Version of Record: https://www.sciencedirect.com/science/article/pii/S1746809422000052
Manuscript_efc42709bdac4e63dd14148ccad696a5

Automatic Motion Artifact Detection in Electrodermal Activity Data Using Machine Learning

Md-Billal Hossain[1], Hugo F. Posada-Quintero[1], Youngsun Kong[1], Riley McNaboe[1] and Ki H. Chon[1,*]

[1]*Department of Biomedical Engineering, University of Connecticut, 260 Glenbrook Road, Unit 3247 Storrs, CT 06269-3247, USA*

[*]Corresponding Author. Email: ki.chon@uconn.edu

Abstract

Background and Objective: Electrodermal activity (EDA) has gained popularity in recent years for diverse applications such as emotion and stress recognition; assessment of pain, fatigue, and sleepiness; and diagnosis of depression and epilepsy. However, presence of motion artifacts (MA) hinders accurate analysis of EDA signals. This study presents a machine learning framework for automatic motion artifact detection on electrodermal activity signals.

Methods: We extracted several statistical and time frequency features from EDA and investigated machine learning algorithms to automatically detect noisy EDA segments. To avoid incorrect adjudication due to the aperiodic nature of EDA signals, we collected both clean and MA-corrupted EDA from immobile and moving hands, respectively. The MA-corrupted EDA data were annotated by three experts as either MA-corrupted or clean using the criteria recommended in the literature, as well as the correlation between MA and the reference EDA.

Results: We performed a subject-independent validation strategy to evaluate the performance of the machine learning models. The best-performing model classified the MA and clean EDA segments with 94.7% accuracy. A comparison of our motion artifact detection approach with two previously published

methods showed that our best performing method outperformed them and retained its accuracy on entirely different, unseen data from a separate study, indicating the method's generalizability.

Conclusions: The current work can provide accurate and autonomous adjudication of MA-corrupted EDA signals. Given the lack of accurate MA detection methods for EDA signals, this work may lead to more applications of EDA as a physio-marker.

INTRODUCTION

Electrodermal activity (EDA) refers to the change in electrical conductance of the skin as a response to the opening of sweat glands in the human body [1]–[3]. EDA represents sudomotor activities innervated by C nerve fibers of the autonomic nervous system and, hence, has the potential to be used as a noninvasive measure of the sympathetic nervous function and cognitive arousal [4]–[7]. Over the last decade, there has been significant advancement in EDA-based research. As an unobtrusive and noninvasive measurement, EDA has been used for assessment of the sympathetic nervous system in various applications such as emotional arousal [8]–[11], stress [12]–[15], pain [16]–[19], decision making [20], autism [21], and panic disorder [22].

Despite its popularity and potential to be used as a noninvasive surrogate of the sympathetic function, EDA has some limitations. Similar to most other physiological signals, in ambulatory environments EDA is often affected by noise and motion artifacts. The artifacts are usually defined as changes in the recorded biosignal which do not stem from the signal source [3]. EDA can be affected by unstable electrode contact [3], [23], environmental temperature and humidity [3], [24], and an individual's activity [1], [3], [25], [26]. Because of these factors, EDA can be severely affected and high-frequency perturbations in the signal may or may not always resemble skin conductance response (SCRs). Regardless of where the EDA is recorded, it can be affected by motion artifacts and depending on their severity, the entire recorded data can be unusable. For example, due to low quality signal, several weeks

of EDA data have been discarded in [27]. A significant amount of data were also discarded in many others studies[10], [23], [28].

While there has been an increase in EDA research, especially in the last decade, there has been only a limited number of works aiming to detect motion artifacts in EDA. Even though visual inspection is recommended for motion artifact detection in the widely used textbook by Boucsein [3], it is cumbersome and time-consuming to visually inspect and mark motion-corrupted EDA data, especially for continuous monitoring via wearable devices, hence, impractical for this purpose. Many researchers used typical signal processing techniques such as exponential smoothing or lowpass filtering to avoid visual inspection. These techniques may smooth the high variations in the signal, however, sometimes they distort the original physiological responses [26] or make artifacts seem like genuine data [29].

A simple EDA data quality assessment method is proposed in [25] using some simple decision rules. While this method works well for spiky and large-amplitude motion artifacts, it fails in several other challenging cases. There have been some supervised [30], semi-supervised [31], and unsupervised [26] machine learning based approaches. Most of the methods were developed on specific manually annotated data and lack generality. Moreover, since there is no reference signal, EDA is annotated based on visualization, which can be difficult to discern, and the annotations may vary from person to person. In addition, there is also risk of labeling sympathetic-induced spiky SCR as motion artifacts on certain occasions, and vice versa.

In this study, we designed experimental protocols to mimic a wide range of typical motion artifacts encountered in EDA data. We collected two channels of simultaneous EDA from the left and right arm. While one arm was immobile throughout the experiment, the other arm was occasionally moving to induce motion artifacts in the corresponding EDA channel. The arm at rest provided a clean or reference EDA signal, which was collected to assist the independent annotators to judiciously adjudicate the target

EDA channel. Finally, we extracted several statistical, model-based, and time-frequency features and used different machine learning algorithms to classify the noisy and clean EDA data segments adjudicated by the three independent expert observers. Our study presents a standard and expertly annotated EDA motion artifacts database, which we will make public to be used by other researchers in this area. Fig. 1 shows the overall framework of this study.

Preliminary results of this study have been accepted for presentation at the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) [32]. The conference paper used only a portion of the entire data and proposed a correlation-based automatic annotation criterion to adjudicate the EDA signals as clean or noisy. In the present study, we used manual annotation on the EDA data since it is more realistic and accurate. We also updated our method and incorporated more subjects and performed more rigorous analysis using many machine learning and deep learning methods.

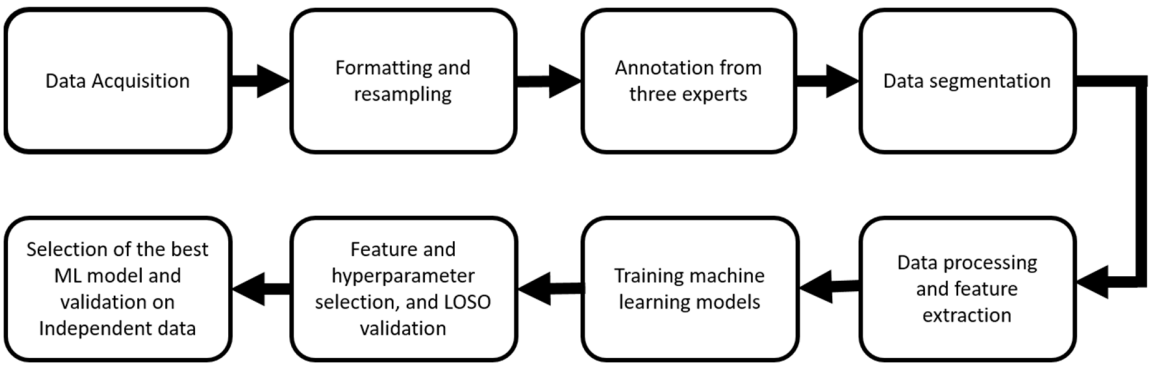

Figure 1. Block diagram of the proposed motion artifact detection framework.

