Interpreting softmax predictions of sleep stage classification models

Iris A. M. Huijben, Lieke W. A. Hermans, Alessandro C. Rossi, Sebastiaan Overeem, Merel M. van Gilst, and Ruud J. G. van Sloun *Member, IEEE*

***Abstract*— Sleep staging is the process of assigning sleep stage annotations to windows of sleep recordings. Since the advent of machine learning, and in particular deep learning, automation of this labor-intensive job has attracted considerable attention. Commonly used models are convolutional classification networks that - by means of a softmax function - provide a probability for each of the different sleep stages, from which the stage with the largest probability is selected. Recently, it was proposed to use these softmax predictions as a means to get more insights into the continuous dynamics of the sleeping brain [1]. Plotted over time, these probabilities are called a *hypnodensity* (graph), as opposed to the conventional *hypnogram* that displays only one stage for every 30 seconds of data. In this work we investigate how to interpret the proposed hypnodensity by introducing a signal and annotation model. We model the selection of a sleep stage as a function of the contribution of five underlying (abstract) signals aggregating the characteristics belonging to one sleep stage. We conclude that a hypnodensity displays the distribution from which (discrete) labels were implicitly drawn during annotation of the training set. Moreover, we found that - despite the additional information available in a hypnodensity with respect to a hypnogram - potentially relevant information is still being lost due to the non-linear normalizing softmax layer, and the label-dependency of supervised training. As a solution, we propose to consider pre-softmax predictions and unsupervised training.**

***Index Terms*— Contrastive Predictive Coding, hypnodensity, hypnogram, sleep, softmax**

I. INTRODUCTION

EVEN though we spend a large part of our lives asleep, there is only a marginal understanding about the processes that happen in our brain during the night. Until the late thirties of the previous century, it was widely believed that sleep is a passive state of the body [2], opposed to the active state of wakefulness. Due to this belief, and the fact that sleep could not quantitatively be measured without disturbing the sleeper, little knowledge about sleep was gained until that time. The decades that followed were marked by new discoveries about the sleeping brain, mainly thanks to the discovery of the electroencephalogram (EEG), a recording that measures electrical activity of the brain via electrodes on the scalp. EEG recordings, in combination with other bio-physiological sensor modalities like elektromyography (EMG), electrooculography (EOG), electrocardiography (ECG), measures of respiratory effort etc. - summarized under the term polysomnography (PSG) - have remained the standard for clinical sleep research ever since.

Quickly after the discovery of the EEG, patterns in the sleeping brain were described. At that time, sleep was described by high-amplitude slow waves (0.5-2 Hz), and spindles (bursts of high frequency), while wakefulness was related to low-amplitude alpha rhythms (8-13 Hz) [2]. These patterns still form the basis for the way we describe sleep nowadays. The current standard for sleep analysis comes from the American Academy of Sleep Medicine (AASM) [3]. The AASM standard distinguishes different states through which a sleeping brain transitions (multiple times) during the night: rapid eye movement (REM) sleep, non-REM sleep (subdivided into N1, N2, and N3), and wakefulness. Given a PSG recording, a sleep technician (often manually) labels each window of 30 seconds with one of the five possible states to create a *hypnogram*; a visual representation of assigned sleep stages over the full night. The hypnogram has served as a powerful tool to assess the quality of someone's sleep. It typically reveals a cyclic pattern of alternating sleep stages in 'healthy sleepers' [2], and disturbed cyclicity or limited time spent in a certain stage can thus be an indicator for the presence of 'abnormal' sleep mechanisms. Moreover, presence of many transitions between states indicates unstable sleep [2]. It should be noted, however, that no two nights of sleep are equivalent, and both inter- and intra-personal differences occur. The terms 'healthy sleeper' and 'abnormal sleep' should thus be used with caution.

Over the last two decades, the immense increase in compute power and data availability has spurred a strong effort towards the development of machine-learning-based sleep staging algorithms that predict a hypnogram from a PSG recording (or recordings using fewer sensors) [4]–[6]. Such algorithms may alleviate the burden of manual data annotation, and might improve upon the intra-rater agreement between (human) scorers, which was found to be around 83% across 2500 scorers, as reported by the AASM inter-scorer reliability program [7]. Beyond automation of labor-intensive processes, machine learning can also be used to reveal intricate patterns in the data that are currently overlooked with hypnograms. While a hypnogram has proven clinical utility, it remains a strongly compressed representation of the measurements, possibly suppressing information that might be of additional clinical relevance. It is for example to be expected that transitioning between two states yields a more gradual pattern than represented in the hypnogram.

Recent developments in automating sleep medicine [1] could be the answer to the quest to reveal said continuities in sleep recordings. The authors of [1] propose to use predicted probabilities for each of the five AASM stages of a trained sleep stage classifier. Plotting these over the night results in, what the authors call, a *hypnodensity* graph. Such plots reveal moments where the probability mass is spread over multiple stages, giving rise to gradual transitions, something that is not captured in the hypnogram. This hypnodensity graph has the potential to induce a paradigm shift in sleep medicine, providing doctors with a more detailed representation of the nocturnal recording that might provide insights in (yet unexplained) phenomena (see fig. 5 for an example). It was, for example, already shown that the hypnodensity graph contains information regarding sleep stage dissociations typically found in people with narcolepsy [1].

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible. This work was supported by Onera Health, the Dutch Research Council (NWO) through the Research Program Rubicon ENW 2018-3 under Project 019.183.EN.014., and the project 'OP-SLEEP'. The project 'OP-SLEEP' is made possible by the European Regional Development Fund, in the context of OPZuid. (Corresponding author: Iris A. M. Huijben.)

Iris A. M. Huijben, Lieke W. A. Hermans, Sebastiaan Overeem, Merel M. van Gilst, and Ruud J. G. van Sloun are with the Department of Electrical Engineering, Eindhoven University of Technology, 5612 AP Eindhoven, The Netherlands (e-mail: {i.a.m.huijben; l.w.h.hermans; s.overeem; m.m.v.gilst; r.j.g.v.sloun}@tue.nl).

Iris A. M. Huijben, and Alessandro C. Rossi are with Onera Health, 5617 BD Eindhoven, The Netherlands (e-mail: alessandro.rossi@onerahealth.com).

Merel M. van Gilst and Sebastiaan Overeem are with Sleep Medicine Center Kempenhaeghe, 5591 VE Heeze, The Netherlands.

The main question that arises now, is how to interpret the probabilities in a hypnodensity graph. In the absence of a ground truth, it is in general uncertain whether these probabilities indeed reflect the underlying continuous processes of sleep. An 80-20% prediction for the occurrence of stage N2 and N3, can for example be interpreted as an 80% certain classification for N2. Alternatively, it can be argued that 80% of the characteristics present in the window belong to N2, while the remainder belongs to N3. The former explanation relates to *model uncertainty*, while the latter is concerned with unbiased estimation and disentanglement of *mixtures* of stages. Both explanations are not necessarily mutually exclusive and could also occur simultaneously. Said mixtures of stages, i.e. when characteristics belonging to multiple stages are simultaneously present in a recorded data window, may occur in practice through co-existence of distinct phenomena in different cortical areas (known as local sleep) [8], or due to temporal compression induced by using 30-second sleep windows.

In this paper we investigate how different parts of the hypnodensity-prediction model contribute to the prediction, in an attempt to facilitate interpretation, and better understand the assumptions and limitations. Specifically, we investigate the respective roles of the label(-generating) distribution, the training strategy of the encoder (supervised vs unsupervised), and the non-linear softmax activation. Experiments are performed both with synthetic data (for which the ground truth signal model is available), and with full-night PSG recordings of healthy sleepers. The main contributions can be summarized as follows:

- We propose a synthetic data set inspired by PSG recordings to faciliate controlled experiments, in which the label distribution can be altered, and the ground-truth signal model is known. We define the signal model as a non-linear mixed measurement of underlying signals (scaled with their respective mixture coefficients) that each represent a ficticious sleep stage.
- Through simulation experiments we show that our supervised classification model with a softmax activation predicts a probability distribution that corresponds to the label distribution. This finding enhances interpretability of a hypnodensity graph in sleep analysis. Moreover, we show that predictions of classifications models trained on unsupervised encodings are more label-agnostic, and therefore better reflect data characteristics like the mixture coefficients in a (non-linearly) mixed measurement.
- We show that the final softmax activation of a sleep stage classification model, leads to loss of potentially relevant information in a hypnodensity graph. E.g., the pre-softmax prediction for deep sleep (N3) much better captures the continuity of slow waves (compared to the post-softmax prediction), which is in turn known to correspond to deep sleep.

