Physiological Measurement

PAPER • **OPEN ACCESS**

An unsupervised automated paradigm for artifact removal from electrodermal activity in an uncontrolled clinical setting

To cite this article: Sandya Subramanian *et al* 2022 *Physiol. Meas.* **43** 115005

IOP Publishing Physiol. Meas. 43 (2022) 115005 https://doi.org/10.1088/1361-6579/ac92bd

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
24 May 2022

REVISED
28 August 2022

ACCEPTED FOR PUBLICATION
16 September 2022

PUBLISHED
3 November 2022

PAPER

An unsupervised automated paradigm for artifact removal from electrodermal activity in an uncontrolled clinical setting

Sandya Subramanian[1,*], **Bryan Tseng**[2], **Riccardo Barbieri**[3,4] **and Emery N Brown**[2,4,5,6]

[1] Department of Bioengineering, Stanford University, Stanford, CA, United States of America
[2] Picower Institute for Learning and Memory, Cambridge, MA, United States of America
[3] Department of Electronics, Informatics and Engineering, Politecnico di Milano, Milano, Italy
[4] Department of Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Boston, MA, United States of America
[5] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[6] Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: sandyas@stanford.edu

Keywords: electrodermal activity, artifact detection, artifact removal, unsupervised learning, clinical data

Supplementary material for this article is available online

Abstract

Objective. Electrodermal activity (EDA) reflects sympathetic nervous system activity through sweating-related changes in skin conductance and could be used in clinical settings in which patients cannot self-report pain, such as during surgery or when in a coma. To enable EDA data to be used robustly in clinical settings, we need to develop artifact detection and removal frameworks that can handle the types of interference experienced in clinical settings while salvaging as much useful information as possible. *Approach*. In this study, we collected EDA data from 70 subjects while they were undergoing surgery in the operating room. We then built a fully automated artifact removal framework to remove the heavy artifacts that resulted from the use of surgical electrocautery during the surgery and compared it to two existing state-of-the-art methods for artifact removal from EDA data. This automated framework consisted of first utilizing three unsupervised machine learning methods for anomaly detection, and then customizing the threshold to separate artifact for each data instance by taking advantage of the statistical properties of the artifact in that data instance. We also created simulated surgical data by introducing artifacts into cleaned surgical data and measured the performance of all three methods in removing it. *Main results*. Our method achieved the highest overall accuracy and precision and lowest overall error on simulated data. One of the other methods prioritized high sensitivity while sacrificing specificity and precision, while the other had low sensitivity, high error, and left behind several artifacts. These results were qualitatively similar between the simulated data instances and operating room data instances. *Significance*. Our framework allows for robust removal of heavy artifact from EDA data in clinical settings such as surgery, which is the first step to enable clinical integration of EDA as part of standard monitoring.

Introduction

Artifact detection and removal are required for any physiological data collection, especially in uncontrolled and 'messy' situations like in the hospital or at home (Jiang *et al* 2019). As sensors become more ubiquitous and optimized for comfort and convenience over signal quality, ensuring data quality is increasingly the responsibility of analysis algorithms that can quickly detect and correct artifacts. Specifically, robust artifact removal is required for any physiological modality to become clinical standard, since artifact removal must be integrated into hardware systems to ensure high quality data for clinicians (Jiang *et al* 2019). Some of this artifact is clearly identifiable by eye and attributable to obvious sources such as patient movement, accidental removal or repositioning of sensors, or interference from other equipment (Jiang *et al* 2019). However, automating what

can be seen by eye can prove to be challenging. Common methods for artifact removal in simpler situations, such as thresholding, may not be sufficient for complex clinical environments (Uriguen and Garcia-Zapirain 2015, Mannan *et al* 2018). In addition, artifact rejection strategies must be optimized for minimal collateral damage in terms of removal of true data, especially in cases where temporal dependencies exist (Uriguen and Garcia-Zapirain 2015, Mannan *et al* 2018). Temporal dependencies may also warrant special considerations in methods development, for example favoring removal of multiple smaller sections of data rather than a single continuous section (Mannan *et al* 2018).

Electrodermal activity (EDA) is one such physiological measure that is inexpensive and convenient to collect, but is not yet clinical standard because there are not rigorous tools to process and analyze it (Boucsein 2012). EDA tracks the changing electrical conductance of the skin due to the activity of sweat glands, which are part of the body's sympathetic 'fight or flight' reflex (Boucsein 2012). It has immense potential as a physiological marker to track sympathetic activation in situations such as pain or stress (Amin and Faghih 2022). In clinical settings, it could be used as a non-invasive marker of physiological pain processing in situations in which patients cannot communicate for themselves, such as under anesthesia, during surgery, or when in a coma (Subramanian *et al* 2020a, 2020b, 2021a). Tracking sympathetic nervous system activation and regulation would be of clinical utility in the operating room to dose pain medication or in the ICU to measure brainstem function and nociceptive reflexes (Subramanian *et al* 2020a, 2020b, 2021a). Developing frameworks and methodologies to process EDA, including artifact detection and removal specific to clinical situations, would bring it one step closer to being used in the clinic.

Supervised learning tools have been used successfully in a number of clinical applications, including radiology and pathology (Biagetti *et al* 2018). However, in the case of artifact detection, creating a labeled training set is a non-trivial task that is not part of the clinical workflow. It would require significant manual labor to label each small increment of time as artifact or true data. Previous studies using advanced supervised machine learning methods, including deep learning, have relied on such expert labeled data instances (Kelsey *et al* 2017, Zhang *et al* 2017, Gashi *et al* 2020, Posada-Quintero and Chon 2020, Llanes-Jurado *et al* 2021, Hossain *et al* 2021, 2022a, 2022b). The timescale of artifact is often a fraction of a second, so to minimize the amount of excess data labeled as artifact, the increments of time must be very small, increasing the manual labor of labeling. Different types of artifact may also require specific labeled training sets. Instead, unsupervised methods do not require labeled training sets, since they assign data to groups based on detecting patterns in the data (Goldstein and Uchida 2016).

In this paper, we develop a fully automated pipeline for removing artifact from EDA involving three unsupervised learning methods (isolation forest, K-nearest neighbor distance, and 1-class support vector machine) (Manevitz and Yousef 2001, Liu *et al* 2008, Hu *et al* 2016) and threshold selection process. These three methods were chosen because they are commonly used for anomaly detection. Specifically, we use EDA collected during surgery in the operating room, where there is maximal artifact due to interference from surgical cautery equipment. This is one of the most intense clinical situations, so by showing that we can robustly remove artifact in this scenario, we can demonstrate that our method is adaptable for any clinical situation, which moves EDA one step closer to being clinical standard to track pain and physiological stress in the operating room and ICU. To feed into the automated pipeline, we defined 12 features in half-second windows based on our own intuition and guidance from existing literature.

EDA data were collected continuously during lower abdominal surgery in 69 human subjects. The source of most artifact was surgical electrocautery, which causes large visible deflections in the data every time it is turned on and off, which can be over 150 times in an average surgery at short, irregular intervals. Each time the cautery is turned on, it typically only remains on for a few seconds. While the cautery-induced deflections are clearly visible, to complicate matters, there are periods of intact but shifted (down typically) EDA between the deflections. Finally, the magnitude, sharpness, and direction of artifactual deflections vary across subjects.

