

Attention on Sleep Stage Specific Characteristics

Citation for published version (APA):

Huijben, I. A. M., Overeem, S., Van Gilst, M. M., & Van Sloun, R. J. G. (2024). Attention on Sleep Stage Specific Characteristics. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2024* Article 10782554 Institute of Electrical and Electronics Engineers.
<https://doi.org/10.1109/EMBC53108.2024.10782554>

Document license:

TAVERNE

DOI:

[10.1109/EMBC53108.2024.10782554](https://doi.org/10.1109/EMBC53108.2024.10782554)

Document status and date:

Published: 17/12/2024

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Attention on Sleep Stage Specific Characteristics

Iris A.M. Huijben^{1,2}, *Student Member, IEEE*, Sebastiaan Overeem^{1,3},
Merel M. van Gilst^{1,3}, and Ruud J.G. van Sloun¹, *Member, IEEE*

Abstract—Manual sleep stage classification relies on visual inspection of 30-second windows comprising multi-sensor measurements. The ability of neural networks to model complex relations has made them a popular, faster, alternative. However, it often remains unclear which parts of the data predominantly contributed to the model’s decision. This is especially ambiguous in sleep staging, where the coarse labeling per 30-second windows may assign mixtures of class-specific features to a single class. To boost the transparency of deep neural classifiers, we propose a dynamic discrete attention module that actively selects the subset of the input space aligned with the class label. The module can be combined with a typical classification network, and may additionally serve as a data-driven tool to discover sleep stage specific features in polysomnography data. We validate the method on synthetic and patient data. We observe that only a small subset of data from the 30-second window is required to retain accurate classification, and that the attention mechanism boosts performance. Analysis of the dynamic attention masks, moreover, shows clear sleep stage adaptive channel selection.

Index Terms—Sleep Staging, Subset selection, Attention, Gumbel-Softmax

I. INTRODUCTION

A classification task requires extraction of useful information from an abundant amount of data, where useful information is defined as the information needed to differentiate some classes. In sleep staging, a sleep expert is asked to pay attention to specific visual features in a 30-second window of a polysomnography (PSG) measurement, the clinical multi-sensor standard to measure a patient’s sleep. For example, only the presence of a K-complex in one of the electroencephalography (EEG) channels demands a scoring of a window as Non-Rapid-Eye-Movement 2 (N2) sleep. The recent popularity of neural networks that can quickly perform automated sleep staging on a new recording raises the question whether such models base their decisions on similar features. Is the full data window used, or are these models also able to distinguish classes on fractions of the data, as sleep experts do?

A multitude of *post-hoc* explanation methods for neural networks has been proposed [1], of which some found their way to sleep analyses as well [2], [3]. Instead, in this work, we focus on identification – *during training* – of a subset of data that is sufficient for accurate classification. This serves two purposes. First, it gives insights into which data segments (in time and across sensors) were (solely) paid attention to by

the model to make its decision. This is especially useful in sleep staging where ambiguous windows contain features of different sleep stages [4], [5]. Second, analyzing the selected subset of data may lead to discovery of new patterns that belong to different sleep stages and/or sleep disorders.

Closest to our approach is the work by Phan *et al.* [6] who use the inherent attention weights of a transformer-based sleep staging model to indicate the importance of different points in time of single-channel measurements. The interpretation of such attention weights as ‘data importance’ masks has, however, been challenged, as the attention mask itself also adds information to the classification model [7]. We differentiate from [6], as we propose discrete selection (taking place both over sensors and time) via hard – instead of soft – attention, through a neural agent that executes the discrete selection policy by conditioning on the full data frame. Also, our module does not require a specific transformer-based architecture for the classification model, and prevents leakage of mask information to the classifier by only feeding it with the selected subset of data. As opposed to learning a *static* EEG channel selection mask [8], [9], our method is *dynamic*, i.e. subset selection follows an active policy that is dependent on the incoming data.

II. METHOD

A. Data

Synthetic data: To investigate our model in a controlled setting, we created a (very) simplified representation of PSG data by generating 200 ‘sleep cycles’ of 90 minutes, measured by two ‘channels’. We defined three classes (A–C), each having its own characteristic feature: A) Artificial spindles in the first channel B) Delta frequency in the first channel C) Artificial eye movements in the second channel¹. Additive Gaussian noise $\sim \mathcal{N}(0, 0.1^2)$ was added, and all signals were sampled at $f_s = 100$ Hz. The 200 traces were randomly divided into a training set, validation set and hold-out test set (100/50/50).