II. Method

A. Signal model

A typical PSG recording $\mathbf{X} \in \mathbb{R}^{\text{ch}\times L}$ contains time series (of L samples) from multiple channels (e.g. multiple EEG channels, EMG, and EOG). We model the data generation/measurement as a non-linear generative mixing process of C signals $\mathbf{S} \in \mathbb{R}^{C\times L}$, where each signal $\boldsymbol{s}_c \in \mathbb{R}^L$ in $\mathbf{S}$ aggregates typical characteristics associated with a specific sleep stage (with $c \in \{\text{W, N1, N2, N3, R}\}$, where W, N1-N3, and R respectively denote wakefulness, non-REM1-3, and REM sleep). Moreover, we model the amplitudes $\boldsymbol{a}_c$ of these signals to vary over time, i.e. characteristics belonging to a certain sleep stage can be fully absent in some moments, while present (with a certain amount) at other moments. The resulting signal model yields:

$$\mathbf{X} = h(\mathbf{A} * \mathbf{S}), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}_{\geq 0}^{C\times L}$ contains the (slow) time-varying amplitudes, which we refer to as *mixture coefficients*, $h : \mathbb{R}^{C\times L} \to \mathbb{R}^{\text{ch}\times L}$ is a non-linear spatial mixing function, and the $*$ symbol denotes an element-wise multiplication. We introduce $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$, a set of non-overlapping data windows $\mathbf{X}_w \in \mathbb{R}^{\text{ch}\times l}$ of length $l = 30\times f_s$, where w denotes the window's index, and f_s the sampling frequency (in Hz). Concatenation (over the last dimension) of all matrices in $\mathcal{X}$ yields $\mathbf{X}$. Analogously, we define $\mathcal{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots\}$, with $\mathbf{A}_w \in \mathbb{R}_{\geq 0}^{C\times l}$ the (unnormalized) amplitudes for all C classes in data window w. We denote the normalized counterpart, i.e. the amplitudes of all classes that sum to one at every moment in time, with $\tilde{\mathbf{A}}_w$ (or $\tilde{\mathbf{A}}$ after concatenation of the separate windows).

B. Annotation model

We introduce an annotation model $g(\tilde{\mathbf{A}}_w)$ that assigns a label $\boldsymbol{y}_w \in \{0,1\}^C$ (i.e. a one-hot embedding of a selected sleep stage) to a data window[1]. Concatenation of these labels for all windows, results in $\mathbf{Y} \in \{0,1\}^{C\times L/l}$. The annotation model can be decomposed into two parts: selection of a (deterministic) *rule set*, and (stochastic) *decision making* or application of the rules. Given our signal model from section II-A, the normalized amplitudes $\tilde{\mathbf{A}}_w$, or *(normalized) mixture coefficients*, represent the contributions of characteristics belonging to different sleep stages. The conversion from the mixture coefficients to selection of one sleep stage is in practice non-linear and fuzzy. Non-linearity arises from the rules prescribed by the AASM standard. E.g. when a K-complex is detected, the window should be classified as N2. Fuzziness is caused by the stochastic nature of human decision making, which finally gives rise to a label distribution. This distribution presents the conditional posterior probability for selecting each of the class labels (i.e. sleep stages) for a given window. Figure 1 depicts the described signal and annotation model. Note that in practice, a technician selects a sleep stage directly given the raw data, rather than first creating mixture coefficients, and subsequently choosing the appropriate label. Though, the process of *disentangling* the raw data into characteristics that describe varies sleep stages, could be considered an implicit process that takes place during decision making.

Given the non-linear decision rules of the AASM standard, we define the following non-linear expression for the label distribution:

$$p[\boldsymbol{c}|\tilde{\mathbf{A}}_w] = \sigma_\tau\{\text{avg}_l(\tilde{\mathbf{A}}_w)\} \propto \exp\{\text{avg}_l(\tilde{\mathbf{A}}_w)/\tau\}, \tag{2}$$

where σ_τ denotes a tempered softmax function with temperature parameter $\tau \in \mathbb{R}_{\geq 0}$, one-hot$(\cdot)$ converts a scalar to a one-hot embedding of length C, and $\text{avg}_l(\cdot)$ returns the average over l samples. When $\tau \to 0^+$, the label distribution becomes one-hot, i.e. all probability mass is placed on one sleep stage, and the softmax function converts into an argmax function. For $\tau \to \infty$ the distribution converges to a uniform distribution.

This model is a simplified version of the true (but unknown) label distribution that follows from applying the AASM standard. Nevertheless, we use it to assess the influence of non-linearly converting mixture coefficients to a label. The temperature parameter facilitates modelling both the AASM intention to guide decision making towards a deterministic process ($\tau \to 0^+$), and stochastic decision making ($\tau > 0$), that leads to inter- and intra-rater disagreement.

---

[1] In section II-F we also define soft labels, which are non-discrete and live in the real domain. Since these only serve theoretical purposes, we chose to define the labels over a discrete domain in the general definition.

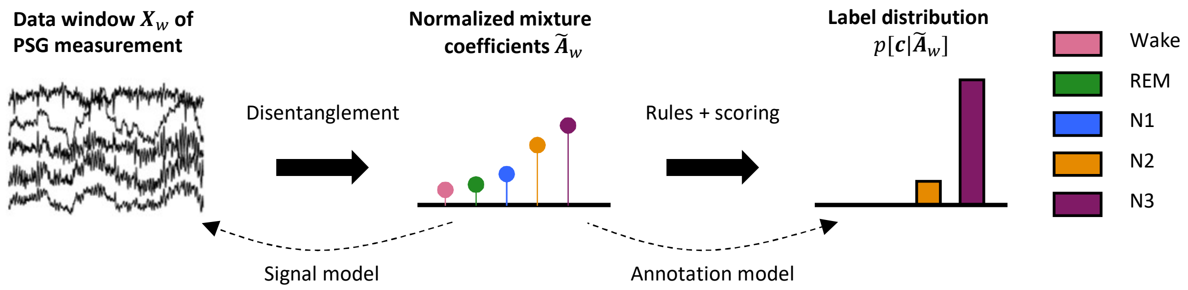

Data window X_w of PSG measurement
Normalized mixture coefficients $\widetilde{A}_w$
Label distribution $p[c|\widetilde{A}_w]$
Disentanglement
Rules + scoring
Signal model
Annotation model
Wake
REM
N1
N2
N3

Fig. 1: We distinguish a signal and annotation model. The signal model assumes that all channels are a non-linear mixture of some implicit (or hidden) signals that each characterize one of the sleep stages. The annotation model converts the contribution of each of these 'sleep stage signals' to one selected sleep stage. This model is characterized by a set of rules (e.g. the AASM standard), and application thereof, i.e. scoring by a technician. Due to variability in annotations, we know that scoring is a stochastic process, that gives rise to a label distribution that expresses the probability for a given window to be classified as either of the sleep stages.

C. *Problem formalization*

We define a model $f_\theta(\cdot)$, parameterized by θ, that maps input $\mathbf{X}_w$ to a prediction $\hat{\boldsymbol{y}}_w = f_\theta(\mathbf{X}_w) \in \mathbb{R}^C$, of which the desired output $\boldsymbol{y}_w$ is generated through the annotation model (e.g. by applying the AASM standard). Training/updating parameters θ is done by maximizing the log-likelihood of the annotated labels, given a (training) set of data that originates from a data-generating distribution p_{data}. The corresponding optimization problem reads:

$$\theta^* = \underset{\theta}{argmax}\{\mathbb{E}_{(\mathbf{X}_w, \boldsymbol{y}_w)\sim p_{data}} \log p(\boldsymbol{y}_w|\mathbf{X}_w, f_\theta)\}. \tag{3}$$

Equation 3 is typically solved by reparameterizing $f_\theta(\cdot)$ as a (deep) neural network, of which trainable parameters θ are optimized using stochastic gradient descent. Under perfect optimization and sufficient model capacity, $\mathbf{Y} \leftarrow \hat{\mathbf{Y}}$. Thus the model's training, and therefore also its predictions on new data, is influenced by the (distribution of the) annotations.

The authors of [1] optimize eq. (3) using AASM annotations, and interpret a prediction $\hat{\boldsymbol{y}}_w$ as a discrete probability distribution over C classes, yielding a hypnodensity graph when depicted over time. This probability distribution could reflect probabilities related to the chance that a window was (manually) annotated with a certain label, in which $\hat{\mathbf{Y}}$ would have converged to the label distribution. On the other hand, the probabilities could also reflect the earlier introduced mixture coefficients (in which they would have converged to $\hat{\mathbf{Y}} \leftarrow \tilde{\mathbf{A}}$). Lastly, a combination of both options could also be thinkable. In this paper we investigate which of these options is most likely the case in order to facilitate interpretation of the hypnodensity graph.

D. *Data acquisition and preprocessing*

1) Polysomnographic data: We used a dataset of nocturnal PSG recordings, collected as part of the Healthbed study, which main aim was development of technologies for sleep analyses. The study protoctol (W17.128) was approved by the medical ethics committee of Maxima Medical Center, Veldhoven, the Netherlands. The dataset includes one clinical video-PSG recording for each subject, made according to the AASM recommendations in Sleep Medicine Center Kempenhaeghe Heeze, the Netherlands. The data analysis protocol for our study (CSG_2021_007_00) was approved by the medical ethics committee of Sleep Medicine Center Kempenhaeghe (11/11/2019).