Existing unsupervised methods for artifact removal are specific to the data instances for which they were built, none of which included critical clinical situations (Chen *et al* 2015, Taylor *et al* 2015, Zhang 2017). None had the degree of artifact that surgical cautery interference produces. None were clinical EDA data instances. In contrast, the artifact detection and removal pipeline we developed was able to successfully remove even heavy cautery artifact from all subjects' data. In addition, our computational process was able to sufficiently attenuate artifact while preserving as much of the remaining true data as possible, including small snippets of real data in between sections of artifact. Previous work in this area was published in (Subramanian *et al* 2021b); however, it was semi-automated and required manual selection of hyperparameters. In contrast, this current work presents a fully automated pipeline.

In the remainder of this paper, we detail the development and validation of this pipeline. In Methods, we discuss the details of the data collection, subject cohort, the features used, how the unsupervised learning algorithms were implemented, and how the artifact threshold was selected. We detail how we constructed simulated data for comparison of methods. In Results, we show each subject's data before and after artifact

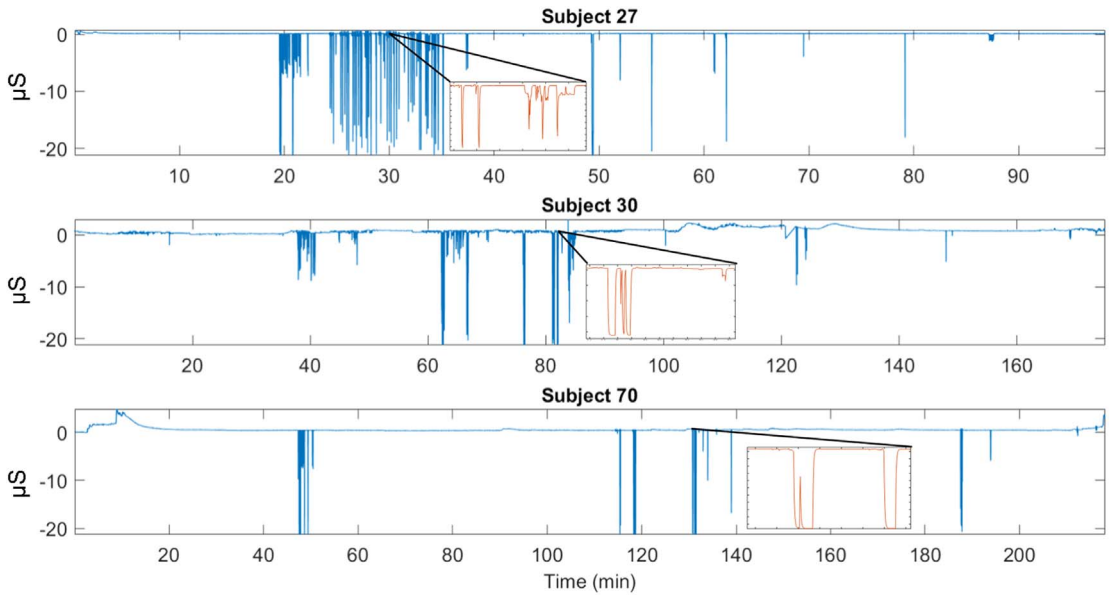

Figure 1. Examples of raw operating room EDA data from three representative subjects.

removal and detail the specific parameters used and fraction of data labeled artifact. We also show a side-by-side comparison with existing artifact removal methods for simulated data. Finally, in Discussion, we address the implications of this work and our future directions.

Methods

Data

In this study, we use EDA data recorded from 70 subjects (38 females and 32 males, ages 29–77), collected under protocol approved by the Massachusetts General Hospital (MGH) Human Research Committee (IRB 2017P002591). The research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements. All participants provided written informed consent to participate in the study and for study results to be published. All subjects were undergoing laparoscopic urologic or gynecologic surgery at MGH. The EDA data were recorded from the most proximal phalanges of two fingers of each subject's left hand at 256 Hz using the Thought Technology Neurofeedback System (Neurofeedback Expert System, Thought Technology Lt .). The electrodes were placed as soon as the patient entered the operating room before induction of anesthesia and only removed after extubation at the end of surgery. The data were fed in real-time to a laptop located at the head of the operating table, near the anesthesiologist, and monitored the whole time by a member of the study team to ensure signal quality. The distribution of ages and surgery durations is shown in Figure S1(available online at stacks.iop.org/PMEA/43/115005/mmedia) in the Supplementary Material. Figure 1 shows an example of the raw data from three subjects. The main sources of artifact were movement at the beginning and end, including positioning, and use of surgical cautery. Each instance of turning cautery on or off caused a visible deflection in the data. Due to logistical concerns, EDA data collection from one subject (Subject 31) was ended before the onset of cautery, and therefore that subject was excluded from this analysis. EDA data from the remaining 69 subjects were analyzed using Matlab 2020b.

Features

The 12 features we used are listed in table 1. These features are a combination of those used by other existing methods (Chen *et al* 2015, Taylor *et al* 2015, Zhang 2017) as well as additional ones that we discovered were useful based on experimentation. We computed these features for each nonoverlapping 0.5 s window (128 samples) for each data instance to match with the timescale of most artifacts. These feature vectors were then fed as inputs into the automated pipeline.

The automated pipeline

Figure 2 is a schematic summarizing our methodology for artifact detection.