Polysomnography data: We used a dataset of nocturnal PSG recordings from the Healthbed study [10]. The study protocol (W17.128) was approved by the medical ethics committee of Maxima Medical Center, Veldhoven, the Netherlands, and the data analysis protocol (CSG_2019_007_00) was

This work was supported by Onera Health, and the project ‘OP-SLEEP’. The project ‘OP-SLEEP’ is made possible by the European Regional Development Fund, in the context of OPZuid. Correspondence: i.a.m.huijben@tue.nl.

¹Department of Electrical Engineering, Eindhoven University of Technology, the Netherlands.

²Onera Health, Eindhoven, the Netherlands.

³Sleep Medicine Center Kempenhaeghe, Heeze, the Netherlands.

¹A spindle was generated as the real part of a Morlet wavelet with a central frequency $\sim u[11, 16]$ Hz, a duration $\sim u[1, 4]$ s, an amplitude $\sim u[0.8, 1.0]$ s, and a sign $\in \{-1, +1\}$. With $u[a, b]$ being a realization of a uniform random variable between a and b . Delta waves followed: $u(0.5, 1.0) \cdot \sin\{2\pi f + 2\pi u[0, 1]\}$, with $f \sim u[0.5, 4]$ Hz. Eye movements were generated using a skewed saw-tooth signal with a central frequency $f \sim u[0.5, 2.0]$, and an amplitude $\sim u[0.4, 0.5]$.

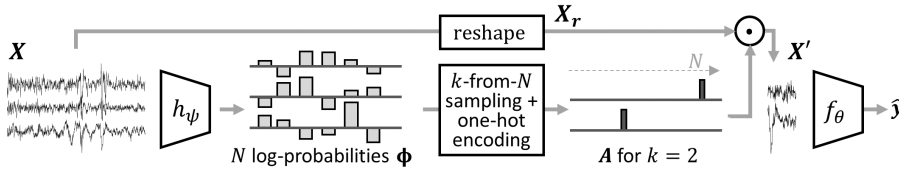


Fig. 1: Hyper-network h_ψ predicts N logits Φ to select k mini-windows from \mathbf{X} . Classifier f_θ receives these mini-windows and makes a prediction.

approved by the medical ethics committee of Sleep Medicine Center Kempenhaeghe. The dataset includes clinical video-PSG recordings and corresponding sleep stage labels of 96 healthy subjects, acquired following the clinical guidelines [4]. We selected EEG (F4/F3, C4/C3, O2/O1), two chin EMG and two EOG channels. Given the redundancy between odd and even EEG and EMG channels², we trained the model such that it requires the two EOG channels, and either the odd or even subsets of EEG and chin EMG data. Channels were filtered with a zero-phase 5th order Butterworth band-pass filter (10-49 Hz for chin EMG, and 0.3-49 Hz for the rest), and an additional band-stop filter (49-51 Hz) for better powerline suppression. Data was down-sampled to $f_s = 128$ Hz, and normalized within-patient and per channel, such that 95% of the samples were mapped between -1 and +1. The 96 subjects were randomly divided into a training, validation, and hold-out test set (75/10/11). Odd and even recordings of each subject were assigned to the same set.

B. Model design

We define a data window $\mathbf{X} \in \mathbb{R}^{ch \times T}$ with ch channels of length T , and a subset containing k single-channel mini-windows of length $T' (< T)$ as $\mathbf{X}' \in \mathbb{R}^{k \times T'}$. We set $T = 30 \times f_s$, and $T' = 5 \times f_s$. A classification model $f_\theta(\mathbf{X}')$, parameterized by θ , returns $\hat{\mathbf{y}} \in \{\mathbb{R}_{\geq 0}^C : |\hat{\mathbf{y}}| = 1\}$, with the predicted class being $\hat{y}^* := \operatorname{argmax}_C(\hat{\mathbf{y}})$.