The study included 96 (60 females) healthy subjects, with an age between 18 and 64. The exclusion criteria were: 1) any diagnosed sleep disorder, 2) a Pittsburgh Sleep Quality Index [9] $\geq$ 6, or Insomnia Severity Index [10] $>$ 7, 3) indication of depression or anxiety disorder measured with the Hospital Anxiety and Depression Scale [11] (score $>$ 8), 4) pregnancy, shift work, use of any medication except for birth control medicine, and 5) presence of clinically relevant neurological or psychiatric disorders or other somatic disorders that could influence sleep.

Visual sleep staging on windows of 30 seconds was performed according to AASM criteria [3] by an experienced and certified sleep technician (BH) from Sleep Medicine Center Kempenhaeghe. In a previous institutional sleep scoring reliability check, inter-scorer reliability of BH compared to other technicians was assessed at 85.6% on average (range 83-88%).

From the full PSG recordings, we selected EEG (F4, C4, O2, F3, C3, O1), chin EMG (Chin2, Chin1), and EOG (E2, E1) derivations, the same modalities as in [1]. Since the EEG and EMG derivations contain redundancy among the left and right hemisphere, the data was virtually doubled by considering an 'odd' and 'even' recording per subject[2]. For simplicity, we also added only one of the two EOG recordings per subset, even though these recordings can not be considered fully redundant. As an example; channel data $\mathbf{X}^{(k;even)} \in \mathbb{R}^{5\times L}$ - the selected even data from the PSG of subject k - thus contained the F4, C4, O2, E2, and Chin2 derivations.

Following [1], all derivations were filtered with a zero-phase (i.e. two-directional) 5th order Butterworth band-pass filter, with cut-off frequencies of 0.2 and 49 Hz. It was followed by another zero-phase 5th order Butterworth notch filter between 49 and 51 Hz, to better suppress powerline interference. All channels were originally recorded with a sampling rate of 512 Hz, but (after filtering) down-sampled to 128 Hz to reduce computational complexity. Channels were normalized within-patient and per channel, yielding mean subtraction, followed by normalization such that amplitudes of 95% of the samples were mapped between -1 and +1.

Finally, we randomly split the full dataset in a training, validation and hold-out test set, comprising respectively 150, 20, and 22 recordings (each recording being either even or odd). Even and odd recordings from the same subject were in all cases assigned to the same subset.

2) Synthetic data: Next to the real PSG recordings, we created synthetic data, for which we can alter the label distribution, and for which the slow-varying amplitudes $\mathbf{A}$ (i.e. the mixture coefficients) are known. We followed the signal model as introduced in eq. (1), and modelled each channel in $\mathbf{X}$ as a non-linear combination of a

---

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

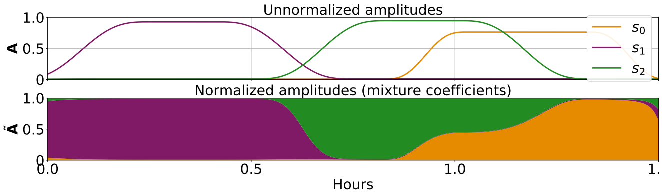


Fig. 2: A randomly generated example of the synthetic data generator. Top: the amplitudes of the three generated signals over time. Bottom: The normalized amplitudes that sum to one and reveal the ground-truth normalized mixture coefficients of the three signals.

set of ($C = 3$) independent signals, where each signal represents a (fictitious sleep) stage. We define the data of 'subject k' as $\mathbf{X}^{(k)} = h(\mathbf{A}^{(k)} * \mathbf{S}^{(k)})$.

Each signal $\boldsymbol{a}_c^{(k)} * \boldsymbol{s}_c^{(k)}$ was generated as a (discretized) sinusoidal signal, with a frequency between $f_c - 0.5$ and $f_c + 0.5$ Hz, a random phase, and an envelop that is described by a smoothened square wave (sw). More specifically, per subject we defined three independent signals, with $c \in \{0, 1, 2\}$:

$$\boldsymbol{s}_c^{(k)}[n] = \sin\{2\pi(\frac{f_c + u[-\frac{1}{2}, \frac{1}{2}]}{f_s})[n] + 2\pi u[0, 1]\}, \tag{4}$$

with $u[a, b]$ being a realization of a uniform random variable between a and b, and $\{f_0, f_1, f_2\} = \{2.5, 6, 11\}$ Hz. Each source's length equals $\lambda = 5400$ s, thereby mimicking the length of one average sleep cycle. The data were sampled with a frequency $f_s = 100$ Hz, and n thus ranges from 0 to $5.4e5$ samples. The amplitude of the c^{th} signal of subject k, was defined as:

$$\boldsymbol{a}_c^{(k)}[n] = \frac{\text{Hanning}_\nu \circledast \text{sw}[\text{n}; \Phi]}{|\,\text{Hanning}_\nu\,|}, \tag{5}$$

where $\circledast$ denotes a convolutional operator, $\nu = u[\frac{1}{20}, \frac{1}{4}]\lambda f_s$ is the length of the applied Hanning window, and the square wave's parameters are given by $\Phi = \{\text{period} = \lambda, \text{sampling\_freq} = f_s, \text{duty\_cycle} = \frac{1}{2}, \text{phase} = 2\pi u[0, 1], \text{min\_value} = \frac{1}{100}, \text{max\_value} = u[\frac{1}{2}, 1]\}$.

From the earlier definition of $\mathbf{X}^{(k)}$ it can be seen that mixing function $h(\cdot)$ is subject-independent. This results in a simplified but valid model, since certain sleep stage characteristics are in practice also measured more in certain channels than in others for all subjects (e.g. slow waves are mainly recorded in the frontal EEG electrodes). Only small deviations - resulting from inter-patient differences - are not captured by choosing one shared setting. For brevity, we use $\boldsymbol{as}_c$ to denote $\boldsymbol{a}_c^{(k)} * \boldsymbol{s}_c^{(k)}$ here. As an arbitrary choice, we defined the non-linear mixing in $h(\cdot)$ as:

$$h(\mathbf{A}^{(k)} * \mathbf{S}^{(k)}) = \begin{bmatrix} 0.3\ \boldsymbol{as}_0 * \boldsymbol{as}_0 + 0.7\ \boldsymbol{as}_2 \\ 0.6\ \boldsymbol{as}_0 + 0.4\ \boldsymbol{as}_1 * \boldsymbol{as}_2 \\ 0.4\ \boldsymbol{as}_0 + 0.5\ \boldsymbol{as}_1 + 0.1\ (\boldsymbol{as}_2)^2 \end{bmatrix}.$$

We finally generated 200 random 'subjects', which were split into a training, validation and a hold-out test set of sizes 75, 25, and 100, respectively. Figure 2 shows a randomly generated example, with unnormalized amplitudes in the top, and normalized amplitudes (i.e. the ground truth mixture coefficients) in the bottom. From the top figure it can be seen that the heights, phases, and the steepnesses of the amplitudes differ per signal, caused by the injected stochasticity in the data generating process. As a result of this stochasticity, we see that at any moment zero to three signals/sleep stages might co-exist. Absence of characteristics belonging to any of the sleep stages, might in practice occur when electrodes become disconnected.

E. Model design

In this section we introduce the base model as used in all experiments. For some experiments, slight deviations from the presented settings were chosen, which will be explained in the corresponding experimental section, where needed.

An encoder converts a data window $\mathbf{X}_w$ to a latent representation: $\boldsymbol{z}_w = \text{Enc}(\mathbf{X}_w) \in \mathbb{R}^F$, with F the number of features in the resulting embedding. All (non-overlapping) embeddings are part of $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots\}$. Data and embeddings from subject k are denoted with $\mathcal{X}^{(k)} \subset \mathcal{X}$, and $\mathcal{Z}^{(k)} \subset \mathcal{Z}$, respectively.

The architecture of the encoder follows standard practice in supervised classification model design [12]. $\text{Enc}(\cdot)$ comprises three consecutive blocks, where each block contains a 1D convolutional layer over the time dimension, activated by a LeakyReLU (negative slope of 0.01), followed by a 1D max pooling layer, and finally a dropout layer ($p = 0.1$). After the third full block, a fourth 1D convolutional layer is added, followed by average pooling that reduces the temporal dimension to size 1, creating a 1D embedding of size F. All convolutional layers have a bias term, and use strides and dilations of 1. The number of channels differs for the real data $(16, 32, 64, 128)$ vs synthetic data $(4, 8, 16, 32)$ setup, to account for the higher complexity of real data. The used kernels are of size $(15, 9, 5, 3)$ for the four convolutions, and the max pooling layers use kernels of size 5 (with stride 5).

Next, we introduce a standard multi-class classification model of the form $\hat{\boldsymbol{y}}_w = \sigma(\mathbf{W}\boldsymbol{z}_w + \boldsymbol{b})$, with trainable parameters $\mathbf{W} \in \mathbb{R}^{C \times F}$, and $\boldsymbol{b} \in \mathbb{R}^C$, and σ the softmax function that maps all class predictions between 0 and 1, with a total sum of 1.