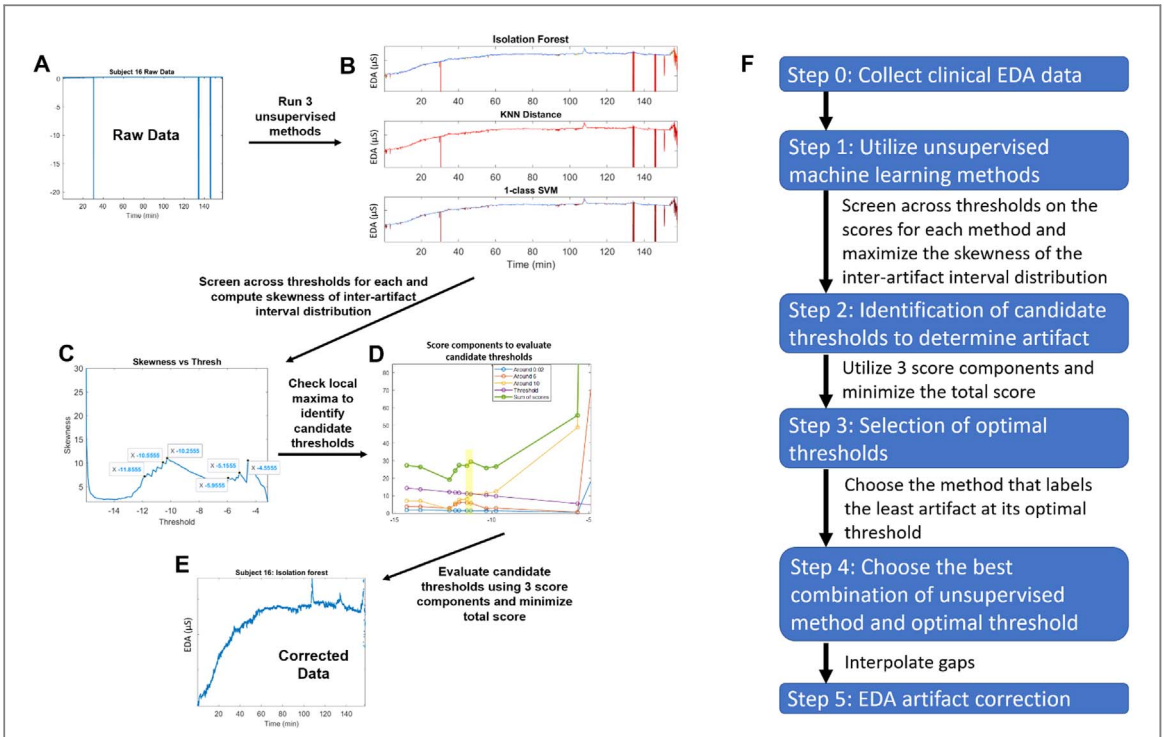

Figure 2. Schematic of methodology. (A) Starting from the raw EDA data, (B) each of the three unsupervised methods yields a score for each 0.5 s window of data, where a higher score is more likely artifact. (C) Screening across thresholds for artifact for all three scores, the labeled artifacts at each threshold can be described by an associated inter-artifact interval distribution. The skewness of that distribution can be computed and plotted by threshold. Selecting a set of local maxima from skewness versus threshold curve, we can arrive at a set of candidate thresholds to evaluation for each unsupervised method. (D) The candidate thresholds can be 'scored' based on the characteristics of the resulting 'cleaned' EDA data in terms of local standard deviation. Three specific characteristics were compared to quantitative limits set from examining artifact-free data and the optimal threshold can be chosen as the one that most closely meets all limits. (E) The best unsupervised method and optimal threshold combination is chosen as the one which has the lowest proportion of labeled artifact while still meeting all other conditions. (F) A flowchart summarizing the methods

Table 1. The 12 features for each 0.5 s window used as inputs for our unsupervised methods.

	Feature description
1	Standard deviation of signal
2	Difference between max and min of signal
3	Mean of first derivative
4	Median of first derivative
5	Standard deviation of first derivative
6	Min of first derivative
7	Max of first derivative
8	Mean of level 4 Haar wavelet coefficients
9	Median of level 4 Haar wavelet coefficients
10	Standard deviation of level 4 Haar wavelet coefficients
11	Min of level 4 Haar wavelet coefficients
12	Max of level 4 Haar wavelet coefficients

Step 1: Utilization of unsupervised machine learning methods

We used three existing unsupervised learning methods to compute scores for each half second segment of data based on the 12-feature vectors, isolation forest (Liu *et al* 2008), K-nearest neighbor (KNN) distance (Hu *et al* 2016), and 1-class support vector machine (SVM) (Manevitz and Yousef 2001). Each isolation forest consisted of 100 decision trees, and the isolation scores were computed as the median of 10 such forests. KNN distance was computed using Euclidean distance and $K = 50$. A 1-class SVM was trained on 90% of the data, based on the 90% with the lowest KNN distance as a conservative estimate of true data and excluding the 10% of data points with the greatest KNN distance. All three unsupervised learning methods yielded a score for each window of data quantifying the degree of abnormality. The higher the score, the more likely that segment of data was artifact. The isolation forest scores were made negative to match the directionality of the other two.

Step 2: Identification of candidate thresholds

The next step of the process was to determine the appropriate threshold to define artifact for each method for each subject. To do this, first we identified a set of candidate thresholds to evaluate. The process used to select these thresholds relied on specific insight about how the unsupervised methods label artifact. For each data instance, as the threshold on any of the unsupervised method scores is decreased, the portions of data that are labeled artifact increase in discrete 'jumps' with more subtle changes in between. The most 'correct' labeling of artifact is likely to occur at one of these discrete jumps, since each jump represents the additional labeling of one similar 'cluster' of data as artifact, whereas gradual changes represent a continuous spectrum of subtle differences within similar 'clusters'. True artifact is highly similar to each other and distinctly different from true data; therefore, there should be no need to rely on subtle differences. To identify the discrete jump that represents the most 'correct' labeling of artifact, we took advantage of the fact that each discrete jump dramatically changes the inter-artifact interval distribution by introducing long gaps between subsequent artifact labels. Therefore, the skewness (3rd moment) of the inter-artifact interval distribution was computed across thresholds for each unsupervised method (DeGroot and Schervish 2012). Since discrete jumps in labeled artifact skew the inter-artifact interval distribution, the jumps can be identified by local maxima in skewness. The candidate thresholds for each unsupervised method were identified using the *findpeaks* algorithm in Matlab to identify local maxima in the skewness versus threshold curve, using minimum peak prominence of 0.1. Peak prominence is defined as the height of a local maximum above the higher of the two neighboring troughs on either side of the peak.

Step 3: Selection of the optimal thresholds

Each of the candidate thresholds was assessed by detecting and removing artifact using that threshold and then computing three metrics on the corrected signal: the maximum standard deviation in any half-second window ($localSTD_{\max}$), the ratio of the maximum standard deviation in any half-second window to the 90th percentile standard deviation in any half-second window ($localSTD_{\max}/localSTD_{90}$), and the ratio of the maximum standard deviation in any half-second window to the median standard deviation in any half-second window ($localSTD_{\max}/localSTD_{med}$). Then, the three metrics were converted into score components by computing their differences from the limits of 0.48, 6, and 10 respectively, and penalizing the distance above each limit by twice as much as the distance below, as described in the formula below. The limits were chosen by examining characteristics of artifact-free EDA data.

$$score_1 = |localSTD_{\max} - 0.48| * 100^* \\ \times (1.5 + 0.5 * sign(localSTD_{\max} - 0.48))$$

$$score_2 = \left| \frac{localSTD_{\max}}{localSTD_{90}} - 6 \right| * \left(1.5 + 0.5 * sign\left(\frac{localSTD_{\max}}{localSTD_{90}} - 6\right)\right)$$

$$score_3 = \left| \frac{localSTD_{max}}{localSTD_{med}} - 10 \right| * \left(1.5 + 0.5 * sign\left(\frac{localSTD_{max}}{localSTD_{med}} - 10\right)\right)$$

The final score for each candidate threshold was computed as the sum of the three score components and the value of the threshold itself (penalizing higher thresholds). The optimal threshold for each unsupervised method was chosen as the candidate threshold with the minimum final score. The proportion of data labeled artifact at this threshold was recorded.

Step 4: Choosing the best combination of unsupervised method and optimal threshold

The best unsupervised method for each data instance was chosen as the one that labeled the smallest proportion of artifact at its optimal threshold (implying that it also satisfied all previous conditions). The goal is to select the combination of unsupervised method and optimal threshold that is the most precise in its selection of artifact without compromising true data.

Step 5: Correcting EDA after artifact detection

After identifying and removing the artifact while preserving as much of the true data as possible using the optimal threshold and best unsupervised method, any 'islands' of true data that were shifted upward or downward due to artifactual deflection were translated back based on computing the linearly interpolated mean of the data at that time. Islands were defined as being shorter than 20 s in duration and more than 0.12 μS from the linearly interpolated mean of the neighboring EDA data. After translating the 'islands' back, the gaps created by removed artifact were filled using linear interpolation once more to create continuous data. This is why the duration of the longest continuous artifact was relevant. Using linear interpolation to fill in a few seconds of data at a time will likely not affect downstream analysis; however, interpolating a few minutes at a time could.

