Detection of Artifacts in Ambulatory Electrodermal Activity Data

SHKURTA GASHI, Università della Svizzera italiana (USI), Switzerland
ELENA DI LASCIO, Università della Svizzera italiana (USI), Switzerland
BIANCA STANCU, Unit8 SA, Switzerland
VEDANT DAS SWAIN, Georgia Institute of Technology, USA
VARUN MISHRA, Dartmouth College, USA
MARTIN GJORESKI, Jozef Stefan Institute, Slovenia
SILVIA SANTINI, Università della Svizzera italiana (USI), Switzerland

Recent wearable devices enable continuous and unobtrusive monitoring of human's physiological parameters, like e.g., electrodermal activity and heart rate, over long periods of time in everyday life settings. Continuous monitoring of these parameters enables the creation of systems able to predict affective states and stress with the goal of providing feedback to improve them. Deployment of such systems in everyday life settings is still complex and prone to errors due to the low quality of the collected data impacted by the presence of artifacts. In this paper we present an automatic approach to detect artifacts in electrodermal activity (EDA) signals collected in-the-wild over long periods of time. To this end we first perform a systematic literature review and compile a set of guidelines for human annotators to label artifacts manually and we use these labels as ground-truth to test our automatic approach. To evaluate our approach, we collect physiological data from 13 participants in-the-wild and two human annotators label 107.56 hours of this data set. We make the data set publicly available to other researchers upon request. Our model achieves a recall of 98% for clean and shape artifacts classification on data collected in-the-wild using leave-one-subject-out cross-validation, which is 42 percentage points higher than the baseline. We show that state of the art approaches do not generalize well when tested with completely in-the-wild data and identify only 17% of the artifacts present in our data set, even after manual adaption. We further test the robustness of our approach over time using leave-one-day-out and achieve very similar performance. We then introduce a new metric to evaluate the quality of EDA segments that considers the impact of not only artifacts in the shape of EDA but also artifacts generated by environmental temperature changes or user's high intensity movement. Our results imply that we can eliminate the need for human annotators or significantly reduce the time they need to label data. Also, our approach can be used in an online manner to automatically detect artifacts in EDA signals.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Wearable Sensing, Data Quality, Electrodermal Activity, Artifacts Detection, Ambulatory

Authors' addresses: Shkurta Gashi, Università della Svizzera italiana (USI), Lugano, Switzerland, shkurta.gashi@usi.ch; Elena Di Lascio, Università della Svizzera italiana (USI), Lugano, Switzerland, elena.di.lascio@usi.ch; Bianca Stancu, Unit8 SA, Zürich, Switzerland, bianca.stancu28@gmail.com; Vedant Das Swain, Georgia Institute of Technology, Atalanta, Georgia, USA, vedantswain@gatech.edu; Varun Mishra, Dartmouth College, Hanover, New Hampshire, USA, varun@cs.dartmouth.edu; Martin Gjoreski, Jozef Stefan Institute, Ljubljana, Slovenia, martin.gjoreski@ijs.si; Silvia Santini, Università della Svizzera italiana (USI), Lugano, Switzerland, silvia.santini@usi.ch.

2474-9567/2020/6-ART44 $15.00
https://doi.org/10.1145/3397316

ACM Reference Format:
Shkurta Gashi, Elena Di Lascio, Bianca Stancu, Vedant Das Swain, Varun Mishra, Martin Gjoreski, and Silvia Santini. 2020. Detection of Artifacts in Ambulatory Electrodermal Activity Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 44 (June 2020), 31 pages. https://doi.org/10.1145/3397316

1 INTRODUCTION

The activation of sweat glands in the human body causes changes in the electrical conductivity of the skin. This phenomenon – usually referred to as Electrodermal Activity (EDA) or Galvanic Skin Response (GSR) [3] – can be measured using inexpensive, body-worn devices. Even though not without controversy [39], it is widely accepted that humans' affective states and stress levels can be inferred from EDA measurements [3, 38]. This is because EDA is linked to the physiological arousal induced by the Sympathetic Nervous System (SNS) in response to humans' distress or eustress [3].