To acquire \mathbf{X}' , k mini-windows must be selected from $N = ch \frac{T}{T'}$ mini-windows in \mathbf{X} . To this end we introduce selection matrix $\mathbf{A} \in \{0, 1 : \sum_j \mathbf{A}_{i,j} = 1 \forall i\}^{k \times N}$, and apply it on reshaped data window $\mathbf{X}_r \in \mathbb{R}^{N \times T'}$: $\mathbf{X}' = \mathbf{A}\mathbf{X}_r$.

To model \mathbf{A} , we follow the Deep Probabilistic Sampling (DPS) framework [12], which reparameterizes \mathbf{A} as a realization of k samples without replacement from a categorical distribution with probabilities $\pi \propto \exp \Phi$, with Φ being unnormalized log-probabilities, dubbed ‘logits’. To make \mathbf{A} dynamic, i.e. dependent on \mathbf{X} , we use a variant of Active-DPS [13] for an active selection policy, by conditioning Φ on \mathbf{X} through $\Phi := h_\psi(\mathbf{X}) \in \mathbb{R}^N$, with h_ψ a hyper-network with trainable parameters ψ , see fig. 1. We leverage Gumbel-top- k sampling [14], [15] to acquire \mathbf{A} from Φ . It perturbs the unnormalized logits with i.i.d. Gumbel noise samples $e \sim \text{Gumbel}(0, 1) \in \mathbb{R}^N$. The k highest perturbed logits yield a sample of k draws without replacement according to probabilities π [15]–[17].

Since mini-window selection is discrete, it hampers gradient backpropagation of the classification loss to train the parameters in ψ . A gradient estimator is, therefore, required. The authors of [15] showed that iterative sampling without

replacement from the Gumbel-Softmax (GS) distribution [18], [19] is a valid relaxation for top- k sampling from a categorical distribution. A single (soft) GS sample takes the following form: $\operatorname{softmax}_\tau(\Phi + e) \in \mathbb{R}_{\geq 0}^N$, with temperature parameter τ . For $\tau \rightarrow 0^+$, the soft sample converts into a hard sample from the categorical distribution. Even though this GS estimator is strictly only needed for gradient back-propagation, using the soft GS samples in the forward pass as well, improved optimization. However, since the classifier should learn to deal with hard selections, we annealed τ from 10 to 0.1 (0.5 for the synthetic data) using exponential decay in the first half of training, after which its value remained constant. During evaluation we select the k mini-windows with the top- k highest logits. More details on iterative application of soft sampling without replacement can be found in [20].

Both hyper-network h_ψ and classifier f_θ followed a typical convolutional neural network structure with four 1D convolutional layers. The first three convolutions were followed by: LeakyReLU \rightarrow Pooling(5) \rightarrow dropout(0.1). In the hyper-network, the last convolutional layer preceded: Tanh \rightarrow Flatten \rightarrow Linear(N) \rightarrow LeakyReLU \rightarrow Linear(N), and poolings were average poolings. In the classifier, the first convolutional layer was adapted such that each mini-window that originated from the same channel was filtered with the same channel-dependent set of convolutional kernels. This removed dependency on the order of mini-window selection, and it made the number of classifier parameters independent of k . Moreover, the last convolutional layer was followed by AdaptiveAvgPool1D(1) \rightarrow Linear(C) \rightarrow LogSoftmax. Lastly, the first three poolings were max pooling operations³. We trained the model by minimizing the cross-entropy between predicted class \hat{y}^* and label y . Training was done in batches of 128 windows, and took maximally 1000 iterations, with one iteration being a push-through of all training windows. On synthetic data we used 500 iterations. The best model was selected based on the lowest validation loss. We used the Adam optimizer [21], with a base learning rate of 1e-4, and – in the second half of training – reduced this rate a factor 10 every time the validation loss did not improve for at least 50 iterations (25 for the synthetic case). All models were run with three different random seeds.

C. Performance metrics

For synthetic data, we have access to the exact location of class-specific features. As such, we analyze the *Precision* with

³The output channels in the convolutional layers of both the hyper-network and classifier were set to: [16, 32, 64, 128], and kernel sizes were: [15, 9, 5, 3]. For synthetic data, model capacity was reduced by halving the number of output channels in both networks. Also the third pooling layer used a factor 3, instead of 5, to deal with the lower sampling frequency. The model implementation is provided at <https://github.com/IamHuijben/MiniWindowSelection.git>.