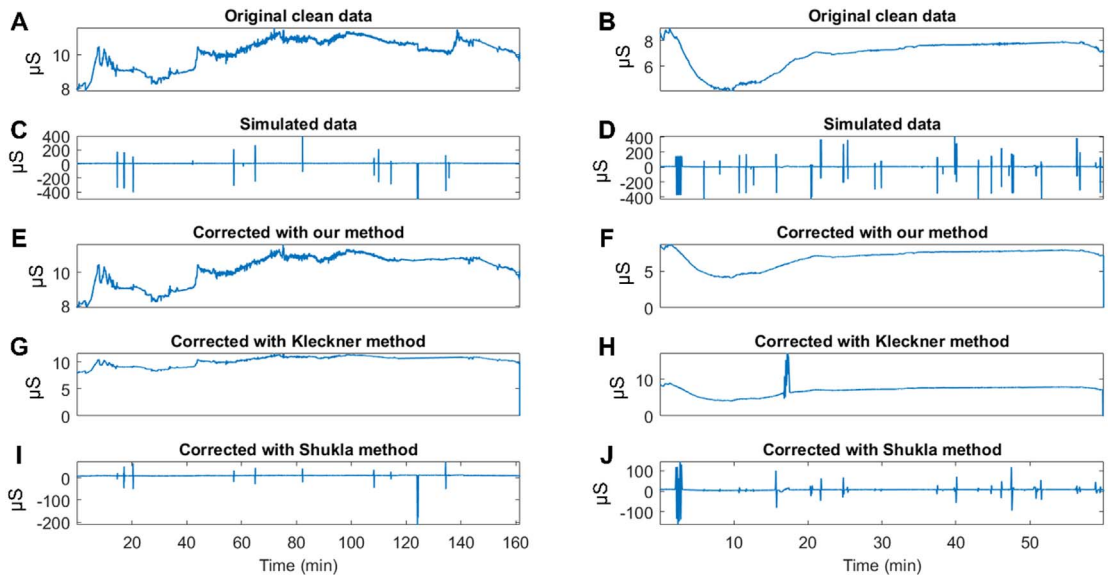

Figure 3. Two example simulated data instances starting from the original 'ground truth' clean data for each, then the simulated data instances with artifacts artificially introduced, and finally the corrected simulated data instances after artifact correction using our method, the Kleckner *et al* method and the Shukla *et al* method.

Validation

We used our pipeline as well as two other methods on the 69 data instances collected during surgery. These two methods were chosen because they were relatively recent methods that were built upon the two major schools of thought with respect to automated and unsupervised EDA artifact correction. The two other methods were a heuristic method which thresholds the signal value and derivative and removes 5 s of data on either side of any identified artifact (Kleckner *et al* 2018) and a wavelet decomposition-based method (Shukla *et al* 2018). The Kleckner method (Kleckner *et al* 2018) only covers artifact detection, but does not provide a method to fill in sections labeled as artifact. We used linear interpolation to fill in those regions similar to our own pipeline. Since there is no 'ground truth' with which to quantify the performance of all three methods on the true surgical data, we quantified this comparison using simulated data as detailed in the following sections.

Step 1: Creating simulated data

We started with the corrected surgical EDA data returned by our pipeline. We used this as the ground truth signal from which to start. We randomly selected 50 subjects' corrected EDA in which to insert artifacts to create 50 simulated EDA data instances. Then, we created a 'database' of artifacts by aggregating all of the sections of EDA labeled as artifact by any of the three methods across all 69 surgical data instances. This database contained over 29,000 artifacts of varying shapes and durations; some were likely not truly artifact since the algorithms are imperfect in their labeling of artifact. For each of the 50 corrected EDA data instances randomly selected, the following process was followed to construct a simulated EDA data instance with artifact:

1. Randomly select a number N between 50 and 150 for the number of artifacts to introduce to the clean data.
2. Randomly select N artifacts from the artifact database.
3. Randomly select N locations from 1 to the length of the EDA data instance at which to introduce each artifact.
4. For each (artifact, location) pair, replace the segment of clean EDA data starting at chosen location and of the same duration as the chosen artifact with the artifact. The artifact was inserted so that the mean value of the artifact was the same as the mean value of the clean EDA segment it replaced.

Examples of two simulated data instances constructed in this manner are shown in figure 3.

Step 2: Evaluating the performance of artifact removal methods on simulated data

We detected and removed the artifact from each simulated EDA data instance using each of the three methods being compared: our pipeline, the Kleckner *et al* heuristic method (Kleckner *et al* 2018), and the Shukla *et al* wavelet-based method (Shukla *et al* 2018). We compared the performance of the three methods both in terms of

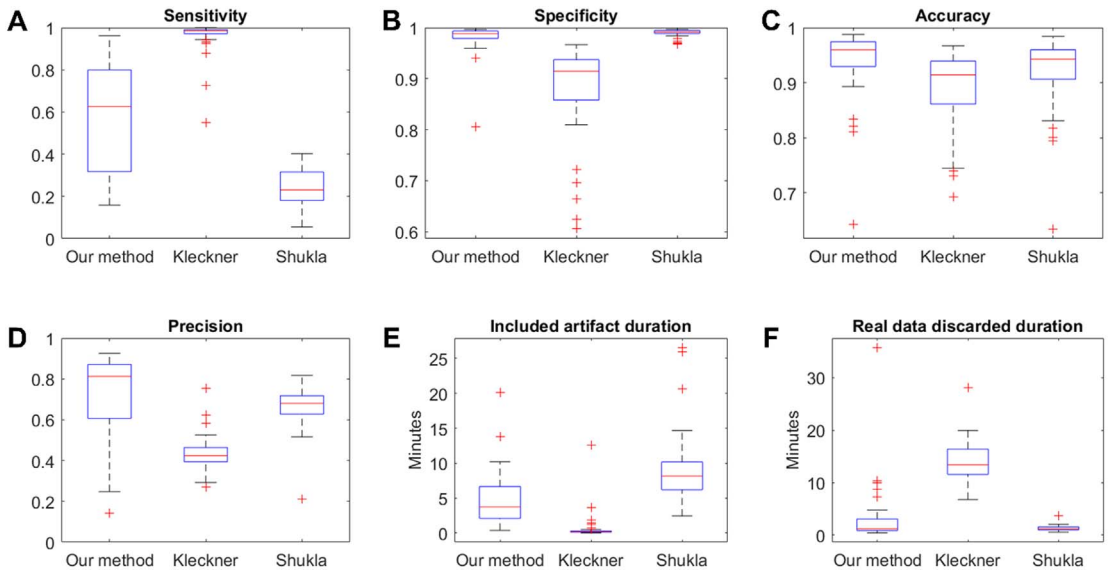

Figure 4. Results of all three artifact removal methods across all 50 simulated data instances, specifically with respect to artifact detection. The six measures shown are sensitivity, specificity, overall accuracy, precision, the duration of artifact left behind, and the duration of true EDA data discarded as artifact.

artifact detection alone, as well as artifact correction. The 50 corrected EDA data instances prior to the introduction of any artifacts were treated as ground truth for comparison. Artifact detection was evaluated by measuring sensitivity, specificity, overall accuracy, and precision for each simulated EDA data instance. Some of these were also quantified in terms of actual duration of data rather than just proportion. Artifact removal was evaluated by computing the sum of squared errors (SSE) from the ground truth in artifact regions, in non-artifact regions, overall, and by computing the artifact power attenuation (APA) (Molavi and Dumont 2012). A greater APA over regions of true artifact and lower SSE in any case are desirable. (Computing the APA over regions of identified rather than true artifact is less optimal but unavoidable when ground truth is unknown.) In this case, the quality of the artifact removal is more important than artifact detection alone, since the artifacts introduced into each simulated EDA data instance from the artifact 'database' included non-artifacts falsely labeled as artifact by at least one of the methods. The quality of the artifact removal indicates how similar the final EDA data is to the ground truth.