The possibility to infer a person's emotions and other "*internal states*" like cognitive load or distress is a fundamental premise towards the creation of ubiquitous computing systems that can "*sense, interpret, adapt, and potentially respond appropriately to human emotion*" [42]. Traditional research on EDA was however conducted mainly in laboratory or semi-controlled settings, and using equipment inadequate to be carried in daily-life situations for long periods of time. This has left open the question about whether results obtained in laboratory experiments would find application in real-life scenarios.

The recent availability of truly robust and unobtrusive, mainly wrist-worn devices able to measure EDA in-the-wild has thus spurred new vigor to this topic and research has moved from the lab to real-life settings [13, 20, 28, 29, 52, 53, 55, 56, 62]. For instance, Hernandez et al. [28] use EDA data of call center employees to detect stressful calls and provide them timely interventions to improve their productivity. Taylor et al. [62] instead investigate a multi-task approach to predict daily mood, stress and health of college students using EDA data and other sensory inputs.

The use of EDA data collected in ambulatory settings is however hampered by a major issue: data quality. Even when collected in laboratory settings, EDA signals – as well as other physiological signals including heart rate – are affected by artifacts. These are defined as "*changes in the recorded biosignal which do not stem from the signal source in question*" [3] and may be caused by the recording procedure or by "*physiological responses in systems other than the electrodermal one*" [3]. In ambulatory settings, artifacts may become so predominant to make the entire recorded signal unusable. Wang et al. [65], for instance, discard all EDA data collected over several weeks due to the low quality of the signals. Other researchers also report that they discard significant amounts of EDA data in the data cleaning phase [13, 20, 26, 29]. Shukla et al. [59] also show that artifacts may mask the existence of correlations in the data.

While detecting and removing artifacts is necessary and crucial step in any EDA data analysis pipeline, the current means available to do so are surprisingly limited. The widely adopted textbook by Boucsein stresses that the "*detection of artifacts in the EDA signal necessitates a visual inspection of the data sequence*" [3]. Visual inspection is however cumbersome and time-consuming, and thus seriously impractical for data sequences collected for many users and continuously over weeks or months. Signal processing techniques – like, e.g., low-pass filtering [28, 37] – can be applied to avoid visual inspection. However, these approaches modify the entire EDA trace thereby distorting also genuine physiological responses [70] or, conversely, cause true artifacts to be transformed into genuine-looking data [35].

To cope with the limitations of visual inspection and signal processing approaches, recent research focused on devising methods to automatically detect and remove individual artifacts using, e.g., rule-based techniques [36], or supervised [61], semi-supervised [66] and unsupervised [70] machine learning approaches. These approaches have mainly been tested using data from a specific context – i.e., in a laboratory setting where users perform a predefined set of tasks [61] – or a specific demographic group of users [36], which might be difficult to generalize to new contexts or even new users. In addition they do not consider the impact of user's movement and environment temperature change to differentiate data segments with high or low quality, which we believe is critical to allow researchers to develop comprehensive methods to judge whether or not specific segments should be included in their data analysis pipeline. To enable the creation and deployment of EDA-based systems in-the-wild, it is however necessary to first devise effective methods to recognize the quality of the data being collected, which is the focus of the work presented in this paper.

In this paper we present our findings in automatic detection of artifacts in EDA signals collected in-the-wild over long periods of time. The main goal of our work is to propose an automatic approach to detect artifacts in the shape of EDA signals, which we refer to as *shape artifacts* and to estimate the overall quality of EDA signals by considering artifacts caused by user's movement and environmental temperature, which we refer to as *thermoregulation responses*. To detect shape artifacts we build upon ensemble and deep learning techniques, which have shown great success in many tasks, like e.g., activity recognition [49] or affect detection [62], but have not previously been explored for this problem. To quantify thermoregulation responses we propose a new metric – EDA_{QI} – that considers not only the amount of shape artifacts as in existing work in literature [10, 35, 35, 59, 61, 66], but also artifacts that resemble physiological responses. To evaluate our approach, we collect a data set from 13 users in-the-wild over the long term and two human annotators label 107.56 hours of this data set. This represents a larger data set than those used in similar studies [35, 36, 59, 61, 70]. Our results show that shape artifacts can be identified with a recall of 98% using deep neural networks and ensemble classifiers, which is 42 percentage points higher than the baseline classifier. Further we test the generizability of our approach to a new user and show that we can effectively identify artifacts in completely unseen data.