F. Experimental setup

In this study, we investigate three factors that contribute to the hypnodensity predictions. This section elaborates on these factors.

a) Changing the label generator: As explained in section II-C, definition of the labels $\mathbf{Y}$ influences model optimization. We therefore investigate the effect of the label distribution on the predicted class probabilities. This is done in the synthetic setup, as the underlying mixture coefficients $\tilde{\mathbf{A}}$, from which labels can be sampled, are unknown for real data. Nevertheless, learning about the effect of annotations on synthetic data might provide insights in the labelling process and the consequences thereof for the real use case. We generated two different set of labels that followed the distribution as described in eq. (2). First, we set $\tau \to 0^+$ and generated *argmax* labels. Second, *sampled* labels were generted by sampling from eq. (2), using $\tau = 1$. Moreover, we defined a third set of *soft* labels (opposed to discrete/hard labels), that directly correspond to the mixture coefficients for each class (averaged over l samples in the window). Note, however, that such soft labels are purely theoretical and also not caught in the model for the label distribution, as given in eq. (2).

We hypothesize that the (unambiguous) argmax labels enhance class separability (in the latent space), and therefore facilitate highly-accurate classification performance with low-entropy predicted probabilities. Both the sampled and soft labels, on the other hand, exhibit ambiguity and are therefore expected to increase the entropy of the predictions, and lower the classification accuracy.

b) Supervised vs unsupervised encoding: As altering the label distribution is expected to influence the model's predictions in varies ways (see previous paragraph), we wonder what happens to the predictions when training (the largest part of) the model in an unsupervised setup, i.e. without access to label information.

We hypothesize that in this case, the predictions will reflect data characteristics, rather than label characteristics.

We compare the fully supervised setting, where the model is trained using input-label pairs, to a setting in which the full encoder is trained unsupervised. For the latter, we leverage Contrastive Predictive Coding (CPC) [13], a recently proposed framework for self-supervised learning, which has been found useful to model EEG data [14]. CPC is able to model *slow features*, i.e. slowly varying data characteristics, like the amplitudes $\mathbf{A}$ in our signal model. As such we hypothesize that a classifier (with its design as described in section II-E) - trained on resulting ‘unsupervised embeddings’ - will predict the normalized mixture coefficients $\tilde{\mathbf{A}}$.

CPC leverages contrastive learning, which builds upon the idea to teach the model that ‘similar data points’ should be embedded closely together, while ‘dissimilar data points’ should be repelled. In the framework of CPC, a similar data point (or positive sample) is defined as a future embedding, with respect to a current causal embedding (i.e. incorporating past information as well). Negative samples, on the other hand, are drawn from a random moment within or between (i.e. from a different) recordings. We use within-subject sampling, and randomly draw three negative samples per positive sample. The set $\mathcal{X}'^{(k)} \subset \mathcal{X}^{(k)}$ comprises these three negatives for subject k, and is renewed in every training iteration.

In order to make the fairest comparison to the fully supervised base model (as defined in section II-E), we do not make any changes to the encoder's design (except for the dropout rates in the CPC encoding trained on synthetic data, for which lower values appeared more beneficial: $(0.1, 0.0, 0.0)$), neither to initialization of its parameters. Moreover, we slightly simplify the original CPC objective by omitting the auto-regressive module, since our supervised classifier also classifies each window independently. Our unsupervised training objective, being an expectation over dataset $\mathcal{X}$, reads:

$$\mathcal{L} = \frac{1}{J}\sum_{j=1}^{J}\mathcal{L}(j), \text{ with} \tag{6}$$

$$\mathcal{L}(j) = -\mathop{\mathbb{E}}_{\mathcal{X}}\Big[\log\frac{\exp(\boldsymbol{z}_{w+j}^{T}\mathbf{V}_j\boldsymbol{z}_w)}{\sum_{\boldsymbol{z}\in\mathcal{Z}'^{(k)}}\exp(\boldsymbol{z}^{T}\mathbf{V}_j\boldsymbol{z}_w)+\exp(\boldsymbol{z}_{w+j}^{T}\mathbf{V}_j\boldsymbol{z}_w)}\Big],$$

$J = 10$ the number of future windows, $\mathcal{Z}'^{(k)} \subset \mathcal{Z}^{(k)}$ $(|\mathcal{Z}'^{(k)}| = 3)$ a set of embeddings of drawn negative samples $\mathcal{X}'^{(k)}$, $\boldsymbol{z}_w$ the current embedding, $\boldsymbol{z}_{w+j}$ the future embedding at index $w + j$, and $\mathbf{V}_j \in \mathbb{R}^{F\times F}$ a trainable mapping between both embeddings.

c) Pre- vs post-softmax predictions: The most widely used final layer in multi-class classification models, is described by the softmax function σ, as given by:

$$\boldsymbol{r} = \sigma(\boldsymbol{q}) = \frac{\exp \boldsymbol{q}}{\sum_c \exp q_c}, \tag{7}$$

where $\boldsymbol{q} \in \mathbb{R}^C$ is the vector of unconstrained pre-softmax predictions per class, and $\boldsymbol{r} \in \{\mathbb{R}^C : 0 \leq r_c \leq 1, \sum_c r_c = 1\}$ the vector of post-softmax probabilities, analogous to the model prediction $\hat{\boldsymbol{y}}_w$, as introduced in section II-C. We investigate the effect of the non-linearity as introduced by said softmax function, by comparing pre-softmax (in short, pre-σ) class predictions in $\boldsymbol{q}$, to post-softmax (post-σ) values in $\boldsymbol{r}$.

Training details: All supervised models were trained using the standard categorical cross-entropy (or negative log-likelihood) loss, and batches of 128 training pairs. Unsupervised encodings were trained with the CPC objective as given in eq. (6), and batches of size 64. The Adam optimizer with default settings [15] was used in all experiments, with a learning rate of 1e-4 for most experiments. Only the supervised classifier, and CPC encoding on synthetic data were trained with learning rates of 1e-3 and 5e-4, respectively. All models were maximally trained for 500 epochs, where one epoch defines one push trough of each data window in the training set. The classifiers trained after CPC encoding, were maximally trained for 250 epochs. In each experiment, the model with the lowest validation loss was finally selected. All experiments were run with the same seed for randomization.

G. Performance assessment

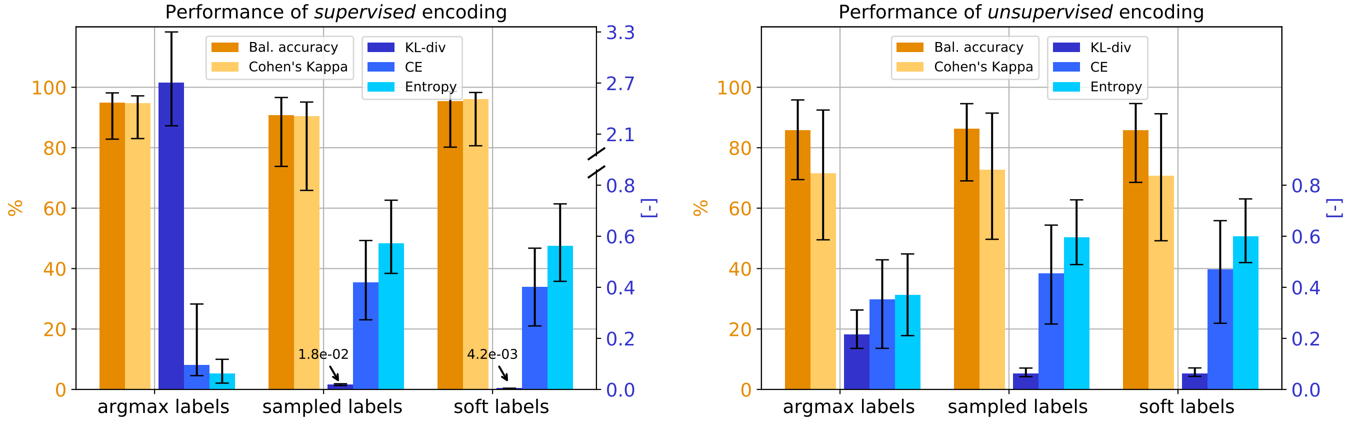
In order to assess classification performance, we follow the vast majority of literature on automatic sleep staging, and use class-balanced accuracy (bal. acc.) and Cohen's Kappa κ [16], where the latter corrects for agreement by chance, which can be particularly present for class-imbalanced data. These metrics have successfully served as interpretable metrics since both equal 1 (or 100%) in an ideal classification model. Notwithstanding their relevance and frequent use, information is lost on how high the predicted probability mass was distributed across classes. A high value for these ‘hard metrics’, can therefore only indicate good performance in the classification problem, but does not inform us about the (spread of) probabilities in the hypnodensity.