We also computed the APA for each of the true surgical EDA data instances using all 3 methods to compare the performance of the methods on the true data versus the simulated data.

Results

Figure 3 shows two examples of simulated data instances from generation through correction with all three artifact correction methods (our method, Kleckner *et al* and Shukla *et al*). Similar figures for all of the simulated data instances can be found in the Supplementary Material, figures S37–61. Figure 3 shows that the process we used to generate simulated data successfully introduced significant artifacts into the original cleaned data. In addition, it shows that our method and the Kleckner method are qualitatively similar at removing artifact, whereas the Shukla method leaves behind significant artifact. With respect to artifact detection alone, figure 4 shows that our method achieves the highest median accuracy of 96%, while the Shukla method achieves a median of 94% and the Kleckner method a median of 91%. Keeping in mind that some of the introduced 'artifacts' were likely not true artifacts, figure 4 also shows that our method achieves varying levels of sensitivity with a median of around 63%. The Kleckner method achieves high median sensitivity of 98%, while the Shukla method achieves only around 23% median sensitivity. However, with respect to specificity, this trend is reversed, with both our method and the Shukla method achieving 99% median specificity while the Kleckner method achieves only 91%.

To further understand the performance of the Kleckner method in comparison to ours, we also examined the precision of all the methods as well as quantified the sensitivity and specificity in terms of actual durations since the proportion of artifact in most data instances is relatively small. Figure 4 shows the results from the three artifact correction methods across all 50 simulated data instances specific to artifact detection. Figure 4 shows that the Kleckner method has the lowest median precision of only 42%, while our method achieves the highest

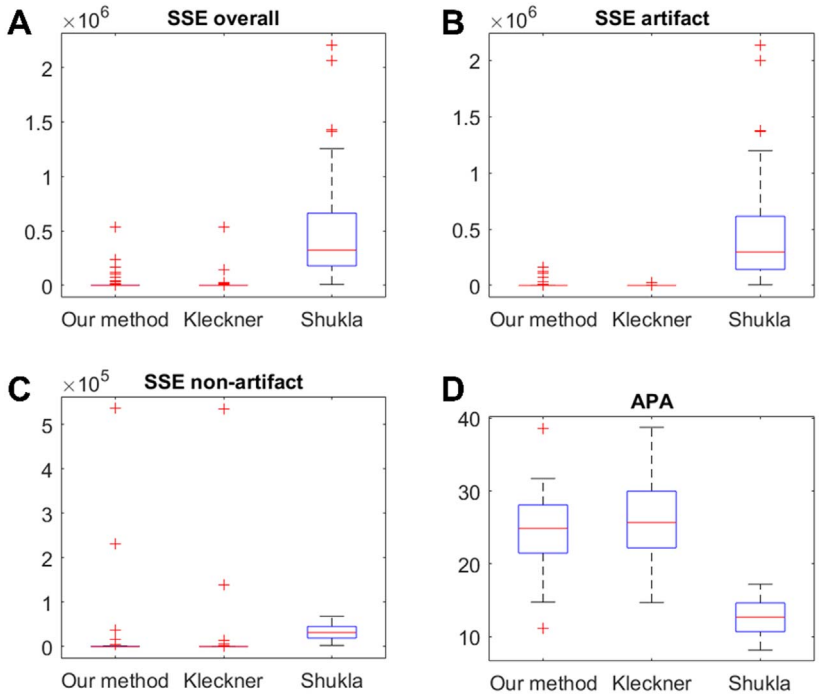

Figure 5. Results of all three artifact removal methods across all 50 simulated data instances, specifically with respect to artifact correction. SSE = sum of squared errors, APA = artifact power attenuation. The four measures shown are the sum of squared errors overall, the sum of squared errors in the true artifact regions, the sum of squared errors in the true non-artifact regions, and the artifact power attenuation for the ground truth artifact regions (not the regions that were identified by each method as artifact).

median precision of 81% and the Shukla method 68%. The Kleckner method grossly overestimates the artifact, allowing it to achieve high sensitivity but lower precision and specificity. When quantifying in actual duration, the Kleckner method leaves very little artifact behind (0.15 min); however, it mislabels a median of 13.4 min of true data as artifact per data instance. In contrast, our method leaves behind about 3.7 min of artifact per data instance but only mislabels about 1.2 min of true data per data instance. Finally, the Shukla method leaves behind the most artifact (8.1 min) and mislabels a similar duration as our method (1.2 min).

To fully understand the performance of the different methods, artifact detection alone is not enough, especially since all of the introduced 'artifact' may not have been true artifact in this case. The clearest indicator of performance is the similarity between the final corrected signal and the original ground truth signal to which artifacts were artificially added for each data instance. Figure 5 shows the results of the three methods across all 50 simulated data instances specific to artifact correction figure 5 shows that despite a marked difference in sensitivity, our method achieves very similar performance in terms of error compared to the Kleckner method, in terms of the overall error (our method median = 150, Kleckner method median = 290) as well as error in artifact regions (our method median = 30, Kleckner method median = 20) and non-artifact regions (our method median = 25, Kleckner method median = 140). In contrast, the Shukla method, which was the least sensitive, has a much greater error because it leaves behind significant artifact (median overall error = 3.2e5, median error in artifact regions = 3.0e5, median error in non-artifact regions = 3.2e4). In terms of APA, which was calculated in this case for the ground truth artifact regions, not the regions identified as artifact by each method, figure 5 shows that the Kleckner method and our method have similar APA (our method median = 24.8, Kleckner method median = 25.6), while the Shukla method is far less (Shukla method median = 12.7). Finally, in terms of the total proportion of labeled artifact and the longest continuous artifact, figure 6 shows that the Kleckner method achieves the highest in both since it labels the most data as artifact (Kleckner median proportion labeled artifact = 0.15, Kleckner median longest continuous artifact = 103 s). Our method achieves higher precision and similar APA for true artifact regions while labeling a median of only 5% of the data as artifact and achieving a median longest continuous artifact length of only 24 s. The Shukla method leaves behind significant artifact and therefore only labels a median of 2% of the data as artifact.