METHODS AND MATERIALS

A) Data Collection

Twenty subjects aged 20-35 (10 male, 10 female), participated in this EDA motion artifact detection study. A pair of stainless steel electrodes were placed on the index and middle fingers of each hand to collect two simultaneous EDA signals. The EDA data were collected using ADInstrument's galvanic skin response (GSR) modules. The GSR module supplies a 75 Hz square wave, low-impedance, low-voltage (22 mv rms) signal between the electrodes and measures the corresponding skin conductance. The output of the GSR module was then fed to one of the four channels of the PowerLab, a high-performance data acquisition system which can digitize the signal with sampling frequency as high as 25 kHz. The output from the PowerLab was connected to a computer and processed through Labchart 7 software which offers real time visualization of data. We designed the data collection protocols so that the right hand occasionally moved, to mimic regular motion artifacts people could create in their daily lives, and the left hand was immobile in order to provide a reference EDA. The experimental protocols were reviewed and verified by the Institutional Review Board (IRB) for human subject research at the University of Connecticut. Written consent was collected from the subjects before participating in the experiment.

Table I shows the summary of the protocol of the study. As shown in Table I, the experimental protocol consisted of two parts. In part I, there was no significant motion other than some orthostatic and cognitive stress. This data helped compare EDA collected from both hands, by observing if the same SCRs are present in EDA data from both hands. Part II was designed for inducing motion artifacts in one channel of the EDA data, as the subjects were performing some light and regular movements to mimic daily life movements. Fig. 2. shows representative pictures of data collection and the instruments used. Figs. 2. (a), (b), and (c) show the ADInstrument modules (PowerLab and GSR modules), a subject typing on a keyboard, and the placement of electrodes on the fingers, respectively.

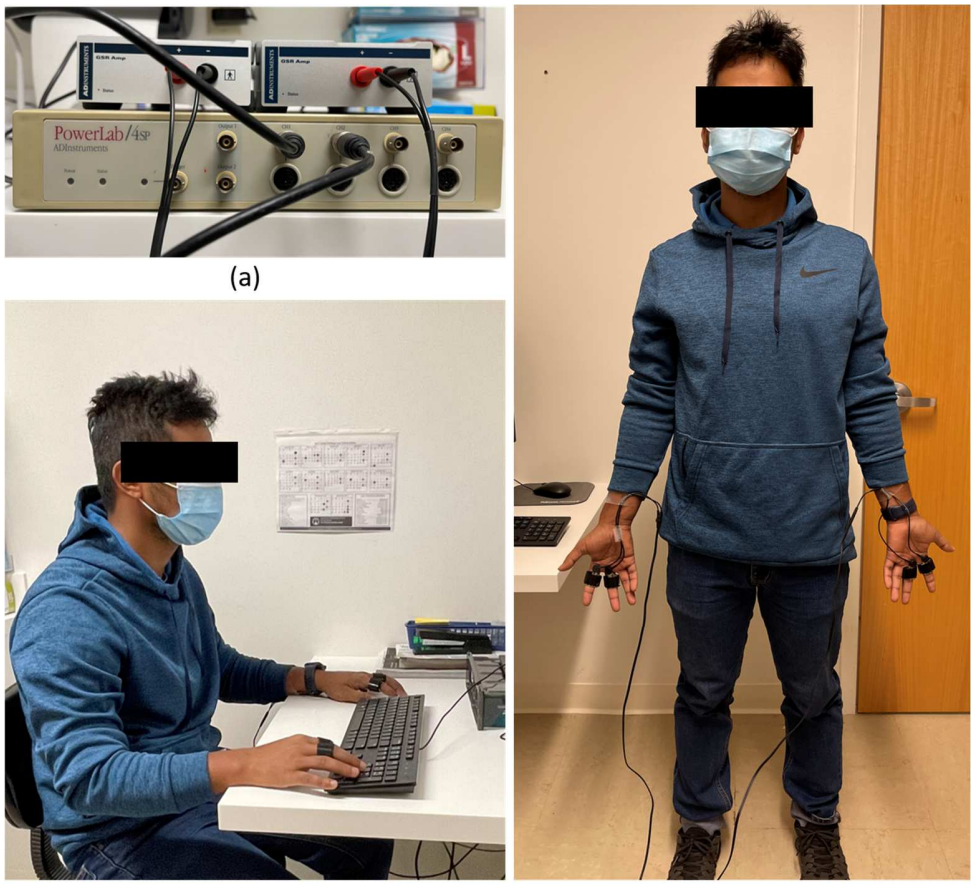

Figure. 2. Data collection: (a) ADInstrument PowerLab and GSR module (b) Subject typing with one hand (c) Subject wearing electrodes on both bands

TABLE I. DATA COLLECTION PROTOCOL SUMMARY

Duration (second)	Activity	Remarks
Part I (Stress Test)		
120	Flat table, relaxing with eyes closed	Baseline
30	Start table tilt	Orthostatic Stress
120	Subject remains in tilted position	
150	Return table to flat position, subject relaxes	
120	Perform Stroop test	Cognitive Stress

Part II (Motion Artifact Test)		
60	Sitting up in a chair with no movement	Motion Artifact Induction
60	Sitting down and typing on the computer	
60	Sitting down and holding a mouse, clicking the mouse	
60	Standing up with the arm next to torso	
60	Standing up swinging arm by the side as if to simulate walking	
60	Standing up with arm straight out, moving arm up and down continuously	
60	Standing up with arm straight out moving the arm at the elbow allowing the wrist to come into the chest and then straightening out and repeating.	
60	Standing up with arm straight, completing a circle with the fingertips by rolling the shoulder.	

B) Data Labelling

To validate our motion artifact detection algorithm, we systematically adjudicated the dataset as artifacts and clean signal. Three independent observers annotated the EDA signal, and we took the majority opinion for the annotations. No fixed window was defined for the annotation, as the observers were able to mark the start and end of artifact segments. We first studied the literature for the existing guidelines to label the artifacts in EDA signals, and then modified or added new guidelines to address common issues in manual annotations. Since the EDA signal does not exhibit periodicity or any regular structure like other biosignals (e.g. electrocardiogram (ECG) and photoplethysmography (PPG)), manual adjudication of clean versus noisy EDA can be a rather difficult task. Using only existing guidelines such as in [25], [26], [30] might not be sufficient to identify all the artifacts in EDA signals. Moreover, sometimes there is also risk of considering fast-varying SCR in EDA signals as artifacts. To avoid this, we exploited the reference clean EDA signal. When there were no movements or motion artifacts, the reference and the noisy channel EDA are almost identical in shape, although different in amplitude. Therefore, if there was

a significant correlation (≥ 0.95) between noisy and reference EDA channels, the observer annotated that portion as clean. Thus, we avoided any mislabeling of clean EDA as noisy. Fig. 3 shows two simultaneous EDA channels when there are no movements. The red circles show some challenging cases where traditional criteria such as fast rising/falling time might consider them as noisy. However, as can be seen from the figure, these waveforms are consistent in both channels meaning that they are not motion artifacts. We considered existing guidelines in the literature as well as the correlation between reference and noisy EDA channels. Table II shows the guidelines we considered for labeling the EDA segments.