Thus, instead of selecting the class with the highest predicted probability, we can also interpret the prediction, given a data window, as parameters of a categorical distribution $\mathrm{Cat}(\hat{\boldsymbol{y}}_w)$, This distribution can be compared against a target distribution $\mathrm{Cat}(\boldsymbol{y})$. By slight abuse of notation, we define $\hat{y}_c$ and y_c as, respectively, the predicted and target probability for class c in window w. In case of discrete labels (e.g. according to the AASM standard), the target probabilities equal either zero or one. The Kullback-Leibler (KL) divergence provides a measure to compare both distribution as follows:

$$\begin{aligned}\mathrm{KL}\big(\mathrm{Cat}(\boldsymbol{y}_w);\mathrm{Cat}(\hat{\boldsymbol{y}}_w)\big) &= \mathrm{CE}\big(\mathrm{Cat}(\boldsymbol{y}_w),\mathrm{Cat}(\hat{\boldsymbol{y}}_w)\big) - \mathrm{H}\big(\mathrm{Cat}(\boldsymbol{y}_w)\big)\\ &= -\sum_c y_c \log \hat{y}_c + \sum_c y_c \log y_c,\end{aligned} \tag{8}$$

where CE and H denote cross-entropy and entropy, respectively, and a KL divergence of zero implies perfectly matched distributions. For soft labels (only available in the synthetic setting), this thus implies that the model is able to predict the mixture coefficients of the distinct signals. Using the discrete one-hot labels (for which $\mathrm{H}(\boldsymbol{y}_w) = 0$ and $y_j = 0$ for all j not being the annotated class), the KL divergence is equal to the CE loss between the discrete label and the prediction. To prevent confusion, we use KL-div to denote the metric that compares the prediction to the soft label, while we use CE to denote the match with the one-hot discrete label. To characterize the spread of the predicted probabilities we can evaluate the entropy of the prediction (i.e. $\mathrm{H}(\hat{\boldsymbol{y}}_w)$), which equals zero when all probability mass is centered in one of the C bins, and $\log C$ in case of a uniform prediction. All aforementioned metrics will be averaged across patients, and one standard deviation (also across patients) will be provided as well.

Lastly, for the real PSG measurements, we assess the correlation between the predictions for N3 and the power of slow waves (0.5-2 Hz) in the frontal EEG lead (F3 or F4). It is known that slow waves (positively) relate to the depth of sleep [2]. As such, we can use it as a surrogate for the contribution of the deepest sleep phase N3, to the total mixture of characteristics belonging to different stages. We thus consider a positive (linear) relation between the predicted N3 contribution and delta power as an indication that the model was well able to predict the amount of contribution of N3 characteristics in the data window.



(a)

(b)

Fig. 3: Performance of fully-supervised (a) and semi-supervised (b) classifiers trained on the synthetic dataset, with varying label generating processes. The orange metrics denote 'hard metrics', concerned with the (discrete) predicted class, while the blue metrics denote 'soft metrics', related to the predicted probabilities. The median and the 25th and 75th percentiles (denoted with the bars) are reported for all metrics. The effect of using different label types (during training) is largest for the fully-supervised model, where argmax labels clearly facilitate accurate, and low-entropy classifications, but at the cost of mixture prediction (seen from the high KL divergence with the soft label).

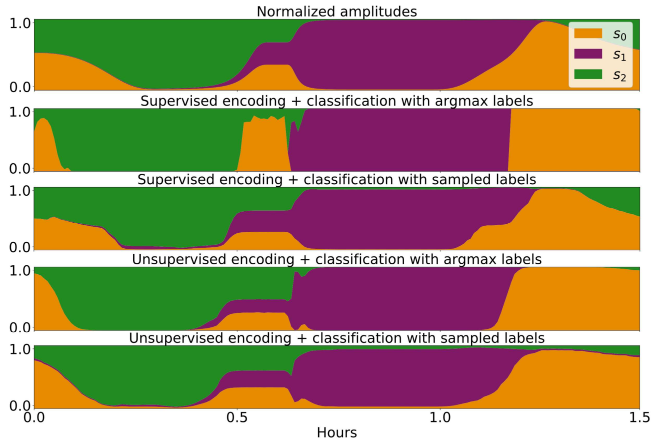


Fig. 4: Ground truth and different predicted hypnodensities of a representative test set sample from the synthetic dataset. The prediction of the supervised model, trained with perfect argmax labels is of lower entropy that the other three setups, and has a higher mismatch with the true mixture of signals. For both type of encodings, training with sampled labels results in the best mixture coefficient prediction.

III. RESULTS

This section provides the results of the three sub-experiments as described in section II-F. Results on altering the label generator are presented in section III-A. Section III-B and III-C discuss results regarding supervised vs unsupervised encoding, and the effect of the softmax function, respectively.

A. Changing the label generator

Figure 3a summarizes both the 'hard' and 'soft' metrics for the supervised classifiers (on the synthetic test set) trained with the three introduced label generators. For all cases, the hard metrics (orange) were computed using the argmax labels, while the soft metrics (blue) were evaluated using the soft labels. Training with the argmax labels induced predictions with high accuracy and low entropy, possibly explained by high separability of the classes thanks to the unambiguity of the argmax labels. However, the KL divergence with soft labels was very high, implying bad mixture coefficient prediction. Note the interesting similarity between all metrics for sampled vs soft labels, indicating that (discrete) labels sampled from a distribution induce similar training behavior as soft labels that represent this same distribution. Both resulting models hardly gave up on classification performance with respect to training with argmax labels, while their prediction of the mixture coefficients was much better (seen from the low KL divergence with soft labels). This makes us believe that a supervised model predicts probabilities that correspond to the label distribution. We confirm this hypothesis by training and testing with different label distributions, for which results are discussed in app. I.

B. Supervised vs unsupervised encoding

The experiment with different label generators was repeated, while performing the full encoding procedure unsupervised. Figure 3b shows the results. Remarkably, training the classifier with argmax labels resulted in much better mixture prediction (lower KL-div), a direct result of being less reliant on annotations. Classification performance, on the other hand, decreased compared to the fully supervised setup, seen from the lower bal. acc. and κ. Figure 4 visually compares (for one representative synthetic test set sample) the predicted hypnodensities for both encoding strategies, and argmax vs sampled labels. In line with fig. 3a, rows II and III visually confirm that the supervised model indeed heavily relies on the type of labels (e.g. argmax labels induce low-entropy predictions), while the CPC-encoded models (rows IV-V) were not that much affected by the annotations. However, the influence of annotations when training on unsupervised embeddings, in some cases depended on how long the classifier was trained. Appendix II discusses this in more detail.

A visual comparison of real data hypnodensity predictions from the two distinct encoding strategies are shown in fig. 5. At first glance the general trend of both hypnodensities looks similar, which is remarkable since the supervised vs unsupervised model achieved balanced accuracies of 84.4% and 61.2% for this subject, indicating that hard metrics do not well reflect the underlying continuous predictions in a sleep stage classification model. Interestingly, occasions where the (annotated) hypnogram showed rapid transitioning behavior (e.g. at 6.2 hours), were also characterized by high entropy predictions

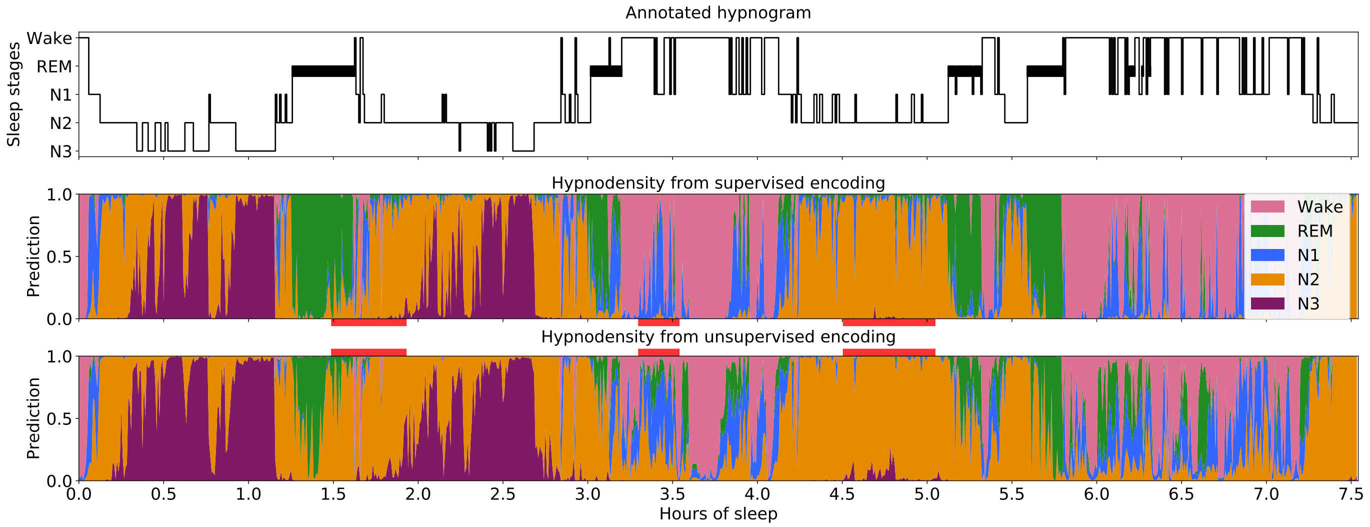
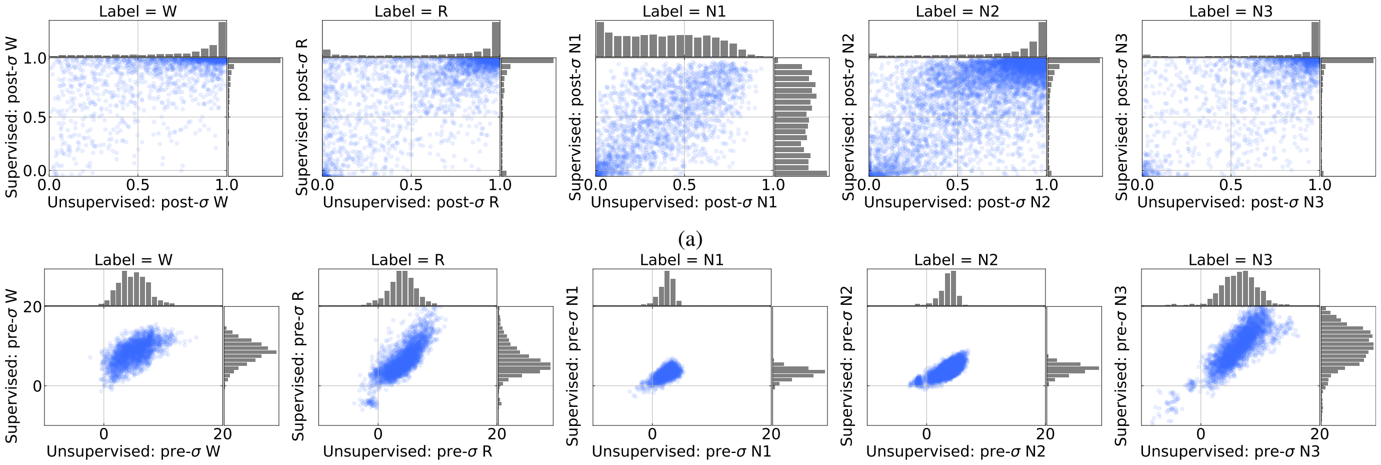


