HHS Public Access

Author manuscript

IEEE Trans Biomed Eng. Author manuscript; available in PMC 2021 June 01.

Published in final edited form as:

IEEE Trans Biomed Eng. 2020 June ; 67(6): 1696–1706. doi:10.1109/TBME.2019.2943062.

Adaptive Sedation Monitoring from EEG in ICU Patients with Online Learning

Wei-Long Zheng [Member, IEEE],
Department of Neurology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, Massachusetts, USA.

Haoqi Sun,
Department of Neurology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, Massachusetts, USA.

Oluwaseun Akeju,
Department of Anaesthesia, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, Massachusetts, USA.

M. Brandon Westover
Department of Neurology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, Massachusetts, USA.

Abstract

Sedative medications are routinely administered to provide comfort and facilitate clinical care in critically ill ICU patients. Prior work shows that brain monitoring using electroencephalography (EEG) to track sedation levels may help medical personnel to optimize drug dosing and avoid the adverse effects of oversedation and undersedation. However, the performance of sedation monitoring methods proposed to date deal poorly with individual variability across patients, leading to inconsistent performance. To address this challenge we develop an online learning approach based on Adaptive Regularization of Weight Vectors (AROW). Our approach adaptively updates a sedation level prediction algorithm under a continuously evolving data distribution. The prediction model is gradually calibrated for individual patients in response to EEG observations and routine clinical assessments over time. The evaluations are performed on a population of 172 sedated ICU patients whose sedation levels were assessed using the Richmond Agitation-Sedation Scale (scores between −5 = comatose and 0 = awake). The proposed adaptive model achieves better performance than the same model without adaptation (average accuracies with tolerance of one level difference: 68.76% vs. 61.10%). Moreover, our approach is shown to be robust to sudden changes caused by label noise. Medication administrations have different effects on model performance. We find that the model performs best in patients receiving only propofol, compared to patients receiving no sedation or multiple simultaneous sedative medications.

corresponding author (M.B. Westover: mwestover@mgh.harvard.edu).

Keywords

Sedation monitoring; EEG; online learning; Richmond Agitation-Sedation Scale; intensive care unit; level of consciousness

I. Introduction

SEDATIVE medications are usually provided to critical ill patients in intensive care units (ICUs) to provide comfort and facilitate clinical care [1]. However, accurate sedation monitoring is challenging, especially for deeply sedated patients. Drug dosage and infusion rates are determined empirically in most cases without brain monitoring of sedation depth. Inappropriate sedation management can lead to oversedation and undersedation [2]. Oversedation is associated with prolonged mechanical ventilation, increased cost, increased duration of ICU stay, increased healthcare costs, and delirium [3]. Oversedation is particularly common, and previous studies suggest that improvements in sedation monitoring can improve ICU outcomes [2]. Conversely, undersedation is associated with pain, anxiety, and agitation. Therefore, accurate sedation monitoring is important for improving patient outcomes.

There are numerous scoring systems proposed for sedation assessment in clinical scenarios, e.g., the Ramsay scale, Glasgow Coma Scale (GCS), and Richmond Agitation-Sedation Scale (RASS) [4], [5]. Dale and colleagues found that implementing regular RASS assessments can decrease delirium, duration of mechanical ventilation, and ICU length of stay [6]. However, these sedation scoring systems have limitations. Because they are based on behavioral observations by medical staff, clinical sedation scoring systems can suffer from subjectivity. Further, it is usually only practical for staff to assess sedation levels every 1-2 hours at most, providing only intermittent snapshots of sedation depth.

Physiological monitoring is an attractive adjunct to clinical assessments of sedation levels, because it offers information that is both continuous and quantitative. Various approaches have been proposed based on the electroencephalogram (EEG). EEG enables the ability to non-invasively measure cortical activity patterns and is commonly used for seizure detection in the ICU [7]. The role of EEG signals for sedation monitoring in the ICU is less well established. Many quantitative EEG parameters have been proposed for monitoring depth of anesthesia in patients undergoing surgery, e.g., the bispectral index (BIS) and state entropy (SE) [8]-[11]. However, these methods were designed for evaluating general anesthetic depth in operating rooms and have not been validated for monitoring ICU patients likely because of between-patient variability afforded by various critical illnesses, medication effects, and significant interference from noise and artifacts [12]-[16]. Several groups including ours have previously explored EEG-based methods for determining the level of consciousness in ICU patients [17]-[20]. However, prior studies do not directly address the problem of robustness to between-patient differences in EEG signatures of sedation.

Many supervised learning algorithms have been utilized to solve healthcare problems [21]-[23]. Deep learning approaches in particular have been succesfully applied to diverse problems in EEG interpretation [24]-[29]. However, supervised learning approaches

generally assume that the data are stationary and separable, and that the data distributions of training and test datasets are identical [30]. However, many scenarios do not satisfy this assumption, especially in clinical settings. Clinical data often contain heterogenous characteristics e.g., patient profiles, disease states, drug effects, device setups etc. In these settings models constructed on the whole patient population may not generalize well to each subpopulation. However, training a specific model for each patient is high cost and sufficient data is usually lacking. Even in cases where it is possible to build a patient-specific model, performance can degrade over time due to changes in patient state or in the clinical environment.

In this study, we aimed to build an EEG-based sedation monitoring model for patients in the ICU. Direct application of supervised learning in this task (Figure 1 A) is challenging due to the following factors: 1) EEG signals are non-stationary; 2) EEG signals corresponding to the same subjective sedation level differ widely across patients; 3) EEG signal quality varies widely across patients and over time. In effect, the EEG data are generated from a continually evolving underlying stochastic process. Therefore, the sedation monitoring model should be adaptively updated in response to the continuously evolving target distribution (as shown in Figure 1 B).

In the field of machine learning, domain adaption or transfer learning has made progress in tackling the problem of distribution shift between training and test data [30]-[33]. The goal of supervised learning is to learn a function f based on the training dataset $\mathbf{D}^{tr} = (\mathbf{x_1^{tr}}, \mathbf{y_1^{tr}}), \ldots (\mathbf{x_m^{tr}}, \mathbf{y_m^{tr}}) \subseteq \mathcal{X} \times \mathcal{Y}$, which denotes the domain with features X and corresponding labels Y. The function f is expected to generalize well on test data $\mathbf{D}^{te}$. However, peformance suffers when training and test data have different distributions, i.e. when $P_{XY}^{tr} \neq P_{XY}^{te}$.

Various domain adaptation approaches have been proposed to deal with the covariate shift problem, where $P_{Y|X}^{tr} = P_{Y|X}^{te}$ but $P_X^{tr} \neq P_X^{te}$. However, in the problem that we address, both the marginal distribution P_X and conditional distribution $P_{Y|X}$ may change across patients and over time. Classic domain adaptation approaches do not address this problem. In the case we study there are various phenotypes in the patient population [34] and it is difficult to separate the whole dataset into clear discrete domains. Moreover, the target domain is not a single stationary domain, but instead involves a continually evolving target distribution [35]-[37].

In our study, we make the following assumptions: 1) some labeled data are available from other patients; 2) EEG samples from the target patient arrive sequentially and labels are available intermittently (e.g. every 1-2 hours, at times when clinical staff perform routine ICU RASS assessments); 3) the target distribution evolves over time. We propose an approach to sedation monitoring using an online learning algorithm called Adaptive Regularization of Weight Vectors (AROW) [38]. AROW has several attractive properties: large margin training, confidence weighting, and the capacity to handle nonseparable data. Due to the non-stationary characteristics of EEG and the continually changing noise background, different sedation states may not have stationary seperable boundaries. The

capacity for online updating helps the model adapt to changing EEG characteristics. We evaluate the performance of our model on EEG data from 172 sedated ICU patients. The experimental results indicate superior performance compared with conventional models without sequential updating. Because sedation labels are obtained using RASS assessments based on patient responses, label noise might exist due to interrater variability. The proposed approach is also shown to be robust to sudden changes caused by label noise.

II. Dataset

EEG data were collected using Sedline EEG monitors (Masimo Corporation, Irvine, CA, USA) from 195 mechanically ventilated adult ICU patients admitted to Massachusetts General Hospital (MGH), Boston, USA. The data were collected under a protocol approved by the local Institutional Review Board (IRB). Sedative and analgesic drugs were administered to patients per routine care. Propofol, Ketamine, Midazolam, and Hydromorphone accounted for the majority of such drugs received. The EEG sampling rate was 250 Hz. Recording electrodes were positioned on the forehead at positions FP1, FP2, F7, and F8 with the reference approximately 1 cm above Fpz and the ground at Fpz, according to the international 10-20 electrode position naming convention. At the time of EEG setup, impedances were adjusted below 5 $k\Omega$. The four frontal EEG channels were re-referenced to create a bipolar montage FP1-F7, FP2-F8, FP1-FP2, and F7-F8. The data of 10 patients were excluded by visual inspection of EEG signal quality. An additional 13 patients were excluded because they lacked any recorded RASS assessments during the EEG recording period. The final cohort analyzed included data from 172 patients, all of whom had at least one RASS assessment during EEG recording. EEGs were recorded as long as patients were on mechanical ventilation, and varied from several hours to several hundred hours, as shown in Figure 2. Patient characteristics are summarized in Table I.

The Richmond Agitation-Sedation Scale (RASS) was used to annotate the sedation and agitation levels of patients in the ICU [5]. The RASS is a ten-point scale, with four levels of agitation from +1 to +4, one level indicating a calm and alert state (0), and five levels of sedation from −1 to −5. RASS levels are determined based on patients responses to verbal or noxious stimulation. In our study, RASS assessments were performed and recorded as part of routine clinical care by ICU nurses approximately once every two hours. Because our study is focused on levels of sedation, we included EEG data with RASS scores between −5 and 0 in the analysis. In total, the dataset contained 3557 RASS assessments from 172 ICU patients. EEG data within an hour time interval centered on a given RASS assessment time (30 minutes before, 30 minutes after) was assumed to have the same RASS label (Figure 3). In addition, the time and dosages of sedative and analgesic medications were also recorded. And we investigate how the model performs under different medication conditions: no medications, propofol only, and multiple medications.

