Computers in Biology and Medicine 171 (2024) 108205



Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed





U-PASS: An uncertainty-guided deep learning pipeline for automated sleep staging

Elisabeth R.M. Heremans [a,*], Nabeel Seedat [b], Bertien Buyse [c], Dries Testelmans [c], Mihaela van der Schaar [b], Maarten De Vos [a,1]

[a] KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[b] University of Cambridge, Cambridge CB2 1TN, UK
[c] UZ Leuven, Department of Pneumology, Herestraat 49, B-3000 Leuven, Belgium

ARTICLE INFO

Keywords:
Uncertainty estimation
Deep learning
Automatic sleep staging
Electroencephalography

ABSTRACT

With the increasing prevalence of machine learning in critical fields like healthcare, ensuring the safety and reliability of these systems is crucial. Estimating uncertainty plays a vital role in enhancing reliability by identifying areas of high and low confidence and reducing the risk of errors. This study introduces U-PASS, a specialized human-centered machine learning pipeline tailored for clinical applications, which effectively communicates uncertainty to clinical experts and collaborates with them to improve predictions. U-PASS incorporates uncertainty estimation at every stage of the process, including data acquisition, training, and model deployment. Training is divided into a supervised pre-training step and a semi-supervised recording-wise finetuning step. We apply U-PASS to the challenging task of sleep staging and demonstrate that it systematically improves performance at every stage. By optimizing the training dataset, actively seeking feedback from domain experts for informative samples, and deferring the most uncertain samples to experts, U-PASS achieves an impressive expert-level accuracy of 85% on a challenging clinical dataset of elderly sleep apnea patients. This represents a significant improvement over the starting point at 75% accuracy. The largest improvement gain is due to the deferral of uncertain epochs to a sleep expert. U-PASS presents a promising AI approach to incorporating uncertainty estimation in machine learning pipelines, improving their reliability and unlocking their potential in clinical settings.

1. Introduction

Machine learning has ushered in a new era in healthcare, providing opportunities for remote monitoring and computer-aided diagnosis. Although machine learning models are performing with human-expert-level accuracy in controlled tasks [1–6], machine learning models have seen limited integration into clinical practice. The challenges to embedding these models in healthcare have shifted to providing guarantees of reliability, explainability and estimates of uncertainty [7,8]. These factors are essential to build trust with clinicians and promote adoption of machine learning in clinical settings [9–11].

Uncertainty estimates are needed at all stages of the machine learning workflow: in data acquisition, at training time, and at deployment time. In data acquisition and in training, it is important to evaluate the usefulness of the information that is utilized to train the machine learning model. Feeding too much data in a machine learning model is sub-optimal both financially and in terms of model performance [12,13]. Medical data is often expensive to acquire, requiring specialized equipment to measure each parameter, and expert knowledge to interpret the data. Moreover, feeding uninformative data to the model can deteriorate the performance because it adds noise without adding information [14]. The same principles apply to active learning [15,16], a human-in-the-loop training strategy in which the model actively queries an expert to label extra data points which would be most informative to improve the model. At deployment time, typical deep learning models always make a prediction, no matter how uncertain they are. Such systems are not usable in high-stakes medical applications: if the machine learning model does not indicate when a prediction is uncertain, a clinician cannot rely on it, as all outputs are equally likely to be incorrect [17]. Consequently, we propose that

* Corresponding author.
E-mail addresses: elisabeth.heremans@kuleuven.be (E.R.M. Heremans), maarten.devos@kuleuven.be (M.D. Vos).

[1] This research is funded by a Ph.D. fellowship from the Research Foundation - Flanders (FWO), Belgium for E. Heremans (FWO project number 1SC2921N), by the Cystic Fibrosis Trust, UK for N. Seedat and by the Flemish Government, Belgium (AI Research Program) for M. De Vos.

https://doi.org/10.1016/j.compbiomed.2024.108205
Received 16 October 2023; Received in revised form 16 February 2024; Accepted 20 February 2024
Available online 23 February 2024
0010-4825/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

uncertainty estimation provides a principled approach to solving the aforementioned challenges in order to successfully integrate a machine learning workflow in the clinic. We demonstrate how an uncertainty-guided machine learning pipeline can help in the challenging task of sleep staging, a clinical annotation task characterized by high levels of uncertainty, and known for its relatively low inter-rater agreement [10,18].

The gold standard for diagnosing sleep-wake disturbances is by performing a polysomnography (PSG) study, which involves a fully monitored overnight stay in the hospital. Various physiological signals are recorded, including electroencephalography (EEG), electro-oculography (EOG) and electromyography (EMG) signals. These full-night recordings are segmented into 30-second segments, which are then labeled by trained clinicians. Each segment is labeled as one of five sleep stages according to clinical standards [19,20]. This so-called sleep staging process (see Fig. 1a) is labor-intensive and subject to both intra- and inter-rater variability [21], so automated approaches have been researched to help humans in this task. Numerous deep learning methods for sleep staging have recently been developed [3–6]. These state-of-the-art sleep staging models achieve a similar performance to human experts. However, state-of-the-art models are mostly validated on healthy adults, and typically perform poorly on diseased subjects [10,21]. Disease influences brain activity and, consequently, PSG recordings, creating uncertainty in the data. Moreover, models trained on data acquired with one measurement setup or device may not generalize well to data obtained using a different measurement protocol, resulting in the well-known distribution shift problem [22–27]. Hence, measurement factors can also contribute to uncertainty. Leveraging uncertainty estimation tools can help identify these uncertainties, which are likely to cause poor-quality predictions. As such, uncertainty estimates enhance model transparency and allow to actively query a sleep expert to label uncertain samples. Hence, by integrating uncertainty estimation into sleep staging, our uncertainty-guided deep learning pipeline for sleep staging (U-PASS) improves both trust and performance of the model.

A variety of uncertainty estimation methods have been studied in the context of machine learning and deep learning [28,29]. Hüllermeier et al. [29] stressed the importance of distinguishing between two types of uncertainty, aleatoric and epistemic, and discussed attempts to quantify these in machine learning. These reviews [28,29], like most work in uncertainty, focused on estimating uncertainty at deployment time. A number of recent works have also investigated the concept of uncertainty within sleep staging. Guillot et al. [30] investigated machine learning algorithm performances in comparison to inter-rater agreement among human experts, and found that machine learning algorithms trained on multi-scored datasets can outperform clinical experts. Stephansen et al. [31] introduced hypnodensity graphs to report sleep stages along with their probabilities, as an alternative to the widely used hypnograms. These hypnodensity graphs were shown to correlate with the average scoring of expert panels [32,33], and sleep staging models were shown to benefit from training on probabilistic, 'smoothed', labels based on these average scores [34]. Alternative methods to measure uncertainty in sleep stage scoring were based on the entropy of the output predictions [35,36], the margin between the probability output of the most likely and second-most likely class [37], and Monte-Carlo dropout [38]. These works addressed the important issue of uncertainty quantification in sleep staging, but were largely focused on post-hoc uncertainty quantification at deployment time, and did not discriminate between different types of uncertainty. Going further than quantifying uncertainty, Kang et al. [36] explored whether sleep experts could successfully correct the labels of uncertain samples. A few studies also investigated uncertainty in the context of active learning in sleep staging [39]. MacAs et al. [37,40] explored an uncertainty-based active learning approach for sleep staging, and Grimova et al. [41] proposed weighing uncertainty with density measures.

In contrast to previous works mentioned here, we integrate different types of uncertainties throughout the whole machine learning pipeline. The goal of this work is to take a leap forward, leveraging the uncertainty estimates to maximize performance and reliability of machine learning models and to combine them with human capabilities. Our framework aims to improve predictions at every stage of the machine learning process. It should be model-agnostic, as it needs to be applicable to various clinical machine learning models and problems. We fulfill these goals by building an uncertainty-guided machine learning pipeline. Our pipeline follows a data-centric approach, and integrates uncertainty into data acquisition, training, active learning, and model deployment. We show how this greatly improves machine learning predictions and overall model reliability at different levels:

1. we improve the model through the quality of the training data,
2. we finetune the model to make better predictions on difficult examples using feedback from the clinician, and
3. we reduce mistakes by deferring predictions on ambiguous data to the clinician.

Combining the capabilities of AI and clinicians, our pipeline represents a crucial step towards the integration of machine learning in clinical settings.

2. Methods

2.1. Data and study cohorts

This study uses PSG recordings of 90 patients [23]. It was recorded at the sleep laboratory of the University Hospitals Leuven (UZ Leuven) from January 2021 to September 2022. We included all patients over 60 years of age with suspicion of sleep apnea. This is a common sleep disorder in which breathing repeatedly stops during sleep. For each patient, the PSG recording includes EOG, EMG and a total of 6 EEG channels, all measured at 500 Hz. In this study, we only use the C4-A1, F4-A1 and O2-A1 EEG derivations, as well as the chin EMG and the left–right EOG derivation, all resampled at 100 Hz. 45 recordings are used as a train set, and 45 recordings are used as the test set and for active learning. The dataset was manually annotated by a certified expert sleep scorer according to the AASM standard [19]. The sleep study was conducted in accordance with the Declaration of Helsinki, and the protocol with registration number S64190/B3222020000148 was approved by the Ethics Committee Ethische Commissie Onderzoek UZ/KU Leuven.