Fig. 5: The ground truth hypnogram (top), and predicted hypnodensities by the supervised (middle) and unsupervised model (bottom). The general trend looks similar, but differences are visible (indicated with the red bars), e.g. the unsupervised model in general shows smoother transitions when transitioning in and out of REM and N3. Note that the unsupervised model does not just predict a smoothened version of the supervised prediction: hard transitions (e.g. when leaving N3 at 1.2 hours) are still predicted as well. Balanced accuracy of these two models (with respect to the annotated labels) are respectively 84.4% and 61.2%. This difference in accuracy seems to imply much worse classification performance for the latter, while the hypnodensities suggest less drastically dissimilar underlying dynamics.



(a)

(b)

Fig. 6: Scatter plots that compare label-conditional post-softmax class predictions (e.g. $p(W|\text{label} = W)$ of fully supervised classifiers (y-axes) vs classifiers trained on unsupervised encodings (x-axes) (a), and the corresponding pre-softmax counterparts (b). Given the more linear relation between both models' pre-softmax predictions (as opposed to post-softmax predictions), it can be implied that the non-linear softmax behaves in different regimes for both models, a direct result of the difference in pre-softmax ranges.

from both models. Despite the similar trend, differences could still be noted (indicated with the horizontal red bars), e.g. low amounts of N2 or N3 were sometimes predicted by the unsupervised model, while the supervised counterpart suppressed these low contributions. This difference in spread of predicted probabilities over the classes was also reflected in average entropy of predictions over the full test set, which was found to be $\text{H} = 0.30 \pm 0.06$ for the supervised, and $\text{H} = 0.42 \pm 0.09$ for the unsupervised model. Note that the unsupervised model is still able to predict abrupt transitions (e.g. at 1.2 hours), so it can not simply be considered a smoothened version of the supervised prediction. In terms of hard metrics on the full test set, the supervised model (bal. acc. $= 82.0 \pm 0.05\%$, $\kappa = 79.0 \pm 0.08\%$) performed better than the unsupervised model (bal. acc. $76.0 \pm 0.09\%$; $\kappa = 71.3 \pm 0.13\%$).

Figure 6a plots the class-conditional probabilities - i.e. $p(\text{N3}|\text{label} = \text{N3})$ etc. - for both models against each other for the full test set. Each dot represents one 30 seconds window from one subject. It can be seen that conditional probabilities (i.e. post-softmax or post-σ values, visualized in the upper row) for Wake, N2 and REM sleep, tend to be higher for the supervised model, something that can also visually be seen in the hypnodensities depicted in fig. 5. Moreover, the histograms clearly indicate distinct probabilistic prediction behavior for N1, compared to the other stages, as probabilities close to one were rarely assigned by either of the models.

C. Pre- vs post-softmax predictions

Interestingly, comparing the *pre*-softmax predictions between both models (see fig. 6b), it can be seen how both models predicted

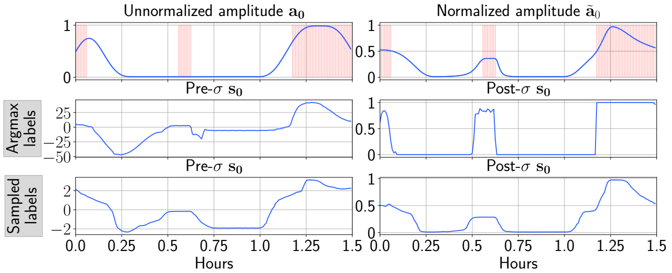


Fig. 7: Pre- and post-softmax predictions for signal s_0 by supervised models, trained with argmax (middle row) or sampled labels (bottom row), of the same example as given in fig. 4. The pink markers indicate times where signal 0 is the source with the largest amplitude (i.e. the locations where the argmax label denotes source 0).

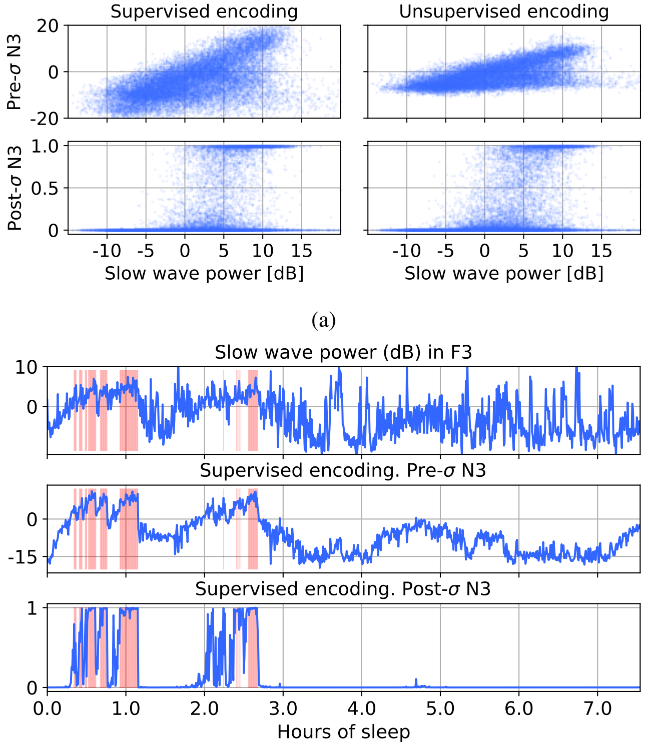
.

more similar values, compared to the post-softmax counterparts. This indicates that the non-linear softmax activation, has a different effect on both models. We demonstrate the effect of the softmax function, using synthetic data, in fig. 7. This figure shows that the softmax function operates in a (more) linear regime when the range of pre-softmax values is small (here for sampled labels), while it operates in a highly non-linear regime for a larger input range (here for argmax labels). From fig. 6b we see how the range of pre-softmax values for supervised vs unsupervised encoding also differs (mainly visible for REM and N3), therefore contributing to differences in post-softmax probability predictions between the models.

As discussed in section II-G, we can use the amount of slow wave power in the frontal EEG lead as a proxy for the (continuous) contribution of deep sleep (N3). Figure 8a visually compares both pre- and post-softmax predictions to slow wave power for the full test set. It can clearly be seen that the former better follow a linear relation with the slow wave power, than the latter. For both models, a tail is visible in the pre-softmax scatter plots (upper row), where a low value for N3 is predicted, while high delta power was computed. A recheck confirmed that this tail was not caused by a low-quality measurement for one of the patients in the test set, but rather was present for multiple patients. A possible explanation can be that low-frequency content, slightly above 0.5 Hz (so included in the delta range), enters the spectrum during wake episodes as a consequence of movement artifacts. Figure 8b depicts the pre/post-softmax predictions for N3 and delta power over time, for the same subject as depicted in fig. 5. In this subject we indeed also found an episode of high slow wave power (at 3.7 hours), which was annotated as Wake. This figure also clearly shows how (similarly as in the synthetic setup) the softmax outputs tended to follow the annotations (in pink), whereas the pre-softmax outputs better captured the continuity of deep sleep.

IV. Discussion

In this work, we investigated different aspects that contribute to post-softmax probabilities of a convolutional classification model for automatic sleep staging. Recent work [1] proposed to use these probabilities to reveal information from nocturnal recordings, which might be dismissed in the hypnogram (in which only one sleep stage is selected per data window). When plotted over time, these probabilities are referred to as a hypnodensity (graph). We foresee great potential for the hypnodensity to move sleep medicine to a new era, as it opens up a legion of research directions about sleep disorders that are known to be related to sleep stage dissociations.



(a)

(b)

Fig. 8: (a) The frontal EEG delta power against the predicted pre-softmax values, and post-softmax probabilities, over the full test set. The pre-softmax predictions correspond much better with slow wave power for both supervised and unsupervised encodings. (b) Illustrative example that shows the pre- and post-softmax predictions over time for the supervised model. The pink lines indicate windows that were annotated as N3. The pre-softmax prediction better follows the continuity of slow wave power, while the post-softmax prediction better matches the annotations for classification.

However, to correctly leverage hypnodensity graphs in future researches, it is of crucial importance to clearly understand their meaning. In order to shed light on the interpretation, a controlled set of experiments was performed using synthetic data, followed by analyses on real PSG recordings. A signal model was introduced, in which each channel comprised a non-linear measurement of (underlying) signals, that each aggregate characteristics belonging to one of the different sleep stages. The presence or contribution of these signals was denoted with the (normalized) mixture coefficients. In our annotation model, a set of scoring rules (e.g. the AASM standard) and application thereof, finally results in a label distribution, representing the probability for any of the sleep stages to be scored (see fig. 1). It should be remarked that, even though the synthetic data were inspired by PSG recordings, these data were simplified and subject to design choices, hampering guarantees regarding full generalizability to real data. Nevertheless, it has helped in investigating the interpretation of the hypnodensity graph, of which conclusions are discussed below.

We found that a hypnodensity, predicted by a supervised classification model, reflected the label distribution (see section I). In practice, it thus represents the scoring disagreement in the data set used for training the model, or in other words, the probability with which a given data window was classified as being one of the different sleep stages. This finding is line with the observation that the hypnodensity graph corresponds well with *inter*-rater disagreement across multiple scorers [1]. Given the fact that the (supervised) hypnodensities on real PSG data showed to have a non-zero entropy on average, we conclude that the label distribution in our dataset, which was annotated by one (experienced) scorer, also yielded non-zero entropy, implying the presence of *intra*-rater variability as well. We can interpret this as if a scorer (unconsciously) ‘samples’ from an (implicit) distribution over possible stages, analogously to the synthetic case where discrete labels were sampled from the distribution defined by the mixture coefficients.

We make a critical note regarding our conclusion that post-softmax predictions reflect the label distribution, as from the field of uncertainty research, it has been known that softmax outputs of (very) deep neural networks tend to become poorly calibrated or overconfident [17], [18], resulting in predictions that are of lower entropy than the corresponding label distribution. Nevertheless, our (rather shallow) convolutional architecture did not show to suffer from this overconfidence, seen from the fact that the normalized mixture coefficients were predict almost perfectly (see fig. 3a and 4), when the labels of synthetic data were sampled from a distribution, parameterized by these coefficients. It is, however, something to take into account when designing deeper neural networks for hypnodensity predictions.

Moreover, it is important to realize that a (supervised) hypnodensity directly reflects the annotation quality of the data set used for training the model. E.g. a group of less experienced technicians is likely to have rather high disagreement in their scorings, resulting in higher-entropy hypnodensities, when the model is trained on their annotated data compared to using annotations from more experienced technicians. On the other hand, from synthetic experiments it was found that training the full encoding unsupervised, gave rise to predictions that were more label-agnostic, and had more tendency to predict data characteristics (i.e. the mixture coefficients in our signal model, see fig. 4). For the real PSG data, differences were also found between predicted hypnodensities from supervised and unsupervised encodings, where the unsupervised predictions yielded higher entropy (also visible in fig. 5). In line with the synthetic results, we thus conclude that an unsupervised hypnodensity might reveal characteristics that directly relate to the data themselves, which might be absent in the (more label-dependent) supervised hypnodensity. Nevertheless, a ground truth for the mixture coefficients belonging to either of the sleep stages is unavailable in practice, hampering direct validation of the (unsupervised) hypnodensity.

In an attempt to get some more insights in the ‘ground-truth’ underlying mixture coefficients in PSG data, the problem could also be considered a variant of independent component analysis (ICA) [19], which is a demixing technique that guarantees identifiability of the underlying sources (up to permutations and scaling), under certain assumptions. The sources should be independent and linearly-mixed, the number of measurements should at least equal the number of sources, and maximally one source may be Gaussian distributed. Thanks to the multi-channel setup of an EEG recording, ICA has been a popular technique to decompose the recording in EEG, ECG, EOG, and or EMG components [20]–[23], but direct application to predict the mixture coefficients seemed inappropriate for three reasons. First, the mixture of sleep-stage dependent signal characteristics is likely non-linear, second, we can not be certain that the resolved signals indeed correspond to the five (abstract) sleep stage signals, and third, ICA is unidentifiable with respect to scaling, so considering the amplitude ratios of the resolved signals does not guarantee to reveal the true mixture ratio.

Luckily, the quest for demixing algorithms that relax (some of) the aforementioned assumptions of ICA has spurred research efforts. The authors of [24] e.g. empirically showed that *single-channel* ICA can approximately separate different sources, provided that these sources are reasonably spectrally disjoint. More recently, efforts have been made in proving identifiability for *non-linear* ICA, under additional assumptions on the sources (e.g. temporal non-stationarities [25], temporal dependencies [26], or temporal sparsity [27]), or availability of an auxiliary variable (e.g. class labels), which was researched in a VAE [28], contrastive learning [29], and hidden Markov Model [30] setting. The CPC model [13] - as used in our experiments for unsupervised encoding - belongs to the class of contrastive learning methods, and might be a suitable candidate for (approximately) solving non-linear ICA, given its empirical successes so far [14], [31], and recent results that show how contrastive learning objectives invert data generating processes [32]. In the current work, it was observed how predictions of a classifier trained on unsupervised (CPC) encodings, in some cases depended on the number of training iterations (see section II), pushing the hypnodensity more towards the label distribution, and further away from predicting data characteristics. This finding opens up a new research direction, in which one may investigate how CPC embeddings can be converted to AASM class probabilities, without relying (again too much) on the labels during classifier training.

In our last experiment we looked at the pre-softmax (opposed to post-softmax) predictions, to see whether these contained information regarding the continuous dynamics (e.g. mixture coefficients over time) of a measurement. It was found that both in the synthetic case (see fig. 7) and for real data (see fig. 8b), these pre-softmax class predictions indeed showed much more gradual/continuous patterns over time, compared to their post-softmax counterparts. In fact, the pre-softmax predictions for N3 (for both types of encoding) were found to correlate better with slow wave power in the frontal EEG lead, than the post-softmax predictions (fig. 8a). We conclude, that while the non-linear effect of the softmax function might be desirable in a classification problem, it (partially) discards continuous patterns in the data that might be of medical relevance as well. Other work has presented a finding in the same direction [33], by showing that class separability (related to discreteness) in an automatic sleep staging model highly increased after the last softmax activation. In future work, one might therefore consider other normalizing functions in the classifier, that may replace this (non-linear) softmax function [34].

In this research, we did not research influences of design choices like data window length (which was fixed to 30 s), depth and width of the models, and the number and type of measurement channels used. Regarding the latter, the authors of [33] visually depicted (supervised) hypnodensities predictions on one EEG channel only, and showed that predictions did not drastically differ dependent on the chosen channel. However, visual inspection seemed to reveal that their single-channel hypnodensities yielded higher entropy than the presented hypnodensities in this work and in [1]. We leave it for future research to investigate the relation between the number of recorded channels and the entropy of predicted hypnodensities.

To conclude, the hypnodensity graph as proposed by [1] (i.e. predicted by a supervised model), informs us about the distribution from which (discrete) labels were (implicitly) drawn by the technicians that annotated the data set used for training the model. We found that, despite the revelation of ample information in a hypnodensity (with respect to a hypnogram), potentially relevant information gets lost in the former, due to the final (non-linear) normalizing softmax layer and the label-dependency of supervised training. On the other hand, consideration of pre-softmax class predictions and/or unsupervised sleep staging models, might finally empower the sleep medicine community with a better understanding about the underlying dynamics of the sleeping brain.

APPENDIX I

We here show the results of an additional synthetic experiment that tests whether the post-softmax probabilities of our supervised classifier indeed follow the label distribution. To that end, we create three different sets of *discrete* labels, on which we train three separate models. The label sets are all sampled from the label distribution as defined in eq. (2), with $\tau = [1, \frac{1}{2}, \frac{1}{4}]$. For all models, we subsequently compute the KL divergence between the soft predictions (i.e. the post-softmax class probabilities) and the distributions, defined by different values of τ.

Figure 9 shows a heat map of these results. The x-axis denotes the value of τ of the distribution from which (discrete) labels were drawn during training, and the y-axis indicates the τ of the distribution that is used to compare the predictions against (by means of the KL divergence). This cross-comparison shows that the KL divergences are lowest when the evaluation distribution matches the label distribution (seen from the low values on the diagonal).