Shape artifacts detection could eliminate the need for researchers to manually inspect the data by, for instance being embedded in the EDA processing pipeline to automatically find segments of data that need visual inspection or to completely remove them. This would significantly reduce the amount of time needed by researchers to visually inspect the amount of data being recently collected in ambulatory studies. EDA quality quantification on the other hand could help researchers to improve their detection tasks by providing information on which data is more reliable. This would then be used to inform the real-time EDA-based monitoring systems to not send interventions in inopportune moments or with inappropriate content. To summarize, this paper presents the following contributions:

- We collect a data set from 13 participants in-the-wild using wearable sensors and two human annotators label 107.56 hours of this data set. To the best of our knowledge, this is the largest data set with annotations of EDA shape artifacts. We provide our data set to other researchers upon request to reproduce the results or perform further analysis.
- We present an automatic approach that identifies shape artifacts in EDA with 98% recall, which is 42 percentage points increment from the baseline. We show that state of the art approaches do not generalize well when tested with completely in-the-wild data or identify only 17% of the artifacts present in our data set, even after manual adaptation. We further train and test existing approaches with in-the-wild data and show that our approach still achieves 8 percentage points increment for shape artifacts detection.
- To enable manual investigation and labeling of EDA traces, we first compile a set of labeling guidelines, through a systematic literature review. We then develop an open-source dashboard – the *EDArtifact* – following standard dashboard design principles and overcoming limitations of existing dashboards, which we make publicly available to other researchers.
- We show statistical evidence that three factors: presence of shape artifacts, change in temperature environment and vigorous physical activity of the user significantly impact the features used commonly in EDA analysis. We propose a new metric – EDA_{QI} – that reflects the quality of EDA based on these three factors.

The structure of the paper is as follows. We present motivation and methodology of our approach in Section 2. We describe the data set we collected in Section 3. We expand on our framework on EDA shape artifacts detection in Section 4. We then discuss the classification results in Section 5. Following that, we explain our approach to devise ground-truth in Section 6. In Section 7 we define and discuss the EDA quality index. We present implications of our approach in Section 8, and in Section 9 we discuss limitations and possible directions for future work. We describe related work on shape artifacts detection and overall EDA quality estimation in Section 10. We present concluding remarks in Section 11.

2 MOTIVATION AND METHODOLOGY OF OUR APPROACH

Our main goal is to develop an automatic approach to recognize *artifacts* in electrodermal activity (EDA) signals collected in ambulatory settings over long periods of time. In this section we further elaborate the definition of EDA and artifacts, and discuss our approach for automatic recognition of artifacts.

EDA indicates the physiological arousal of a person, which measures the changes in electrical potential on the skin surface due to eccrine sweat gland activity [3]. The main characteristic of EDA are peaks, known as skin conductance responses (SCRs) [3], which occur as a reaction to a stimuli. Measurement of EDA in ambulatory settings creates uncertainty on what causes the peaks in the signal and hence, makes the signal vulnerable to the presence of artifacts. As of the standard textbook for EDA data collection and analysis [3] "*artifacts are defined as changes in the recorded bio-signal, which do not stem from the signal source in question. Instead, they may result from the recording procedure or from physiological responses in systems other than the electrodermal one*". Among the factors that influence the presence of artifacts in ambulatory EDA signals are the recording procedure (e.g., the stability of electrodes) [3, 26], the influence of environment temperature (e.g., temperature rise or drop) [3, 57], and the user's physical activity [3, 16, 36, 45, 70]. These factors cause artifacts that could resemble or not SCRs. For this reason we divide artifacts into two sub-groups described as follows:

- *Shape artifacts* refer to artifacts that do not resemble physiological responses. Figure 5 presents several examples of shape artifacts. Improper placement of electrodes or their movement for instance causes abrupt changes in the signal that cannot be generated by the electrodermal system itself and do not conform to the specifications of physiological responses. We compile a set of guidelines – specific to the SCRs shape and for the EDA in general – for human annotators to manually identify shape artifacts. We then derive features based on these heuristics to capture the characteristics of SCRs and use these features as input to a classification pipeline to distinguish shape artifacts from clean physiological responses.
- *Thermoregulation responses* refer to physiological responses that are similar to EDA responses but are not caused by the electrodermal system. High physical activity or even increase in environmental temperature rises user's sweating – i.e., to dissipate the body heat generated – and hence, leading to physiological responses in EDA signal caused by thermoregulation rather than the electrodermal system [16]. Figure 6 shows an example of thermoregulation responses in the EDA signal. Such artifacts might be misinterpreted as physiological responses elicited by for instance an emotion, reducing the reliability of an emotion recognition system, as in [68]. Since thermoregulation responses are frequently encountered in practical applications, they represent a serious problem that needs to be addressed but has so far gained little attention in EDA analysis. To identify these artifacts, we suggest to use the accelerometer sensor to detect the parts of the signal with high intensity movement and skin temperature sensor data to identify the parts when the environment temperature is not constant. We use the skin temperature as a proxy of environment temperature because the latter is essentially a low pass version of the skin temperature [18].

We therefore suggest to consider not only shape artifacts but also thermoregulation responses before further analysis of the EDA signal. Thereby, we propose an EDA quality metric that considers and weights the user's movement intensity, the environment temperature changes and the presence of shape artifacts.

3 DATA SET

To evaluate our approach for shape artifacts detection, we run an in-the-wild study at our institution to collect data while users perform their daily activities in a totally unconstrained setting. We collect in total 2260.03 hours of data from 13 participants – to which we refer as *collected data* – and we annotate 107.56 hours of this data from 13 participants – to which we refer as *annotated data*. For our analysis we use the annotated data and we provide both the collected and annotated data to other researchers upon request[1]. We provide below details about the participants, the procedure we followed and the data we collected and annotated.

3.1 Procedure and Participants

We recruit 16 participants (6 females and 10 males) for approximately 30 days between October and December 2018. The majority of participants are of age in the range 20-30 and one 30-40. The participants are either students, researchers or lecturers at our institution. To recruit the participants, we present the study during one of the lectures of a course held at our institution. We therefore inform the participants about the purpose and duration of the study, the devices used and the data collection procedure to be followed. Participants who volunteered to participate signed an informed consent form – in accordance with the Institutional Review Board at our institution – and received their data to analyse as part of the assignments of the course. We ask the participants whether or not they provide permission to analyze their data for research purposes after being anonymized. Out of the 16 participants, two declined permission to use their data and from one participant we did not receive any data. Thereby, we end up having data from 13 participants.

3.2 Collected Data

To collect physiological data from participants, we use the E4 wristband presented in [18]. The E4 is a lightweight, unobtrusive wristband equipped with four sensors that measure: the electrodermal activity (EDA), the blood volume pulse (BVP), the 3-axis acceleration (ACC), and the skin temperature (ST) [18]. Participants were requested to wear the E4 on their non-dominant hand and to wear it everyday during their wake hours. Our recruitment and data collection efforts led to a data set with 13 participants, 2260.03 hours of data over 36 unique days.

3.3 Annotated Data

Our annotated data set contains labels from two human annotators for 107.56 hours of the collected data. Each 5-second segment is manually labeled with shape artifact or clean by two human annotators. Each human annotator spent on average 48 hours to label the data. The size of our data set is larger than data sets used in similar studies shown in Table 1. We create ground-truth for each 5-second EDA segment based on the agreement of two human annotators, which we describe in details in Section 6. Our labeling efforts lead to 67318 samples in the clean class, 8267 samples in the artifacts class and 2642 disagreements.

3.3.1 Subset 1. To test the performance of our automated procedure to new users, we annotate the same amount of data from all 13 participants. We select four hours of data from each participant and we provide that to human annotators for labeling. To select the data, we first look at the amount of data collected by each user for each day. Figure 11 in Appendix A shows the amount of data in hours collected by each user over the period of data collection. Figure 12 in Appendix A shows the quantity of data collected from all users per hour. Given that the majority of participants were wearing the device during November 12 to November 15, as show in Figure 11, and mostly during the morning (e.g., from 09:00 until 13:00) and evening (e.g., from 19:00 to 22:00), as shown in Figure 12), we randomly select four hours of data from 13 participants either from the morning or evening period collected during November 12 to November 15. For participants who did not have any data during that week,

[1] Please contact the last author to make a request for the collected or annotated data sets.

Table 1. Comparison between our and existing data sets. * – The approaches presented in [10, 66] use the same data set as in [61]. N/A stands for not available.

Paper	**Participants**	**Setting**	**Collected Data**	**Annotated Data**
Kleckner et al. [36]	20	Home	181 hours	32 hours
Taylor et al. [61], [10, 66]*	32	In-the-lab	N/A	1560 samples
Zhang et al. [70]	21	In-the-lab & in-the-wild	23 hours	23 hours
Kesley et al. [34]	10	Hospital	N/A	376 samples
Shukla et al. [59]	15	Driving	N/A	9056 samples
Our data set	13	**In-the-wild**	**2260.03 hours**	**107.56 hours**

we randomly select the data before or after that week. We report the annotated data for Subset 1 in Table 7 in Appendix A.

3.3.2 Subset 2. To test the robustness of our automated approach through time, we annotate six days of data from three subjects. We pick the data on a unique day of the week because the data from the same user on the same day may contain similar activities, i.e., a user may perform similar physical activities on Mondays. With this procedure we try to ensure to have data collected from multiple and different ambulatory settings. The inclusion criteria based on the month of data collection was set to ensure the presence of data with diversity in the environment temperature. With this procedure we end up with one session collected in October, four sessions collected in November and one session in December. We report the annotated data in Subset 2 in Table 8 in Appendix A.

4 ELECTRODERMAL ACTIVITY SHAPE ARTIFACTS DETECTION FRAMEWORK

To recognize EDA *shape artifacts* we setup a binary classification framework. The framework is a pipeline of signal processing and machine learning techniques that distinguishes shape artifacts from clean parts of the signal. We first apply a common procedure for pre-processing EDA data similar to the one described in [61]. To test our approach we obtain ground-truth labels from the agreement of two human annotators, as described in Section 6. We then segment the EDA signals into 5-second non-overlapping segments. We extract features from these segments based on the labeling guidelines and features suggested in the literature [61, 70]. We provide these features as an input to several models. We test the performance of our model using leave-one-user-out and leave-one-day-out validation procedures. In this section we describe the pre-processing, segmentation, feature extraction, classification, evaluation procedure and classifier hyper-parameter tuning we followed to perform the analysis.

4.1 Pre-processing

In the first stage we use raw signals collected from the EDA sensor located on user's wristband. We filter the signal using first order Butterworth low-pass filter with a cut-off frequency of 0.6Hz similar to [32] to remove the high frequency noise fluctuations. Even after filtering there is still a considerable amount of artifacts remaining in the data, as shown in Section 3.3, confirming our initial assumptions that traditional filtering techniques might not be effective to remove artifacts that are common in natural settings. As a part of our pre-processing steps, we further decompose the mixed EDA signal into the tonic and phasic components using the cvxEDA approach presented in [24], as a common procedure in the literature [13, 20, 29]. In this work we use only the mixed EDA signal and phasic component because the tonic component does not contain information about the SCRs but only the overall trend, which is also present in the mixed EDA signal. We then discard the EDA segments with a

Table 2. Overview of the 37 features used in this work. Statistical features are extracted from both the EDA-mixed and phasic signals, SCR features from the phasic component and wavelets features from the wavelet coefficients applied in the whole EDA signal.

Feature group	**Features**
Statistical* (22 features)	minimum, maximum, mean, median, standard deviation, dynamic range, slope, mean and standard deviation of the first and second derivative
SCR (6 features)	number of peaks, peaks' amplitude, rise time, half-recovery time, width and area under the curve
Wavelets (9 features)[3]	mean, median, standard deviation of wavelets coefficients extracted at three different time scales 4Hz, 2Hz and 1Hz

slope in range [-0.002, +0.002] because we consider them as flat responses, which do not contain for instance SCRs and hence do not contain shape artifacts. We choose the slope range empirically by visualizing several parts of the signal. Removal of flat responses leads to an imbalanced data set with 7729 points in the artifacts class and 8057 in the clean class, which is still comparable or even larger than existing similar data sets e.g., in [33, 61, 66, 70]. In Appendix B we present results with (Table 12) and without flat responses (Table 9).

4.2 Segmentation

Given that a SCR response lasts between 1 to 5 seconds [3], we divide the EDA traces in 5 second non-overlapping segments as a common procedure in shape artifacts detection in [36, 61, 70]. The EDA segments might contain artifacts or clean responses. This segmentation procedure allows us to compare our approach to the existing work in [36, 61, 70].

4.3 Feature Extraction

We extract 37 features from each 5 second non-overlapping window of EDA that could appropriately characterize the shape artifacts. Table 2 provides the list of the considered features. To identify artifacts we extract most of the features suggested in the existing literature [58, 61, 70] and we further extract the features related to the SCR to be able to distinguish between valid and invalid SCRs. We separate the features into three high level groups, as suggested in [58], namely, statistical, SCR and wavelet features. We normalize each feature before providing as input to the classifiers, using the minmax scaler[2], as a common pre-processing procedure in [25, 44, 58].

4.3.1 Statistical Features. We first extract 11 statistical features from both EDA-mixed and phasic component of EDA. We extract the common statistical features including features such as the minimum, maximum, mean, median, standard deviation as suggested in [56, 58]. To account for changes in the signal we derive the dynamic range – defined as the difference between the maximum and minimum value in the 5 second segment – [14, 20], the slope of the signal (estimated with linear regression [29]), mean and standard deviation of the first and second derivative as in [61, 69]. We expect to observe higher dynamic range in artifacts segments i.e., when EDA rises or drops too quickly, lower minimum value or higher maximum value i.e., when EDA is out of range.

4.3.2 SCR Features. We extract 6 SCR features from the phasic component of EDA, including the number of peaks, peaks' rise time, half-recovery time, amplitude, width and area under the curve. These features specifically show the characteristics of the SCRs and could have an impact on detecting shape artifacts. SCR features have been considered in previous work for other detection problems [29, 56, 58], but not specifically for automatic

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

identification of shape artifacts. We expect to observe changes in all of these features between shape artifacts and clean segments. For artifacts we expect a higher number of peaks within a 5-second segment when EDA changes too quickly, shorter half-recovery time when SCR drops quickly and shorter rise time when SCR rises quickly. We extract SCR features using the open-source *EDAExplorer* [4] tool, presented in [61]. We set the parameters for finding features related to the shape of EDA based on SCR specifications presented in [12] and discussed in Section 6. In particular, we set the rise time or the maximum number of seconds before the apex of a peak to 3 seconds, the half-recovery time or the maximum number of seconds after the apex of a peak to 10 seconds, as in [12], and the minimum amplitude of the peak at least 0.01 as in [29]. In some cases the half-recovery of the peak is not found, i.e., when there are overlapping peaks, some of the SCR features, such as e.g., decay time, area under the peak, contain missing data. We fill the missing data with a special value such as -1, as suggested in [31].

4.3.3 Wavelets Features. Given that EDA signal exhibits non-stationary behavior [58], we extract wavelet features that might be indicative of sudden changes in EDA [61]. We extract the coefficients of a Discrete Wavelet Transform (DWT) with the Haar wavelet applied in the EDA signals at three different time scales: 4 Hz, 2 Hz and 1 Hz as in [61, 70]. We use a Haar wavelet transform since it computes the degree of relatedness between adjacent points in the signal and it is suitable for detecting edges and sharp changes [61], which are typical characteristics of changes quickly kind of shape artifacts. From the obtained DWT coefficients we then extract the mean, median and standard deviation over a 5-second window.