²EEG recordings of the left and right hemispheres are denoted with odd, respectively, even numbers in the 10-20 electrode positioning [11].

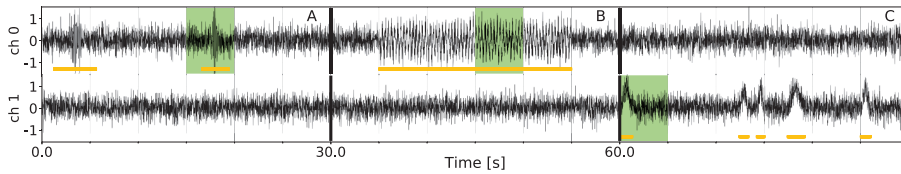


Fig. 2: Three windows from the synthetic test set: $k = 1$ mini-window per window was selected (denoted in green) which coincides with the presence of characteristic features, denoted in yellow.

TABLE I: Synthetic data: Precision of features in selected mini-windows ($k = 1$) per class and class-average F1 scores, for three different seeds (left, middle, right), either with random mini-window selection or our active selection policy.

Policy	Precision A	Precision B	Precision C	Avg. F1 score
Random	24.2 27.0 26.2	41.7 42.3 42.9	21.8 24.3 23.6	58.9 59.7 58.7
Active	74.7 71.6 49.9	99.1 95.5 0.00	99.1 99.3 99.0	95.8 94.6 95.8

which such features are selected, defined as the percentage of selected mini-windows from windows with class label $c \in \{A, B, C\}$, that contained a feature of class c .

For PSG data, such feature localization labels are unavailable. We, therefore, analyze the classification performance between the ground-truth sleep stage label and the predicted class label on the test set, expressed in *F1 score*, as a function of the number of selected mini-windows. We stress that achieving the highest possible sleep staging performance is not the goal of this work. We use the F1 score as a proxy for information loss with respect to the classes, when using only k mini-windows vs using the full window. We report means and 1 standard deviation (SD) across test set patients, randomized seeds and odd vs even measurements. We compensate the SD for the three seeds and the double measurement per patient.

III. RESULTS

A. Synthetic data

Table I shows the feature precisions per class and class-average F1 scores for $k = 1$. The F1 scores are near-perfect when using the active mini-window selection policy (5.2% of the windows in the generated test dataset contained no characteristic features), while random selection heavily reduced these scores. For active selection, the feature precision scores are much higher than in the random case, for at least 2 of the 3 classes. The fact that a third class can have low precision, while maintaining a high F1 score shows redundancy that is present in the classification problem. Our model, trained with seed 3, shows that neural networks can also learn to use negative discrimination, as the precision for class B features was zero, while the average F1 score remained high. Section II-A shows an example of three windows with the selected mini-windows in green, and the presence of characteristic features in yellow. Our model is clearly able to selectively pick mini-windows that contain such features.

B. Polysomnography data

Figure 3 shows the class-average F1 score as a function of the number of selected mini-windows. Most spread originated from inter-patient differences, opposed to seeding or odd vs even measurements. Using our active mini-window selection

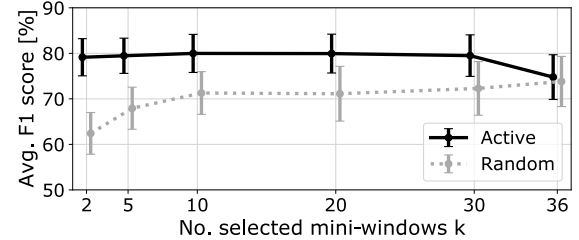


Fig. 3: F1 score averaged across sleep stages (mean \pm SD) degrades quickly when randomly selecting only few mini-windows (low k), while it remains high with active selection.

TABLE II: F1 scores per sleep stage (mean \pm SD across test set subjects) for $k = 2$ selected mini-windows. Active mini-window selection improves F1 scores across all sleep stages.

Policy	F1 N1	F1 N2	F1 N3	F1 REM	F1 W
Random	16.5 \pm 5.9	79.4 \pm 5.9	72.4 \pm 16.8	72.1 \pm 10.8	69.0 \pm 10.2
Active	55.2 \pm 8.0	88.9 \pm 3.0	82.8 \pm 13.7	83.4 \pm 9.8	84.4 \pm 8.1

policy, performance hardly dropped for decreasing k , while random selection heavily degraded classification performance. When selecting all mini-windows ($k = 36$), performance is, by definition, independent of the selection strategy, but interestingly this resulted in worse classification than using fewer mini-windows with an active selection policy. The latter suggests that performance benefits from the hard attention mechanism introduced by the active selection policy.