III. Methodology

We propose a clinician-in-the-loop framework for sedation monitoring, which leverages both human and machine intelligence to improve the accuracy of sedation monitoring (Figure 4). The framework contains two components: offline initialization and online calibration. In the

offline initialization phase, a generic model is trained using all available data from previously monitored patients. For a new ICU patient, monitoring commences with this generic, patient-independent model. Calibration is performed sequentially with each arriving RASS scores (e.g. every two hours), based on the mismatch (if any) between the model's current RASS prediction and the RASS score provided by the clinician based on direct behavioral assessment. Between clinical RASS assessments, the system continuously provides estimates of the patient's level of sedation. In this way, a personalized model able to track distribution shift over time is constructed for EEG-based sedation monitoring.

A. Preprocessing and Feature Extraction

The raw EEG signals were filtered using a bandpass filter between 0.5 and 20 Hz and a notch filter of 60 Hz. To reduce computation complexity, EEG signals were down-sampled to 62.5 Hz. We segmented EEG recordings into 4 second epochs using a sliding time window without overlap.

Since ICU EEG signals were often contaminated by nonphysiologic artifacts, we identified some typical types of artifacts and removed EEG segments with low quality. The criteria for artifact detection were as follows: 1) abnormally high amplitude values above 500 μV; 2) small standard deviation of the signal ($< 0.2\mu V$) for more than 2 s within the 4 second epoch; 3) overly fast amplitude change with more than 900 μV within 0.1 s; 4) staircase-like spectral patterns (commonly caused by ICU machines such as cooling blankets or pumps). To detect such noise patterns, two increasing and decreasing staircase-like convolution filters were applied to the spectra between 5 and 20 Hz with a threshold of 10 for artifacts.

Features were extracted from the multitaper spectrogram computed from each preprocessed 4 sec EEG segment as a feature set for model training. Multitaper spectral estimation optimizes a tradeoff between spectral resolution and variance, producing high quality spectral features [39]-[41]. The time-bandwidth product (TW) and number of tapers (K) were set to 2 and 3, respectively. The feature dimension of each channel was 120, and we used four signals, from FP1-F7, FP2-F8, FP1-FP2, F7-F8 in the bipolar montage, thus the total feature dimension of each EEG epoch was 480. Feature smoothing was performed using a Kalman filter to remove high frequency temporal fluctuations on timescales much faster than changes in sedation state. The Kalman filter provides Bayesian inference for hidden states given noisy observations (EEG features) with linear and Gaussian assumptions [42]. The hidden variables can be thought of as the underlying EEG features or 'sedation signature', whereas the observed EEG data provides only a noisy view of the underlying state. The Kalman filter smoothes the noisy observations in order to estimate the underlying noise-free sedation signature. For Kalman filtering, we used off-line inference with constant parameters. The transition and observation matrices were identity matrices and their variances were set to 0.0001 and 1, respectively.

B. Adaptive Regularization of Weight Vectors

Online learning represents a family of machine learning algorithms which update model parameters sequentially [43], [44]. Traditional machine learning algorithms operate in batch mode, in an offline fashion. For batch training, the data and corresponding labels must be

prepared as a complete training dataset in advance, and the model is subsequently deployed without further modification. In contrast, online learning is designed to update models incrementally from sequential incoming data. Therefore, online learning is efficient and scalable for large-scale data processing in streaming data environments. In the sedation monitoring setting, EEG signals are time series, which are naturally acquired in a sequential manner from recording devices. RASS assessments are routine tasks performed by clinicians as part of care in the ICU. Therefore, the labels are available without additional annotation cost. Therefore, online learning is suitable for the task of monitoring sedation in the ICU.

At time t, the online model receives data $\boldsymbol{x}_t \in \mathbb{R}^d$ and makes a prediction $\hat{y}_t \in \mathcal{Y}$ with the current prediction function. After the prediction, the true label $y_t \in \mathcal{Y}$ becomes available as feedback and the loss $l(y_t, \hat{y}_t)$ can be measured. For binary classification, we have $\mathcal{Y} = \{-1, +1\}$ and $l(y_t, \hat{y}_t) = 0$ if $y_t = \hat{y}_t$ and 1 otherwise. The model updates its parameters using $(\boldsymbol{x}_t, y_t)$ and remains unchanged until the next time a labeled sample of data arrives. In the present study, we use a linear function as the prediction function with a weight vector $\boldsymbol{w}$: $\hat{y} = sign(\boldsymbol{w} \cdot \boldsymbol{x})$.

To sequentially update the model weights, we apply the Adaptive Regularization of Weight Vectors (AROW) algorithm [38]. In AROW, the weight vector $\boldsymbol{w}$ is assumed to be Gaussian $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. The smaller $\Sigma_{p,p}$ is, the more confidence the model has in the mean value μ_p. The model predicts the label according to $sign(\boldsymbol{\mu} \cdot \boldsymbol{x})$ using the average weight vector $\mathbb{E}[\boldsymbol{w}] = \boldsymbol{\mu}$. The loss function of AROW at time step t is:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \Sigma) = & D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \| \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) \\ & + \lambda_1 l_h^2(y_t, \boldsymbol{\mu} \cdot \boldsymbol{x}_t) + \lambda_2 \boldsymbol{x}_t^\top \Sigma \boldsymbol{x}_t, \end{aligned} \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence between the distribution of previous parameters and current parameters, $l_h^2(y_t, \boldsymbol{\mu} \cdot \boldsymbol{x}_t) = (max\{0, 1 - y_t(\boldsymbol{\mu} \cdot \boldsymbol{x}_t)\})^2$ is the squared-hinge loss, and $\lambda_1, \lambda_2 \geq 0$ are two tradeoff hyperparameters, $\lambda_1 = \lambda_2 = 1/(2r)$.

The loss function consists of three terms. The first term constrains the parameters of two consecutive updates not to change too dramatically. This is desirable because sedation states evolve slowly and the current parameters contain discriminative information about previous samples. The second term aims to minimize the prediction error obtained with the new / updated parameters. The last term is the prediction confidence. During training, the parameter variance Σ should be reduced so that we get higher confidence for estimating classifier parameters.

For implementing the AROW algorithm, it is benecifial to rewrite the Kullback-Leibler divergence in the cost function (Equation 1) to obtain

$$\begin{aligned}\mathscr{L}(\boldsymbol{\mu}, \Sigma) = & \frac{1}{2} log(\frac{det\Sigma_{t-1}}{det\Sigma}) + \frac{1}{2} Tr(\Sigma_{t-1}^{-1}\Sigma) \\ & + \frac{1}{2}(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu})^{\top}\Sigma_{t-1}^{-1}(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}) - \frac{d}{2} \\ & + \frac{1}{2r} l_h{}^2(y_t, \boldsymbol{\mu} \cdot \boldsymbol{x}_t) + \frac{1}{2r}\boldsymbol{x}_t^{\top}\Sigma\boldsymbol{x}_t \,.\end{aligned} \tag{2}$$

The loss function can be further decomposed into two terms: $\mathscr{L}_1(\boldsymbol{\mu})$ and $\mathscr{L}_2(\Sigma)$, which are only associated with $\boldsymbol{\mu}$ and Σ, respectively. Therefore, the optimization of $\boldsymbol{\mu}$ and Σ can be performed independently. Thus we can update the parameters alternately as follows:

1. Update the mean parameters:

$$\boldsymbol{\mu}_t = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \mathscr{L}_1(\boldsymbol{\mu}), \tag{3}$$

2. If $\boldsymbol{\mu}_t \neq \boldsymbol{\mu}_{t-1}$, update the confidence parameters:

$$\Sigma_t = \underset{\Sigma}{\operatorname{argmin}} \mathscr{L}_2(\Sigma) \,. \tag{4}$$

Taking the derivative of $\mathscr{L}(\boldsymbol{\mu}, \Sigma)$ with respect to $\boldsymbol{\mu}$ and setting it to zero, we get

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} - \frac{1}{2r}[\frac{d}{dz} l_h{}^2(y_t, z) \mid_{z = \boldsymbol{\mu}_t \cdot \boldsymbol{x}_t}]\Sigma_{t-1}\boldsymbol{x}_t \,. \tag{5}$$

Substituting the derivative of the squared-hinge loss in Equation 5 and assuming $1 - y_t(\boldsymbol{\mu}_t \cdot \boldsymbol{x}_t) \geq 0$, we get

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \frac{y_t}{r}(1 - y_t(\boldsymbol{\mu}_t \cdot \boldsymbol{x}_t))\Sigma_{t-1}\boldsymbol{x}_t \,. \tag{6}$$

We reorganize Equation 6 to obtain

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \frac{max(0, 1 - y_t\boldsymbol{x}_t^{\top}\boldsymbol{\mu}_{t-1})}{\boldsymbol{x}_t^{\top}\Sigma_{t-1}\boldsymbol{x}_t + r}\Sigma_{t-1}y_t\boldsymbol{x}_t \,. \tag{7}$$

The update for the confidence parameters is made only if $\boldsymbol{\mu}_t \neq \boldsymbol{\mu}_{t-1}$. By setting the derivative of $\mathscr{L}(\boldsymbol{\mu}, \Sigma)$ in Equation 2 with respect to Σ to zero, we get

$$\Sigma_t^{-1} = \Sigma_{t-1}^{-1} + \frac{\boldsymbol{x}_t\boldsymbol{x}_t^{\top}}{r} \,. \tag{8}$$

Using the Woodbury identity, we can rewrite the update for Σ in non-inverted form:

$$\Sigma_t = \Sigma_{t-1} - \frac{\Sigma_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\top\Sigma_{t-1}}{r + \boldsymbol{x}_t^\top\Sigma_{t-1}\boldsymbol{x}_t}. \quad (9)$$

The final form of AROW for binary classification is presented in Algorithm 1. We extend AROW to multiclass classification using the one-vs-rest strategy.