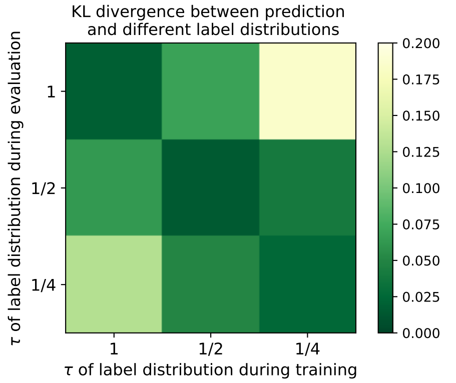
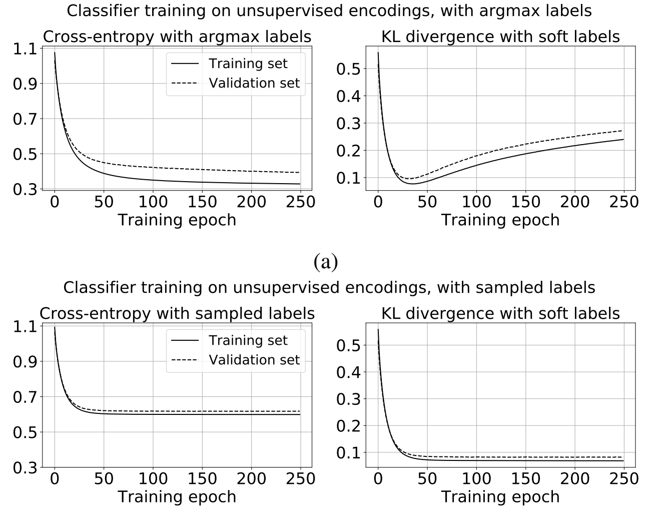


Fig. 9: Cross-comparison of KL divergences (evaluated over the full synthetic test set) between post-softmax model predictions and different evaluation distributions. The different models were trained with labels that were sampled from a distribution with distinct values for τ. A low KL divergence (i.e. dark green) indicates that the model's prediction on average corresponds well with the evaluation distribution.

APPENDIX II

In this section, we visualize how classification and mixture coefficient prediction may be opposite objectives when the label generating distribution is distinct from the distribution, defined by the mixture coefficients. When training a classifier on an unsupervised encoded latent space, using argmax labels, it can be seen from fig. 10a how the cross-entropy with these labels continues to decrease during training, while the KL divergence with the soft labels (which equal the mixture coefficients) diverges. In this situation the label- and mixture distribution were different, and predicting both at the same time was hampered. The training epoch where the final model is selected thus influences both the hard and soft metrics, as given in fig. 3b. Note that, even though the KL divergence starts to diverge again, it is still far lower than the reported KL divergence of 2.7 for the supervised model in fig. 3a.

On the other hand, when using labels that are sampled from the distribution defined by the mixture coefficients, it can be seen from fig. 10b that the classification task and mixture prediction task did not counteract anymore. Both cross-entropy with the sampled discrete labels, and KL divergence with the soft labels converged. Note that the initial phase of classifier training with argmax labels, results in a minimum KL divergence that is on par with the KL divergence found using sampled labels. Since the KL divergence between predictions and ground truth soft labels is unknown in practice, all models in this work were selected based on lowest validation (cross-entropy) loss with the discrete labels.



(a)

(b)

Fig. 10: Progression of cross-entropy with discrete labels (i.e. the training objective), and KL divergence with the ground truth soft labels during training a classifier on unsupervised embeddings, with argmax (a) and sampled (b) labels.

REFERENCES

[1] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. Sorensen, P. Jennum, and E. Mignot, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications*, vol. 9, no. 1, pp. 1–15, 2018. [Online]. Available: http://dx.doi.org/10.1038/s41467-018-07229-3

[2] M. H. Kryger, T. Roth, and W. C. Dement, "Principles and practice of sleep medicine fifth edition," 2011.

[3] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, B. V. Vaughn *et al.*, "The AASM manual for the scoring of sleep and associated events," *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, vol. 176, p. 2012, 2012.

[4] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia," *Computer Methods and Programs in Biomedicine*, vol. 176, pp. 81–91, 2019.

[5] L. Fiorillo, A. Puiatti, M. Papandrea, P. L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, and F. D. Faraci, "Automated sleep scoring: A review of the latest approaches," p. 101204, 12 2019.

[6] P. Chriskos, C. A. Frantzidis, C. M. Nday, P. T. Gkivogkli, P. D. Bamidis, and C. Kourtidou-Papadeli, "A review on current trends in automatic sleep staging through bio-signal recordings and future challenges," *Sleep Medicine Reviews*, vol. 55, p. 101377, 2021. [Online]. Available: https://doi.org/10.1016/j.smrv.2020.101377

[7] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring," *Journal of clinical sleep medicine*, vol. 9, no. 1, pp. 81–87, 2013.

[8] L. Nobili, L. De Gennaro, P. Proserpio, F. Moroni, S. Sarasso, A. Pigorini, F. De Carli, and M. Ferrara, "Local aspects of sleep: Observations from intracerebral recordings in humans," *Progress in Brain Research*, vol. 199, pp. 219–232, 2012.

[9] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The pittsburgh sleep quality index: a new instrument for psychiatric practice and research," *Psychiatry research*, vol. 28, no. 2, pp. 193–213, 1989.

[10] C. M. Morin, G. Belleville, L. Bélanger, and H. Ivers, “The insomnia severity index: psychometric indicators to detect insomnia cases and evaluate treatment response,” *Sleep*, vol. 34, no. 5, pp. 601–608, 2011.

[11] R. P. Snaith, “The hospital anxiety and depression scale,” *Health and quality of life outcomes*, vol. 1, no. 1, pp. 1–4, 2003.

[12] I. Goodfellow, *Deep Learning*, 2018, vol. 12, no. 8.

[13] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” 7 2019. [Online]. Available: http://arxiv.org/abs/1807.03748

[14] H. Banville, O. Chehab, A. Hyvärinen, D. A. Engemann, and A. Gramfort, “Uncovering the structure of clinical EEG signals with self-supervised learning,” *Journal of Neural Engineering*, pp. 1–32, 2020.

[15] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[16] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *ICML*, 2017, pp. 1321–1330.

[18] D. Ulmer and G. Cinà, “Know your limits: Monotonicity & softmax make neural classifiers overconfident on OOD data,” in *UAI*, 2020.

[19] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[20] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski, “Removing electroencephalographic artifacts by blind source separation,” *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.

[21] C. A. Joyce, I. F. Gorodnitsky, and M. Kutas, “Automatic removal of eye movement and blink artifacts from EEG data using blind component separation,” *Psychophysiology*, vol. 41, no. 2, pp. 313–325, 2004.

[22] W. Zhou and J. Gotman, “Removal of EMG and ECG artifacts from EEG based on wavelet transform and ICA,” in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1. IEEE, 2004, pp. 392–395.

[23] F. Poree, A. Kachenoura, H. Gauvrit, C. Morvan, G. Carrault, and L. Senhadji, “Blind source separation for ambulatory sleep recording,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 2, pp. 293–301, 2006.

[24] M. E. Davies and C. J. James, “Source separation using single channel ICA,” *Signal Processing*, vol. 87, no. 8, pp. 1819–1832, 8 2007.

[25] A. Hyvärinen and H. Morioka, “Unsupervised feature extraction by time-contrastive learning and Nonlinear ICA,” in *Advances in Neural Information Processing Systems*, no. Nips, 2016, pp. 3772–3780.

[26] A. Hyvarinen and H. Morioka, “Nonlinear ICA of temporally dependent stationary sources,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 2017.

[27] D. Klindt, L. Schott, Y. Sharma, I. Ustyuzhaninov, W. Brendel, M. Bethge, and D. M. Paiton, “Towards nonlinear disentanglement in natural data with temporal sparse coding,” in *ICLR*, 2021.

[28] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen, “Variational Autoencoders and Nonlinear ICA: A Unifying Framework,” in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 2207–2217.

[29] A. Hyvarinen, H. Sasaki, and R. E. Turner, “Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, vol. 89, 2018, pp. 859–868. [Online]. Available: http://arxiv.org/abs/1805.08651

[30] H. Hälvä and A. Hyvarinen, “Hidden Markov Nonlinear ICA: Unsupervised Learning from Nonstationary Time Series,” in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 939–948.

[31] O. Henaff, “Data-efficient image recognition with contrastive predictive coding,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 4182–4192.

[32] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel, “Contrastive Learning Inverts the Data Generating Process,” *arxiv*, 2021. [Online]. Available: http://arxiv.org/abs/2102.08850

[33] P. Krauss, C. Metzner, N. Joshi, H. Schulze, M. Traxdorf, A. Maier, and A. Schilling, “Analysis and Visualization of Sleep Stages based on Deep Neural Networks,” *Neurobiology of sleep and circadian rhythms*, vol. 10, no. May, p. 100064, 2021. [Online]. Available: https://doi.org/10.1101/2020.06.25.170464

[34] A. De Brébisson and P. Vincent, “An exploration of softmax alternatives belonging to the spherical loss family,” in *ICLR*, 2016, pp. 1–9.