For the operating room data, there is no ground truth. Figure 7 shows examples of the performances of all three methods on two operating room data instances. Similar figures for all of the operating room data instances can be found in the Supplementary Material, figures S2–36. Figure 7 shows that the Shukla method performs similarly compared to simulated data instances; it leaves many significant artifacts behind. The Kleckner method seems to perform worse on the operating room data than on the simulated data instances and also leaves behind

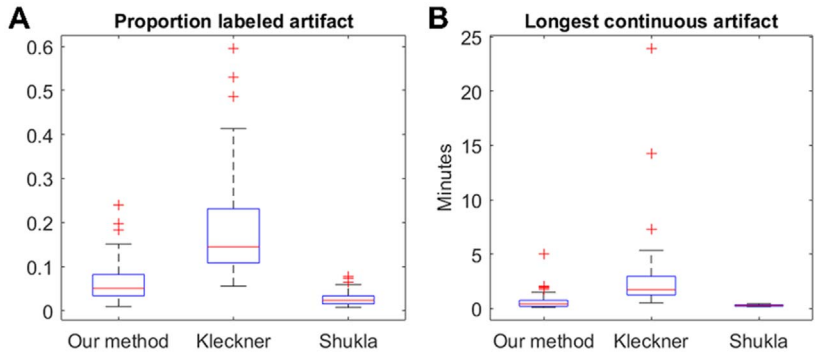

Figure 6. Results of all three artifact removal methods across all 50 simulated data instances, in terms of the proportion of signal labeled as artifact and the longest continuous single artifact.

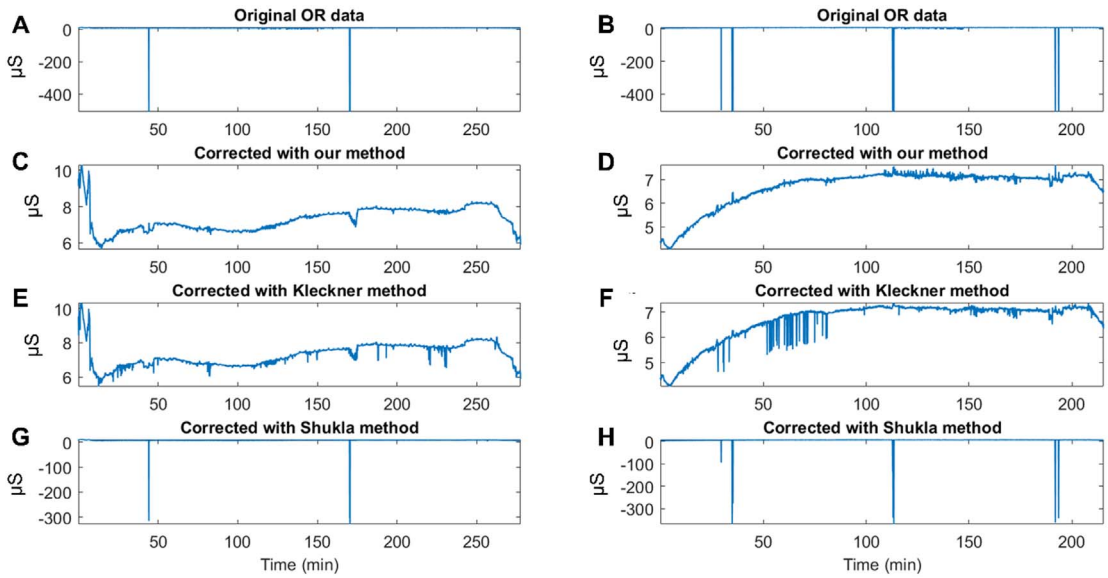

Figure 7. Two example operating room data instances starting from the raw EDA data for each, and then the corrected operating room data instances after artifact correction using our method, the Kleckner *et al* method and the Shukla *et al* method.

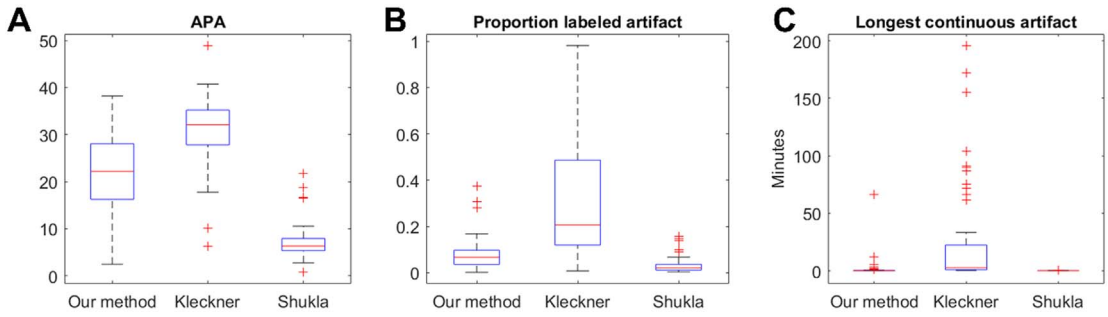

Figure 8. Results of all three artifact removal methods across all 70 operating room data instances, in terms of the artifact power attenuation of the regions identified by each method as artifact, the proportion of signal labeled as artifact, and the longest continuous single artifact.

some artifact. In terms of the quantifiable measures, figure 8 shows that the trends for proportion of labeled artifact and longest continuous artifact are similar to the simulated data instances. The Kleckner method labels the most and longest artifacts (Kleckner median proportion labeled artifact $=$ 0.21, Kleckner median longest continuous artifact $=$ 165 s), while our method labels less (our method median proportion labeled

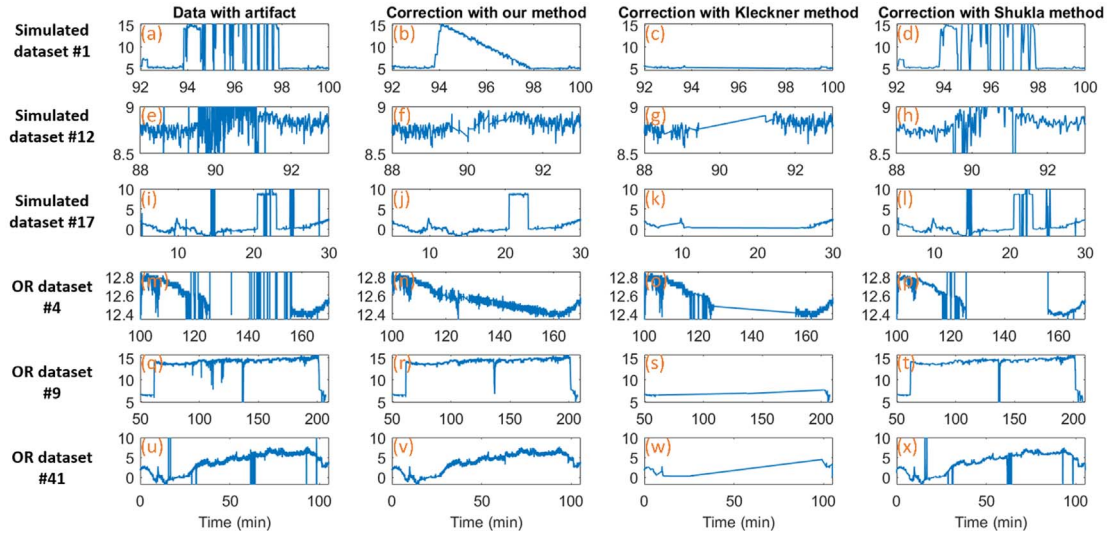

Figure 9. Excerpts of simulated and operating room data instances with artifact and after artifact removal with all three methods. All y-axes are in μS.

artifact = 0.07, our method median longest continuous artifact = 20 s), and the Shukla method the least (Shukla median proportion labeled artifact = 0.02, Shukla median longest continuous artifact = 16 s). The APA of our method and the Kleckner method is qualitatively similar, though our method seems to be slightly lower (our method median APA = 22.2, Kleckner method median APA = 32.1). Similar to the simulated data instance results, the Shukla method had the lowest APA (Shukla method median APA = 6.3). It is important to note that the APA for the operating room data instances is calculated differently than for the simulated data instances since there is no ground truth, so it can only be calculated for each method based on the regions that were identified as artifact rather than known true artifact.