Algorithm 1 The AROW algorithm for binary classification

input : Tradeoff parameter r.
output : Weight vector $\boldsymbol{\mu}_T$ and corresponding covariance matrix Σ_T.

1: Initialize μ_0 and Σ_0 with labeled data from preivous patients.
2: **For** $t = 1, \ldots, T$

- Receive a training sample $\boldsymbol{x}_t \in \mathbb{R}^d$.
- Compute margin and confidence: $m_t = \boldsymbol{\mu}_{t-1} \cdot \boldsymbol{x}_t, v_t = \boldsymbol{x}_t^\top\Sigma_{t-1}\boldsymbol{x}_t$.
- Receive true label y_t, and loss $l_t = 1$ if $sign(m_t) \neq y_t$.
- if $m_t y_t < 1$, update using Equations 7 and 9:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \alpha_t\Sigma_{t-1}, y_t\boldsymbol{x}_t$$

$$\Sigma_t = \Sigma_{t-1} - \beta_t\Sigma_{t-1}\boldsymbol{x}_t\boldsymbol{x}_t^\top\Sigma_{t-1}$$

$$\beta_t = \frac{1}{\boldsymbol{x}_t^\top\Sigma_{t-1}\boldsymbol{x}_t + r}$$

$$\alpha_t = max(0, 1 - y_t\boldsymbol{x}_t^\top\boldsymbol{\mu}_{t-1})\beta_t.$$

3: **Return**

C. Model Evaluation

We utilized leave-one-patient-out cross validation (LOPOCV) to evaluate the performance of the proposed approach. For each round of LOPOCV, the data from one patient were used as the test data and the data of the rest patients were used as training data. In this way, we were able to evaluate performance in individual patients.

Standard supervised learning models have fixed training and test datasets. In contrast, the training and test data in our problem change over time in the online learning framework. At each time step (i.e. the time that each new RASS measurement arrives), the prediction of the current online model is evaluated against the next newly arriving RASS label. The mean accuracy (ACC) and mean absolute error (MAE) after a given number of updates were calculated by taking the average performance across individual patients. Since RASS assessments are not perfect and may vary across clinicians due to interrater variability, accuracies with tolerance of one level difference are calculated. In clinical practice, one level

error in sedation monitoring is acceptable. The linear model without any update was used as the baseline method for comparison.

We performed the evaluations on how the model performs with numbers of RASS assessments seen. For the baseline model without any update, the model M_i^0 was trained on data from all patients other than patient i and used to predict the RASS values $\hat{y}_{i,j}^0$, where $j = 1, 2, \ldots, n_i$, and n_i is the number of RASS assessments obtained for patient i. The error values were computed by comparing the predictions and the observed values $y_{i,j}$: $e_{i,j}^0 = e(y_{i,j}, \hat{y}_{i,j}^0)$; here $e(x, y)$ is 0 when $x = y$ and 1 otherwise, or when considering one level difference accuracy, it is 0 when $|x - y| \leq 1$ and 1 otherwise. We calculated the average performance for each patient, $E_i^0 = \frac{1}{n_i}\sum_{j=1}^{n_i} e_{i,j}^0$, and the average performance across all patients, $E^0 = \frac{1}{N}\sum_{i=1}^{N} E_i^0$, where N was the number of patients.

For the update model, M^k, $k = 1, 2, \ldots, n_i$, we investigated how well the online model performs, on average, after adapting k times, starting from the baseline model M^0. For patient i, the model which has undergone k sequential adaptations leading up to and including RASS assessment j is denoted as $M_{i,j}^k$ and the prediction is $\hat{y}_{i,j}^k$. In general, if patient i has n_i RASS assessments available, we repeat the entire process up to $m_i = n_i - k$ times to get the average performance after k adaptations:

$$E_i^k = \frac{1}{n_i - k}\sum_{j=k+1}^{n_i} e^k(y_{i,j}, \hat{y}_{i,j}^k). \tag{10}$$

All the possible update steps on each RASS assessment are considered and the mean performance is calculated by averaging over all the different possible starting points within each patient's data and over all patients. The average performance across all patients after k adaptations is calculated: $E^k = \frac{1}{N}\sum_{i=1}^{N} E_i^k$, where E^k is a function of the number of RASS adaptations, k.

D. Model Comparison

Another online learning algorithm called Multiclass Soft Confidence-Weighted Learning (SCW) was also included for comparison [45]. SCW is an extension of the confidence weight (CW) algorithm [46], which softens the aggressiveness of CW learning. The optimization function of CW is

$$\begin{aligned}(\boldsymbol{\mu}_t, \Sigma_t) = \underset{\boldsymbol{\mu}, \Sigma}{\operatorname{argmin}} & D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \| \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) \\ s.t. \quad & \ell^{\phi}(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\boldsymbol{x}_t, y_t)) = 0,\end{aligned} \tag{11}$$

where ϕ is a positive constant and $\ell^{\phi}(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\boldsymbol{x}_t, y_t)) = max(0, \phi\sqrt{\boldsymbol{x}_t^{\top}\Sigma\boldsymbol{x}_t} - y_t(\boldsymbol{\mu} \cdot \boldsymbol{x}_t))$, which constrains models to provide correct predictions.

In the update strategy of CW, the parameters are constrained to correctly classify the current sample. This is an excessively agressive constraint in applications like ours in which there can be label noise. To overcome this limitation, SCW was proposed to soften the constraints. The optimization objective of SCW is

$$(\boldsymbol{\mu}_t, \Sigma_t) = \underset{\boldsymbol{\mu}, \Sigma}{\operatorname{argmin}} D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \| \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) + C\ell^{\phi}(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\boldsymbol{x}_t, y_t)), \tag{12}$$

where C is the tradeoff parameter. The above learning algorithm is denoted "SCW1". The second term in Equation 12 can be modified as a squared penalty (denoted "SCW2") as follows:

$$(\boldsymbol{\mu}_t, \Sigma_t) = \underset{\boldsymbol{\mu}, \Sigma}{\operatorname{argmin}} D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \| \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) + C\ell^{\phi}(\mathcal{N}(\boldsymbol{\mu}, \Sigma); (\boldsymbol{x}_t, y_t))^2. \tag{13}$$

The optimization of SCW has a closed-form solution. We refer readers for the details of SCW to the original paper [45]. In our experiments, the value of the parameter C was set to 1. In AROW, the value of tradeoff parameter r was set to 1.

IV. Experimental Results

In this section, we present experimental results characterizing model performance, model robustness to label noise, and medication effects on model performance.

A. Performance Comparison Across Models

Figure 5 shows the performance of the baseline method without updating, compared with the performance of AROW over time. The evaluation metrics are calculated based on each RASS assessment. Different patients have varying numbers of the RASS assessments. The more RASS assessments, the fewer patients we have for performance estimation. In order to robustly estimate the performance trend, only the result of 23 RASS adaptations is shown. From Figure 5, we see that the performance of AROW initially is identical to the baseline method, and consistently improves with further RASS assessments. For any given patient, the initial baseline model is trained on the labeled data from other patients, and performs only modestly well due to the large between-individual variability. As time passes, the online learning model is adaptively updated with sequentially arriving EEG data and RASS assessments. Performance of the online learning model improves with more RASS labels seen (from about 20% to 45% accuracy, or from about 61% to 75% one-level difference accuracy). The precision with which we can estimate the mean performance curve decreases (confidence bands become wider) for larger numbers of RASS measurements, as the number of patients with that number of assessments decreases.

We evaluate the overall performance of our proposed approach in terms of accuracy and mean absolute error by averaging over the RASS assessments available in time and all the patients. Figure 6 shows the average performance of the four different models: the baseline model without update, SCW1, SCW2, and AROW. Their mean accuracies are 22.59%,

29.04%, 25.41%, and 36.52%, respectively, while their mean one-level difference accuracies are 61.10%, 64.15%, 62.49%, and 68.76%, respectively. Their mean absolute errors are 1.43, 1.29, 1.36, and 1.17, respectively. The online learning models enhance performance, and AROW achieves a significant improvement over the baseline without update ($p < 0.01$, one-way ANOVA) as well as SCW1 and SCW2 ($p < 0.05$, one-way ANOVA). The experimental results indicate the efficiency of online learning models with the property of adapting to the changing target distributions.

We investigate the classification accuracies for individual RASS labels in Figure 7. The values are the average accuracies over all RASS assessments available in time and all the patients. The online learning models perform better than the baseline method on all RASS labels, especially on RASS −5 (from 9.06% to 22.07%). RASS −5 receives the lowest classification accuracy among all sedation levels. Misclassifications by the model are generally to nearby RASS levels.

In order to help interpret the model, we average the learned weights of each feature over all the patients and rank the features according to the mean weights. The top 100 features are selected according to their corresponding ranked weights and grouped into 5 Hz frequency bins. From Figure 8, more discriminative features are located within 10-15 Hz (25%), followed by the 15-20 Hz frequency range (19%). This is physiologically reasonable, because oscillations within this range generally correspond to wakefulness and effects induced by propofol, the most commonly used sedative in our cohort. Moreover, the model provides not only the predictions, but also the confidence. The third term $\boldsymbol{x}_t^\top \Sigma \boldsymbol{x}_t$ in Equation 1 denotes the prediction confidence, which consists of two parts: the variance of the estimated weights and the EEG variability compared with the previous EEG features. The lower the term is, the higher prediction confidence the model has. In general, the prediction confidence becomes higher after updating the model with the available labeled data from the target patients. Without calibration, the prediction confidence will gradually decrease over time, which provides clinicians additional information them to gauge how much confidence to place in the predictions of the sedation monitoring system.