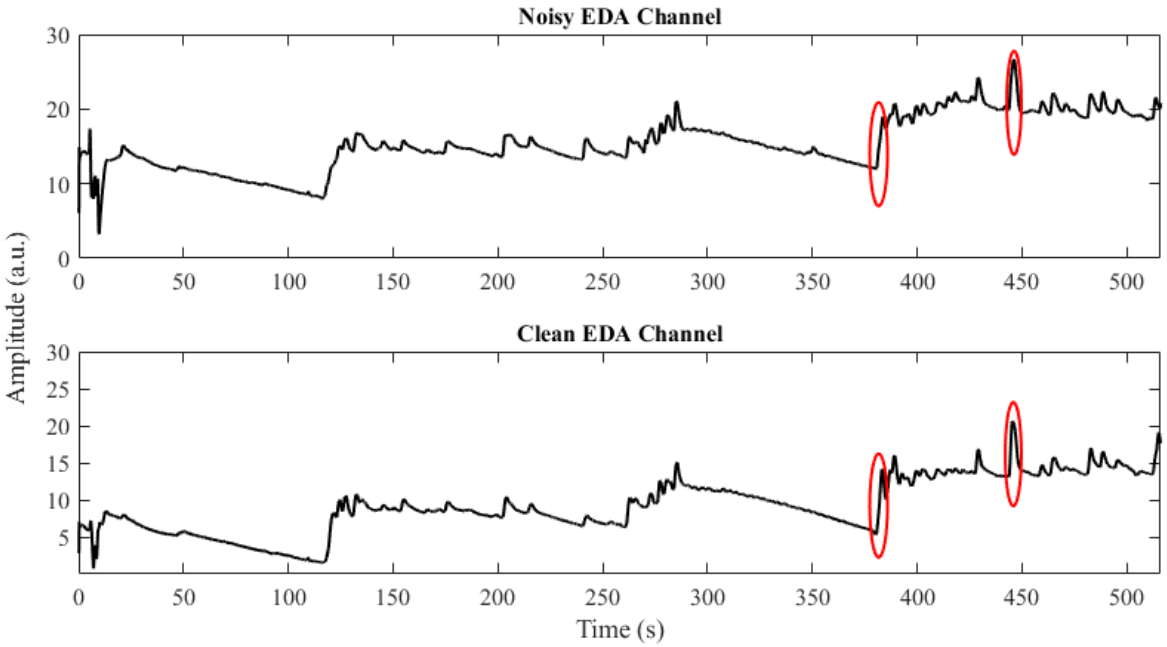

Figure 3. Two simultaneous EDA channels collected from two hands without any movements.

Table II. Guidelines for annotation of EDA signal

Index	Criteria
1	EDA out of range (EDA range -10S to 40S)
2	Quick change in EDA (if EDA changes faster than $\pm 10S$)

3	EDA peak decays (EDA is considered noisy if EDA peak does not follow an exponential decay except when there are two peaks within a short period of time)
4	Correlation of reference and noisy EDA channels (considered clean only if the correlation coefficient is ≥ 0.95)

C) EDA Motion Artifact Detection Framework

We developed a binary classification framework to automatically detect the EDA segments with motion artifacts. The framework consists of a set of signal processing and machine learning techniques that distinguishes motion artifacts from clean EDA signals. We first downsampled the EDA data from 1,000 to 8 Hz, a sampling frequency used by the previous literature for this purpose [30], [33]. To validate our approach, we obtained ground truth labels from three experts (first, second, and third authors). As mentioned earlier, we used majority voting for the final labeling when there was disagreement among the three observers. EDA signals were then divided into 5-second non-overlapping segments, as were used in most of the previous literature [25], [26], [30]. We extracted several common features previously used [30] and proposed a set of new features for the classifications. We performed a feature selection and trained several machine learning models using the features. The performance of our machine learning models was validated using the leave-one-subject-out (LOSO) validation method. In the following subsections, we describe the feature extraction methods, classification, hyperparameter tuning, and evaluation methods.

1) Feature Extraction

To appropriately characterize the EDA artifacts, we extracted 40 different features from each 5-second segment of EDA. We extracted most of the features mentioned in the literature [26], [30]. We further extracted features from EDA modeling using the autoregressive (AR (2)) model, and phasic and

tonic decomposition of EDA. Table II shows the list of features we computed in this study. The features can be broadly classified into three groups. All features were standardized before feeding into the classifiers.

(i) Statistical Features: We extracted 19 different statistical features from EDA, its derivatives, and its phasic component. We extracted commonly used statistical features such as mean, standard deviation, minimum, maximum, and dynamic range (difference between the maximum and the minimum value) as suggested in the literature [30], [34]. To characterize the change in the signal, we computed the mean and standard deviation of the first and second derivative of the EDA signal. We used the phasic decomposition of EDA to compute some statistical features that could identify the high frequency and spiky noise. We used a popular algorithm called CVXEDA [35] for the phasic-tonic decomposition of EDA. CVXEDA models the EDA signal as a sum of phasic and tonic components, and a Gaussian noise that incorporates prediction errors, measurement error, and artifacts. This method then uses the convex optimization technique to decompose the signal into phasic and tonic components, minimizing the prediction errors. As shown in Fig. 4, when EDA is affected by motion artifacts, there is a higher number of large amplitude false phasic drivers in the phasic component of EDA. Therefore, we computed the energy, number of peaks, and mean value as features to discriminate the noisy portions from the clean EDA segments. In addition, we computed some common statistical features such as mean, standard deviation, maximum, and minimum of the absolute value of the first and second derivatives of EDA.

(ii) AR model-based features: The motivation behind including Autoregressive (AR) modelling is that when EDA data are corrupted by noise, there will be more residual noise in the AR model than in the clean data; similarly, AR parameters show either decreased or increased values

for the noisy vs. clean data. AR modeling is a common and often parsimonious approach for extracting informative features from a time series signal [36], [37].

