CCL*	Number of sleep cycles*
Sleep stage transition index	SSTi	Number of transitions per hour of recording
Awakening index	WKNi	Number of transitions to W per hour of recording
N1 transition index	N1Ti	Number of transitions to N1 per hour of recording
N2 transition index	N2Ti	Number of transitions to N2 per hour of recording
N3 transition index	N3Ti	Number of transitions to N3 per hour of recording
REM transition index	REMSSTi	Number of transitions to REM per hour of recording
REM awakenings index	REMWKNi	Number of transitions from REM to W per hour of recording
N3 awakenings	N3WKN	Number of transitions from N3 to W
Snooze time	ST	Number of minutes W at the end of the hypnogram
Sleep onset latency	SOL	Number of minutes from the start until the first non-W epoch
N3 onset latency	N3OL	Number of minutes from the start until first N3 epoch
REM onset latency	REMOL	Number of minutes from the start until first REM epoch
Wake after sleep onset	WASO	Number of minutes spent in W after sleep onset
Long awakenings	WKNL	Amount of Wake periods of at least 5 minutes
Time in bed	TIB	Total time of hypnogram.

*Sleep cycles were computed as the number of periods in the hypnogram ≥ tr minutes with at least pr % REM in this period, followed by a period ≥ tn minutes with a minimum of pn % NREM. The parameters were chosen as: tr = 10 minutes, tn = 30 minutes, and pr = pn = 55%.

weights were used during training to account for class imbalance and differences in the number of assessments per expert. Probabilities predicted by the model were binarized at a threshold of 0.5 to obtain class predictions. The model was implemented using Python and scikit-learn [13].

Model performance was evaluated by means of its accuracy on the test set. The confusion matrix was inspected to understand model performance across the classes. Due to the subjective nature of the target variable, we did not expect the classes to be perfectly separable. Therefore, the accuracy of the model was compared across certainty levels assigned by the expert. Here we assume that errors on low-certainty evaluations are less severe compared to high-certainty evaluations. Similarly, for the hypnograms that were evaluated by multiple experts, we inspected the model performance with respect to the agreement between experts. Pairwise agreement between models and experts was computed using Cohen's Kappa. All model evaluation methods were computed over the test set.

To evaluate which features drove inference, the model coefficients of the included features were evaluated. Standardization of features before fitting the model allowed us to compare the magnitude of the model coefficients across features. Therefore, features with positive model coefficients contribute to *normal* and *abnormal-healthy* evaluation, whereas features with negative model coefficients contribute to *abnormal-disordered* evaluation.

Convolutional NN.

In addition to the logistic regression model, we implemented a NN for the same task because of its ability to automatically learn suitable abstract representations from data. Because they are not limited to manually engineered features, they can be used to identify additional (more complex) patterns. More specifically, CNN was selected for its proven effectiveness in time-series classification [14]. All hypnograms were one-hot encoded as a binary matrix where each row represents one of the five stages and each column an epoch of the hypnogram. The one-hot encoded hypnograms were used as input to the CNN. The model was trained using the same target variable, train-test split, and sample weights as the LR model.

We considered CNN architectures with a small number of convolutional layers, followed by a global average pooling (GAP) layer that takes the average value for each channel across the temporal dimension. By using a GAP layer, we enable the use of class activation maps (CAM) to deal with the challenge of explainability in NNs [15] and allow a visual interpretation of the results, and features driving NN inference.

A suitable number of layers and nodes was selected by experimenting with several combinations and inspecting the results. The output of the model was obtained using a fully connected final layer with a sigmoid activation function to map the outcome to probabilities. The model was trained using binary cross-entropy loss. Implementation was performed using Python and Keras [16].

The same evaluation methods used for the LR model were also used for the CNN, including accuracy and confusion matrix. In addition, CAM was implemented as described by Zhou et al. [15], where a heatmap was plotted as background for a group of randomly sampled hypnograms from both classes to visually analyze and evaluate which hypnogram characteristics were used by the model. These characteristics were considered drivers of expert interpretation.

Results

The three experts assessed 612, 351, and 405 hypnograms, respectively. Only 1%–4% of the hypnograms were assessed as *normal* by the experts. Therefore, as mentioned in the methods section, it was decided to group the evaluations as *healthy* (including the assessments *normal [hypnogram]* and *abnormal [hypnogram, but considered] -healthy*) and *disordered* (*abnormal [hypnogram, considered]-disordered)*. Similarly, the lowest certainty score was only used 13 times in total, therefore, certainty was grouped as *low* (1–2), *medium* (3), and *high* (4–5).