Figure 9 shows an example of sedation monitoring. In the beginning, both the baseline model and the adaptive model perform poorly. With online calibration, the performance of the adaptive prediction model improves. The baseline model usually has larger bias from the observed labels than the adaptive model.

B. Model Robustness to Label Noise

Because RASS assessments are based on behavioral assessments, they can vary even when the EEG and patient state remain the same due to the interrater variability. We therefore investigated how well the AROW model performs in response to various levels of label noise. To do this, we randomly changed the RASS labels of some EEG samples (one level difference [up or down] from the recorded label), used these in the AROW updates, and compared subsequent model predictions with the incoming unchanged labels. Table II shows the results. In our experiments the percentage of samples with label noise increases from 10% to 50%. As label noise increases, performance of the online model degrades, but not

dramatically: ACC: 33.74% → 25.50% ($p < 0.01$), one level difference ACC: 68.20% → 65.23% ($p = 0.14$), MAE: 1.20 → 1.33 ($p = 0.02$). These results suggest that performance of the adaptive model degrades gracefully in response to label noise, and is therefore robust to realistic amounts of human error.

C. Medication Effects on Model Performance

Sedative and analgesic medications strongly modulate EEG patterns. However, these effects have been investigated primarily in healthy volunteers and in patients undergoing elective surgery. [41], [47], [48]. Medication-specific effects in critically ill patients remain largely unexplored.

In this section, we investigate some aspects of medication effects on model performance. Multiple sedative and analgesic medications are often administrated to the same patient at the same time in the ICU. Figure 10 summarizes drugs administration information across the study cohort for drugs commonly used in our cohort which have the potential to influence EEG patterns. These include four anesthetic drugs (propofol, dexmedetomidine, ketamine, midazolam), and three analgesics (hydromorphone, fentanyl, morphine). The most commonly administered drugs were propofol (76.16%), hydromorphone (35.47%), and dexmedetomidine (22.67%). Figure 11 illustrates the number of patients with different drug combination. In total 26 drug combinations were used in our cohort. The most common combinations of sedating drugs were propofol only, no sedating drug, propofol +hydromorphone, propofol+hydromorphone+dexmedetomidine, propofol+fentanyl, and propofol+dexmedetomidine with numbers of patients of 47, 31, 29, 10, 9, and 9, respectively.

Given that the large majority of patients received propofol, we hypothesized that model performance might be preferentially attuned to propofol EEG signatures and thus might perform better in patients receiving only propofol. We therefore investigated model performance in three groups of patients, receiving: 1) none of the 7 medications, 2) propofol only, and 3) multiple medications. Figure 12 shows model performance in these three groups. The patient numbers for these groups are 31, 47, and 90, respectively. Among the 172 patients, three patients received dexmedetomidine only and one patient received hydromorphone only, which are excluded from the experiment here. The mean accuracies/ mean absolute errors of the three groups are 26.37%/1.49, 41.30%/0.99, and 36.93%/1.16, respectively. The model performance in the propofol only group is significantly higher than in the no medications ($p < 0.01$, Wilcoxon rank sum test).

V. Discussion

In this work we propose a clinician-in-the-loop framework for sedation monitoring using online learning. Our adaptive model leverages labeled data from previously monitored patients in the offline training phase, and during the online calibration phase automatically adapts to incoming labeled data from an individual patient. Since RASS assessments are performed routinely in the ICU as part of clinical care, these intermittent labels can be obtained without additional cost. The proposed sedation monitoring framework leverages both human and artificial intelligence. Rather than aiming for full automation without the

need for human input (a problematic goal [49]), our work aims to help clinicians perform an important clinical task with greater precision. By augmenting intermittent clinical RASS assessments with continuous, quantitative brain monitoring, our work aims to help reduce the problem of over- and undersedation in the ICU, which in turn holds promise for improving neurologic and medical health outcomes for ICU patients.

Learning robust representations of different sedation states is challenging due to inter-individual variability and heterogeneous treatments and disease states in the ICU. Although our adaptive approach improves upon patient-independent modeling, other approaches may further improve performance. One promising future direction is to construct a library of models for different patient ‘types’, and to thereby transfer knowledge from patients that are similar to target patients based on clinical characteristics and demographics rather than directly from the entire patient population. This field of research, named computational phenotyping, has been developed to deal with the problem of inter-patient heterogeneity [34]. A single model built on the entire patient population may not generalize well to specific patients, wherease patient-specific models do not take advantage of common knowledge in the patient population. Computational phenotyping aims to learn common characteristic of patients relevant for clinical tasks (e.g., sedation monitoring) in a data-driven manner. The common EEG representations of different sedation states and how features differ in different patient cohorts merits further investigation [50].

A limitation of our study is the frequency of EEG annotation. We assume EEG data within an hour time interval centered on a given RASS assessment time had the same RASS label. However, the dynamics of sedation states might fluctuate within this period. Therefore, more accurate labels with a higher time resolution would be helpful for model training. In the present study we use only labeled data from target patients for model calibration. However, a much larger amount of unlabeled EEG data is available between RASS assessments. How to efficiently leverage this unlabeled data for knowledge transfer is an important unexplored topic. The unlabeled data may provide meaningful information for estimating the marginal distribution P_X in target patients. Transductive domain adaptation has proved useful in various applications, and might be used in the context of sedation monitoring to achieve further performance improvements in patient-specific monitoring [30].

Previous studies of using EEG to monitor level of consciousness focus primarily on relatively healthy subjects (compared with ICU patients) undergoing general anesthesia for elective surgery [51], [52]. Nevertheless some studies have explored EEG-based methods for determining the level of consciousness in ICU patients [17]-[19]). However, in most of these studies performance is evaluated on small groups of patients. How well these EEG features generalize for sedation monitoring in the general ICU population requires further investigation. Although we have analyzed some simple aspects of medication effects on our model performance, we did not include drug information in our model training. Both drug type and dose strongly impact EEG features, thus taking drug-specific EEG features into account in model training should be explored to improve model performance in future work.

VI. Conclusion

We have presented an adaptive learning framework for EEG-based sedation monitoring in ICU patients using online learning. Due to the heterogeneous clinical environment and disease factors and individual variability across patients, generic models trained from the entire patient population are usually suboptimal for monitoring individual patients. In addition, statistical features of an individual's EEG may evolve over time, further necessitating continuous adaptation. In our framework, the sedation monitoring model for an individual is adaptively updated as additional labeled EEG samples arrive during the course of RASS assessments performed as part of routine ICU care. The proposed adaptive method achieves substantially higher performance than the non-adaptive baseline method. The experimental results also demonstrate our model's robustness to label noise, which simulates interrater variability inherent in clinical settings. We explored medication effects on model performance and found that models achieve the best performance under only propofol administration in comparison with no medications and multiple medications administration, likely due to propofol being the most common sedative used in our patient cohort.

Acknowledgments

This work was supported in part by grants from NIH-NINDS (1K23NS090900, 1R01NS102190, 1R01NS102574, 1R01NS107291).

References

[1]. Reade MC and Finfer S, "Sedation and delirium in the intensive care unit," New England Journal of Medicine, vol. 370, no. 5, pp. 444–454, 2014. [PubMed: 24476433]

[2]. Jackson DL, Proudfoot CW, Cann KF, and Walsh TS, "The incidence of sub-optimal sedation in the ICU: a systematic review," Critical Care, vol. 13, no. 6, p. R204, 2009. [PubMed: 20015357]

[3]. Hughes CG, McGrane S, and Pandharipande PP, "Sedation in the intensive care setting," Clinical Pharmacology: Advances and Applications, vol. 4, p. 53, 2012. [PubMed: 23204873]

[4]. Carrasco G, "Instruments for monitoring intensive care unit sedation," Critical Care, vol. 4, no. 4, p. 217, 2000. [PubMed: 11094504]

[5]. Sessler CN, Gosnell MS, Grap MJ, Brophy GM, O'neal PV, Keane KA, Tesoro EP, and Elswick R, "The richmond agitation–sedation scale: validity and reliability in adult intensive care unit patients," American Journal of Respiratory and Critical Care Medicine, vol. 166, no. 10, pp. 1338–1344, 2002. [PubMed: 12421743]

[6]. Dale CR, Kannas DA, Fan VS, Daniel SL, Deem S, Yanez ND III, Hough CL, Dellit TH, and Treggiari MM, "Improved analgesia, sedation, and delirium protocol associated with decreased duration of delirium and mechanical ventilation," Annals of the American Thoracic Society, vol. 11, no. 3, pp. 367–374, 2014. [PubMed: 24597599]

[7]. Friedman D, Claassen J, and Hirsch LJ, "Continuous electroencephalogram monitoring in the intensive care unit," Anesthesia & Analgesia, vol. 109, no. 2, pp. 506–523, 2009. [PubMed: 19608827]

[8]. LeBlanc JM, Dasta JF, and Kane-Gill SL, "Role of the bispectral index in sedation monitoring in the ICU," Annals of Pharmacotherapy, vol. 40, no. 3, pp. 490–500, 2006. [PubMed: 16492796]

[9]. Sackey PV, "Frontal EEG for intensive care unit sedation: treating numbers or patients?" Critical Care, vol. 12, no. 5, p. 186, 2008. [PubMed: 18983711]

[10]. Haenggi M, Ypparila-Wolters H, Bieri C, Steiner C, Takala J, Korhonen I, and Jakob SM, "Entropy and bispectral index for assessment of sedation, analgesia and the effects of unpleasant stimuli in critically ill patients: an observational study," Critical Care, vol. 12, no. 5, p. R119, 2008. [PubMed: 18796156]