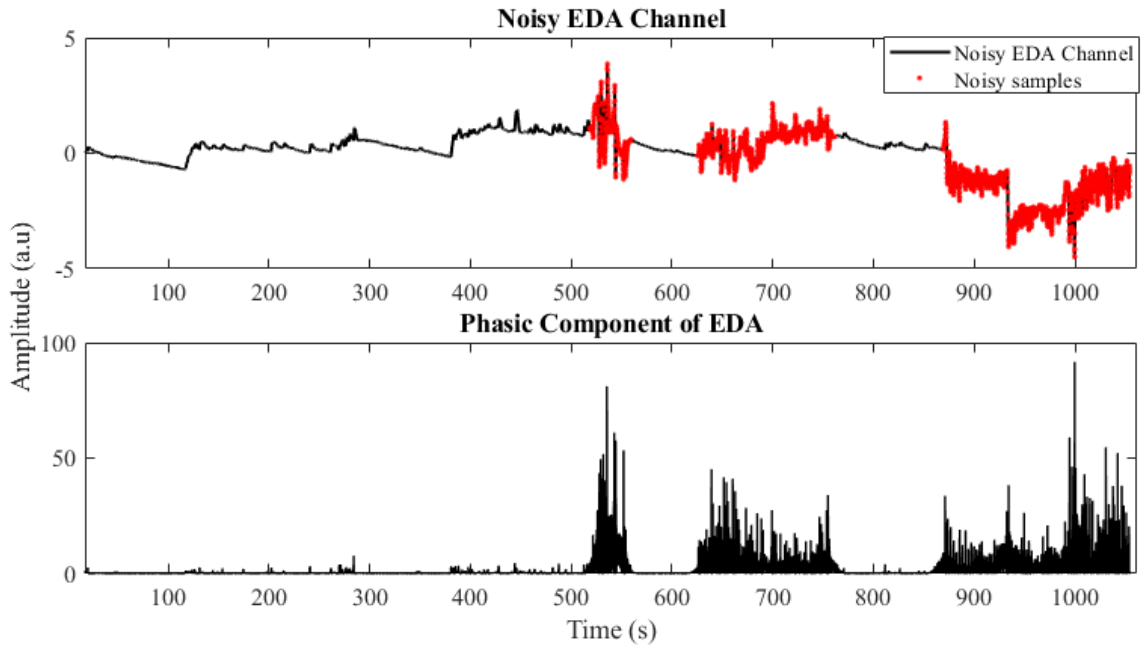

Figure 4. Noisy EDA channel and corresponding phasic component.

(iii) Time frequency features: To capture the non-stationary characteristics, we used time frequency decomposition, namely, wavelet transformation and variable frequency complex demodulation (VFCDM). Wavelet decomposition has been used previously in the literature and is well suited to characterize the edges and sharp changes in signals. VFCDM, on the other hand, is a high resolution time frequency technique [38] that has been successfully used for many biosignal applications [39]–[42]. The difference between wavelet and VFCDM is that VFCDM can decompose the main signal into several frequency bands with equal lengths of the original signal. We decomposed the EDA signal using the three-level wavelet transform with the 'Haar' wavelet and computed the mean and standard deviation of the approximate and details signals. Using VFCDM modes, we computed four intermediate frequency bands between 1Hz, 2Hz, 3Hz, and 4Hz and computed the mean and variance in each band. By doing

so, we may have included some redundant features. However, at this point we are not concerned about including too many features because we later used a feature selection algorithm to reduce the dimensionality curse or overfitting.

Table III: Summary of the features computed

Index	Category	Specific features
1-3	AR(2) Modelling	AR parameters and AR noise variance
4-10	Raw EDA	Mean, median, standard deviation, maximum, minimum, range, and Shannon entropy
11-20	Absolute value of first and second derivative of EDA	Mean, standard deviation, and max and min
21-28	Wavelet decomposition	Mean, median, standard deviation, and range of the wavelet coefficients
29-32	Phasic component	Energy, number of peaks, mean value, and Shannon entropy
33-40	VFCDM decomposition	Mean and standard deviation of the four intermediate reconstructed signals

2) Classification

We investigated several machine learning algorithms to automatically identify motion artifacts in EDA signals. Particularly, we used machine/deep learning techniques such as the multi-layer perceptron (MLP), Linear Discriminant Analysis (LDA), k—Nearest Neighbors (KNN), Linear Support Vector Machine (l-SVM), Support Vector Machine with radial basis function kernel (k-SVM), Random Forest [43], Adaboost [44], GradBoost [45], and XGBoost [46]. We explored all these classifiers to learn which one can best recognize

the artifacts from the clean EDA. Since we have an imbalanced dataset in which the clean EDA is the majority class, the accuracy metric might not be sufficient to validate the performance of the machine learning models [47]. Therefore, we included two baseline 'dummy' classifiers that always predicts either the majority class or the minority class and compared their performance with the machine learning models. We compared the performance in terms of sensitivity, specificity, accuracy, and F1 score. Moreover, we compared our motion artifact detection approach with two state of the art methods, EDAQA [25] and the method proposed in [30]. EDAQA is based on some simple thresholds on the EDA range, its derivative, and temperature, while the other method uses the support vector machine (SVM). The machine learning approach in [30] calculated features from raw EDA data, filtered EDA and its derivative, and wavelet coefficients. For all the classifications we used Scikit-learn library in Python [48].

3) Feature Selection

Since we computed 40 features, we used a feature selection algorithm to reduce the number of features in order to avoid the curse of dimensionality. We used random forest (RF) as a feature selection algorithm [43]. Feature selection using RF is an embedded method that combines the quality of filter and wrapper methods. Since it is quite straightforward to compute how much each variable contributes to the decision tree, RF is a highly accurate, generalizable, and easily interpretable feature selection algorithm.

4) Evaluation Procedure

In order to evaluate the performance of the machine learning models, we used a LOSO validation strategy, also used in [26], [49]–[51]. For each fold, we used all the subjects except one for training the classifiers and the remaining one for testing. We did that for every subject and the performance metrics were averaged over all the subjects. Since for each fold the training and testing data are different, due to interpersonal variances [14], the machine learning models are not biased by subject characteristics and thus, are more generalizable.

We computed LOSO validation accuracy, sensitivity, specificity, and F1 score [52], to evaluate the performance of the machine learning models. The accuracy is the percent correctness of differentiation between clean EDA and signals with motion artifacts. The sensitivity and the specificity are the measures of goodness of the model to correctly determine the positive and negative classes. In our case, the positive class is the motion artifact and the negative class is the clean EDA.

5) Hyperparameter Tuning

Within each training, we performed hyper-parameter tuning for the classifiers using a group 5-fold cross validation strategy where each time 5 subjects of the training data were left for validation and the rest of the subjects were used for training. This was performed repeatedly for a set of hyperparameters of choices and the best parameter was chosen based on the average accuracy on the validation data. Again, we used a subject-independent validation scheme for the hyperparameter selection so that our machine learning models would not become subject-biased.