Finally, figure 9 shows excerpts of artifact-heavy regions in three simulated and three operating room data instances along with the corrected data using each of the three artifact removal methods on the same *y*-axis scale. Across all of the excerpts, but especially in the case of the operating room data instances, two major trends are clear. First, the Shukla method leaves behind significant large artifacts. Second, the Kleckner method removes large sections of data completely, sometimes tens or hundreds of minutes at once, especially areas around large artifacts. This agrees with the quantitative results which show the Kleckner method achieving high sensitivity and low precision and the Shukla method achieving low sensitivity.

Further detailed results for each data instance are in the Supplementary Material. Table S1 summarizes the results from all three artifact correction methods on all 50 simulated data instances in terms of artifact detection. Table S2 does the same in terms of artifact correction, while table S3 summarizes the proportion of labeled artifact and maximum single continuous artifact. Table S4 summarizes the proportion of labeled artifact, longest single continuous artifact, and artifact power attenuation for all 70 operating room data instances.

Discussion

When comparing the performances of the three artifact correction methods, it is clear that the Shukla *et al* method was the worst performing. Not only did it miss detecting significant large artifacts, as evidenced by low sensitivity and precision (figure 4), it also did a poor job of correcting the artifacts that were detected, as demonstrated by high error and low APA (figure 5). Both quantitatively and qualitatively, this method did not satisfactorily remove artifact from the data. The Kleckner *et al* method performed much better, achieving the highest sensitivity with moderate specificity and accuracy (figure 4). However, the true nature of the method is revealed in the precision, which is below 50% for more than half of the data instances, and the duration of mislabeled true signal, which is between 10 and 20 min for most data instances, which is longer than the duration of true artifact itself (figure 4). The proportion of labeled artifact and longest continuous artifact are consistently far greater for the Kleckner method than for either of the other two methods (figure 6). These results are unsurprising given that the method removes a generous 5 s on either side of any identified artifact as well. While ensuring high sensitivity in detection of artifact, this results in large swaths of true data being mislabeled as artifact, decreasing precision (figure 4). In many cases, this method would lead to unnecessary loss of data, especially affecting downstream analysis techniques which depend on temporal continuity.

This is important to consider, since relevant information about sympathetic activity is contained in the dynamic pulse-like phenomena in EDA (Boucsein 2012, Subramanian *et al* 2020c). Any method that modifies long, continuous chunks of data will likely affect the readout of dynamic activity in that timeframe. In contrast, short regions of missing data can be interpolated since they are only likely to contain a few pulses, and the missing data can be accounted for in estimation of uncertainty (Barbieri *et al* 2005).

Our method achieved intermediate sensitivity and specificity, but high overall accuracy and precision (figure 4). The duration of artifact left behind was comparable to the duration of true signal mislabeled as artifact (figure 4). Based on the sensitivity and proportion of labeled artifact alone, one might expect that the error of our method would have also been intermediate between that of the other two methods. However, the error of our method was very similar to that of the Kleckner method, which achieved a much higher sensitivity (figure 5). Our method actually achieved better overall error and error in non-artifact regions compared to the Kleckner method (figure 5). Our pipeline also achieved comparable APA to the Kleckner method (figure 5). All of these results suggest that while our method may not have removed all of the artificially added 'artifact', the artifact that our method left behind was not very different from the ground truth. It neither contributed significantly to the error nor decreased the APA. This suggests that perhaps it is not even true 'artifact', which is a definite possibility since the 'artifacts' were drawn from a database that could have included mislabeled true signal. The over-labeling of artifact by the Kleckner method explains why our method achieves lower error in non-artifact regions. Our method labels a far lower proportion of artifact than the Kleckner method in shorter duration chunks (figure 6), but in doing so, still achieves comparable error and artifact attenuation (figure 5), resulting in much higher precision and lower duration of true data removed (figure 4). Unlike the Shukla method, which also removes far less artifact than the Kleckner method (figure 6), our method still sufficiently attenuates artifacts and does not leave large artifacts behind resulting in large errors.

When examining the artifact removal in the true operating room data instances for which there is no ground truth, the relative performance of the three methods has several consistencies to that of the simulated data, suggesting that the same conclusions may apply. For example, the proportion of labeled artifacts and longest single continuous artifacts were far greater for the Kleckner method than for the other two (figure 8). The median values of proportion of labeled artifact and longest continuous artifact were similar between the simulated data instances and the operating room data instances for both our method and the Shukla method. However, it was actually larger for the Kleckner method in the operating room data instances than in the simulated data instances. Figures 7 and 9 also suggest that qualitatively, the performance of the Kleckner method may be worse in the operating room data instances than in the simulated data instances (figure 3). High sensitivity and low precision mean that the Kleckner method overestimates artifact and removes large swaths of data, while the Shukla method retains artifacts (figure 9). The relative APA distributions are also similar between the operating room and simulated data instances, especially the difference between the Shukla *et al* method and the other two methods (figures 5, 8). The difference between our method and the Kleckner method is slightly greater in the operating room data instances compared to the simulated data instances, but still largely overlapping. This could be due to the slightly different computation of APA between the simulated data instances and operating room data instances since the latter has no ground truth.

Conclusions

In this study, we collected EDA data in the operating room during surgery from 70 subjects and demonstrated that by constructing a pipeline which includes unsupervised machine learning methods and a set of 12 literature and physiology-informed features, we were able to remove artifact due to surgical cautery and movement from the EDA. This overcomes a major barrier for EDA to be used clinically, such as to track responses to stress and pain in the operating room and the ICU. Our pipeline is fully unsupervised and automated, requiring neither labeled training data instances nor manual curation at any intermediate step. We also compared other artifact detection methods such as the Kleckner *et al* heuristic method which thresholds the signal and its derivative and the Shukla *et al* wavelet decomposition-based method. We generated 50 simulated data instances from the true by introducing artifacts into cleaned data. We then used all three methods to remove artifact from the simulated data instances and compared the performances to the ground truth. Our method achieved the highest accuracy and precision while balancing sensitivity and specificity.

All three of the methodologies compared in this study, including ours, are completely automated and unsupervised, which is ideal for clinical settings in which manually labeling training data instances would be too time-consuming and laborious. Despite the absence of training data, our methodology successfully removed cautery artifact from the data even when true EDA data was interspersed between sections of heavy artifact, a situation in which the Kleckner method would likely remove the true EDA as well. Even methods that could detect some artifact were not necessarily able to remove it successfully (i.e. Shukla *et al* method). While our study

included several major sources of artifact in the surgical setting (cautery, motion artifact, etc), one limitation is that there may be other sources of artifact linked to specific equipment, such as that used in orthopedic procedures.