Lower k generally adds interpretability. We, therefore, zoom in on the models for $k = 2$. First, we observe that, compared to random selection, a large increase in F1 score was visible for the active policy. Table II shows (for one of the three seeds) that this improvement generalized to all sleep stages, but the largest increase was found for N1, in which the random policy performed extremely bad. It suggests that features of N1 are present only in specific locations in the window. Figure 4 shows how often each mini-window got selected (in the test set), normalized per sleep stage label. This figure confirms specific locations for N1 features by showing high variance across selection patterns in the test set. On the contrary, selection patterns for N3, Wake, and REM were more localized. N3 was mainly classified based on the frontal EEG channel, where one also expects most prominent delta waves, while Wake and REM were distinguished mainly based on EOG and Chin EMG data.

Figure 5 shows two windows from the test set (labeled and classified as N2 and Wake, resp.). In the first window the model only selected mini-windows that occurred before activity appeared around 20 seconds. For all higher k , this finding was consistent, e.g. the $k = 20$ model still only selected

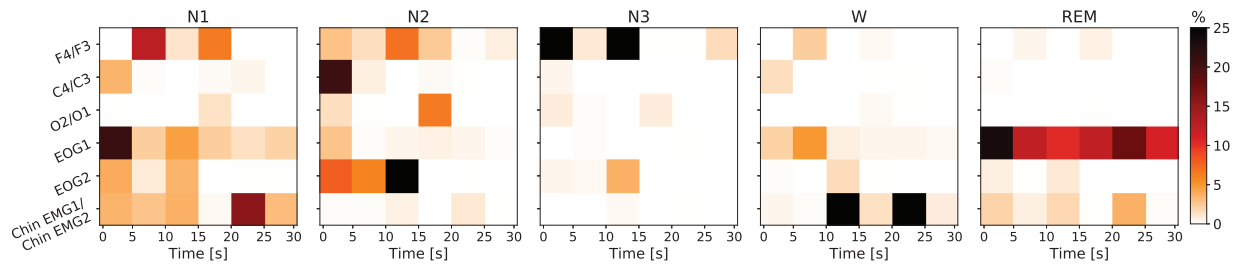


Fig. 4: The heat maps show how often each mini-window was selected as a percentage of all selected mini-windows of the test set windows with a certain sleep stage label. The model was trained for selecting $k = 2$ mini-windows. Different sleep stages use different channels.

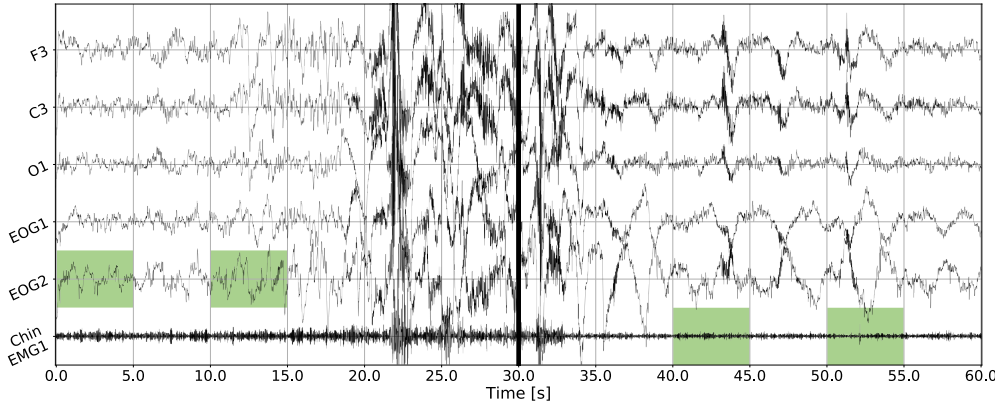


Fig. 5: Example of two windows from a test patient, labeled and predicted as N2 and Wake, respectively. Selected mini-windows are denoted in green.