For the kernel SVM, the parameters C and gamma were selected using grid search from a list of parameter candidates 1, 10, 100, and 1000, and 0.001, 0.01, 0.1, 1, respectively. For all the tree-based machine learning, we optimized the number of estimators from a list of choices. For MLP we applied grid-search cross validation to optimize four different parameters. We varied the number of hidden layers between 1 and 4 and the activation functions were chosen from logistic, tanh, and rectified linear unit. We experimented with three different optimizers, namely, stochastic gradient descent [53], Adam [54], and limited memory Broyden–Fletcher–Goldfarb–Shanno (lbfgs) [55]. The initial learning rate was chosen from 0.0001, 0.001, and 0.01 and we fixed the maximum number of epochs to 200.

RESULTS

In this section, we report on the performance of the machine learning classifications and comparison with an existing method (EDAQA) [25]. First, we describe the performance of the classifiers using LOSO validation and compare it with the baseline classifiers. We then present a comparison of the proposed framework with those of Kleckner et al. [25] and Taylor et al. [30].

Table IV. Machine Learning Performance and Comparison with Baseline Classifiers

Classifier	Sensitivity ($mean \pm sd$ %)	Specificity ($mean \pm sd$ %)	Accuracy ($mean \pm sd$ %)	F1 Score ($mean \pm sd$ %)
Baseline 1	0	100 ± 0	67.54 ± 13.46	0
Baseline 2	100 ± 0	0	35.44 ± 13.46	51.05 ± 13.75
LDA	74.80 ± 16.61	$\mathbf{97.20 \pm 6.06}$	90.21 ± 6.10	82.60 ± 12.84
KNN	83.33 ± 8.84	94.76 ± 6.21	90.21 ± 6.10	85.29 ± 7.85
Linear SVM	92.03 ± 9.86	93.18 ± 6.99	92.81 ± 5.90	88.82 ± 8.32
Kernel SVM	91.28 ± 9.06	93.02 ± 7.08	92.42 ± 6.04	88.21 ± 8.67
Random Forest	90.05 ± 9.05	94.38 ± 6.44	92.98 ± 6.06	88.80 ± 8.87
AdaBoost	89.36 ± 7.09	95.01 ± 6.90	93.17 ± 5.89	88.98 ± 8.86
GradBoost	$\mathbf{92.90 \pm 7.86}$	95.75 ± 6.62	$\mathbf{94.82 \pm 5.95}$	$\mathbf{91.61 \pm 8.37}$
XGboost	92.81 ± 7.75	95.54 ± 6.50	94.66 ± 5.91	91.37 ± 8.41
MLP	89.28 ± 6.97	95.29 ± 6.97	93.34 ± 6.02	89.20 ± 8.65

Table IV shows the machine learning performance using the LOSO validation strategy. As can be seen, all classifiers perform better compared to the baseline classifiers. LDA shows the highest specificity of 97.2%, however, the sensitivity is the lowest among the classifiers. Baseline 1 has the highest specificity while having zero sensitivity and F1 score (always predicts negative class), and baseline 2 has the highest sensitivity while having zero specificity (always predicts positive class). GradBoost has the highest sensitivity, accuracy, and F1 score among the classifiers. Its high accuracy suggests that our model can

distinguish between the clean EDA and EDA with motion artifacts almost 95% of the time. The results suggest that our EDA motion artifact detection is generalizable across the subjects, largely due to LOSO training and testing strategy, as discussed in the Methods section. We also observed that the performance of all the machine learning classifiers is significantly higher than that of the baseline classifiers ($p<0.01$ in t-test).

We compared the classification results using our best performing machine learning model (GradBoost) with the approaches presented in [25] and [30]. Fig. 5 shows a comparison of the performance of the proposed framework with those of the two state of the art methods. Since EDA may vary from device to device, in order to perform a fair comparison we adjusted the threshold value given in [25]. We tried a range of thresholds for each of the rules and chose the combination that provided the highest performance for EDAQA. There is an online platform for motion artifact detection using [30], however, this only supports data collected using an Empatica device or Q sensor. Therefore, to compare

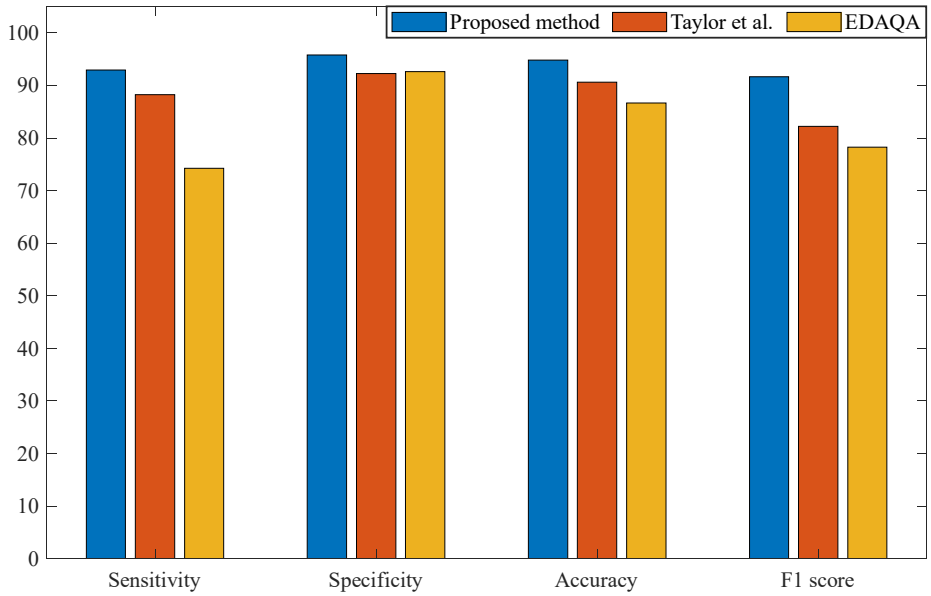

Figure. 5. Comparison of proposed framework with EDAQA

with this approach we computed the features mentioned in the paper and then used SVM for classification as described in the paper. For fair comparison, we trained the classifier using the same LOSO validation strategy using our dataset. As shown in Fig. 5, the proposed method shows significantly higher ($p < 0.05$ in pairwise t-test) sensitivity, accuracy, and F1 score than both compared methods.

A) Results on central nervous system-oxygen toxicity (CNS-OT) Dataset

To further validate the proposed machine learning model and the generalizability of the approaches, we tested on an independent dataset called the central nervous system oxygen toxicity (CNS-OT) dataset. While the details of this database can be found in [56], we present a brief description here. We (the same three reviewers) randomly selected 10 subjects' data from the CNS-OT dataset and annotated them as clean or noisy data. The experimental protocol for this study required the subjects to be immersed in 28±1ºC water to the shoulders, breathing 100% O2 at 35 feet of seawater (oxygen partial pressure 2.06 ATA), while exercising on an underwater cycle ergometer at approximately 100W output, and executing NASA's Multi-Attribute Task Battery-II (MATB-II) cognitive testing software. The experiment continued until symptoms of CNS-OT were observed, or until the maximum time duration of two hours. EDA data were collected using a pair of stainless-steel electrodes placed on the index and ring fingers of each subject's left hand and a galvanic skin response module FE116 (ADInstruments). The sampling frequency of the EDA signal was 100 Hz, which was then downsampled to 8 Hz. Because of the exercise and long-term monitoring, this dataset had more motion artifacts than typical ambulatory EDA data. Fig. 6 shows a representative example of an EDA record from the CNS-OT dataset, in which the red dots mark the noisy portions.