While our methodology used some of the same features as existing methods, we allowed the unsupervised algorithms to 'learn' the difference between artifact and true signal for each data instance rather than hardcoding rules. The selected features, including those that overlap with existing methods, simply highlighted relevant characteristics of the data, based on the physiology of EDA and observations about the nature of cautery-related artifact. A straightforward expansion of this approach for other types of 'clearly visible' artifact in modalities such as ECG and EEG could be implemented using custom feature definition, again informed by the physiology and nature of artifact in those signals. In addition to non-clinical uses, EDA has tremendous potential to be an inexpensive and non-invasive clinical marker of sympathetic activation in situations in which patients cannot express symptoms of pain or stress, such as the operating room and ICU. Our work presents a significant step in that direction.

Acknowledgments

S S, B T, R B, and E N B. thank the Department of Anesthesia at MGH, the operating room staff, and Drs. Douglas Dahl, Marcela Del Carmen, and Annekathryn Goodman for their support in carrying out this study.

Ethical statement

This research was conducted under protocol approved by the Massachusetts General Hospital Human Research Committee (IRB 2017P002591). The research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements. All participants provided written informed consent to participate in the study and for study results to be published.

ORCID iDs

Sandya Subramanian https://orcid.org/0000-0001-5885-1199

References

Amin R and Faghih R T 2022 Physiological characterization of electrodermal activity enables scalable near real-time autonomic nervous system activation inference *PLoS Comput. Biol.* **18** e1010275

Barbieri R, Matten E C, Alabi A A and Brown E N 2005 A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability *Am. J. Physiol. Heart. Circ. Physiol.* **288** H424–35

Biagetti G, Crippa P, Falaschetti L, Tanoni G and Turchetti C 2018 A comparative study of machine learning algorithms for physiological signal classification *Procedia Computer Science* **126** 1977–84

Boucsein W 2012 *Electrodermal Activity.* (New York, NY: Springer)

Chen W, Jacques N, Taylor S, Sano A, Fedor S and Picard R W 2015 Wavelet-based motion artifact removal for electrodermal activity *Conf. Proc. IEEE Eng. Med. Biol. Soc. pp. 6223–6226* 2015, 6223–6

DeGroot M H and Schervish M J 2012 *Probability and Statistics.* (Boston, MA: Pearson Education)

Gashi S, Di Lascio E, Stancu B, Das Swain V, Mishra V, Gjoreski M and Santini S 2020 Detection of artifacts in ambulatory electrodermal activity data *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol* **4** 44

Goldstein M and Uchida S 2016 A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data *PLoS One* **11** e0152173

Hossain M-B, Posada-Quintero H F and Chon K H 2022b A deep convolutional autoencoder for automatic motion artifact removal in electrodermal activity *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2022** 325–328

Hossain M-B, Posada-Quintero H F, Kong Y, McNaboe R and Chon K H 2021 A preliminary study on automatic motion artifacts detection in electrodermal activity data using machine learning *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2021, 6920–6923

Hossain M-B, Posada-Quintero H F, Kong Y, McNaboe R and Chon K H 2022a Automatic motion artifact detection in electrodermal activity data using machine learning *Biomed. Signal Process. Control* **74** 103483

Hu L-Y, Huang M-W, Ke S-W and Tsai C-F 2016 The distance function effect on k-nearest neighbor classification for medical datasets *SpringerPlus* **5** 1304

Jiang X, Bian G-B and Tian Z 2019 Removal of artifacts from EEG signals: a review *Sensors (Basel)* **19** 987

Kelsey M, Palumbo R V, Urbaneja A, Akcakaya M, Huang J, Kleckner I R, Feldman Barrett L, Quigley K S, Sejdic E and Goodwin M S 2017 Artifact detection in electrodermal activity using sparse recovery *Proc. of SPIE.* 10211, pp.102110D–1

Kleckner I R *et al* 2018 Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data *IEEE Trans. Biomed. Eng.* **65** 1460–7

Liu F T, Ting K M and Zhou Z-H 2008 Isolation forest *Proc. 8th IEEE Int. Conf. Data Mining.* pp.413–422

Llanes-Jurado J, Carrasco-Ribelles L A, Alcaniz M and Marin-Morales J 2021 Automatic artifact recognition and correction for electrodermal activity in uncontrolled environments *Research Square.* 10.21203/rs.3.rs-717360/v1 [Preprint Online]

Manevitz L M and Yousef M 2001 One-class SVMs for document classification *Journal of Machine Learning Research* **2** 139–54

Mannan M M N, Kamran M A and Jeong M Y 2018 Identification and removal of physiological artifacts from electroencephalogram signals: a review *IEEE Access* **6** 30630–52
Molavi B and Dumont G A 2012 Wavelet-based motion artifact removal for functional near-infrared spectroscopy *Physiol. Meas.* **33** 259–70
Neurofeedback Expert System, Thought Technology Ltd https://thoughttechnology.com/neurofeedback-expert-system/
Posada-Quintero H F and Chon K H 2020 Innovations in electrodermal activity data collection and signal processing: a systematic review *Sensors* **20** 479
Shukla J, Barreda-Angeles M, Oliver J and Puig D 2018 Efficient wavelet-based artifact removal for electrodermal activity in real-world applications *Biomed. Signal Process. Control* **42** 45–52
Subramanian S, Barbieri R and Brown E N 2020c Point process temporal structure characterizes electrodermal activity *PNAS* **117** 26422–8
Subramanian S, Barbieri R, Purdon P L and Brown E N 2020a Analyzing transitions in anesthesia by multimodal characterization of autonomic state *Conf. Proc. 2020 11th Conf. of the European Study Group on Cardiovascular Oscillations (ESGCO)* pp.1–2
Subramanian S, Barbieri R, Purdon P L and Brown E N 2020b Detecting loss and regain of consciousness during propofol anesthesia using multimodal indices of autonomic state *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2020, pp.824–7
Subramanian S, Purdon P L, Barbieri R and Brown E N 2021a Quantitative assessment of the relationship between behavioral and autonomic dynamics during propofol-induced unconsciousness *PLoS One* **16** e0254053
Subramanian S, Tseng B, Barbieri R and Brown E N 2021b Unsupervised machine learning methods for artifact removal in electrodermal activity *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2021** 399–402
Taylor S, Jacques N, Chen W, Fedor S, Sano A and Picard R 2015 Automatic identification of artifacts in electrodermal activity data *Conf. Proc. IEEE Eng. Med. Biol. Soc. pp. 1934–1937* 2015, 1934–7
Uriguen J A and Garcia-Zapirain B 2015 EEG artifact removal—state-of-the-art and guidelines *J. Neural Eng.* **12** 031001
Zhang Y 2017 Unsupervised motion artifact detection in wrist-measured electrodermal activity data *Dept. Electric. Eng. M.S. Thesis* University of Toledo http://rave.ohiolink.edu/etdc/view?acc_num=toledo1501876131092933
Zhang Y, Haghdan M and Xu K S 2017 Unsupervised motion artifact detection in wrist-measured electrodermal activity data *Proceedings of the 2017 ACM International Symposium on Wearable Computers.* **2017** 54–7