[11]. Hajat Z, Ahmad N, and Andrzejowski J, “The role and limitations of EEG-based depth of anaesthesia monitoring in theatres and intensive care,” Anaesthesia, vol. 72, pp. 38–47, 2017. [PubMed: 28044337]

[12]. Riess M, Graefe U, Goeters C, Van Aken H, and Bone H, “Sedation assessment in critically ill patients with bispectral index,” European Journal of Anaesthesiology, vol. 19, no. 1, pp. 18–22, 2002. [PubMed: 11913799]

[13]. Frenzel D, Greim C-A, Sommer C, Bauerle K, and Roewer N, “Is the bispectral index appropriate for monitoring the sedation level of mechanically ventilated surgical ICU patients?” Intensive Care Medicine, vol. 28, no. 2, pp. 178–183, 2002. [PubMed: 11907661]

[14]. Vivien B, Di Maria S, Ouattara A, Langeron O, Coriat P, and Riou B, “Overestimation of bispectral index in sedated intensive care unit patients revealed by administration of muscle relaxant,” Anesthesiology, vol. 99, no. 1, pp. 9–17, 2003. [PubMed: 12826836]

[15]. Roustan J-P, Valette S, Aubas P, Rondouin G, and Capdevila X, “Can electroencephalographic analysis be used to determine sedation levels in critically ill patients?” Anesthesia & Analgesia, vol. 101, no. 4, pp. 1141–1151, 2005. [PubMed: 16192535]

[16]. Bilgili B, Montoya JC, Layon AJ, Berger AL, Kirchner HL, Gupta LK, and Gloss DS, “Utilizing bi-spectral index (BIS) for the monitoring of sedated adult ICU patients: a systematic review,” Minerva Anestesiol, vol. 83, no. 3, pp. 288–301, 2017. [PubMed: 27314595]

[17]. Ferenets R, Lipping T, Anier A, Jantti V, Melto S, and Hovilehto S, “Comparison of entropy and complexity measures for the assessment of depth of sedation,” IEEE Transactions on Biomedical Engineering, vol. 53, no. 6, pp. 1067–1077, 2006. [PubMed: 16761834]

[18]. Engemann DA, Raimondo F, King J-R, Rohaut B, Louppe G, Faugeras F, Annen J, Cassol H, Gosseries O, Fernandez-Slezak D et al., “Robust EEG-based cross-site and cross-protocol classification of states of consciousness,” Brain, vol. 141, no. 11, pp. 3179–3192, 2018. [PubMed: 30285102]

[19]. Nagaraj SB, McClain LM, Boyle EJ, Zhou DW, Ramaswamy SM, Biswal S, Akeju O, Purdon PL, and Westover MB, “Electroencephalogram based detection of deep sedation in ICU patients using atomic decomposition,” IEEE Transactions on Biomedical Engineering, vol. 65, no. 12, pp. 2684–2691, 2018. [PubMed: 29993386]

[20]. Sun H, Kimchi E, Akeju O, Nagaraj SB, McClain LM, Zhou DW, Boyle E, Zheng W-L, Ge W, and Westover MB, “Automated tracking of level of consciousness and delirium in critical illness using deep learning,” NPJ Digital Medicine, vol. 2, no. 1, pp. 1–8, 2019. [PubMed: 31304351]

[21]. Yu K-H, Beam AL, and Kohane IS, “Artificial intelligence in healthcare,” Nature Biomedical Engineering, vol. 2, no. 10, p. 719, 2018.

[22]. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, ODonoghue B, Visentin D et al., “Clinically applicable deep learning for diagnosis and referral in retinal disease,” Nature Medicine, vol. 24, no. 9, p. 1342, 2018.

[23]. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, and Ng AY, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” Nature Medicine, vol. 25, no. 1, p. 65, 2019.

[24]. Zheng W-L, Zhu J-Y, Peng Y, and Lu B-L, “EEG-based emotion classification using deep belief networks,” in IEEE International Conference on Multimedia and Expo (ICME) IEEE, 2014, pp. 1–6.

[25]. Zheng W-L and Lu B-L, “Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks,” IEEE Transactions on Autonomous Mental Development, vol. 7, no. 3, pp. 162–175, 2015.

[26]. Tjepkema-Cloostermans MC, de Carvalho RC, and van Putten MJ, “Deep learning for detection of focal epileptiform discharges from scalp EEG recordings,” Clinical Neurophysiology, vol. 129, no. 10, pp. 2191–2196, 2018. [PubMed: 30025804]

[27]. Tjepkema-Cloostermans MC, da Silva Lourenco C, Ruijter BJ, Tromp SC, Drost G, Kornips FHM, Beishuizen A, Bosch FH, Hofmeijer J, and van Putten MJAM, “Outcome prediction in postanoxic coma with deep learning,” Critical Care Medicine, 2019, DOI: 10.1097/CCM.0000000000003854.

[28]. Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W, and Ball T, “Deep learning with convolutional neural networks for EEG decoding and visualization,” Human Brain Mapping, vol. 38, no. 11, pp. 5391–5420, 2017. [PubMed: 28782865]

[29]. Yannick R, Hubert B, Isabela A, Alexandre G, Jocelyn F et al., “Deep learning-based electroencephalography analysis: a systematic review,” arXiv preprint arXiv:1901.05498, 2019.

[30]. Pan SJ and Yang Q, “A survey on transfer learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[31]. Zheng W-L and Lu B-L, “Personalizing EEG-based affective models with transfer learning,” in *the* Twenty-Fifth International Joint Conference on Artificial Intelligence AAAI Press, 2016, pp. 2732–2738.

[32]. Jayaram V, Alamgir M, Altun Y, Scholkopf B, and Grosse-Wentrup M, “Transfer learning in brain-computer interfaces,” IEEE Computational Intelligence Magazine, vol. 11, no. 1, pp. 20–31, 2016.

[33]. Li Y, Carlson DE et al., “Extracting relationships by multi-domain matching,” in Advances in Neural Information Processing Systems, 2018, pp. 6799–6810.

[34]. Che Z, Kale D, Li W, Bahadori MT, and Liu Y, “Deep computational phenotyping,” in 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM, 2015, pp. 507–516.

[35]. Hoffman J, Darrell T, and Saenko K, “Continuous manifold based adaptation for evolving visual domains,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 867–874.

[36]. Wulfmeier M, Bewley A, and Posner I, “Incremental adversarial domain adaptation for continually changing environments,” in IEEE International Conference on Robotics and Automation (ICRA) IEEE, 2018, pp. 1–9.

[37]. Bobu A, Tzeng E, Hoffman J, and Darrell T, “Adapting to continuously shifting domains,” in 6th International Conference on Learning Representations Workshop (ICLR), 2018.

[38]. Crammer K, Kulesza A, and Dredze M, “Adaptive regularization of weight vectors,” in Advances in Neural Information Processing Systems, 2009, pp. 414–422.

[39]. Bronez TP, “On the performance advantage of multitaper spectral analysis,” IEEE Transactions on Signal Processing, vol. 40, no. 12, pp. 2941–2946, 1992.

[40]. Babadi B and Brown EN, “A review of multitaper spectral analysis,” IEEE Transactions on Biomedical Engineering, vol. 61, no. 5, pp. 1555–1564, 2014. [PubMed: 24759284]

[41]. Purdon PL, Sampson A, Pavone KJ, and Brown EN, “Clinical electroencephalography for anesthesiologistspart i: background and basic signatures,” Anesthesiology: The Journal of the American Society of Anesthesiologists, vol. 123, no. 4, pp. 937–960, 2015.

[42]. Pnevmatikakis EA, Rad KR, Huggins J, and Paninski L, “Fast kalman filtering and forward–backward smoothing via a low-rank perturbative approach,” Journal of Computational and Graphical Statistics, vol. 23, no. 2, pp. 316–339, 2014.

[43]. Hoi SC, Wang J, and Zhao P, “Libol: A library for online learning algorithms,” The Journal of Machine Learning Research, vol. 15, no. 1, pp. 495–499, 2014.

[44]. Hoi SC, Sahoo D, Lu J, and Zhao P, “Online learning: A comprehensive survey,” arXiv preprint arXiv:1802.02871, 2018.

[45]. Wang J, Zhao P, and Hoi SC, “Exact soft confidence-weighted learning,” arXiv preprint arXiv:1206.4612, 2012.

[46]. Crammer K, Dredze M, and Pereira F, “Exact convex confidence-weighted learning,” in Advances in Neural Information Processing Systems, 2009, pp. 345–352.

[47]. Purdon PL, Pierce ET, Mukamel EA, Prerau MJ, Walsh JL, Wong KFK, Salazar-Gomez AF, Harrell PG, Sampson AL, Cimenser A et al., “Electroencephalogram signatures of loss and recovery of consciousness from propofol,” Proceedings of the National Academy of Sciences, vol. 110, no. 12, pp. E1142–E1151, 2013.

[48]. Sanz-García A, Pérez-Romero M, Pastor J, Sola RG, Vega-Zelaya L, Vega G, Monasterio F, Torrecilla C, Pulido P, and Ortega GJ, “Potential EEG biomarkers of sedation doses in intensive care patients unveiled by using a machine learning approach,” Journal of Neural Engineering, vol. 16, no. 2, p. 026031, 2019. [PubMed: 30703765]

[49]. Topol EJ, “High-performance medicine: the convergence of human and artificial intelligence,” Nature Medicine, vol. 25, no. 1, p. 44, 2019.

[50]. Suresh H, Gong JJ, and Guttag JV, “Learning tasks for multitask learning: Heterogenous patient populations in the ICU,” in 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining ACM, 2018, pp. 802–810.

[51]. Jospin M, Caminal P, Jensen EW, Litvan H, Vallverdú M, Struys MM, Vereecke HE, and Kaplan DT, “Detrended fluctuation analysis of EEG as a measure of depth of anesthesia,” IEEE Transactions on Biomedical Engineering, vol. 54, no. 5, pp. 840–846, 2007. [PubMed: 17518280]

[52]. Li X, Li D, Liang Z, Voss LJ, and Sleigh JW, “Analysis of depth of anesthesia with hilbert–huang spectral entropy,” Clinical Neurophysiology, vol. 119, no. 11, pp. 2465–2475, 2008. [PubMed: 18812265]

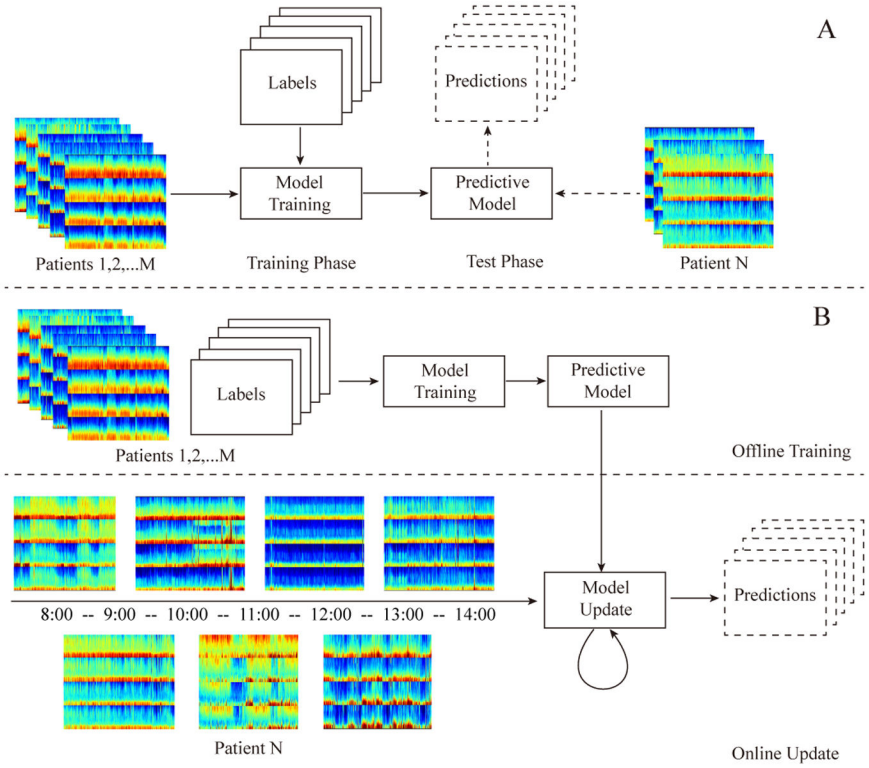

Fig. 1.

(A) Traditional supervised learning approach: training a model on EEG data from a patient population and directly deploying the model for a new patient without any customization / adaptation. (B) Our approach: In light of EEG variability in a continually changing clinical environment, in our approach the initial prediction model is adaptively updated over time, calibrating it to the current patient.

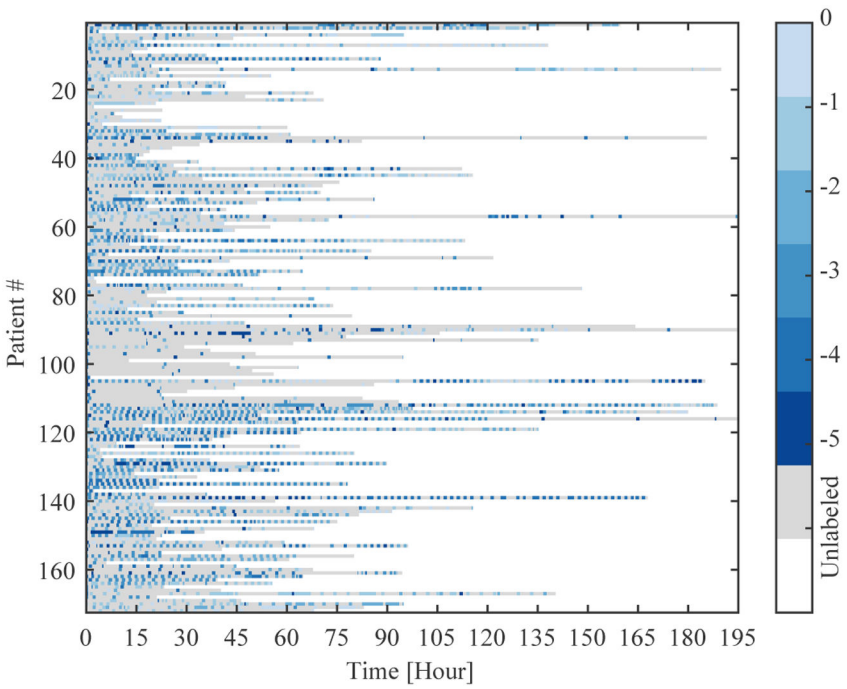

Fig. 2.

EEG durations for the 172 individual patients in the dataset. Two patients had more than 280 hours of EEG recording, extending beyond the time axis shown in the figure. The colorbar indicates RASS level of consciousness assessments, between −5 (comatose) to 0 (alert/awake).

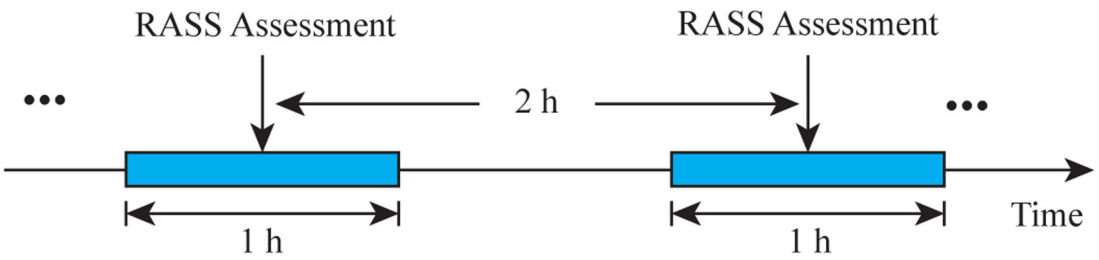

Fig. 3.

Temporal details of RASS assessments. Assessments were performed by nurses, generally once every two hours. For analysis purposes, we assume that the sedation level (RASS score) remains constant for the one hour period centered on the time of the assessment.

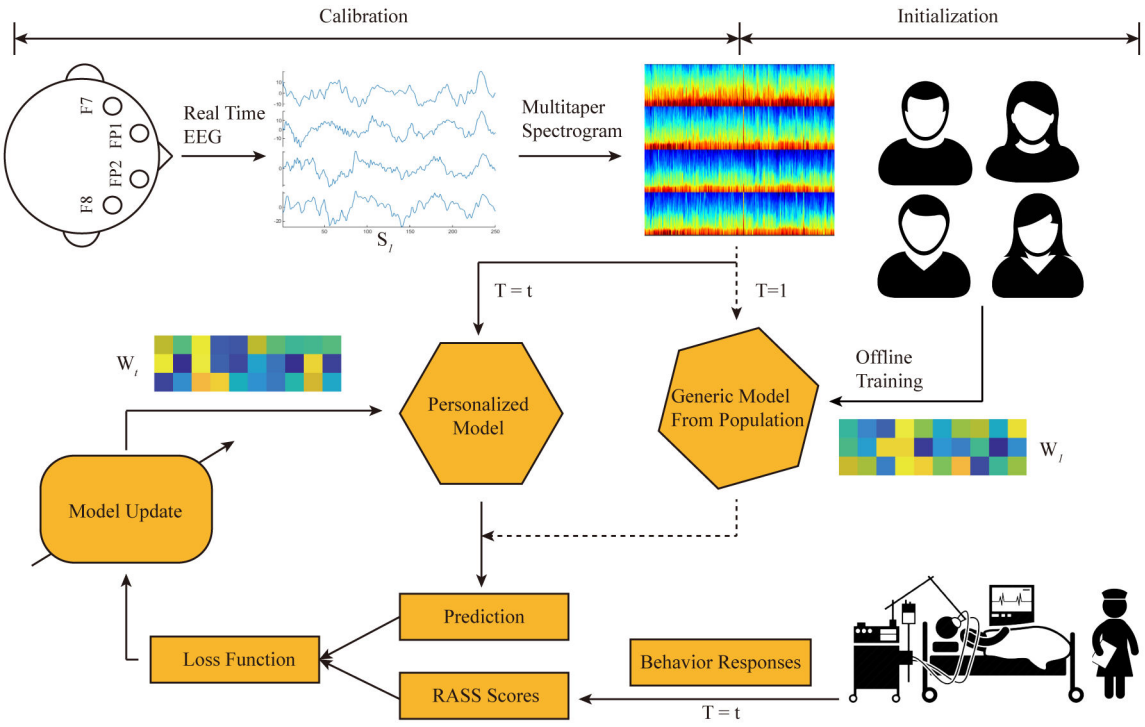

Fig. 4.

The proposed adaptive clinician-in-the-loop framework for sedation monitoring in our study. Our framework includes two components: offline initialization and online calibration. A patient-independent model is trained from the patient population (excluding the current patient) in the offline initialization phase. The model is then adaptively updated for the current patient using sequential EEG data and corresponding RASS scores in the online calibration phase.