Table V. Performance comparison on CNS-OT dataset

Classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)	F1 Score (%)
Proposed Method (Gradboost)	89.58	94.74	93.87	84.62
Taylor et al. [30]	88.45	90.96	90.48	78.07
Kleckner et al. [25] (EDAQA)	88.06	93.51	92.46	81.73

Table V shows the performance comparison of our proposed machine learning framework to the method proposed by Kleckner et al. [25] and Taylor et al. [30] on the independent CNS-OT dataset which was not used for training. It shows that the proposed machine learning method performed better than both EDAQA and the SVM classifier proposed by Taylor et al. [30].

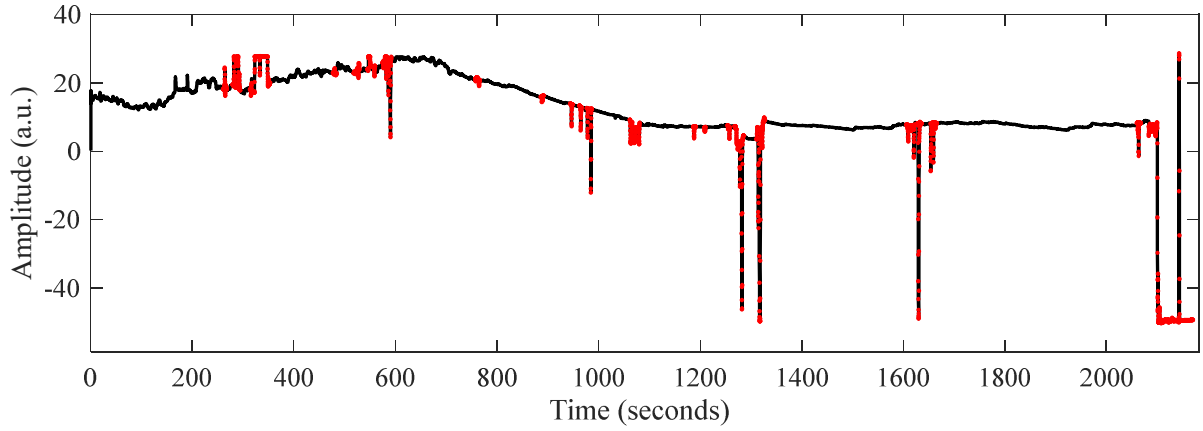

Figure 6. Example record of EDA from CNS-OT dataset

DISCUSSION

This study presents an automatic motion artifact detection framework using adjudicated EDA signals. There have been some prior machine learning approaches for EDA motion artifact detection [26], [30]. However, the machine learning methods were applied to an even smaller dataset than our study. Moreover, these machine learning methods were not tested in a subject-independent fashion, hence, it is not clear whether or not these approaches are generalizable to other datasets.

As discussed earlier, labeling EDA signals as either clean or containing motion artifacts is non-trivial and requires highly expert people who have extensive working experience with EDA signals. Even so, there is always a risk of incorrect annotation of the fast-varying skin conductance response as motion artifacts. Therefore, it is crucial to design a motion artifact detection algorithm on properly annotated data so that the model does not misclassify clean EDA, which might carry important physiological information, as motion artifact. In our study, we collected a reference EDA to assist the annotation of the target EDA channel. Because of the reference EDA, the annotators were aware of the clean EDA signals, hence, incorrect adjudication of the data was minimized. We believe this adjudicated dataset could be beneficial for researchers working on EDA signals, and so we will make this database along with our expert annotation available upon request.

Another contribution of this work is the inclusion of more informative features to better understand EDA signal dynamics. We included autoregressive (AR) modeling on the EDA segments to extract important features that represent the characteristics of the clean EDA and motion artifact

segments. We also included new time frequency features using VFCDM decomposition. Moreover, previous researchers used EDA phasic component features for explaining different physiological behavior. We showed that there is a significant change in the extracted phasic component when there are motion artifacts present in the EDA signal. Thus, EDA features extracted from the phasic component were found to be significant in distinguishing motion-corrupted EDA data.

Since we validated the machine learning models using the LOSO validation technique, we can assume our machine learning model is relatively devoid of subject bias that may arise if the models are validated using the typical separation of random training and testing datasets. The comparison of the machine learning models with two baseline classifiers show that machine learning models consistently perform significantly better. The F1 score of the GradBoost classifier is 42.64% higher than the baseline 2 and 12.68% higher than that of the EDAQA.

The performance on the independent CNS-OT test dataset shows that the Kleckner et al. method [24] (EDAQA) performed better on this data than it did on the EDA motion artifact dataset. This is most likely because the artifacts in the CNS-OT dataset were more prominent when compared to those in the motion artifact dataset, and EDAQA is specifically designed to identify large spiky artifacts. For example, EDAQA is designed to look for outliers of amplitudes that are out of the typical range (0.05-60 μS) and dynamics that change quicker than 10μS/sec. The SVM classifier implemented from [30] showed the lowest accuracy and F1 score among the three compared methods. This could be for two reasons. First, perhaps the features used in SVM were not sufficient to characterize the motion artifacts, or second, the training samples may not have been sufficiently diverse, hence, it did not perform well on an unknown dataset.

For our machine learning framework, the performance decreased slightly for the CNS-OT dataset, although it was still higher than that of EDAQA. We believe this is because the EDA motion artifact dataset

is smaller than the CNS-OT dataset (3,496 segments vs 5,142 segments). Hence, increasing the number of training samples will increase the diversity in training and may improve the accuracy on the independent testing datasets. The fact that the performance of our machine learning method was more accurate when compared to the SVM classifier proposed in [30] indicates that the proposed feature set better characterized the motion artifacts than did the one used by [30].

Note that the sole purpose of having a reference EDA dataset was for the purpose of accurate annotation of clean versus motion-corrupted data. The data collected in this work is limited to some of the most common scenarios of movement, but we cannot claim to have accounted for all types of motion artifact scenarios. However, even with this limited training dataset, when the machine learning approach was tested on an entirely different dataset, the CNS-OT dataset, our results did not differ much, albeit they were slightly decreased in performance. This suggests that, once the machine learning model is built with sufficient training data, no reference data is required during data collection. Note that accelerometer data are often available with wearable devices, including EDA. Hence, the accelerometer data can also be used to discriminate severely corrupted data. However, there are cases when accelerometer data are not useful. For example, poor EDA data due to bad electrode contact with skin, not motion artifacts, would not be detected if only accelerometer data were being relied on. Moreover, in certain cases, although accelerometer data may indicate motion corruption, the EDA may not be as affected. Hence, some portion of useable and good data might be discarded if the motion artifact detection is based solely on accelerometer data.