2.2. Model and training

U-PASS can be applied to any machine learning model and problem, as we designed it to be model agnostic, with no requirements or adaptations to the network architecture or training procedure. In this paper, we use it in a sleep staging context with a state-of-the-art sleep staging model, SeqSleepNet [3]. Taking into account the temporal context to accurately classify segments, this deep neural network uses a sequence-to-sequence classification scheme. It transforms a sequence of adjacent raw data segments into a corresponding sequence of outputs, in this case sleep stages. The inputs are presented as time–frequency images, spectrograms of each 30-second segment. The sequence length is a parameter, which is fixed to 10 in this study. SeqSleepNet is a hierarchical architecture, composed of a block of layers that processes individual segments, and a block that works at the sequence level. The segment processing block comprises a number of frequency filters, followed by a recurrent neural network (RNN) on the segment level. The outputs from the segment processing block are then concatenated into a sequence of inputs for the sequence processing block. A sequence-level RNN transforms this input sequence into an output sequence, which is mapped to a sequence of sleep stages by a softmax layer. The optimization and training parameters in this study are identical to those in the original paper [3]: L2 regularization is applied, and the Adam optimizer is used with a learning rate of 10^{-4}.

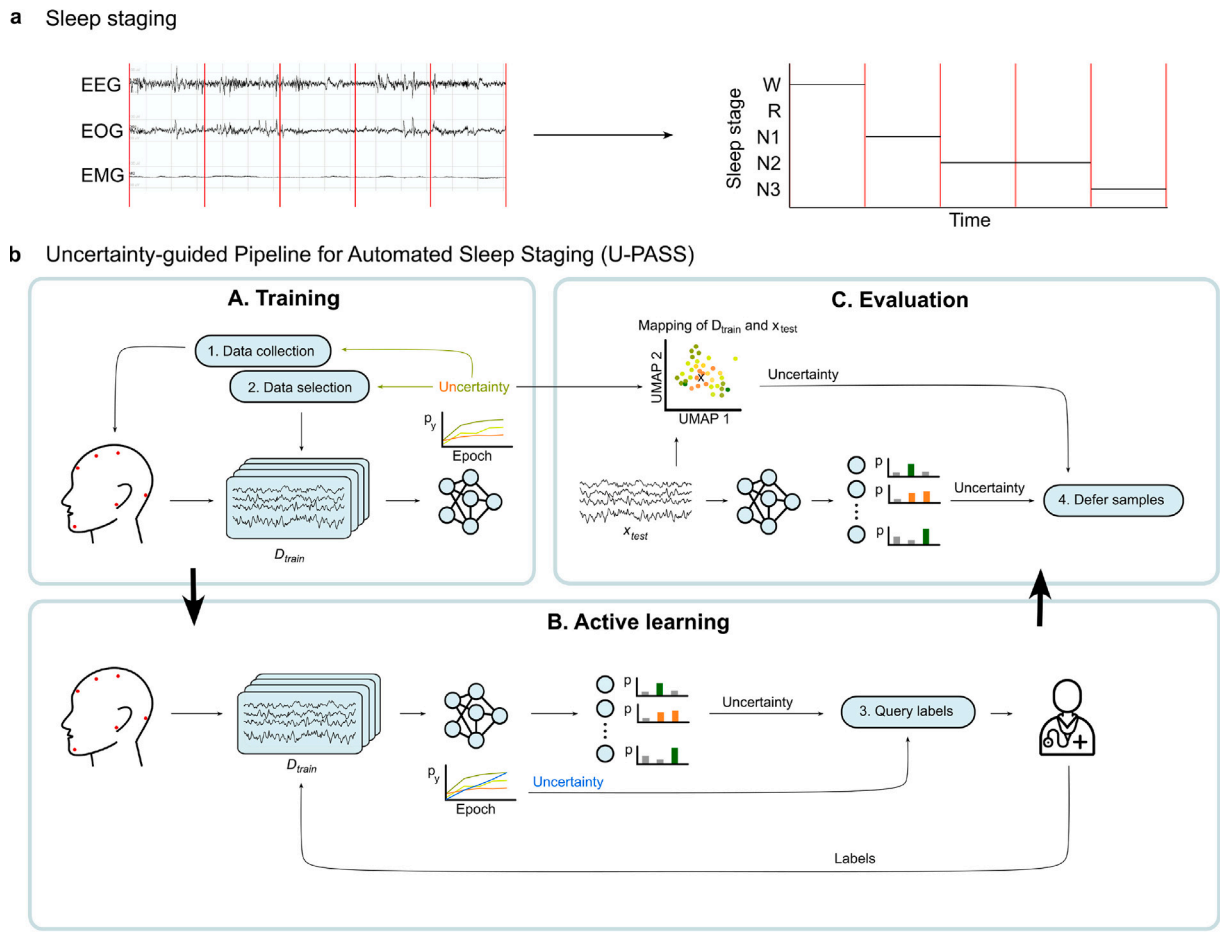


Fig. 1. (a) Illustration showing the classification problem of sleep staging. The recorded physiological signals are segmented in 30-second segments. Each segment is labeled as one of five sleep stages by SeqSleepNet [3], a deep neural network architecture. (b) Full uncertainty-guided pipeline for sleep staging (U-PASS). The pipeline is more generally applicable to machine learning for clinical applications. **A.** At training time, data uncertainty is tracked to (1) select which channels (or features) to use, and (2) remove uninformative data. **B.** After training, active learning is used to finetune the model. Model uncertainty is tracked to (3) identify the most informative samples to query. **C.** At evaluation or deployment time, the most uncertain samples are deferred to the clinician (4). Used symbols: p is the model output, p_y is the model output for the true class y, D_{train} is the training dataset and x_{test} is a test sample.

2.3. Formulation of uncertainty

U-PASS utilizes insights from two recent methods that examine the characterization of uncertainty in training data. These methods analyze the training dynamics, which refers to the behavior of individual samples during the model training process. Data Maps [42] maps instances by tracking predictive confidence and model uncertainty (epistemic uncertainty) during training. Data-IQ [43] on the other hand maps instances using predictive confidence and data uncertainty (aleatoric uncertainty). Both methods use mappings to stratify the training dataset into subgroups of easy, hard and ambiguous data. We further refer to instances with high epistemic uncertainty as *model-ambiguous* samples. These are examples for which the prediction variability of the model is highest during training. They are important to improve the performance and generalizability of a model [42]. On the other hand, instances with high aleatoric uncertainty are further referred to as *data-ambiguous* samples. These are samples for which predictions are the most uncertain on average, and on which the model systematically underperforms [43].

Let us formally define the characterization of model-ambiguous samples and data-ambiguous samples. Consider a typical supervised learning setting for a classification problem, where inputs $x \in \mathbb{R}^d$ need to be assigned to classes $y \in \mathbb{N}$. The goal is to learn a model f_θ which maps inputs to outputs by assigning a probability to each class given the input: $f_\theta(x) = P(Y|X = x, \Theta = \theta)$. Let θ designate a certain instantiation of the model parameters. The model is iteratively trained for e^* training epochs, resulting in e^* different instantiations of the model parameters $\{\theta_1, \theta_2, \ldots, \theta_{e^*}\}$. Let $\Theta \sim P_{emp}(\{\theta_1, \theta_2, \ldots, \theta_{e^*}\})$ be a random variable with an empirical distribution over the model parameters throughout the training process. $\mathcal{P}(x, \theta) = [f_\theta(x)]_y$ is the model's output probability for the true class label. Using the variance as an uncertainty metric, we can decompose the uncertainty by the law of total variance. The epistemic uncertainty of the prediction is $v_{ep} = \mathbb{V}_\Theta[\mathcal{P}(x, \Theta)]$. The variance is evaluated over the model outputs during training, hence epistemic uncertainty captures the variability of model outputs, representing model uncertainty. The aleatoric uncertainty is $v_{al} = \mathbb{E}_\Theta[\mathcal{P}(x, \Theta)(1-\mathcal{P}(x, \Theta))]$ [43]. In the aleatoric uncertainty, the variance is computed over the prediction. Hence, this represents the data uncertainty. Using the empirical distribution of Θ, the uncertainties are approximated as follows [43]:

$$v_{ep} = \frac{1}{e^*} \sum_{e=1}^{e^*} (\mathcal{P}(x, \theta_e) - \overline{\mathcal{P}}(x))^2, \tag{1}$$

$$v_{al} = \frac{1}{e^*} \sum_{e=1}^{e^*} \mathcal{P}(x, \theta_e)(1 - \mathcal{P}(x, \theta_e)), \tag{2}$$

with $\overline{\mathcal{P}}(x) = 1/e^* \sum_{e=1}^{e^*} \mathcal{P}(x, \theta_e)$.

We define model-ambiguous data as data with a high v_{ep}, and data-ambiguous samples as data with high v_{al}. In making these two groups, we use the definitions of 'ambiguity' proposed in Data Maps [42] and

Data-IQ [43], respectively. We can further stratify the training dataset using the concept of predictive confidence, defined in both works as the mean output probability for the ground-truth-class:

$$c = \overline{P}(x), \tag{3}$$

i.e. the confidence in the correct class. This allows us to define easy-to-classify samples as instances with high confidence and low ambiguity and hard-to-classify samples as instances with low confidence and low ambiguity.

2.4. U-PASS pipeline

Based on the uncertainty quantification methods from Section 2.3, we build a full uncertainty-guided pipeline, outlined in Fig. 1b. **A.** First, we tailor and improve the training dataset. Even before collecting a full dataset, we can start training a model using various measured signals and monitor the data uncertainty (2) to pinpoint which signals are effective in decreasing uncertainty. This information can aid in deciding which modalities and channels should be recorded. If a particular signal does not lower the data uncertainty, it should be excluded. Furthermore, there is a trade-off to be made between the cost of data acquisition and the amount of uncertainty that can be tolerated to achieve a clinically useful result. After choosing a measurement setup, a full dataset is collected. We can then train a model on this dataset, and track the data uncertainty to identify the most data-ambiguous samples. These samples are removed from the training dataset, as they confuse the model instead of improving its performance. **B.** After training, the second step in U-PASS is to finetune the model using active learning. The model uncertainty can be used to identify for which recordings finetuning is the most useful, i.e. which recordings have the highest learning potential. Indeed, model uncertainty measures how much predictions of the model vary during training. In this work, we achieve finetuning with a semi-supervised transfer learning technique, using the additional labels supplied through active learning. **C.** Lastly, we can use uncertainty at evaluation or test time. A simple uncertainty measure based on uncertainty of the neighboring training samples can be used to identify uncertain test samples. Instead of making an incorrect prediction on these samples, it is better to defer them to a clinician. The three principled uncertainty-guided steps in U-PASS result in a reliable and robust machine learning pipeline that is tailored for clinical environments. This framework enables effective interaction between clinicians and the model, thus enhancing clinical applicability. We delve into each next.

2.4.1. Data collection and data selection

In the data collection experiments, we train the model on different numbers of channels to identify the best measurement setup, i.e. the measurement setup with the lowest data uncertainty. In the data selection experiments, we train on the best channels based on the data collection experiment, removing different amounts of ambiguous training data in each experiment. In both the data collection and data selection experiments, the model is trained on the full training dataset for 10 training epochs. Early stopping is applied using a validation dataset consisting of 2 recordings, retaining the model with the best performance on the validation set. As the training dataset consists of 45 recordings, each experiment is repeated $\lfloor 45/2 \rfloor = 23$ times, with different recordings in the validation set for every repeat. Accuracies are computed on the separate test set consisting of 45 recordings. The test set accuracies and training set uncertainty measures reported in Fig. 3a are computed as averages over these 23 repeats.

2.4.2. Active learning

The active learning experiments start from the model obtained with the optimal setting based on the channel and data selection experiments. Using active learning to query informative labels, the model is then finetuned to personalize it to individual patients with the semi-supervised adversarial domain adaptation approach from [23]. The training set of 45 patients is used as the source dataset, and each selected individual recording of the test set as the target dataset for adversarial domain adaptation. This experiment is performed 23 times, with each of the trained models (see 2.4.1). The labels for the semi-supervised training approach are acquired through the following active learning scheme. Before every training epoch, 1% of samples (roughly 10) are queried. The labels of these samples are used during that training epoch, along with the labels cumulatively acquired in previous epochs. The basic unsupervised adversarial domain adaptation is thus augmented by adding a supervised loss from the queried labels. The final model obtained after 10 training epochs with querying is retained for evaluation. As the same recordings are unavoidably used for training and testing in this active learning personalization step, evaluation is only performed on the unlabeled portion of those recordings.

The model uncertainty is a measure for how much the model performance varies while training on certain data, and can hence be seen as a measure for learning potential. Therefore, we focus the labeling efforts of the clinician on the recordings with the highest model uncertainty. Section 2.3 describes how to estimate model uncertainty in (1) using the variance of the prediction for the correct class as an uncertainty metric. This is applicable specifically for training data that has ground truth labels available. In our active learning scenario, labels are not available. However, we can still analyze training dynamics to estimate uncertainty, using entropy as an alternative, unsupervised uncertainty metric [44,45]. The definitions of epistemic (model) and aleatoric (data) uncertainty with this uncertainty metric are the following:

$$v_{al}^{unsup} = \mathbb{E}_{\Theta}[H[P(y|x,\Theta)]] \tag{4}$$

$$v_{ep}^{unsup} = H[\mathbb{E}_{\Theta}[P(y|x,\Theta)]] - \mathbb{E}_{\Theta}[H[P(y|x,\Theta)]] \tag{5}$$

with H the entropy metric, and $P(y|x,\Theta)$ the output distribution. The expectations are again approximated with averages over the model parameters as in (1) and (2). We train the model on every recording using unsupervised adversarial domain adaptation, estimate the model uncertainty v_{ep}^{unsup}(5) based on the training dynamics in this unsupervised training step, and select the recordings with the highest v_{ep}^{unsup}(5) for active learning. The querying strategy to select the samples that are most informative is based on the same uncertainty estimate, but any other querying strategy could also fit in the U-PASS pipeline. We shall further refer to this approximation of model uncertainty (5) as the unsupervised model uncertainty, to make a distinction with the model uncertainty defined in (1).

2.4.3. Deferring ambiguous test samples

In the evaluation experiments, we use different models. For the selected recordings with the highest model uncertainty, we use the personalized models obtained in the active learning experiments. Evaluation is performed on the unlabeled portion of those recordings. For the other recordings, we use the optimal non-retrained models obtained from step A in Fig. 1. As uncertainty metrics based on training dynamics cannot be used at deployment time, the evaluation experiments rely on post-hoc uncertainty metrics. Several post-hoc uncertainty metrics are compared, including two that are based on the output of the model: the entropy of the output distribution and the maximum output probability (i.e. the output probability of the predicted class). Two other metrics are distance-based. One metric measures the distance of the test sample to its n closest train samples, while the other uses the distance of the test sample to the n closest training samples for each separate class, and is calculated as the proportion of the smallest distance to the second-smallest distance. The remaining three metrics calculate uncertainty by averaging metrics of the n closest training samples, weighted by their distance to those training samples. One metric uses the average confidence (3) of the closest training samples, the second uses the

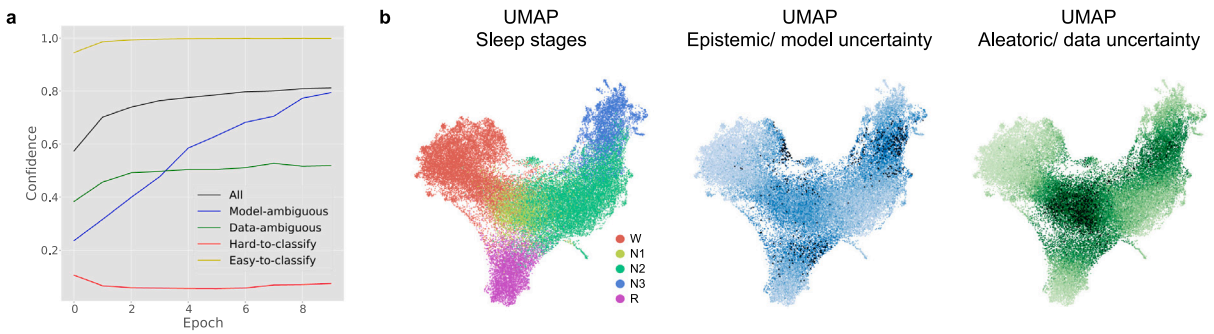


Fig. 2. (a) The average training dynamics for the 1% most data-ambiguous, model-ambiguous, easy-to-classify and hard-to-classify samples. (b) UMAP visualization of the feature distribution resulting from training a sleep staging network. The left plot shows the sleep stages in this distribution. In the middle plot, the tint indicates the model uncertainty, and the 1% most model-ambiguous are shown in black. In the right plot, the tint indicates the data uncertainty, and the 1% most data-ambiguous are shown in black.

average data uncertainty (2), and the third uses the average model uncertainty (1).

Every uncertainty metric for evaluating uncertainty on the test set has its own scale, and it is up to the domain experts to define what level of uncertainty and how many mistakes can be tolerated. Therefore, this experiment ranks test set samples from most to least uncertain, and plots the accuracy corresponding with the z% most uncertain test samples. The uncertainty measures are then compared by averaging these accuracies over the 23 repeats.

3. Results

3.1. Insights from uncertainty

First, we illustrate how the two types of uncertainty are manifested in our data in Fig. 2. Fig. 2a shows the training dynamics for the 1% most model-ambiguous and data-ambiguous samples, as well as for the 1% most easy-to-classify and hard-to-classify samples. Model-ambiguous samples show the greatest learning curve during training, while data-ambiguous samples are characterized by a flat learning curve. The distributions of the two types of ambiguity are visualized in Fig. 2b. The figure shows two-dimensional UMAP embeddings of the training data, with the different colors indicating the five sleep stages in the left figure, and heatmaps showing the model and data uncertainty in the middle and right figure, respectively. In both heatmaps, the 1% most ambiguous samples are highlighted in black. The 1% most model-ambiguous samples are found mostly at the borders and outside the main training data distribution, as well as on class boundaries. Indeed, atypical and out-of-distribution samples are more challenging for the model at first, but the model can learn these samples through more training. The most data-ambiguous samples are centered in sleep stage N1, which is the least well defined and hardest to classify. Data uncertainty is also higher on the class boundaries, where some samples contain characteristics of multiple sleep stages.

3.2. Data collection

U-PASS starts with evaluating the data uncertainty during data collection to assess which signals should be acquired and used. Fig. 3a shows the confidence and data uncertainty in the PSG training datasets, mapped by using Data-IQ [43]. Each 30-second sample is one dot, and the density of samples is represented by the brightness of colors. The figure demonstrates that most samples are in the upper left corner, characterized by high confidence and low data uncertainty (easy-to-classify). Some samples have high data uncertainty (data-ambiguous), and a minority of samples is characterized by low data uncertainty and low confidence (hard-to-classify). The concentration of samples in the data-ambiguous region decreases when going from one channel to three channels, and decreases again from three channels to five channels. This is also evident from the average data uncertainty over the whole training dataset, which is shown under the plots. The average confidence increases accordingly, as samples go from being data-ambiguous to easy-to-classify. As a result, the accuracy on the test set also improves, which is the final proof that reducing the data uncertainty is helping the outcome.

We conclude that tracking the data uncertainty allows to compare the quality of different measurement setups for sleep monitoring. Indeed, a decrease in data uncertainty with the addition of a channel indicates that this channel adds valuable information. It can advise users in selecting the optimal setup that will result in the best accuracy at deployment time.

3.3. Data selection

We can curate our training dataset, by discarding from it the most data-ambiguous training samples. Fig. 3a shows how removing these ambiguous data affects the model's test performance on an independent test set. There is a trade-off to be made between removing samples that confuse the model and removing too many samples so that useful information is discarded from the training dataset. In this case, the optimal point is reached when discarding 1% of training samples.

Data uncertainty in sleep staging can arise from biological sources (e.g. age, pathology) and sources related to the measurement technique (e.g. segmentation, interference) [18], as well as from ambiguity in sleep scoring rules. Fig. 4a shows that data uncertainty per segment is significantly higher for segments on sleep stage transitions than for segments that are not on sleep stage transitions. This source of uncertainty is hence related to the discretization caused by segmentation and the continuous nature of sleep, with features from multiple sleep stages in a single segment. Fig. 4b shows that wake segments have the lowest data uncertainty, and N1 segments have the highest data uncertainty. The fact that different stages of sleep are characterized by different levels of uncertainty is related to how clearly sleep stages are recognized and defined, which is influenced by sleep scoring rules and the biology of sleep. We find evidence for this in the fact that the inter-rater variability between human scorers shows the same trends as data uncertainty [46]. This is further discussed in Section 4 and shown more formally with an experiment on a multi-scored dataset in Appendix C.

3.4. Active learning

Once the model is trained on the improved training set, U-PASS evaluates unsupervised model uncertainty to assess on which recordings the model should be finetuned through active learning. We use active learning to personalize sleep staging models to those individual recordings. Fig. 3b shows the average improvement from active learning to individual PSG recordings. The improvement is shown for the selected 40% recordings with the highest unsupervised model

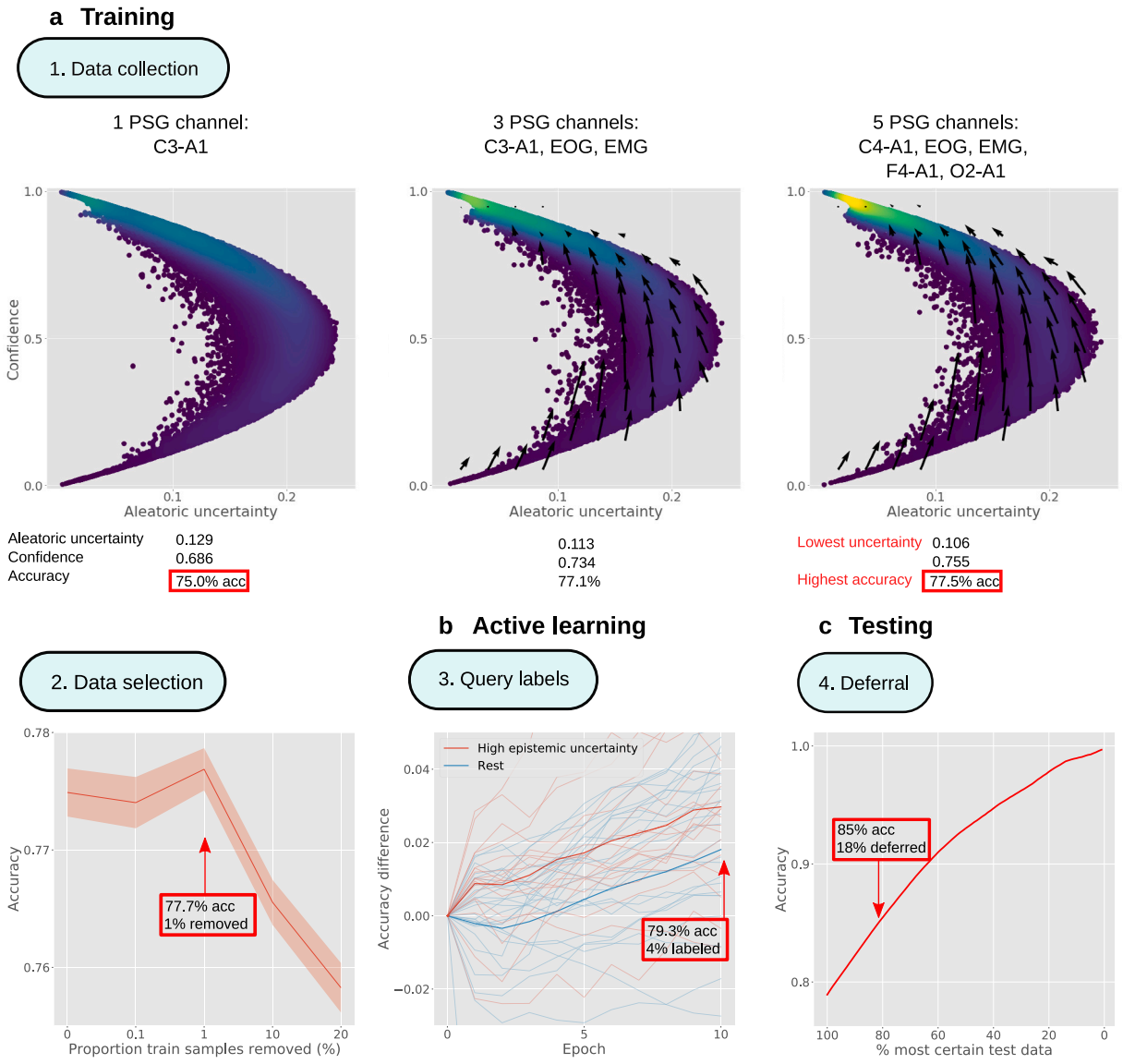


Fig. 3. U-PASS applied to hospital-based polysomnography (PSG) data. The resulting accuracy after every consecutive step is indicated with a red box. (a) During training, the training dataset is tailored using data collection and data selection based on data uncertainty. In the data collection step, samples move to the upper-left corner with high confidence and low data uncertainty when more channels are used. We thus proceed with all five channels. In the data selection step, the optimal test result is achieved when removing the one percent (roughly 440) most data-ambiguous training samples. (b) Active learning is used to adapt the model to individual recordings. 1% of samples is queried in each of the 10 training epochs. We only perform active learning on recordings characterized by high model uncertainty. On the rest of the recordings, it does not show a large improvement. (c) At deployment time, uncertainty-based deferral increases the test accuracy to the desired level of 85%. Samples with uncertainty values below the threshold are referred to a clinician for manual labeling.

uncertainty and for the rest of the recordings. Unsupervised model uncertainty clearly points to the recordings that benefit the most from the active learning adaptation. Fig. 5 shows two examples of how active learning improves the predictions in both labeled and unlabeled samples.

3.5. Deferring uncertain test samples

Lastly, the models are deployed on unseen data. For the 40% selected recordings with highest unsupervised model uncertainty, we use the personalized models. The other recordings are evaluated with the optimal non-retrained models from the data collection and data selection step. U-PASS then uses uncertainty measures to defer test samples to an expert when predictions are uncertain. Fig. 3c shows the accuracy on the 100% to 1% most certain test samples. This is equivalent to the accuracy when removing 0% to 99% of the most uncertain test samples. The higher this accuracy, the better the uncertainty estimation, as the uncertainty metric should retain the most accurate predictions and defer the least accurate ones. Different post-hoc uncertainty metrics are tested. Fig. 3c only shows the best-performing one: the weighted average confidence of the n closest training samples. The performances of all the uncertainty metrics are shown in Appendix A.

The results in Fig. A.1 show that different uncertainty metrics perform best depending on the threshold set on the tolerated uncertainty. If we require 85% accuracy, the confidence of neighboring training samples achieves the best performance, with 18% of samples deferred. For reference, 85% accuracy is a great performance, as the inter-rater agreement in sleep staging is estimated to be 82.6% [46,47]. When we exclude samples that lie on sleep stage transitions, the accuracy improves from 85% to 89%.

As clear from Fig. A.1, the difference between the entropy metric and metrics based on the uncertainty in the training set is mostly quite small, and depends on the uncertainty threshold we set. However, the metrics based on uncertainty in the training set have an important

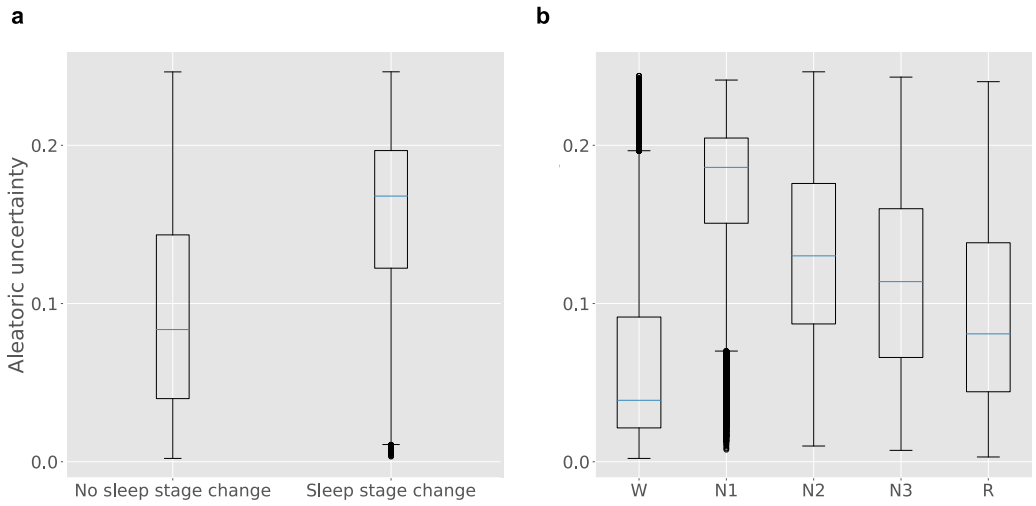


Fig. 4. Parameters causing data uncertainty in the PSG training dataset. (a) The data uncertainty per segment is higher for segments on sleep stage transitions than for segments that are not on sleep stage transitions. (b) The data uncertainty per segment, aggregated per sleep stage, shows that wake segments have low data uncertainty and N1 segments have the highest data uncertainty.

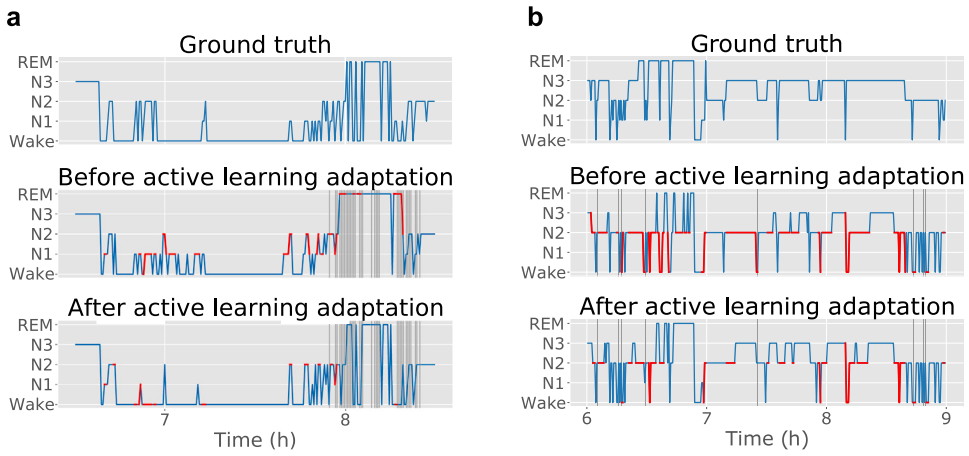


Fig. 5. Hypnograms (sleep stages over time) for snippets of two different recordings, demonstrating how personalization through active learning (AL) improves the performance. Queried samples are shown in gray, labeling errors of the sleep staging model are shown in red. (a) Before AL, the sleep staging model incorrectly labels a lot of segments as N1 sleep around hour 7, and the jumps between REM-sleep and Wake around hour 8 are missed. After AL, most of these labeling errors are corrected. Although the queried samples are centered around hour 8, the mistakes in N1-sleep around hour 7 are partly corrected too. (b) Both the episode of REM-sleep between hour 6.5 and 7, and the episodes of N3-sleep between hour 7 and 8.5 are more accurately predicted after AL. Overall, we see that both the queried samples and the unlabeled parts of the recording are labeled more accurately after AL in both examples.

advantage compared to the decision entropy, as they provide a reason for deferral. Indeed, the distance-weighted confidence of neighboring training samples is a product of the distance from the training samples and the uncertainty of close training samples. Fig. 6 visualizes both factors. The Spearman correlation coefficient shows that accuracy is negatively correlated with distance from the closest training samples ($r(43) = -.40$, $p = .0067$) and positively correlated with the confidence of the closest training samples ($r(43) = .51$, $p = .00032$). The two factors detect, respectively, out-of-distribution samples, and samples lying close to uncertain regions, e.g. decision boundaries.

The effect of applying U-PASS for sleep staging is summarized in the red boxes of Fig. 3, which show the resulting accuracy after each step in the pipeline. This clearly demonstrates how U-PASS leverages uncertainty in training, active learning and deployment to improve the accuracy at every stage. Appendix B includes an additional experiment on the Dreem Open Datasets [30], showing that our pipeline also generalizes to other datasets.

4. Discussion

We have developed and validated U-PASS, an uncertainty-guided pipeline for automated sleep staging. Although we applied U-PASS to sleep staging, the pipeline is generally applicable to any machine learning problem. It is particularly useful for machine learning in healthcare, which has strong requirements for reliability, trust, and benefits from user interaction. By utilizing data uncertainty and model uncertainty at different stages in the data processing pipeline (training, active learning and deployment) and interacting with a clinician, we have shown how U-PASS improves predictions step by step. It can easily be tailored to the accuracy needs of any particular application, by adapting the amount of data queried and deferred in the active learning and deployment phase, respectively.

We have validated our pipeline in PSG data of a real patient population of elderly subjects with suspicion of sleep apnea. Many performant deep learning models have been developed for sleep staging on PSG,

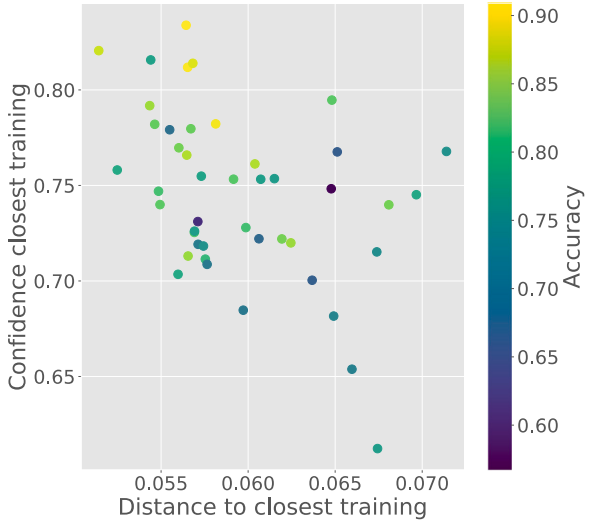


Fig. 6. The accuracy on the test data depends on both the distance from the training data and the uncertainty of those training data. For every patient, this figure shows the accuracy, confidence of the closest training samples and the distance to those training samples. These metrics are averaged over all the samples of the patient's recording, so every patient is represented by one dot.

but they are developed as static, stand-alone input–output machines. This limits their performance and utility in the clinic. Our contribution consists of building a pipeline in which such deep learning models can be integrated, boosting their performance and allowing for the clinician to choose the desired performance level. To showcase the broad applicability of our approach and its resilience in scenarios involving domain shifts, we conducted additional validation of our pipeline on a multi-scored, openly accessible sleep dataset (Appendices B and C).

The performance gains achieved by applying U-PASS to a state-of-the-art sleep staging model learning are shown in Fig. 3. The first two parts of the pipeline, data collection and data selection, are focused on tailoring the training dataset to achieve optimal results. Both steps decrease the data uncertainty in different ways. In the data collection step, data uncertainty is estimated to find out which features (or channels in the case of sleep staging) need to be collected. Once the data acquisition protocol is fixed and the training set is acquired, the data selection step filters out the data with the highest data uncertainty, as these can deteriorate model performance. In our experiments, the data collection step shows the largest improvement by going from one EEG channel to five channels. The data selection attains its maximum improvement when 1% of the training data are discarded. This step only increases the accuracy by a modest relative 0.3%. Despite the minor magnitude of the improvement, we systematically observe improvements with data pruning in other datasets. Appendix B shows the result on a public dataset. The current data selection approach is still rather rudimentary and may benefit from further refinement. First, it does not account for class imbalance. In the case of sleep staging, this particularly concerns class N1, the smallest class with samples often found on class boundaries. Caution is advised to avoid compromising N1 sensitivity and future work could explore balancing the removed samples across classes. Second, in the context of sleep staging, the current data selection approach does not consider temporal dependencies between consecutive epochs. It remains an open question how to optimize data selection for time series with temporal dependencies. Future research could investigate how to take these into account. Future investigations could explore the impact of various factors, such as data type, quality, and volume, as well as the model architecture, on the effectiveness of data selection.

The third part of the U-PASS pipeline includes finetuning a sleep staging model with an active learning approach. Active learning or any other form of additional training can resolve model uncertainty, which is caused by distribution shifts (related to patient populations or measurement equipment) or out-of-distribution data. Another and perhaps even more important source of model uncertainty is the difference in interpretations of scoring rules between different experts or sleep centers [48]. We have developed a recording-wise personalization step to handle out-of-distribution data, or 'difficult-to-score' recordings and samples. Adaptation to a new scorer is not the focus of this study, as our experiments are performed on a dataset labeled by a single scorer, but this could also be resolved through additional learning. Variability due to hardware and different patient populations is addressed in Appendix B, which shows that the pipeline also handles domain shifts between the training and test set.

The last step of U-PASS, deferring ambiguous samples, allows us to choose a tolerated level of uncertainty or a desired accuracy for a specific task. This comes at a labeling cost, similar to the active learning step of the U-PASS pipeline. The difference between the two steps is the type of uncertainty they tackle. Active learning can only resolve model uncertainty, coming from a lack of knowledge of the model. Pure data uncertainty, on the other hand, is inherent to the data and can hence not be resolved through learning. The only possible mechanism to cope with high data uncertainty at deployment time is deferral to an expert. In theory, doctors do not know more than a machine learning model in cases of pure data uncertainty. However, they can take the right course of action, whether that be ignoring the data, performing a new measurement, or fixing a broken electrode. We conclude that both active learning and deferring help to increase the accuracy to the desired level at a certain labeling cost. Since they tackle different types of uncertainty, they should be combined for optimal performance and utility.

Along with performance gains, the uncertainty estimation methods in the U-PASS pipeline also provide some interesting clinical insights into the sleep data. Fig. 2b and Fig. 4 shows that the data uncertainty in segments on sleep stage transitions is higher than for others, and that sleep stage N1 is characterized by more uncertainty than other sleep stages. This data uncertainty is by definition inherent to the data and cannot be resolved by more training. Hence, human expert annotations should exhibit the same data uncertainty, and indeed, our results are consistent with the sources of uncertainty identified in inter-rater variability studies [46]. In line with our results, previous studies on automated sleep staging found that hypnodensity plots, showing the model's output probabilities for all sleep stages, also correlated well with the average scoring of expert panels [31–33].

Sleep scoring rules are notoriously ambiguous, leading to high variability between scorers and sleep centers [21,48]. As such, the objectivity of uncertainty metrics can guide clinical practitioners and experts developing sleep scoring guidelines [19] to define better rules. For example, changing the 30-second segmentation to a more fine-grained segmentation should benefit the data uncertainty. Furthermore, the less ambiguous the sleep staging rules are, the less data uncertainty there is. An additional insight we have gained from plotting the sleep features (Fig. 2b) is that sleep stages do not seem to be discretely separated, but rather lie on a continuous spectrum. Recent works have advocated modeling sleep as a continuous process [49–51], which may be a more biologically accurate representation. Hence, uncertainty estimation methods can inform the medical field on how to best describe and define sleep.

In conclusion, U-PASS is a machine learning pipeline that integrates uncertainty-informed curation of the training set, uncertainty-based active learning to incorporate a clinician's feedback and deferral of uncertain decisions to a clinician. As such, all the facets of uncertainty and benefits of uncertainty estimation are harmonized in a single framework and combined with the clinician's capabilities. This optimizes the machine learning pipeline and unlocks its potential in a clinical setting by adding safety guardrails to the process. Not only does U-PASS give agency to clinicians, but it also has the potential to provide them with

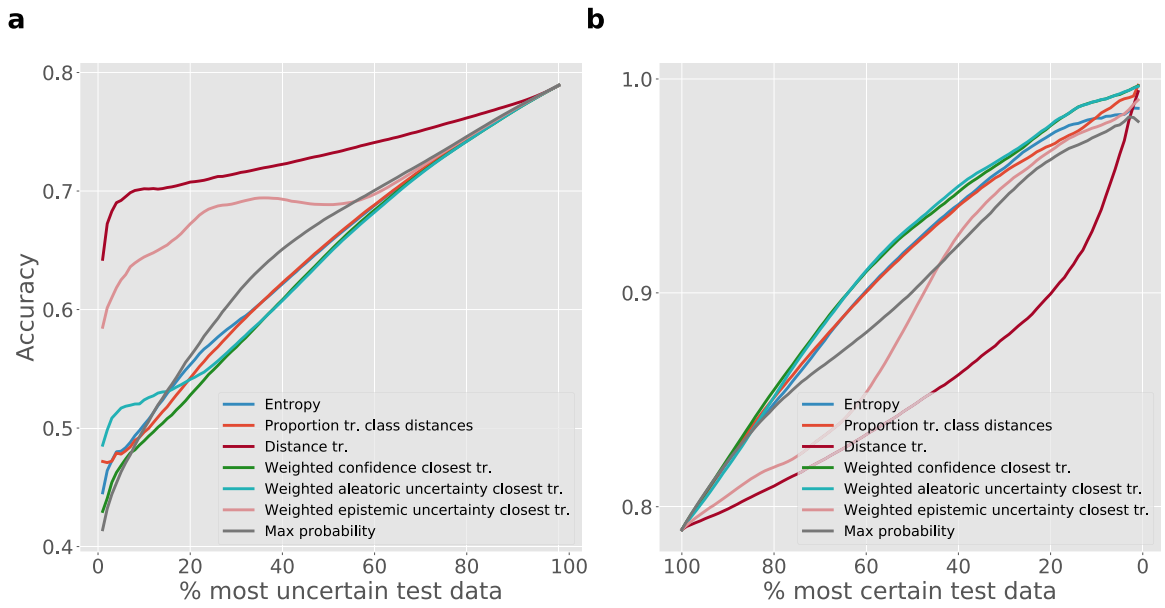


Fig. A.1. Extended results showing the performance of the different measures of uncertainty for deferring uncertain test samples. (a) The accuracy for the x% most uncertain samples. The better the uncertainty estimate, the lower this figure. (b) The accuracy for the x% most certain samples. The better the uncertainty estimate, the higher this figure. Depending on the threshold and scenario, different uncertainty measures perform best.

valuable insights into their data, offering a deeper understanding of sleep biology and potentially other clinical applications. Overall, U-PASS represents a promising approach to enhancing the reliability and safety of machine learning systems in critical fields such as healthcare.

CRediT authorship contribution statement

Elisabeth R.M. Heremans: Conceptualization, Data curation, Funding acquisition, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Investigation. **Nabeel Seedat:** Conceptualization, Methodology, Writing – review & editing. **Bertien Buyse:** Data curation, Supervision, Writing – review & editing. **Dries Testelmans:** Data curation, Supervision, Writing – review & editing. **Mihaela van der Schaar:** Conceptualization, Supervision, Writing – review & editing. **Maarten De Vos:** Conceptualization, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Extended deferral results

See Fig. A.1.

Appendix B. U-PASS in Dreem Open Datasets

The U-PASS pipeline is applied to the Dreem Open Datasets (DOD) [30] to demonstrate its generalizability in different datasets. The chosen model architecture for sleep staging is still SeqSleepNet, and the consensus scores of this multi-scored dataset are used to train the model. It should be noted that a potential improvement here could be to train the model on smoothed labels, using probabilities obtained by averaging the scores over the expert panel [34]. The results for this experiment are shown in Fig. B.1. For the data collection and data selection experiments (Fig. B.1a), the network architecture is trained on 28 randomly selected patients from DOD-O, the OSA cohort of the DOD. For the active learning personalization experiments (Fig. B.1b) and evaluation experiments (Fig. B.1c), we use the remaining 28 OSA patients from DOD-O and the 25 healthy participants from DOD-H. The DOD-H is a dataset with a different patient population, acquired with different equipment, at a different institution compared to the DOD-O dataset. This dataset is included for testing to verify that the framework still works in case there is a domain shift between the training set and test set.

The results on the DOD are in line with the results on the dataset from Leuven (Fig. 3), with a few interesting differences. In the data collection experiment (Fig. B.1a, left), we only test a maximum of three channels, because these are the only ones that are present in both DOD-O and DOD-H. In the active learning experiments (Fig. B.1b), the recording-wise personalization improves the accuracy more for the DOD-H patients than for the DOD-O patients. This is in line with our expectations, as active learning and other forms of training resolve model uncertainty, which is expected to be higher in DOD-H. Differences in measurement equipment and in patient characteristics typically lead to domain shift, and, consequently, model uncertainty and lower accuracies. The average unsupervised model uncertainty proves that epistemic uncertainty is indeed larger in the DOD-H dataset. At test time (Fig. B.1c), we decide not to defer any samples from the DOD-O dataset, because the accuracy is near-perfect (greater than 85%) after the active learning step. For the DOD-H dataset, we defer 7% of samples to get to 85% accuracy. We also test what would happen if we choose an uncertainty threshold on the DOD-O data and apply it to the DOD-H data. When selecting an uncertainty threshold that results in 90% accuracy (with a deferral of 13% of samples) for the DOD-O test data, this threshold results in a deferral of 15% of samples for the DOD-H dataset, with a resulting accuracy of 87.5% on the remaining data. Although the uncertainty threshold set in one dataset cannot guarantee the same accuracy in another dataset in case of domain shift, it still achieves a very decent result, with a final accuracy close to the one obtained in the original dataset. This shows that our pipeline is indeed robust to domain shift.

Appendix C. Inter-rater variability

In general, inter-rater disagreement results from a mix of aleatoric (data) and epistemic (model) uncertainty. A study by Van Gorp et al.

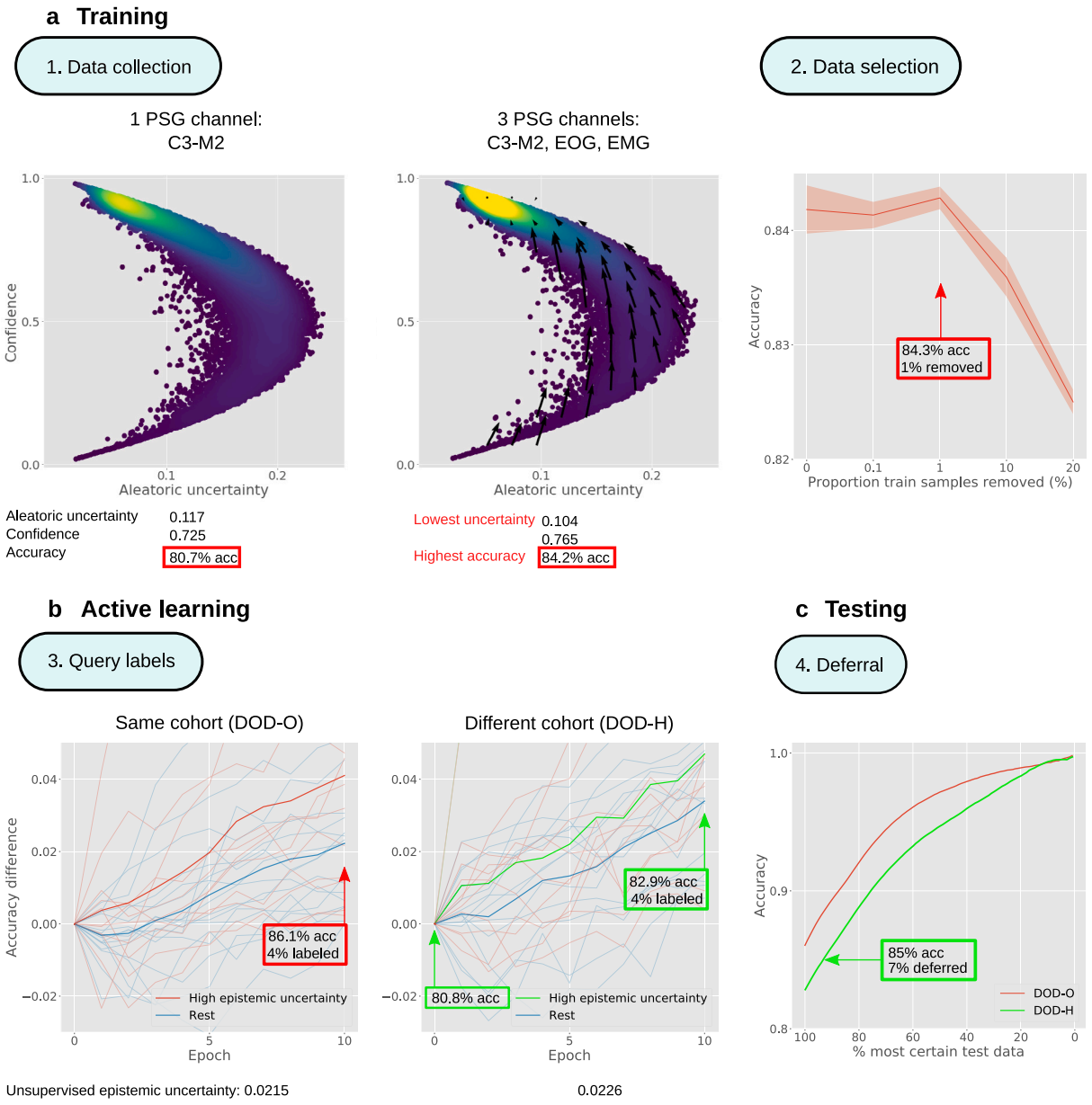


Fig. B.1. U-PASS applied to the Dreem Open Datasets (DOD). The resulting accuracy after every consecutive step is indicated with a box. (a) During training, the training dataset is tailored using data collection and data selection based on data uncertainty. In the data collection step, samples move to the upper-left corner with high confidence and low data uncertainty when more channels are used. We thus proceed with three channels. In the data selection step, the optimal test result is achieved when removing the one percent most data-ambiguous training samples. (b) Active learning is used to adapt the model to individual recordings. 1% of samples is queried in each of the 10 training epochs. Active learning shows a larger improvement on DOD-H than on DOD-O, which is as expected because of the domain mismatch. In both datasets, we only perform active learning on the 40% recordings characterized by high model uncertainty. (c) At deployment time, uncertainty-based deferral increases the test accuracy in DOD-H to the desired level of 85%. Samples with uncertainty values below the threshold are referred to a clinician for manual labeling. In the DOD-O dataset, this accuracy is already achieved without deferral, so we do not defer any samples.

[18] showed that both machine learning models and humans suffer from both aleatoric and epistemic uncertainty. However, assuming that all the raters are well-trained, certified scorers, the contribution of aleatoric uncertainty is higher than the contribution of epistemic uncertainty. When there is an ensemble of predictions, we can take the expectation (average) over the different predictions to estimate aleatoric uncertainty, and the variance over the predictions to estimate epistemic uncertainty. This could in theory be done with a panel of scorers as the 'ensemble', but unfortunately, human scorers do not give probabilities but one-hot predictions, which do not allow to calculate these estimates. Therefore, we can only get a measure of total uncertainty (aleatoric and epistemic) by taking the entropy or variance of the average prediction.

The Dreem Open Datasets were scored by 5 different scorers, allowing to quantify the total human uncertainty. Our training procedure for this dataset is explained in Appendix B. For the 28 training recordings, we compared our estimates of uncertainty with the inter-rater variability. The inter-rater variability was calculated as $v_{tot} = P^* \times (1 - P^*)$ where P^* is average rating of the majority voted sleep stage. We obtained per-recording averages of inter-rater variability and uncertainty estimates, and calculated Pearson's correlation coefficient between these values. The best match was found between inter-rater variability and the data uncertainty (2) estimated on our training data, with a correlation coefficient of r=0.64 (see Fig. C.1). The correlation coefficient obtained with model uncertainty (1) was r = 0.56. Model uncertainty is model-dependent, and so, it is not necessarily the same in machine learning

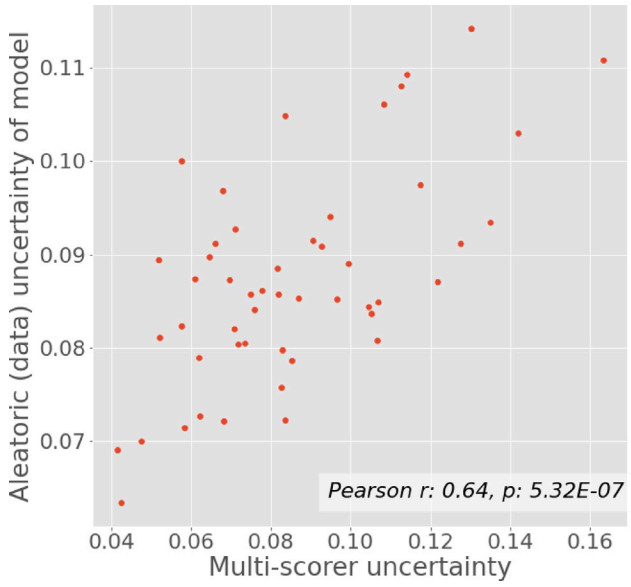


Fig. C.1. The relation between the data uncertainty ('aleatoric uncertainty') and the inter-scorer variability for 28 randomly selected recordings used for training U-PASS on the DOD-O dataset.

models and humans. In conclusion, although this experiment does not exclude the presence of epistemic uncertainty in the human scorers, it does confirm that data uncertainty is model-independent and thus correlates very well between human scorers and machine learning models.

References

[1] S. Yang, F. Zhu, X. Ling, Q. Liu, P. Zhao, Intelligent health care: Applications of deep learning in computational medicine, Front. Genet. 12 (2021) 444.

[2] J. Egger, C. Gsaxner, A. Pepe, K.L. Pomykala, F. Jonske, M. Kurz, J. Li, J. Kleesiek, Medical deep learning—A systematic meta-review, Comput. Methods Programs Biomed. 221 (2022) 106874.

[3] H. Phan, F. Andreotti, N. Cooray, O.Y. Chén, M. De Vos, SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (3) (2018) 400–410, http://dx.doi.org/10.1109/TNSRE.2019.2896659.

[4] H. Phan, O.Y. Chén, M.C. Tran, P. Koch, A. Mertins, M. De Vos, XSleepNet: Multi-view sequential model for automatic sleep staging, IEEE Trans. Pattern Anal. Mach. Intell. (2020) http://dx.doi.org/10.1109/TPAMI.2021.3070057, [Online]. Available: http://arxiv.org/abs/2007.05492.

[5] A. Guillot, V. Thorey, RobustSleepNet: Transfer learning for automated sleep staging at scale, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 1441–1451.

[6] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P.J. Jennum, C. Igel, U-Sleep: resilient high-frequency sleep staging, npj Digital Medicine 4 (1) (2021) 72, [Online]. Available: http://www.nature.com/articles/s41746-021-00440-5.

[7] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nat. Mach. Intell. 1 (1) (2019) 20–23, [Online]. Available: https://www.nature.com/articles/s42256-018-0004-1.

[8] S. Kundu, AI in medicine must be explainable, Nature Med. 27 (8) (2021) 1328, [Online]. Available: https://www.nature.com/articles/s41591-021-01461-z.

[9] A. Qayyum, J. Qadir, M. Bilal, A. Al-Fuqaha, Secure and robust machine learning for healthcare: A survey, IEEE Rev. Biomed. Eng. 14 (2021) 156–180.

[10] H. Phan, K. Mikkelsen, Automatic sleep staging: Recent development, challenges, and future directions, 2021, [Online]. Available: https://arxiv.org/abs/2111.08446v1.

[11] R. Miotto, F. Wang, S. Wang, X. Jiang, J.T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Brief. Bioinform. 19 (6) (2017) 1236–1246, http://dx.doi.org/10.1093/bib/bbx044.

[12] S. Yang, Z. Xie, H. Peng, M. Xu, M. Sun, P. Li, Dataset pruning: Reducing training data by examining generalization influence, in: The Eleventh International Conference on Learning Representations, 2023, [Online]. Available: https://openreview.net/forum?id=4wZiAXD29TQ.

[13] D. Montes de Oca Zapiain, M.A. Wood, N. Lubbers, C.Z. Pereyra, A.P. Thompson, D. Perez, Training data selection for accuracy and transferability of interatomic potentials, npj Comput. Mater. 8 (1) (2022) 1–9, [Online]. Available: https://www.nature.com/articles/s41524-022-00872-x.

[14] Y. Kwon, J. Zou, Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning, in: Conference on Artificial Intelligence and Statistics (AISTATS), 2022, [Online]. Available: https://arxiv.org/abs/2110.14049v2.

[15] B. Settles, Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009, [Online]. Available: http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

[16] V.L. Nguyen, M.H. Shaker, E. Hüllermeier, How to measure uncertainty in uncertainty sampling for active learning, Mach. Learn. 111 (1) (2022) 89–122, [Online]. Available: https://link.springer.com/article/10.1007/s10994-021-06003-9.

[17] H. Kilian, P. Lorenzo, V. der Plas Dries, M. Wannes, D. Jesse, K. Hendrickx, L. Perini, D. Van der Plas, W. Meert, J. Davis, Machine Learning with a Reject Option: A survey, 2021, [Online]. Available: https://arxiv.org/abs/2107.11277v1.

[18] H. van Gorp, I.A. M. Huijben, P. Fonseca, R.J. G. van Sloun, S. Overeem, M.M. van Gilst, C. author Hans van Gorp, Certainty about uncertainty in sleep staging: a theoretical framework, Sleep 2022 (2022) 1–8, [Online]. Available: https://academic.oup.com/sleep/advance-article/doi/10.1093/sleep/zsac134/6604464.

[19] R. Berry, R. Brooks, C. Gamaldo, S. Harding, R. Lloyd, S. Quan, M. Troester, B. Vaughn, AASM Scoring Manual Updates for 2017 (Version 2.4), J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med. 13 (2017).

[20] A. Kales, A. Rechtschaffen, A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects, United States Government Printing Office, Washington DC, 1968.

[21] N. Norori, Q. Hu, F.M. Aellen, F.D. Faraci, A. Tzovara, Addressing bias in big data and AI for health care: A call for open science, Patterns 2 (10) (2021) 100347.

[22] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.K. Kwoh, X. Li, C. Guan, ADAST: Attentive Cross-Domain EEG-Based Sleep Staging Framework With Iterative Self-Training, IEEE Trans. Emerg. Top. Comput. Intell. 7 (1) (2023) 210–221.

[23] E.R.M. Heremans, H. Phan, P. Borzée, B. Buyse, D. Testelmans, M. De Vos, From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging, J. Neural Eng. 19 (3) (2022) 036044, [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/ac6ca8 https://iopscience.iop.org/article/10.1088/1741-2552/ac6ca8/meta.

[24] H. Phan, O.Y. Chén, P. Koch, Z. Lu, I. Mcloughlin, A. Mertins, M. De Vos, Towards more accurate automatic sleep staging via deep transfer learning, IEEE Trans. Biomed. Eng. 68 (6) (2021) 1787–1798.

[25] S. Pathak, C. Lu, S.B. Nagaraj, M. van Putten, C. Seifert, STQS: Interpretable multi-modal Spatial-Temporal-seQuential model for automatic Sleep scoring, Artif. Intell. Med. 114 (2021) 102038.

[26] H. Seo, S. Back, S. Lee, D. Park, T. Kim, K. Lee, Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG, Biomed. Signal Process. Control 61 (2020) 102037.

[27] H. Sun, J. Jia, B. Goparaju, G.B. Huang, O. Sourina, M.T. Bianchi, M.B. Westover, Large-scale automated sleep staging, Sleep 40 (10) (2017) [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29029305/.

[28] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, V. Makarenkov, S. Nahavandi, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Inf. Fusion 76 (2020) 243–297, http://dx.doi.org/10.1016/j.inffus.2021.05.008, [Online]. Available: http://arxiv.org/abs/2011.06225.

[29] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Mach. Learn. 110 (3) (2021) 457–506, [Online]. Available: https://link.springer.com/article/10.1007/s10994-021-05946-3.

[30] A. Guillot, F. Sauvet, E.H. During, V. Thorey, Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (9) (2020) 1955–1965.

[31] 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, E. Mignot, Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy, Nature Commun. 9 (1) (2018) [Online]. Available: https://pmc/articles/PMC6283836/pmc/articles/PMC6283836/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6283836/.

[32] I.A. Huijben, L.W. Hermans, A.C. Rossi, S. Overeem, M.M. van Gilst, R.J. van Sloun, Interpretation and further development of the hypnodensity representation of sleep structure, Physiol. Meas. 44 (1) (2023) 015002, [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/aca641 https://iopscience.iop.org/article/10.1088/1361-6579/aca641/meta.

[33] P. Anderer, M. Ross, A. Cerny, R. Vasko, E. Shaw, P. Fonseca, Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing, Front. Sleep 2 (2023) 1163477.

[34] L. Fiorillo, D. Pedroncelli, V. Agostini, P. Favaro, F.D. Faraci, Multi-scored sleep databases: how to exploit the multiple-labels in automated sleep scoring, Sleep 46 (5) (2023) 1–12, http://dx.doi.org/10.1093/SLEEP/ZSAD028.

[35] H. Phan, K. Mikkelsen, O.Y. Chén, P. Koch, A. Mertins, M. De Vos, SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification, IEEE Trans. Biomed. Eng. 69 (8) (2021) 2456–2467, http://dx.doi.org/10.1109/TBME.2022.3147187, [Online]. Available: http://arxiv.org/abs/2105.11043.

[36] D.Y. Kang, P.N. DeYoung, J. Tantiongloc, T.P. Coleman, R.L. Owens, Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine, npj Digit. Med. 4 (1) (2021) 1–9, [Online]. Available: https://www.nature.com/articles/s41746-021-00515-3.

[37] M. MacAs, N. Grimova, V. Gerla, L. Lhotska, E. Saifutdinova, Active learning for semiautomatic sleep staging and transitional EEG segments, in: Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 2621–2627.

[38] L. Fiorillo, P. Favaro, F.D. Faraci, DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 2076–2085.

[39] J. Sebek, H. Schaabova, V. Krajca, Active learning approach for eeg classification using neural networks: A review, in: 2019 7th E-Health and Bioengineering Conference, EHB 2019, Institute of Electrical and Electronics Engineers Inc., 2019.

[40] M. Macaš, N. Grimová, V. Gerla, L. Lhotská, Semi-automated sleep EEG scoring with active learning and HMM-based deletion of ambiguous instances, Proceedings 31 (1) (2019) 46, [Online]. Available: https://www.mdpi.com/2504-3900/31/1/46.

[41] N. Grimova, M. Macas, V. Gerla, Active learning for semi-automated sleep scoring, in: N. Maglaveras, I. Chouvarda, P. de Carvalho (Eds.), Precision Medicine Powered By PHealth and Connected Health, Springer Singapore, Singapore, 2018, pp. 139–143.

[42] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N.A. Smith, Y. Choi, Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics, in: EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Association for Computational Linguistics (ACL), 2020, pp. 9275–9293, [Online]. Available: https://arxiv.org/abs/2009.10795v2.

[43] N. Seedat, J. Crabbé, I. Bica, M. van der Schaar, Data-IQ: Characterizing subgroups with heterogeneous outcomes in tabular data, in: A.H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022, [Online]. Available: https://openreview.net/forum?id=qC2BwvfaNdd.

[44] A. Olmin, On Uncertainty Quantification in Neural Networks: Ensemble Distillation and Weak Supervision (Ph.D. dissertation), Linköping University Electronic Press, 2022.

[45] M.H. Shaker, E. Hüllermeier, Aleatoric and epistemic uncertainty with random forests, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12080 LNCS, Springer, 2020, pp. 444–456, [Online]. Available: https://arxiv.org/abs/2001.00893v1.

[46] Y.J. Lee, J.Y. Lee, J.H. Cho, J.H. Choi, Interrater reliability of sleep stage scoring: a meta-analysis, J. Clin. Sleep Med. 18 (1) (2022) 193–202, [Online]. Available: https://jcsm.aasm.org/doi/10.5664/jcsm.9538 https://jcsm.aasm.org/.

[47] R.S. Rosenberg, S.V. Hout, The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring, J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med. 9 (1) (2013) 81, [Online]. Available: https://pmc/articles/PMC3525994//pmc/articles/PMC3525994/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3525994/.

[48] L. Fiorillo, G. Monachino, J. van der Meer, M. Pesce, J.D. Warncke, M.H. Schmidt, C.L. Bassetti, A. Tzovara, P. Favaro, F.D. Faraci, U-Sleep's resilience to AASM guidelines, npj Digit. Med. 6 (1) (2023) 1–9, [Online]. Available: https://www.nature.com/articles/s41746-023-00784-0.

[49] M. Younes, M. Ostrowski, M. Soiferman, H. Younes, M. Younes, J. Raneri, P. Hanly, Odds ratio product of sleep EEG as a continuous measure of sleep state, Sleep 38 (4) (2015) 641–654, [Online]. Available: https://academic.oup.com/sleep/article/38/4/641/2416969.

[50] M. Cesari, A. Stefani, T. Mitterling, B. Frauscher, S.V. Schönwald, B. Högl, Sleep modelled as a continuous and dynamic process predicts healthy ageing better than traditional sleep scoring, Sleep Med. 77 (2021) 136–146.

[51] L.W. Hermans, I.A. Huijben, H. van Gorp, T.R. Leufkens, P. Fonseca, S. Overeem, M.M. van Gilst, Representations of temporal sleep dynamics: Review and synthesis of the literature, Sleep Med. Rev. 63 (2022) 101611.