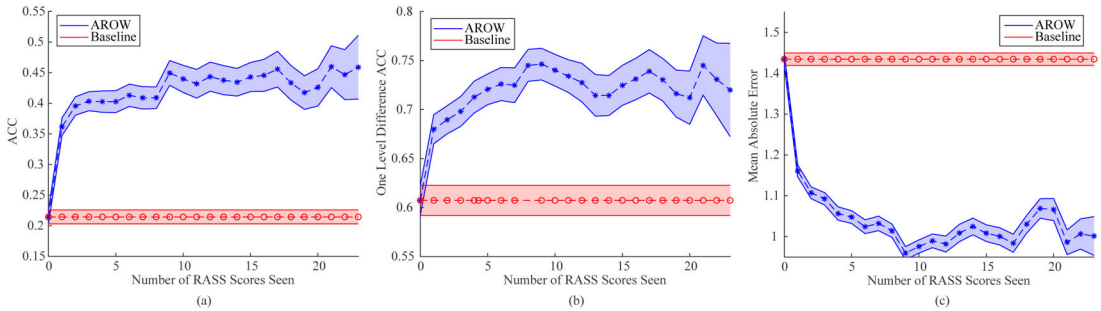

Fig. 5.

Comparison of the baseline method without update versus AROW over time. The x axis denotes the number of RASS scores seen by the models; the y axis denotes different evaluation metrics. The blue and red colors represent AROW and the baseline model, respectively. (a) The mean accuracies, (b) one level difference accuracies, (c) mean absolute errors. The lines denote the mean metrics and the shaded areas denote the standard errors (confidence).

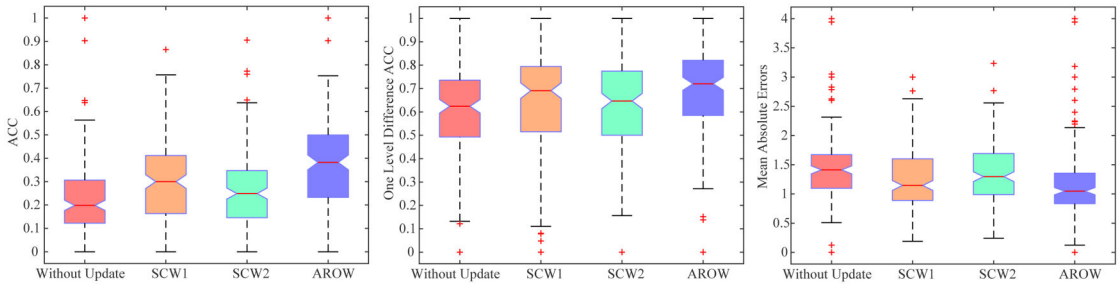

Fig. 6.

Performance comparisons of different models: baseline model without update, Soft Confidence-Weighted Learning 1 and 2 (SCW1, SCW2), and Adaptive Regularization of Weight Vectors (AROW). AROW achieves the highest mean accuracies and lowest mean absolute errors.

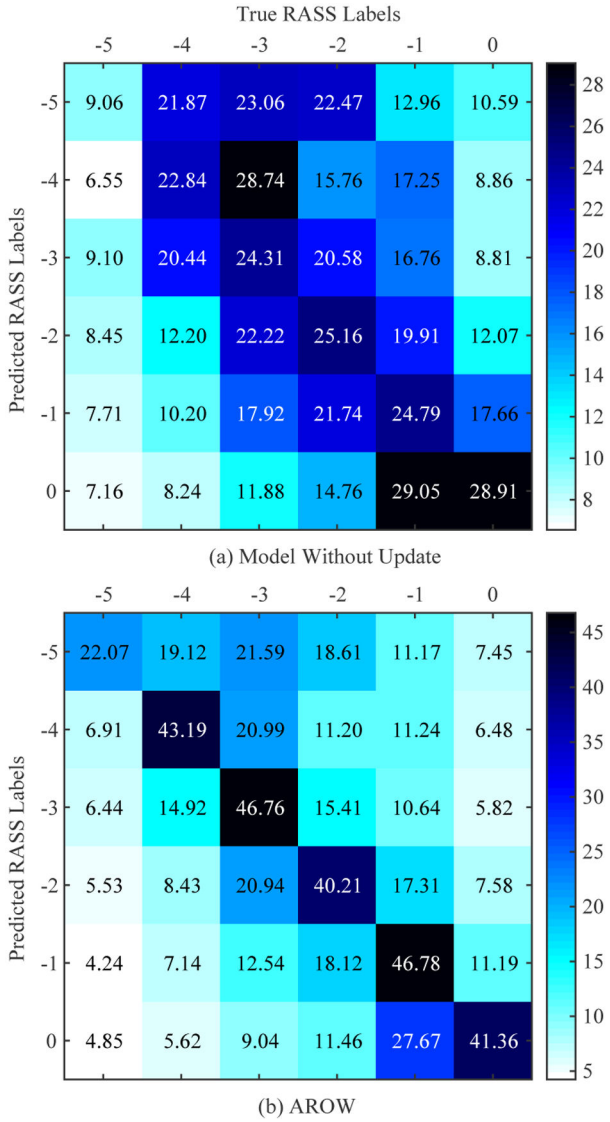

Fig. 7.
The confusion matrices of the baseline method and AROW.

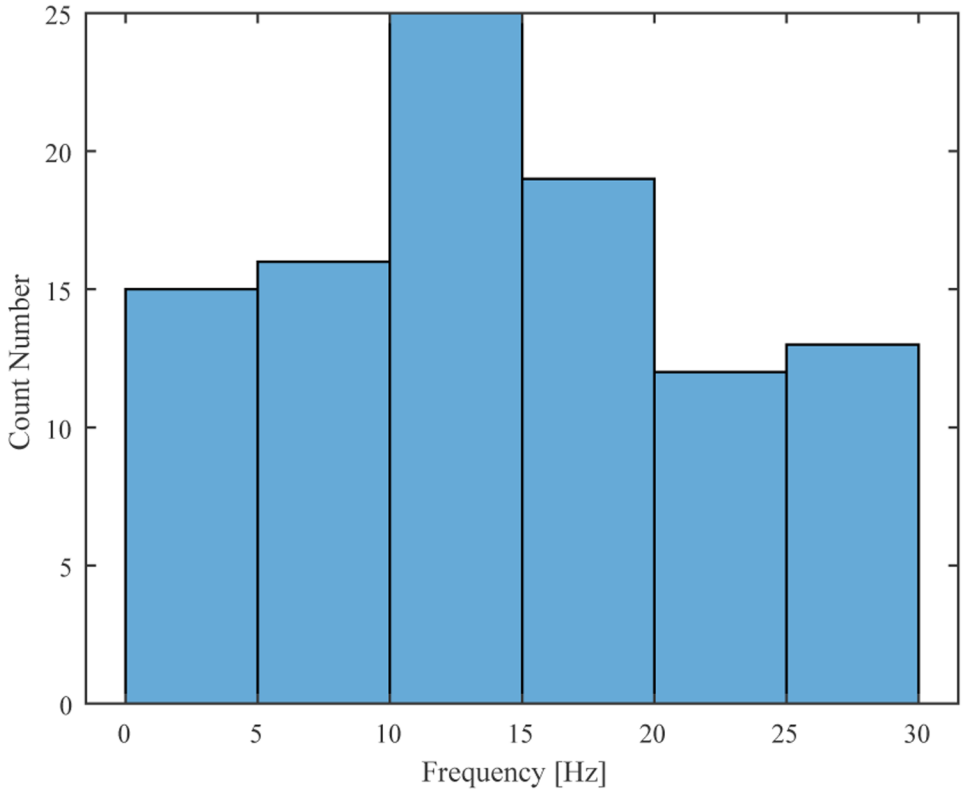

Fig. 8.
The top 100 feature distribution according to the average weights learned from all patients.

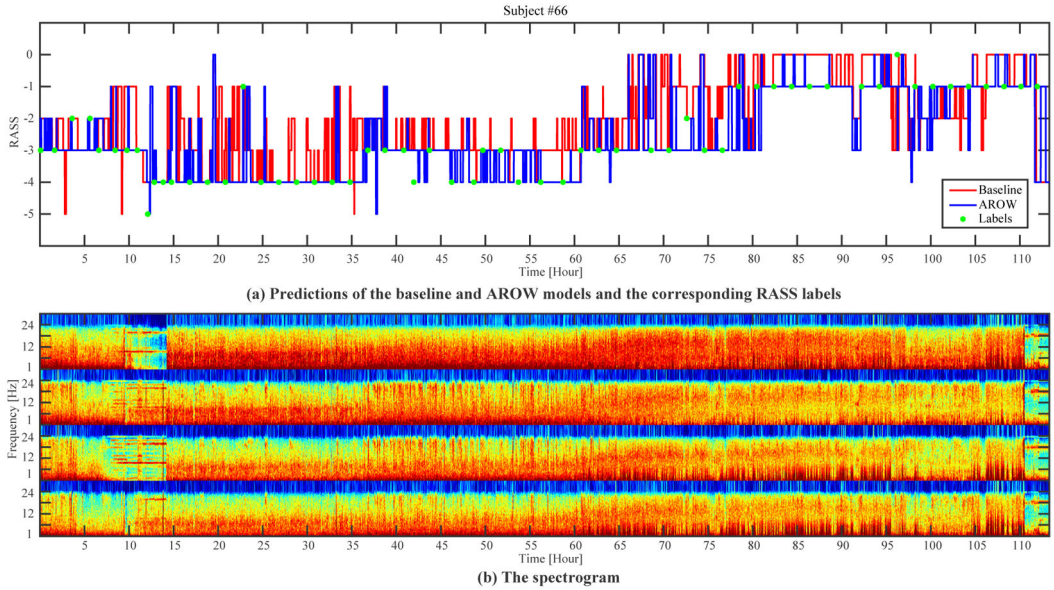

Fig. 9.

Example of sedation monitoring in one patient. (a) The predictions from the baseline model without update and the AROW model as well as the RASS assessments from the patient's nurse ('true labels') over time. (b) The corresponding spectrogram.

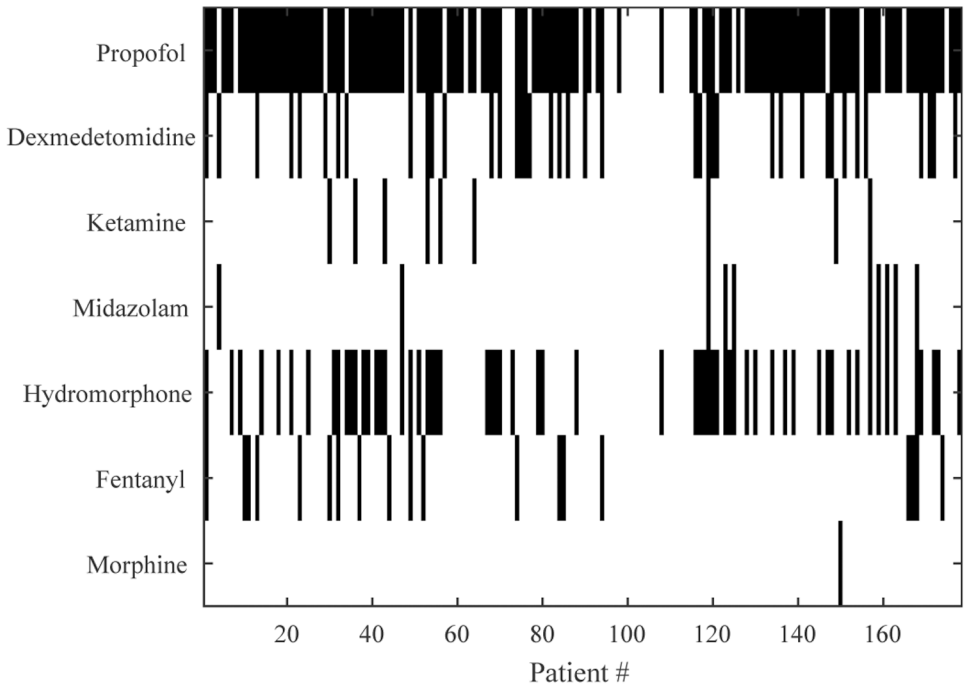

Fig. 10.

Drug information across patients (only seven drugs are shown). Black and white colors denote administration and no administration, respectively.

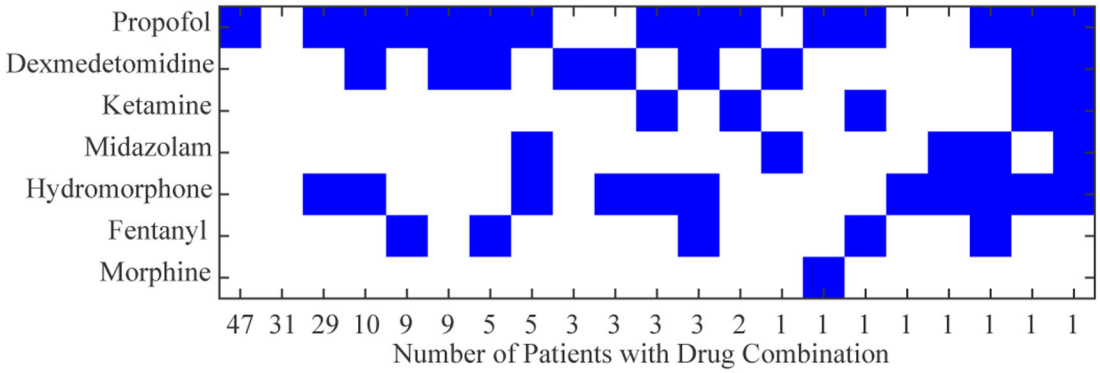

Fig. 11.

Number of patients with different drug combination (only seven drugs are shown). Blue and white colors denote administration and no administration, respectively.

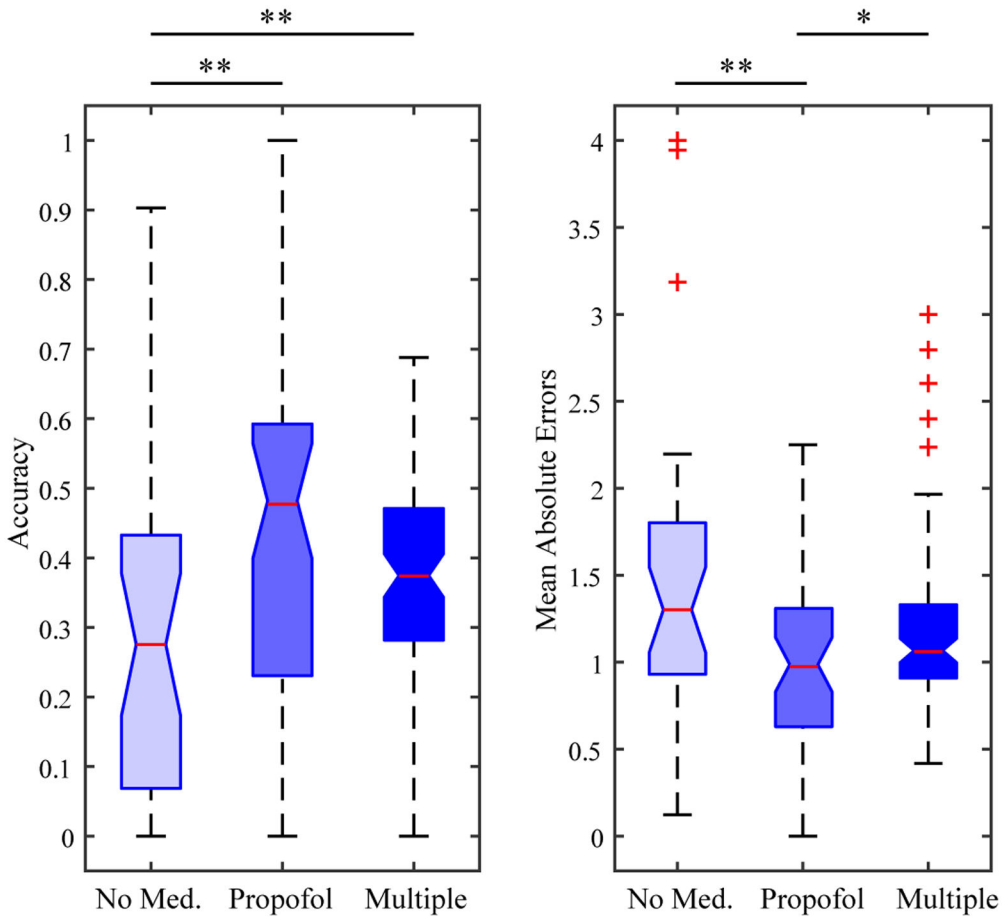

Fig. 12.

Model performance in three patient groups: 1) no medications, 2) propofol only, and 3) multiple medications. * and ** denote $p < 0.05$ and $p < 0.01$, respectively (Wilcoxon rank sum test).

TABLE I

Patient Characteristics

Characteristics	Value
Number of patients	172
Number of assessments	3557
Number of assessments per patient (IQR)	13 (5.5, 28)
RASS, N (% among all assessments)	
0 (Alert and calm)	448 (12.59%)
−1 (Drowsy)	777 (21.84%)
−2 (Light sedation)	725 (20.38%)
−3 (Moderate sedation)	827 (23.25%)
−4 (Deep sedation)	584 (16.42%)
−5 (Unarousable)	196 (5.51%)
Age: year, median (IQR)	60 (50,70)
Male, N (%)	115 (66.86%)
Race, N (%)	
White	149 (86.63%)
Black or African	11 (6.40%)
Asian	2 (1.16%)
More than one race	1 (0.58%)
Unknown or not reported	9 (5.23%)
BMI: kg/m2, median (IQR)	28.31 (23.14, 33.07)
Days in ICU: day, median (IQR)	12 (7, 20)
APACHE II at ICU admission: median (IQR)	22 (15, 28)
CCI at ICU admission: median (IQR)	3 (2, 4)
ICU admission diagnosis (N, %)	
Acute respiratory failure	106 (61.63%)
Diabetes mellitus	45 (26.16%)
Surgery (including gastrointestinal)	44 (25.58%)
Chronic obstructive pulmonary disease	31 (18.02%)
Kidney failure, chronic kidney disease	29 (16.86%)
Sepsis	25 (14.53%)
Liver disease, liver failure	25 (14.53%)
Arrhythmia, congestive heart failure, myocardial ischemia	24 (13.95%)
Solid tumor	14 (8.14%)
Peripheral vascular disease	12 (6.98%)
Use of sedative or analgesic agents (N, %) *	
Propofol	131 (76.16%)
Hydromorphone	61 (35.47%)

Characteristics	Value
Dexmedetomidine	39 (22.67%)
Fentanyl	19 (11.05%)
Midazolam	10 (5.81%)
Ketamine	9 (5.23%)
Morphine	1 (0.58%)

BMI = body mass index, IQR = interquartile range, APACHE = Acute Physiology and Chronic Health Evaluation, CCI = Charlson comorbidity index

* Some patients received more than one sedative or analgesic agent.

TABLE II

Experimental results (median and IQR values) of model performance under conditions of label noise. The percentages of samples with label noise increases from 10% to 50%.

Percentages of Label Noise	10%	20%	30%	40%	50%
Accuracy	0.3413 (0.2271, 0.4487)	0.3216 (0.2194, 0.4226)	0.2991 (0.1994, 0.3928)	0.2793 (0.1940, 0.3685)	0.2504 (0.1759, 0.3390)
1-Level Difference ACC	0.7138 (0.5951, 0.8059)	0.7071 (0.5937, 0.7933)	0.7021 (0.5988, 0.7768)	0.6990 (0.5915, 0.7658)	0.7045 (0.5730, 0.7723)
Mean Absolute Error	1.1134 (0.8534, 1.3875)	1.1535 (0.8928, 1.4435)	1.1699 (0.9377, 1.4481)	1.1953 (0.9649, 1.4855)	1.2298 (1.0095, 1.4633)