A) Limitations

This study has some limitations. One limitation of the study is that the data were collected from only 20 subjects and they may not reflect the overall population. We considered a limited number of movements to mimic common motion artifacts, hence, they might not be sufficient to characterize most

typical types of motion artifact. Moreover, as there are many modes of EDA sensors, depending on the sensors used for the EDA collection, retraining of the machine learning model may be required.

CONCLUSIONS

We presented an automated and accurate EDA motion artifact detection approach based on machine learning. We created an EDA database that is annotated as either clean or noisy data using a reference signal that is devoid of motion artifact for more accurate adjudication. This annotated database will be available to researchers upon request. For machine learning, we used a set of features that have been noted to be useful in the literature as well as new features we found to have good ability to identify motion artifacts. We investigated several machine learning algorithms and evaluated their performance using the subject independent LOSO validation strategy. Our results suggest that the proposed machine learning method performed significantly better than both the baseline classifiers and a recently published promising method, which we considered for comparison with our method. Moreover, the performance of the proposed machine learning method maintained its accuracy on an independent dataset, suggesting that most of the dynamics associated with motion artifacts have been accounted for. As we combine EDA motion artifact dataset along with other EDA datasets, we can then build a more comprehensive and accurate motion artifact detection method using machine and deep learning approaches in the future.

Acknowledgement

This work was supported by Office of Naval Research (ONR) grant #N00014-21-1-2255.

References

[1] "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, 2012, doi: https://doi.org/10.1111/j.1469-8986.2012.01384.x.

[2] J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds., *Handbook of Psychophysiology*, 4th ed. Cambridge: Cambridge University Press, 2016. doi: 10.1017/9781107415782.

[3] W. Boucsein, *Electrodermal Activity*, 2nd ed. Springer US, 2012. doi: 10.1007/978-1-4614-1126-0.

[4] P. H. Ellaway, A. Kuppuswamy, A. Nicotra, and C. J. Mathias, "Sweat production and the sympathetic skin response: Improving the clinical assessment of autonomic function," *Auton. Neurosci. Basic Clin.*, vol. 155, no. 1, pp. 109–114, Jun. 2010, doi: 10.1016/j.autneu.2010.01.008.

[5] C. Setz, B. Arnrich, J. Schumm, R. L. Marca, G. Tröster, and U. Ehlert, "Discriminating Stress From Cognitive Load Using a Wearable EDA Device," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, Mar. 2010, doi: 10.1109/TITB.2009.2036164.

[6] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005, doi: 10.1109/TITS.2005.848368.

[7] H. F. Posada-Quintero and K. H. Chon, "Innovations in Electrodermal Activity Data Collection and Signal Processing: A Systematic Review," *Sensors*, vol. 20, no. 2, Art. no. 2, Jan. 2020, doi: 10.3390/s20020479.

[8] M. M. Bradley and P. J. Lang, "Emotion and motivation," in *Handbook of psychophysiology, 3rd ed*, New York, NY, US: Cambridge University Press, 2007, pp. 581–607. doi: 10.1017/CBO9780511546396.025.

[9] E. Di Lascio, S. Gashi, and S. Santini, "Laughter Recognition Using Non-invasive Wearable Devices," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, New York, NY, USA, May 2019, pp. 262–271. doi: 10.1145/3329189.3329216.

[10] E. Di Lascio, S. Gashi, and S. Santini, "Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, p. 103:1-103:21, Sep. 2018, doi: 10.1145/3264913.

[11] Md. R. Amin and R. T. Faghih, "Identification of Sympathetic Nervous System Activation From Skin Conductance: A Sparse Decomposition Approach With Physiological Priors," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 5, pp. 1726–1736, May 2021, doi: 10.1109/TBME.2020.3034632.

[12] T. Reinhardt, C. Schmahl, S. Wüst, and M. Bohus, "Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST)," *Psychiatry Res.*, vol. 198, no. 1, pp. 106–111, Jun. 2012, doi: 10.1016/j.psychres.2011.12.009.

[13] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, New York, NY, USA, Sep. 2016, pp. 1185–1193. doi: 10.1145/2968219.2968306.

[14] J. Hernandez, R. R. Morris, and R. W. Picard, "Call Center Stress Recognition with Person-Specific Models," in *Affective Computing and Intelligent Interaction*, Berlin, Heidelberg, 2011, pp. 125–134. doi: 10.1007/978-3-642-24600-5_16.

[15] K. Kalimeri and C. Saitis, "Exploring multimodal biosignal features for stress detection during indoor mobility," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, Oct. 2016, pp. 53–60. doi: 10.1145/2993148.2993159.

[16] H. F. Posada-Quintero *et al.*, "Using electrodermal activity to validate multilevel pain stimulation in healthy volunteers evoked by thermal grills," *Am. J. Physiol.-Regul. Integr. Comp. Physiol.*, vol. 319, no. 3, pp. R366–R375, Jul. 2020, doi: 10.1152/ajpregu.00102.2020.