Activity appears within the N2 window (at 20 seconds), but the selected mini-windows inform us on the model's ability to still score the window as N2.

mini-windows before the first 20 seconds. Interestingly, the amplitude of the chin EMG channel in the second window is rather low for a Wake period, which may occur when a patient lies very still. Still our model classified the window as Wake based on this channel, which shows that it carries more information to distinguish wakefulness than only its amplitude.

IV. CONCLUSION

We proposed a discrete attention module with an active selection policy that selects a subset of the data window that is required for decision-making. The model was validated on synthetic data and PSG data. It was shown to facilitate sleep stage classification while only using a small subset of the full window. Further analyses revealed that some sleep stages require selection of specific features in the window (e.g. N1), while others can simply rely on one or two channels (e.g. N3, Wake and REM). This work paves the way towards a better understanding of features related to different sleep stages, and (mis)-classification by sleep staging models of ambiguous PSG windows that contain features of more than one sleep stage. In a broader context, our proposed active selection policy could also be used to reveal disorder-specific features when training a classification model to distinguish different disorders.

REFERENCES

- [1] W. Samek *et al.*, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 3 2021.
- [2] F. Andreotti *et al.*, "Visualising Convolutional Neural Network Decisions in Automated Sleep Scoring," in *CEUR Workshop on AI in Health*, 2018, pp. 70–81.
- [3] M. Dutt *et al.*, "SleepXAI: An Explainable Deep Learning Approach for Multi-class Sleep stage Identification," *Appl. Intell.*, vol. 53, no. 13, pp. 16 830–16 843, 2023.
- [4] M. T. Troester *et al.*, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, 2023.
- [5] I. A. M. Huijben *et al.*, "Interpretation and Further Development of the Hypnosity Representation of Sleep Structure," *Phys. Meas.*, vol. 44, no. 1, 2023.
- [6] H. Phan *et al.*, "SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2456–2467, 2022.
- [7] B. Bai *et al.*, "Why Attention May Not Be Interpretable?" in *ACM SIGKDD Intern. Conf. Knowl. Disc. Data Mining*, 2021.
- [8] T. Strypsteen *et al.*, "End-to-end Learnable EEG Channel Selection for Deep Neural Networks with Gumbel-softmax," *J. Neural Eng.*, vol. 18, no. 4, 2021.
- [9] H. Stenwig *et al.*, "Automatic Sleep Stage Classification with Optimized Selection of EEG Channels," in *Intern. Conf. Mach. Learn. Appl.*, 2022.
- [10] F. B. van Meulen *et al.*, "Contactless Camera-Based Sleep Staging: The HealthBed Study," *Bioengineering*, vol. 10, no. 1, 2023.
- [11] M. H. Kryger *et al.*, *Principles and Practice of Sleep Medicine fifth edition*. Elsevier Health Sciences, 2011.
- [12] I. A. M. Huijben *et al.*, "Learning Sampling and Model-Based Signal Recovery for Compressed Sensing MRI," in *ICASSP*, 2020.
- [13] H. Van Gorp *et al.*, "Active Deep Probabilistic Subsampling," in *ICML*, 2021.
- [14] W. Kool *et al.*, "Ancestral Gumbel-top-k Sampling for Sampling Without Replacement," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 1–36, 2020.
- [15] S. M. Xie *et al.*, "Reparameterizable Subset Sampling via Continuous Relaxations," in *Intern. Joint Conf. Artif. Intell.*, 2019.
- [16] W. Kool *et al.*, "Stochastic Beams and Where to Find them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement," in *ICML*, 2019.
- [17] I. A. M. Huijben *et al.*, "A Review of the Gumbel-max Trick and its Extensions for Discrete Stochasticity in Machine Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1353–1371, 2022.
- [18] C. J. Maddison *et al.*, "The CONCRETE Distribution: A Continuous Relaxation of Discrete Random Variables," in *ICLR*, 2017.
- [19] E. Jang *et al.*, "Categorical Reparameterization with Gumbel-Softmax," in *ICLR*, 2017.
- [20] I. A. M. Huijben *et al.*, "Deep Probabilistic Subsampling for Task-adaptive Compressed Sensing," in *ICLR*, 2020.
- [21] D. P. Kingma *et al.*, "Adam: A Method for Stochastic Optimization," in *ICLR*, 2015.