Most of the hypnograms were evaluated as disordered (82%, 79%, and 72% for experts A, B, and C, respectively). Pairwise Cohen's Kappa agreements between the experts were 0.44 (A–B), 0.50 (A–C), and 0.38 (B–C). The values for MCC were 0.44 (A–B), 0.52 (A–C), and 0.38 (B–C). Figure 3 shows the certainty-level distribution across *healthy* and *disordered* assessments for each expert. The high-certainty bars (red) are larger than the low-certainty bars (pink) in case of disordered evaluation. The exact opposite holds for healthy evaluations. Thus, experts were highly certain of disordered evaluations but in general less certain about healthy evaluations. Note that the inter-rater agreement between the clinical experts themselves also serves as an upper bound on the performance of the learned models: in the best case, the models can reach a Cohen's kappa agreement of 0.38~0.50 with respect to the individual experts.

Table 2 displays the number of disordered and healthy evaluations across the SOMNIA and HealthBed sets per expert. The sleep-disordered patients from the SOMNIA dataset had a hypnogram that was evaluated as disordered in 85%, 87%, and 77% of the cases for experts A, B, and C, respectively. In contrast, the hypnograms of the healthy participants from the HealthBed cohort, were evaluated as disordered in 62%, 46%, and 45% of the cases.

Logistic regression

During hyperparameter optimization, the values for alpha and lambda were selected to be 0.7 and 0.01. The model achieved an overall 77% on the evaluations in the test set. The confusion matrix for the predictions on the test set is shown in Table 3. The model was 80% accurate on hypnograms that were evaluated as healthy by the expert and 76% on the disordered

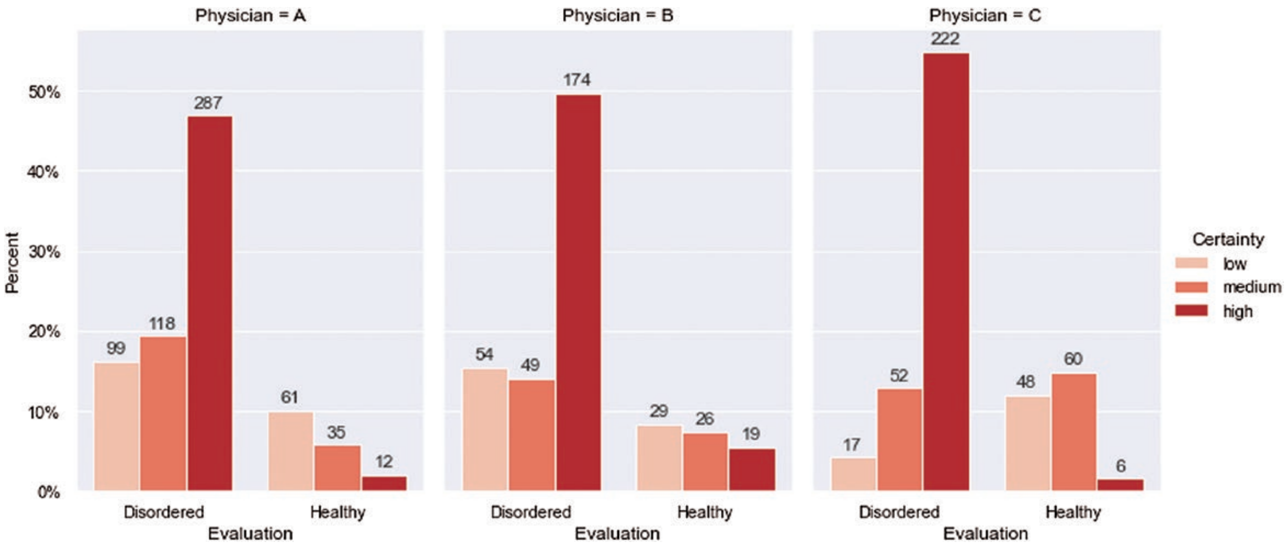


Figure 3. Certainty distribution. Certainty-level distribution across healthy and *disordered* assessments for each expert. For all experts, certainty was high for *disordered* evaluations and low for *healthy* evaluations. Each bar is annotated with the exact count.

Table 2. Number (N) and Percentage of *Disordered/Healthy* Evaluations for Each Expert Across Healthy (HealthBed) and Sleep-Disordered Adults (SOMNIA)

	HealthBed (*N* = 97)			SOMNIA (*N* = 1096)		
Expert	N	P (Disordered)	P (Healthy)	N	P (Disordered)	P (Healthy)
A	68	62%	38%	544	85%	15%
B	68	46%	54%	283	87%	13%
C	67	45%	55%	338	77%	23%

evaluations. Accuracy was higher on the high-certainty hypnograms, with 90% and 100% accuracy for disordered and healthy, respectively, as shown in Table 4. In contrast, accuracy was limited to 38% and 68% when experts indicated low certainty. The logistic model was 90% accurate on the subset of 164 hypnograms where the experts agreed; thereby ignoring contrasting examples in the test set where the same hypnogram was assessed differently by multiple experts. Pairwise Cohen's Kappa agreements of the model with each expert were between 0.47 and 0.52. MCC values were between 0.48 and 0.56. For both measures, this is in line with the inter-rater agreement between the experts themselves.

The 15 features included in the LR model after training (i.e. non-zero coefficients) are presented in Table 5. The intercept of the model was −0.98. The presence of REM, sleep cycles, and consecutive N2 were found to contribute the most to a hypnogram evaluation as *healthy*. On the other hand, a high transition index, presence of (long) W, awakenings, and presence of N1 were the strongest contributors to an evaluation as *disordered*.

Table 3. Confusion Matrix for the Logistic Model. Percentages are Computed Over the True Label

		Predicted	
		Disordered	**Healthy**
True	**Disordered**	256 (76%)	80 (24%)
	Healthy	18 (20%)	71 (80%)

Table 4. Accuracy of the Logistic Model at Each Certainty Level

True evaluation (certainty)	Predicted	
	Disordered	**Healthy**
Disordered (high)	**198 (90%)**	23 (10%)
Disordered (medium)	**37 (62%)**	23(38%)
Disordered (low)	**21 (38%)**	34 (62%)
Healthy (low)	14 (32%)	**30 (68%)**
Healthy (medium)	4 (12%)	**29 (88%)**
Healthy (high)	0 (0%)	**12 (100%)**

Convolutional NN

The CNN was trained using the same labels, train-test split, and sample weights as the logistic model. The final CNN architecture consisted of three convolutional layers, a GAP layer, and a fully connected layer.

The overall accuracy of the model on the test set was 75%, on the disordered and healthy groups, this was 73% and 83%, respectively. On high-certainty hypnograms, the model was 86% and 100% accurate for disordered and healthy, respectively. Accuracy on healthy hypnograms evaluated with medium-certainty was 94%, in contrast, the performance on other low- and medium-certainty evaluations ranged between 35% and 70%. On the subset of 164 hypnograms in the test set where experts agreed, the model correctly predicted the evaluation in 84% of the cases. Pairwise Cohen's Kappa agreements of the CNN model with each expert were between 0.48 and 0.50, which is on par with the LR model and the agreement across clinical experts themselves. The agreement between the CNN and LR model for all hypnograms in the test set was 0.86; all pairwise agreements between model(s) and experts are shown in Table 6. The exact model architecture and detailed results (including confusion matrices) are provided in Supplementary Table S1–S3.

The CAM visualizations for four randomly sampled hypnograms are shown in Figure 4 and provide insight into the characteristics that most contribute to the model predictions. The CNN distinguished between the continuous presence of a stage and transitions. From the heatmaps, it can be observed that the presence of N2, N3, and REM (i.e. percentage of hypnogram in each stage) contribute to a prediction of *healthy*, and W, N1, and fragmentation to a prediction of *disordered*.

Table 5. Input Features Included in the Elastic Net Logistic Regression and Associated Coefficients. The Features are Sorted by Model Coefficient From Positive (*Healthy*) to Negative (*Disordered*)

Feature	Abbreviation	Description	Coefficient
% REM	%REM	Percentage REM in hypnogram	0.25
Sleep cycles	CCL	Number of sleep cycles	0.22
Maximum N2	MaxN2	Maximum consecutive duration of N2 sleep	0.20
N3 Transition index	N3Ti	Number of transitions to N3 per hour of recording	0.19
Maximum N3	MaxN3	Maximum consecutive duration of N3 sleep	0.18
Snooze time	ST	Number of minutes W at the end of the hypnogram	0.18
REM awakenings index	REMWKNi	Number of transitions from REM to W per hour of recording	0.02
Time in bed	TIB	Total time of hypnogram.	−0.04
REM transition index	REMTi	Number of transitions to REM per hour of recording	−0.15
N3 awakenings	N3WKN	Number of transitions from N3 to W	−0.25
% N1	%N1	Percentage of N1 in hypnogram	−0.33
Awakening index	WKNi	Number of transitions to W per hour of recording	−0.36
Long awakenings	WKNL	Amount of Wake periods of at least 5 minutes	−0.50
Maximum wake	MaxW	Maximum consecutive duration of wake.	−0.73
Sleep stage transition index	SSTi	Number of transitions per hour of recording	−0.78

Table 6. Pairwise Cohen's Kappa and MCC Between Experts, Logistic Regression (LR), and CNN

	Expert A	Expert B	Expert C	LR	CNN
Expert A	—	0.44 (0.44)	0.50 (0.52)	0.50 (0.56)	0.49 (0.56)
Expert B		—	0.38 (0.38)	0.47 (0.48)	0.50 (0.52)
Expert C			—	0.52 (0.53)	0.48 (0.51)
LR				—	0.86 (0.86)
CNN					—

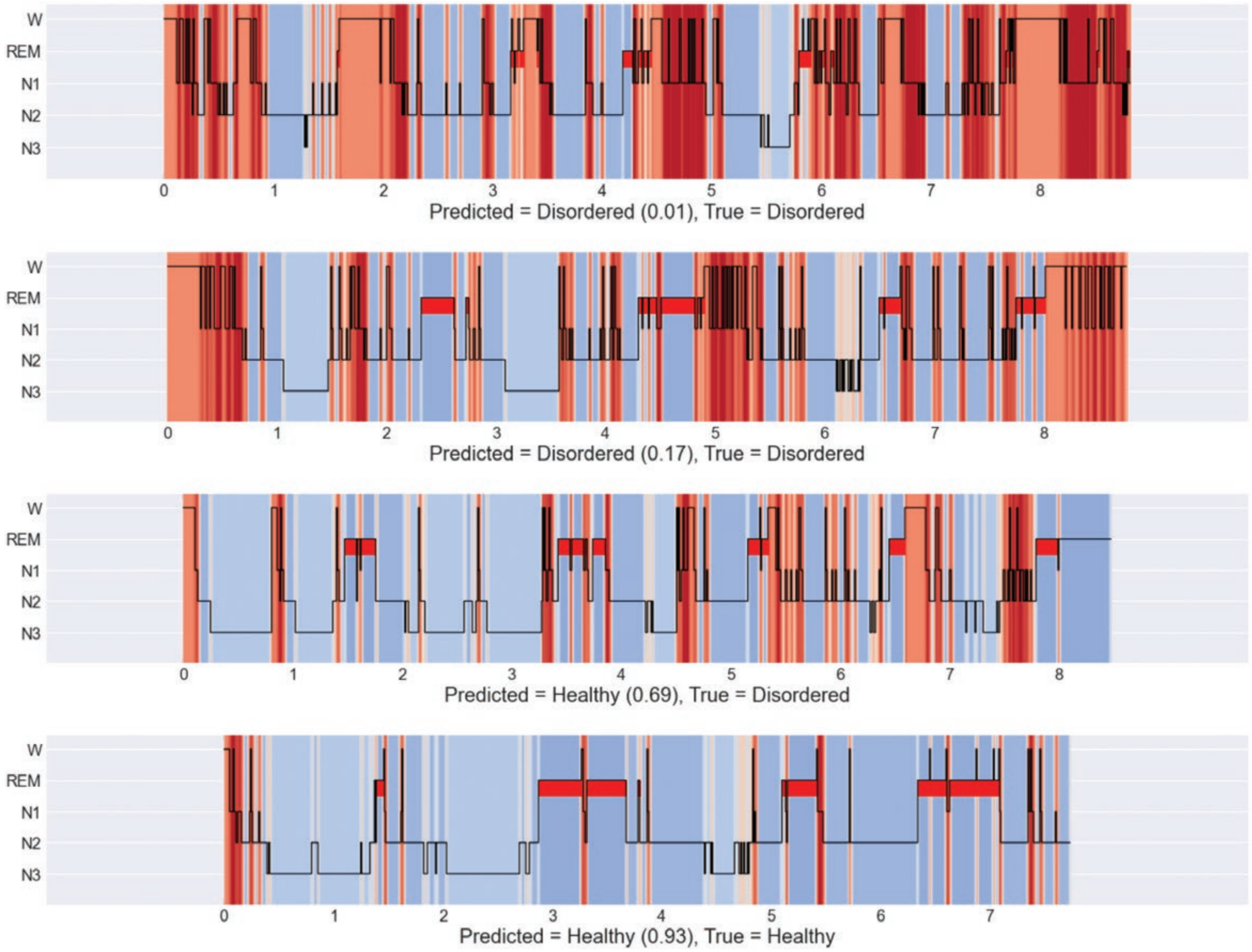


Figure 4. Class activation maps for hypnograms. Four randomly sampled hypnograms and their CAM visualizations. The heatmap illustrates how the CNN looks at the hypnogram; presence of N2, REM, and N3 contribute towards healthy evaluation prediction whereas W, N1, and fragmentation are associated with *disordered* evaluation by the model.

Discussion

The hypnogram remains one of the most encompassing representations of sleep structure and has important clinical use. However, the interpretation of the hypnogram relies on visual pattern recognition developed by physicians as clinical intuition, while quantitative parameters that are traditionally extracted may not fully convey all the information contained. Here, we took an unconventional approach by "forcing" experts to assess isolated hypnograms and used their expert assessment to train classification models to predict these assessments, and subsequently look for salient aspects driving this interpretation.

Both the LR, and the CNN models performed similarly well with agreement versus expert human scorers comparable to the agreement between the expert scorers themselves. Importantly, using two completely different approaches to hypnogram feature extraction, both models agreed on which features contributed to the (subjective) interpretation of the hypnogram as *healthy*, or *disordered*. Overall, the presence of REM, N2, sleep cycles, and N3 were associated with a normal sleep structure by experts. In contrast, wake, N1, and overall sleep fragmentation (i.e. a high amount of sleep stage changes per hour) contributed to an evaluation of an abnormal hypnogram. Earlier work by Amouh [3] also found that sleep stage distribution is a predictor for expert interpretation; however, our work also emphasizes the importance of other features such as sleep fragmentation and sleep cycles.

The agreement between experts on the task is moderate (with a Kappa between 0.38 and 0.50), which illustrates the subjectivity of the labels and the difficulty of the task, especially for the more ambiguous, low-certainty cases. Interestingly, the agreement between the models and experts is similar (between 0.45 and 0.52), which means that the model was able to adequately mimic the experts' choices. Although this level of agreement might seem limited for a machine learning model, we cannot expect it to achieve higher agreement than that of experts amongst themselves. The high accuracy of the model in the high-certainty cases and in cases where all physicians agreed further suggests that the model adequately learned the hypnogram characteristics that physicians use for interpretation, taking into account that there is a significant group of ambiguous cases where the hypnogram is difficult to assess (also by experts) with such a limited context.

Our findings have several implications. The first relates to the performance evaluation of surrogate sleep measurement techniques. Due to the high cost of polysomnography, wearable sleep trackers have drawn the attention of sleep experts. By recommendation of the AASM, further validation is necessary before wearable sleep trackers can be used in clinical practice [17]. However, sleep staging methods are traditionally optimized for epoch-by-epoch performance which does not guarantee that clinically relevant features are correctly captured. It is not uncommon to report on overnight statistics (e.g. [18–21].), but in fact, these statistics often demonstrate the limitation of optimizing for epoch-by-epoch measures such as accuracy or agreement. A typical limitation is that performance with respect to reference PSG is poor for fragmented sleep [18–21]. Our results emphasize the importance of sleep fragmentation in the interpretation of hypnograms. As such, a measuring technique that achieves reasonable epoch-per-epoch agreement with PSG, but presents a "smoothed" picture of sleep architecture because it missed sleep stage transitions and sleep fragmentation might lead to misinterpretation. Despite recent approaches to standardize validation of sleep trackers (e.g. [22, 23].) there is no consensus yet on how to report on overnight statistics. Based on our findings we recommend to include sleep stage distribution and sleep fragmentation in evaluation of automated sleep staging methods.

Besides clinical applications, wearable sleep trackers have also become increasingly popular among consumers, presenting laypeople with hypnograms [24]. However, we show that a hypnogram is difficult to interpret, requires clinical experience and domain knowledge, and is ambiguous especially when isolated from other clinical parameters. Consequently, and regardless of how reliably a consumer sleep tracker performed sleep staging, the hypnograms presented by these trackers might drive wrong interpretations by the consumer. Misinterpretation in combination with overreliance on sleep tracker data can have negative side effects, possibly leading to further disordered sleep, a condition recently coined "orthosomnia" [25]. We here argue for the derivation of features that are clinically relevant, this however, comes with the caveat that the sleep staging of these wearables must be done reliably. Due to their reliance on surrogate measures of sleep, these trackers usually have a larger degree of uncertainty about the correct output. In an ideal world, the sleep tracker would display not only interpretable and relevant measures of sleep to the user, but also the degree of certainty that the tracker has about its beliefs. This way, the user may be reassured when e.g. the tracker reports bad sleep at very low certainty.

Our results also cast a new light on the issue of low inter-rater agreement between human sleep stage scoring. Inter-rater agreement is typically calculated on an epoch-by-epoch basis comparing the expert interpretation, by trained scorers, of the same recordings. The AASM inter-scorer reliability program, which was developed to aid sleep centers in achieving accreditation standards, showed an overall agreement of around 82.6% [2]. However, by only calculating agreement on an epoch-by-epoch basis, all forms of disagreement are counted equally. Our research shows that some forms of disagreement might be more important than others, as these could make a difference in the diagnostic process. For example, sleep stage fragmentation was found to be the largest contributor to an evaluation of a hypnogram as *disordered*. Programs such as the inter-scorer reliability, or the AI/Autoscoring Pilot Certification program announced by AASM on their website, could be expanded to promote inter-scorer, or autoscoring agreement not only in terms of overall sleep stages, but also to increase agreement on hypnogram-derived metrics that might have clinical relevance, such as measures of fragmentation, etc.

In our study, we found that experts were systematically more certain about abnormal sleep structure evaluation, whereas normal, healthy evaluations were often associated with lower certainty. Moreover, based on the hypnogram alone, a sleep disorder was "suspected" in roughly half of the cases for healthy adults who did not have any sleep disorders or other comorbidities. These findings may be explained by first-night effects that impact sleep during the first night of PSG and therefore cause abnormal sleep structure in healthy participants [6]. In addition, it may well be that the notion clinicians have of a normal sleep structure is mostly based on textbook examples and infrequently encountered in their practice. The link between subjective sleep quality and objective measures derived from the hypnogram is not well established and remains an area of interest for further research [26]. Finally, experts in our study may have an inherited bias towards abnormal sleep, as they frequently encounter patients diagnosed with a sleep disorder but rarely see participants with normal sleep. In any case, the ability to identify healthy participants from an isolated hypnogram is limited. This underscores the well-known fact that hypnograms need to be interpreted in a clinical context. Although seemingly obvious, it is important to emphasize this point, since the increasing popularity of wearable sleep trackers will inevitably lead to an abundance of hypnograms with very little, or no additional context.

Regarding the limitations of this study, it considered hypnograms in an artificial research setting where the hypnograms were reviewed without any additional clinical information. By doing so, we aimed to emphasize the structure of the hypnogram itself. Nevertheless, this could be a limiting factor as it might lead to an overemphasis of the more salient features in the hypnogram; when additional information is presented, more specific and localized patterns might be more meaningful and stronger drivers for interpretation. Identifying such features is a topic for future research. This is relevant as disagreement between sleep technicians is known to be lower for certain sleep stages/types of fragmentation, for example, N1 [2].

Another potential limitation of this work is that all hypnograms were scored by technicians from one institution. However, there can be quite some inter-rater disagreement between technicians, which is even higher when these technicians come from different institutions [27]. The impact of institution and inter-rater disagreement on the evaluation of the hypnogram by experts is also left to future work.

Conclusions

Sleep stage distribution, sleep cycles, and fragmentation are strong predictors of how an expert interprets a hypnogram. By comparing hypnograms not solely on an epoch-by-epoch basis, but also on these more specific features that drive the evaluation by experts, performance assessment of (automatic) sleep-staging and surrogate sleep trackers may be improved and their relevance increased. In particular, sleep fragmentation is a feature that deserves more attention as it is often not included in the PSG report, and existing (wearable) sleep trackers have shown relatively poor performance in capturing this aspect well. Finally, our work again emphasizes the limited ability to identify sleep disorders from an isolated hypnogram. Future work could explore how the mentioned salient hypnogram features could be used to further improve inter-rater programs, validate sleep trackers for clinical practice, and support novices in the field of sleep medicine in interpreting hypnograms.

Supplementary Material

Supplementary material is available at *SLEEP* online.

Funding

This work has been done in the IMPULS framework of the Eindhoven MedTech Innovation Center (e/MTIC, incorporating Eindhoven University of Technology, Philips Research, Sleep Medicine Center Kempenhaeghe). This activity is in part funded by the PPS program research and innovation of the Dutch Ministry of Economic Affairs and Climate. The funders had no role in the study design, decision to publish, or preparation of the manuscript.

Acknowledgments

SO, PF, SvdE, and CvdW designed the study. CvdW collected the data and conducted the analysis. SO, PF, and SvdE supervised the study. SO, SD, and MS assessed the hypnograms. FvM managed datasets and HGC provided technical support. All authors provided critical reviews of the manuscript.

Disclosure Statement

Financial Disclosure: At the time of writing and/or while conducting the research CvdW, HvG, and PF were employed and/or affiliated with Royal Philips, a commercial company and manufacturer of consumer and medical electronic devices, commercializing products in the area of sleep diagnostics and sleep therapy. Philips had no role in the study design, decision to publish, or preparation of the manuscript. SO received an unrestricted research grant from UCB Pharma and participated in advisory boards for UCB Pharma, Jazz Pharmaceuticals, Takeda, and Bioproject, all paid to the institution and all unrelated to the present work. *Nonfinancial Disclosure*: None.

Data Availability

The SOMNIA and HealthBed data used in this study are available from the Sleep Medicine Center Kempenhaeghe upon reasonable request. The data can be requested by presenting a scientific research question and by fulfilling all the regulations concerning the sharing of the human data. The details of the agreement will depend on the purpose of the data request and the entity that is requesting the data (e.g. research institute or corporate). Each request will be evaluated by the Kempenhaeghe Research Board and, depending on the request, approval from independent medical ethical committee might be required. Access to data from outside the European Union will further depend on the expected duration of the activity; due to the work required from a regulatory point of view, the data is less suitable for activities that are time-critical, or require access on short notice. Specific restrictions apply to the availability of the data collected with sensors not comprised in the standard PSG set-up, since these sensors are used under license and are not publicly available. These data may however be available from the authors with permission of the licensors. For inquiries regarding availability, please contact Merel van Gilst (M.M.v.Gilst@tue.nl).

References

1. Berry, RB, *et al. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. 2.6th ed. Westchester, IL: American Academy of Sleep Medicine; 2018.
2. Rosenberg R, *et al*. The American Academy of Sleep Medicine inter-scorer reliability program: Sleep stage scoring. *J Clin Sleep Med*. 2013;**9**(1):81–87.
3. Amouh T. *Analysis of Tabular Non-Standard Data With Decision Trees, and Application to Hypnogram-Based Detection of Sleep Profile*. Diss. Catholic University of Louvain: Diss. Catholic University of Louvain. 2011.
4. Swihart BJ, Caffo B, Bandeen-Roche K, Punjabi NM. Characterizing sleep structure using the hypnogram. *J Clin Sleep Med*. 2008;**04**(04):349–355. doi: 10.5664/jcsm.27236
5. Chaparro-Vargas R, Ahmed B, Wessel N, Penzel T, Cvetkovic D. Insomnia characterization: From hypnogram to graph spectral theory. *IEEE Trans Biomed Eng*. 2016;**63**(10):2211–2219. doi: 10.1109/TBME.2016.2515261
6. Ding L, Chen B, Dai Y, Li Y. A meta-analysis of the first-night effect in healthy individuals for the full age spectrum. *Sleep Med*. 2022;**89**:159–165. doi: 10.1016/j.sleep.2021.12.007
7. van Gilst MM, van Dijk JP, Krijn R, *et al*. Protocol of the somnia project: An observational study to create a neurophysiological database for Advanced Clinical Sleep Monitoring. *BMJ Open*. 2019;**9**(11):1–9. doi: 10.1136/ bmjopen-2019-030996
8. Berry, RB, *et al. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. 2.5th ed. Westchester, IL: American Academy of Sleep Medicine; 2018.
9. van Meulen FB, Grassi A, van den Heuvel L, *et al*. Contactless camera-based sleep staging: The healthbed study. *Bioengineering (Basel)*. 2023;**10**(1):109. doi: 10.3390/bioengineering10010109
10. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;**521**(7553):436–444. doi: 10.1038/nature14539
11. Huijben I, Hermans LW, Rossi AC, Overeem S, Gilst MM, van Sloun RJ. Interpretation and further development of the hypnodensity representation of sleep structure. *Physiological Measurement*. 2023;**44**(1):015002. doi: 10.1088/1361-6579/aca641
12. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;**67**(2):768–768. doi: 10.1111/j.1467-9868.2005.00527.x
13. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine learning in python. *J Mach Learn Res*. 2011;**12**:2825–2830.
14. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: A review. *Data Min Knowl Discov*. 2019;**33**(4):917–963. doi: 10.1007/s10618-019-00619-1

15. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. *Learning Deep Features for Discriminative Localization*. In proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV.; 2016; 2921–2929.
16. Chollet F, Others. Keras. 2015. Accessed August 1, 2022. https://github.com/fchollet/keras
17. Khosla S, Deak MC, Gault D, *et al.*; American Academy of Sleep Medicine Board of Directors. Consumer sleep technology: An American Academy of sleep medicine position statement. *J Clin Sleep Med*. 2018;**14**(05):877–880. doi: 10.5664/jcsm.7128
18. Wulterkens BM, Fonseca P, Hermans LW, *et al.* It is all in the wrist: Wearable sleep staging in a clinical population versus reference polysomnography. *Nat Sci Sleep*. 2021;**13**:885–897. doi: 10.2147/nss.s306808
19. Chinoy ED, Cuellar JA, Huwa KE, *et al.* Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep.* 2020;**44**(5). doi: 10.1093/sleep/zsaa291
20. Sridhar N, Shoeb A, Stephens P, *et al.* Deep learning for automated sleep staging using instantaneous heart rate. *npj Digital Med*. 2020;**3**(1):106. doi: 10.1038/s41746-020-0291-x
21. Kotzen K, Charlton PH, Salabi S, Amar L, Landesberg A, Behar JA. SleepPPG-Net: A deep learning algorithm for robust sleep staging from continuous photoplethysmography. *IEEE J Biomed Health Inf*. 2022;**27**(2):924–932. doi: 10.1109/jbhi.2022.3225363
22. Menghini L, Cellini N, Goldstone A, Baker FC, de Zambotti M. A standardized framework for testing the performance of sleep-tracking technology: Step-by-step guidelines and open-source code. *Sleep.* 2021;**44**(2). doi: 10.1093/sleep/zsaa170
23. Nguyen QN, Le T, Huynh QB, Setty A, Vo TV, Le TQ. Validation framework for sleep stage scoring in wearable sleep trackers and monitors with polysomnography ground truth. *Clocks Sleep*. 2021;**3**(2):274–288.
24. De Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc*. 2019;**51**(7):1538–1557. doi: 10.1249/MSS.0000000000001947
25. Baron KG, Abbott S, Jao N, Manalo N, Mullen R. Orthosomnia: Are some patients taking the Quantified Self Too Far? *J Clin Sleep Med*. 2017;**13**(02):351–354.
26. Kaplan KA, Hirshman J, Hernandez B, *et al.*; Osteoporotic Fractures in Men (MrOS), Study of Osteoporotic Fractures SOF Research Groups. When a gold standard isn't so golden: Lack of prediction of subjective sleep quality from sleep polysomnography. *Biol Psychol*. 2017;**123**:37–46. doi: 10.1016/j.biopsycho.2016.11.010
27. Bakker JP, Ross M, Cerny A, *et al.* Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: Hypnodensity based on multiple expert scorers and auto-scoring. *Sleep.* 2022;**46**(2). doi: 10.1093/sleep/zsac154