[17] Y. Kong, H. F. Posada-Quintero, and K. H. Chon, “Pain Detection using a Smartphone in Real Time*,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Jul. 2020, pp. 4526–4529. doi: 10.1109/EMBC44109.2020.9176077.
[18] Y. Kong, H. F. Posada-Quintero, and K. H. Chon, “Real-Time High-Level Acute Pain Detection Using a Smartphone and a Wrist-Worn Electrodermal Activity Sensor,” *Sensors*, vol. 21, no. 12, Art. no. 12, Jan. 2021, doi: 10.3390/s21123956.
[19] H. F. Posada-Quintero, Y. Kong, and K. H. Chon, “Objective Pain Stimulation Intensity and Pain Sensation Assessment Using Machine Learning Classification and Regression Based on Electrodermal Activity,” *Am. J. Physiol.-Regul. Integr. Comp. Physiol.*, Jun. 2021, doi: 10.1152/ajpregu.00094.2021.
[20] A. Bechara, H. Damasio, A. R. Damasio, and G. P. Lee, “Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision-Making,” *J. Neurosci.*, vol. 19, no. 13, pp. 5473–5481, Jul. 1999, doi: 10.1523/JNEUROSCI.19-13-05473.1999.
[21] E. B. Prince *et al.*, “The relationship between autism symptoms and arousal level in toddlers with autism spectrum disorder, as measured by electrodermal activity,” *Autism*, vol. 21, no. 4, pp. 504–508, May 2017, doi: 10.1177/1362361316648816.
[22] A. E. Meuret *et al.*, “Do Unexpected Panic Attacks Occur Spontaneously?,” *Biol. Psychiatry*, vol. 70, no. 10, pp. 985–991, Nov. 2011, doi: 10.1016/j.biopsych.2011.05.027.
[23] J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris, “Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life,” in *Pervasive Computing*, Berlin, Heidelberg, 2010, pp. 156–173. doi: 10.1007/978-3-642-12654-3_10.
[24] F. Shaffer, D. Combatalade, E. Peper, and Z. M. Meehan, “A Guide to Cleaner Electrodermal Activity Measurements,” *Biofeedback*, vol. 44, no. 2, pp. 90–100, Jun. 2016, doi: 10.5298/1081-5937-44.2.01.
[25] I. R. Kleckner *et al.*, “Simple, Transparent, and Flexible Automated Quality Assessment Procedures for Ambulatory Electrodermal Activity Data,” *IEEE Trans. Biomed. Eng.*, vol. 65, no. 7, pp. 1460–1467, Jul. 2018, doi: 10.1109/TBME.2017.2758643.
[26] Y. Zhang, M. Haghdan, and K. S. Xu, “Unsupervised motion artifact detection in wrist-measured electrodermal activity data,” in *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, New York, NY, USA, Sep. 2017, pp. 54–57. doi: 10.1145/3123021.3123054.
[27] R. Wang *et al.*, “Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 43:1-43:26, Mar. 2018, doi: 10.1145/3191775.
[28] J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, and R. W. Picard, “Using electrodermal activity to recognize ease of engagement in children during social interactions,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA, Sep. 2014, pp. 307–317. doi: 10.1145/2632048.2636065.
[29] “Artifact detection in electrodermal activity using sparse recovery.” https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10211/102110D/Artifact-detection-in-electrodermal-activity-using-sparse-recovery/10.1117/12.2264027.short (accessed Jun. 28, 2021).
[30] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, “Automatic identification of artifacts in electrodermal activity data,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2015, pp. 1934–1937. doi: 10.1109/EMBC.2015.7318762.
[31] V. Xia, N. Jaques, S. Taylor, S. Fedor, and R. Picard, “Active learning for electrodermal activity classification,” in *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Dec. 2015, pp. 1–6. doi: 10.1109/SPMB.2015.7405467.

[32] M. B. Hossain, H. F. Posada-Quintero, Y. Kong, R. McNaboe, and K. Chon, “A Preliminary Study on Automatic Motion Artifacts Detection in Electrodermal Activity Data Using Machine Learning,” *ArXiv210707650 Eess*, Jul. 2021, Available: http://arxiv.org/abs/2107.07650
[33] W. Chen, N. Jaques, S. Taylor, A. Sano, S. Fedor, and R. W. Picard, “Wavelet-based motion artifact removal for electrodermal activity,” *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2015, pp. 6223–6226, 2015, doi: 10.1109/EMBC.2015.7319814.
[34] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig, “Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity,” *IEEE Trans. Affect. Comput.*, pp. 1–1, 2019, doi: 10.1109/TAFFC.2019.2901673.
[35] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, “cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2016, doi: 10.1109/TBME.2015.2474131.
[36] A. Hajj-Ahmad, R. Garg, and M. Wu, “ENF-Based Region-of-Recording Identification for Media Signals,” *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 6, pp. 1125–1136, Jun. 2015, doi: 10.1109/TIFS.2015.2398367.
[37] J. Moon, M. B. Hossain, and K. H. Chon, “AR and ARMA model order selection for time-series modeling with ImageNet classification,” *Signal Process.*, vol. 183, p. 108026, Jun. 2021, doi: 10.1016/j.sigpro.2021.108026.
[38] H. Wang, K. Siu, K. Ju, and K. H. Chon, “A High Resolution Approach to Estimating Time-Frequency Spectra and Their Amplitudes,” *Ann. Biomed. Eng.*, vol. 34, no. 2, pp. 326–338, Feb. 2006, doi: 10.1007/s10439-005-9035-y.
[39] H. F. Posada-Quintero, J. P. Florian, Á. D. Orjuela-Cañón, and K. H. Chon, “Highly sensitive index of sympathetic activity based on time-frequency spectral analysis of electrodermal activity,” *Am. J. Physiol.-Regul. Integr. Comp. Physiol.*, vol. 311, no. 3, pp. R582–R591, Jul. 2016, doi: 10.1152/ajpregu.00180.2016.
[40] M.-B. Hossain, S. K. Bashar, J. Lazaro, N. Reljin, Y. Noh, and K. H. Chon, “A robust ECG denoising technique using variable frequency complex demodulation,” *Comput. Methods Programs Biomed.*, p. 105856, Nov. 2020, doi: 10.1016/j.cmpb.2020.105856.
[41] M.-B. Hossain, J. Lázaro, Y. Noh, and K. H. Chon, “Denoising Wearable Armband ECG Data Using the Variable Frequency Complex Demodulation Technique,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Jul. 2020, pp. 592–595. doi: 10.1109/EMBC44109.2020.9175665.
[42] Y. Kong and K. H. Chon, “Heart Rate Tracking Using a Wearable Photoplethysmographic Sensor During Treadmill Exercise,” *IEEE Access*, vol. 7, pp. 152421–152428, 2019, doi: 10.1109/ACCESS.2019.2948107.
[43] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[44] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006. Accessed: Jul. 23, 2021. [Online]. Available: https://www.springer.com/gp/book/9780387310732
[45] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
[46] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[47] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[48] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.
[49] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, “Monitoring stress with a wrist device using context,” *J. Biomed. Inform.*, vol. 73, pp. 159–170, Sep. 2017, doi: 10.1016/j.jbi.2017.08.006.
[50] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, “Voodoo Machine Learning for Clinical Predictions,” *bioRxiv*, p. 059774, Jun. 2016, doi: 10.1101/059774.
[51] Y. Kong, H. Posada-Quintero, and K. Chon, “Sensitive Physiological Indices of Pain based on Differential Characteristics of Electrodermal Activity,” *IEEE Trans. Biomed. Eng.*, pp. 1–1, 2021, doi: 10.1109/TBME.2021.3065218.
[52] “Introduction to Machine Learning with Python [Book].” https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/ (accessed Jul. 26, 2021).
[53] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA, USA: MIT Press, 2004.
[54] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *ArXiv14126980 Cs*, Jan. 2017, Available: http://arxiv.org/abs/1412.6980
[55] “Practical Methods of Optimization, 2nd Edition | Wiley,” *Wiley.com*. https://www.wiley.com/en-us
[56] Hugo F. Posada-Quitero, Bruce J. Derrick, Christopher Winstead-Derlega, Sara I. Gonzalez, M. Claire Ellis, John J. Freiberger, Ki H. Chon, “Time-varying Spectral Index of Electrodermal Activity to Predict Central Nervous System Oxygen Toxicity Symptoms in Divers:Preliminary results,” presented at the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC).