 Adaptive Sedation Monitoring from EEG in ICU Patients with Online Learning
Wei-Long Zheng, Member, IEEE, Haoqi Sun, Oluwaseun Akeju, and M. Brandon Westover*

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.

Index Terms—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 [3]. 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 successfully applied to diverse problems in EEG interpretation [17]–[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 complex.

This work was supported in part by grants from NIH-NINDS (1K23NS090900, 1R01NS102190, 1R01NS102574, 1R01NS107291). Asterisk indicates corresponding author (M.B. Westover: mwestover@mgh.harvard.edu).

W.-L. Zheng, H. Sun, and M. B. Westover are with the Department of Neurology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, Massachusetts, USA.

O. Akeju is with the Department of Anaesthesia, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, Massachusetts, USA.
we study there are various phenotypes in the patient population 
which denotes the domain with features 
widely across patients and over time. In effect, the EEG data 
are generated from a continually evolving underlying stochastic 
data have different distributions, i.e. when 
$P_X \neq P_Y$, however, in the problem that we address, both the 
conditional distribution $P_{Y|X}$ and marginal distribution $P_{X}$ 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 
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.

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 non-separable 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.

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.
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 (comatose) and 0 (alert/awake). 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...
was performed using a Kalman filter to remove high frequency noise above 500 µV from each preprocessed 4 sec EEG segment as a noise-free sedation signature. For Kalman filtering, we used the noisy observations in order to estimate the underlying noisy view of the underlying state. The Kalman filter smoothes and removes 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 (< 0.2µV) 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, F7-F8 in the bipolar montage, thus the total feature dimension of each channel and 2 and 3, respectively. The feature dimension of each channel was 480. Feature smoothing 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.

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 \( x_t \in \mathbb{R}^d \) 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 \( (x_t, y_t) \) and remains unchanged until the next time a labeled sample of data arrives. In the present study, we use a...
Adaptive Regularization of Weight Vectors (AROW) algorithm

The distribution of previous parameters and current parameters, states evolve slowly and the current parameters contain distortion too dramatically. This is desirable because sedation µ may be reduced so that we get higher confidence for estimating µ. The model predicts the label according to sign(µ · x) using the average weight vector E[|w|] = µ. The loss function of AROW at time step t is:

\[ L(µ, Σ) = \frac{1}{2} log\left(\frac{det(Σ_{t-1})}{det(Σ)}\right) + \frac{1}{2}T(Σ_{t-1}^{-1} - Σ_{t-1} + \frac{1}{2}(µ_{t-1} - µ)^\top Σ_{t-1}^{-1}(µ_{t-1} - µ) - \frac{d}{2}} \]

where \( D_{KL} \) is the Kullback-Leibler divergence between the distribution of previous parameters and current parameters, \( l_{h_z}(y, µ, x_i) = (max\{0, 1 - y_i(µ \cdot x_i)\})^2 \) is the squared-hinge loss, and \( λ_1, λ_2 ≥ 0 \) are two tradeoff hyperparameters, \( λ_1 = λ_2 = 1/(2r) \).

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 beneficial to rewrite the Kullback-Leibler divergence in the cost function (Equation 1) to obtain

\[ L(µ, Σ) = \frac{1}{2} log\left(\frac{det(Σ_{t-1})}{det(Σ)}\right) + \frac{1}{2}T(Σ_{t-1}^{-1} - Σ_{t-1} + \frac{1}{2}(µ_{t-1} - µ)^\top Σ_{t-1}^{-1}(µ_{t-1} - µ) - \frac{d}{2}} + \frac{1}{2r}l_{h_z}(y, µ, x_i) + \frac{1}{2r}x_i^\top Σx_i. \]

The loss function can be further decomposed into two terms: \( L_1(µ) \) and \( L_2(Σ) \), which are only associated with µ and Σ, respectively. Therefore, the optimization of µ and Σ can be performed independently. Thus we can update the parameters alternately as follows:

1) Update the mean parameters:

\[ µ_t = arg\min_µ L_1(µ), \]

2) If \( µ_t ≠ µ_{t-1} \), update the confidence parameters:

\[ Σ_t = arg\min_Σ L_2(Σ). \]

Taking the derivative of \( L(µ, Σ) \) with respect to µ and setting it to zero, we get

\[ µ_t = µ_{t-1} - \frac{1}{2r}l_{h_z}(y_t, x_t) = µ_{t-1}Σ_{t-1}x_i. \]

Substituting the derivative of the squared-hinge loss in Equation 5and assuming \( 1 - y_i(µ \cdot x_i) ≥ 0 \), we get

\[ µ_t = µ_{t-1} + \frac{y_t}{r}(1 - y_i(µ \cdot x_i))Σ_{t-1}x_i. \]

We reorganize Equation 6 to obtain

\[ µ_t = µ_{t-1} + \frac{max(0, 1 - y_i x_i^\top µ_{t-1})}{x_i^\top Σ_{t-1}x_i + r}Σ_{t-1}y_i x_i. \]

The update for the confidence parameters is made only if \( µ_t ≠ µ_{t-1} \). By setting the derivative of \( L(µ, Σ) \) in Equation 2 with respect to Σ to zero, we get

\[ Σ_{t-1}^{-1} = Σ_{t-1}^{-1} + \frac{x_i x_i^\top}{r}. \]

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

\[ Σ_t = Σ_{t-1} - \frac{Σ_{t-1} x_i x_i^\top Σ_{t-1}}{r + x_i x_i^\top}. \]

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 µ_T and corresponding covariance matrix Σ_T.

1: Initialize µ_0 and Σ_0 with labeled data from previous patients.
2: For \( t = 1, ..., T \)
   • Receive a training sample \( x_i \in \mathbb{R}^d \).
   • Compute margin and confidence: \( m_t = µ_{t-1} \cdot x_i, v_i = x_i^\top Σ_{t-1} x_i \).
   • Receive true label \( y_i \), and loss \( l_t = 1 \) if \( sign(m_t) ≠ y_i \).
   • If \( m_t y_i < 1 \), update using Equations 7 and 9.

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 individuals. 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 \( M_0 \) was trained on data from all patients other than patient \( i \) and used to predict the RASS values \( \hat{y}_{i,j} \), where \( j = 1, 2, ..., n_i \), 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 \( y_{i,j} \); \( e_{i,j} = e(y_{i,j}, \hat{y}_{i,j}) \); here \( e(x, y) = 0 \) 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, \( E_i = \frac{1}{n_i} \sum_{j=1}^{n_i} e_{i,j} \), and the average performance across all patients, \( E_0 = \frac{1}{N} \sum_{i=1}^{N} E_i \), where \( N \) was the number of patients.

For the update model, \( M^k \), \( k = 1, 2, ..., n_i \), 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 \( M^k_{i,j} \), 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:

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

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

\[
(\mu_t, \Sigma_t) = \arg\min_{\mu, \Sigma} D_{KL}(N(\mu, \Sigma)||N(\mu_{t-1}, \Sigma_{t-1}))
\]

\[
s.t. \quad \ell^0(N(\mu, \Sigma); (x_t, y_t)) = 0,
\]

where \( \phi \) is a positive constant and \( \ell^0(N(\mu, \Sigma); (x_t, y_t)) = \max\{0, \sqrt{x_t^\top \Sigma x_t} - y_t (\mu - x_t)\} \), 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 aggressive 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

\[
(\mu_t, \Sigma_t) = \arg\min_{\mu, \Sigma} D_{KL}(N(\mu, \Sigma)||N(\mu_{t-1}, \Sigma_{t-1}))
\]

\[
+ C \ell^0(N(\mu, \Sigma); (x_t, y_t)),
\]

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:

\[
(\mu_t, \Sigma_t) = \arg\min_{\mu, \Sigma} D_{KL}(N(\mu, \Sigma)||N(\mu_{t-1}, \Sigma_{t-1}))
\]

\[
+ C \ell^0(N(\mu, \Sigma); (x_t, y_t))^2.
\]

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
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2019.2943062, IEEE Transactions on Biomedical Engineering

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.

order to robustly estimate the performance trend, only the result of 23 RASS adaptations is shown. From Figure 3 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 $x_t^T \Sigma x_t$ 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.

<table>
<thead>
<tr>
<th>Percentages of Label Noise</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.3413 (0.2271, 0.4487)</td>
<td>0.3216 (0.2194, 0.4226)</td>
<td>0.2991 (0.1994, 0.3928)</td>
<td>0.2793 (0.1940, 0.3685)</td>
<td>0.2504 (0.1759, 0.3390)</td>
</tr>
<tr>
<td>t-Level Difference ACC</td>
<td>0.7138 (0.3931, 0.8039)</td>
<td>0.7071 (0.3931, 0.7933)</td>
<td>0.7021 (0.3988, 0.7768)</td>
<td>0.6990 (0.3915, 0.7638)</td>
<td>0.7045 (0.3730, 0.7723)</td>
</tr>
<tr>
<td>Mean Absolute Error</td>
<td>1.1134 (0.8534, 1.3875)</td>
<td>1.1535 (0.8928, 1.4435)</td>
<td>1.1699 (0.9377, 1.4481)</td>
<td>1.1953 (0.9649, 1.4855)</td>
<td>1.2298 (1.0095, 1.4633)</td>
</tr>
</tbody>
</table>

As label noise increases, performance of the online model degrades, but not dramatically: ACC: 33.74% → 25.50% ($p < 0.01$),
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. 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 administered to the same patient at the same time in the ICU. Figure 11 summarizes the 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 combinations. 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), our work aims to help clinicians perform an important clinical task with greater precision.

![Fig. 10. Drug information across patients (only seven drugs are shown). Black and white colors denote administration and no administration, respectively.](image)

![Fig. 11. Number of patients with different drug combination (only seven drugs are shown). Blue and white colors denote administration and no administration, respectively.](image)

![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).](image)
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. A single model built on the entire patient population may not generalize well to specific patients, whereas 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.

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.

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. Nevertheless some studies have explored EEG-based methods for determining the level of consciousness in ICU patients. 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 administrations, likely due to propofol being the most common sedative used in our patient cohort.

REFERENCES