4.4 Classification

To automatically recognize shape artifacts, we investigate several machine learning techniques used in the literature. In particular we use deep learning techniques – Feed-forward Deep Neural Network (FDNN) [17] –, linear classifiers – Linear Discriminant Analysis (LDA) and Logistic Regression (LR) –, non-linear classifiers – Support Vector Machine (SVM), Quadratic Discriminant Analysis (QDA) and k-Nearest Neighbours (kNN) – [44], and ensemble learning methods – Random Forest (RF) [5], AdaBoost [2], XGBoost [9] – and Decision Tree (DT) [44]. We explore these different approaches to understand which one recognizes shape artifacts more effectively. Given that both our subsets of data are imbalanced, with the shape artifacts being the majority class in Subset 1 and clean being the majority class in Subset 2, the accuracy metric is not an adequate metric to measure the performance of the models [28]. This is because accuracy values could be high if the model classifies all the data samples as the majority class and ignore the class of interest. For this reason, to compare the performance of the classifiers, we include as a baseline a "dummy" classifier that always predicts the majority class, as also suggested in [67]. We refer to this classifier as Baseline1. We further include another "dummy" classifier that does not consider the distribution of the data, to be able to compare the performance in terms of other metrics such as e.g., recall and F1. We refer to this classifier as Baseline2. We use the default parameters for all classifiers as provided by Scikit-Learn[5] library in Python. We compare the results of the classifier with the baseline and with existing approaches in literature using a t-test, with a significance level 0.01 as suggested in [30].

4.4.1 Feed-forward Deep Neural Network. We build a Feed-forward Deep Neural Network (FDNN) comprised of six stacked fully-connected layers – five hidden and one output – each with *tanh* activation function, and with the number of neurons on each hidden layer as follows (240, 120, 60, 30, 15). We provide as input to the FDNN the features we extract from EDA segments. Since we model our problem as a binary classification, to predict whether a 5-second segment is an artifact or not, we use *sigmoid* activation function in the last layer of the network and *binary_crossentropy* as a loss function, as suggested in [17, 63]. Since our aim is to develop a model that generalizes to unseen data, we add a dropout layer after each hidden layer, as suggested in [17, 43], to

[4] https://eda-explorer.media.mit.edu/

[5] https://scikit-learn.org/stable/

regularize the model and to combat overfitting. We choose these parameters of the network by optimizing the value of dropout rate (0.1, 0.2), number of epochs (50, 100, 300), batch size (5, 10, 32), activation functions (ReLU, tanh) and optimizers (ADAM, rmspop). To tune the model we employ a grid search approach that trains the model for each combination of the explored parameters and evaluates the performance on a held-out validation set [1]. The optimal values of the best model for dropout rate, activation function, optimizer, epochs and batch size are 0.1, tanh, ADAM, 300, 5, respectively. We implement FDNN in Python using Keras[6] framework with TensorFlow as a backend engine and monitor the performance of the model using Keras callbacks and TensorBoard[7].

4.5 Evaluation Procedure

To evaluate the performance of the models, we follow common procedures in machine learning [17, 44]. We first divide the data set into train and test using two procedures as follows: (1) leave-one-subject-out and (2) leave-one-day-out. Given that our final data set is imbalanced, we resample the train set by undersampling the majority class to the minority class. This ensures that our models learn the representations from both classes. To resample the data we use the random undersampling technique[8]. We then divide the train set into train and validation with a threshold of 0.2 and by shuffling the data before splitting. We keep a separate validation set to find which model has the best performance as in [61] and to tune the parameters of the classifier. We describe the two validation procedures in the following sections.

4.5.1 Leave-one-subject-out (LOSO). We perform leave-one-subject-out (LOSO) cross validation to evaluate the performance of the models, used also in [13, 23, 51, 54, 55, 70]. We keep the data of all subjects, except one, in the training set and use the remaining subject for testing the performance of the model. We repeat the same procedure for all the subjects and report the performance of the model as average score across all iterations. Using this validation procedure, shape artifact and clean segments derived from the EDA signals of a single user are not contained in the train and test simultaneously. This is a more realistic protocol because the training and test data are different due to interpersonal variance [28]. Thereby it ensures the generalization of the model to new subjects because the models are not biased by the characteristics of particular subjects [14, 23].

4.5.2 Leave-one-day-out (LODO). We use leave-one-day-out (LODO) validation to test the robustness of our system over time. We first separate the data into training and test sets, we keep the data from all days in the training set except one and the remaining for testing, similar to [28, 40, 60]. We repeat the same procedure until all the days are used once for testing and report the performance of the classifiers as the average of all metrics. Since adjacent segments in time-series are not statistically independent, as discussed in [25], we believe the LODO is an efficient approach to evaluate the performance of the classifiers because it avoids putting the segments of the same EDA trace in both train and test sets. With this approach we ensure the generizability of our results to cases when for instance the users might wear the device differently over the days e.g., tighter in one day and looser in another day. We expect this validation procedure to give better performance for our data set because data of the same participant may be present both in training and testing set, making the classification task easier.

4.6 Evaluation Metrics

While the final goal of our work is to distinguish between shape artifacts and clean EDA segments, we are specifically interested to recognize all the artifacts. This is because their presence in the signal might have a significant impact in the analysis, as demonstrated in [59]. For this reason we consider artifacts as the positive class. To evaluate the performance of the models, we consider accuracy, precision, recall, F1 score [44] and Cohen

[6] https://keras.io/

[7] https://www.tensorflow.org/tensorboard

[8] https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.under_sampling.RandomUnderSampler.html

κ [11] metrics. Accuracy quantifies the fraction of samples correctly classified by the model. Precision measures how many of the samples classified as positive are actually positive, in our case how many of the predicted artifacts are indeed artifacts. Recall quantifies the ability of the classifier to identify all instances in the positive class, in our case the artifacts and is our metric of interest. The F1 metric is the harmonic mean of precision and recall [44]. The Cohen κ measures the agreement between two annotators by considering not only the accuracy, but also the degree to which the agreement is due to chance [11]. We consider the Cohen Kappa to compare the agreement between the classifications generated automatically by our models and the labels generated manually by human annotators, as also suggested in [36]. This metric allows us to understand how well the model performs in comparison to human annotators and whether our model could be used to replace them.

4.7 Hyper-parameter Tuning

Within training with the data of all users except one, we perform hyper-parameter tuning on the validation set using a 10-fold cross validation. We optimize the parameters for XGBoost classifier because it is the classifier that performs best in the validation set. To tune the classifier we employ grid search algorithm, which trains the classifier for each combination of the provided parameters and evaluates the performance on a held-out set [1]. The parameter space we explore for XGBoost are: number of estimators, learning rate, maximum tree depth and gamma, as in [8]. We evaluate the performance of the model, for each combination of the parameters. The model hyper-parameters that provided the highest recall across the iterations of the LOSO procedure were selected to train on the full training set and to make classifications.

5 RESULTS

In what follows we report the classification results obtained applying the data analysis steps described in the previous section to distinguish between shape artifacts and clean segments. We first report the performance of the classifiers using the LOSO and LODO validation procedures and compare it to the baseline. We then compare the performance of our approach with existing approaches in literature.

5.1 Performance of all Classifiers and Comparison to the Baseline

Figure 1 (top) shows the classification results for considered classifiers trained using all the features described in Section 4.4 and applying the LOSO validation procedure in Subset 1. In Appendix B, we include a table with numerical representation of these results. From the Figure 1 (top), we can observe that in general the ensemble methods and deep learning classifiers perform better than the others. The deep neural network – FDNN – and ensemble classifiers – XGBoost and AdaBoost – achieve the highest performance in shape artifacts detection with a recall of 96%, which is 40 percentage points higher than the performance of Baseline2. The RF achieves the highest performance in terms of accuracy of 97% which is 40 percentage points higher than Baseline1. RF and SVM shows the highest precision, with a precision 97%, which is 54 percentage points higher than the Baseline2. XGBoost achieves the highest F1 (96%), which is 53 percentage points increment from the Baseline2. The precision, recall and F1 metrics are equal to 0 for the Baseline1 because this classifier always predicts the clean class and hence, the number of predicted true positive samples is always 0. The high performance in terms of recall and precision shows that we are able to correctly detect 96% of the shape artifacts present in the EDA signals in Subset 1, but also ensure that out of all predicted artifacts only a small portion – 5% – of clean segments are miss-classified as shape artifacts. Further, these results show that our approach generalizes well to the data of new users. This might be because shape artifacts are not user-dependent but rather are similar across different users. The performance of all classifiers is significantly higher than the baseline according to t-test with $p<0.01$.

5.1.1 LODO. Figure 1 (bottom) shows the performance of classifiers trained using all the features and LODO evaluation procedure in Subset 2. In Table 10 in Appendix B, we include a table with numerical representation

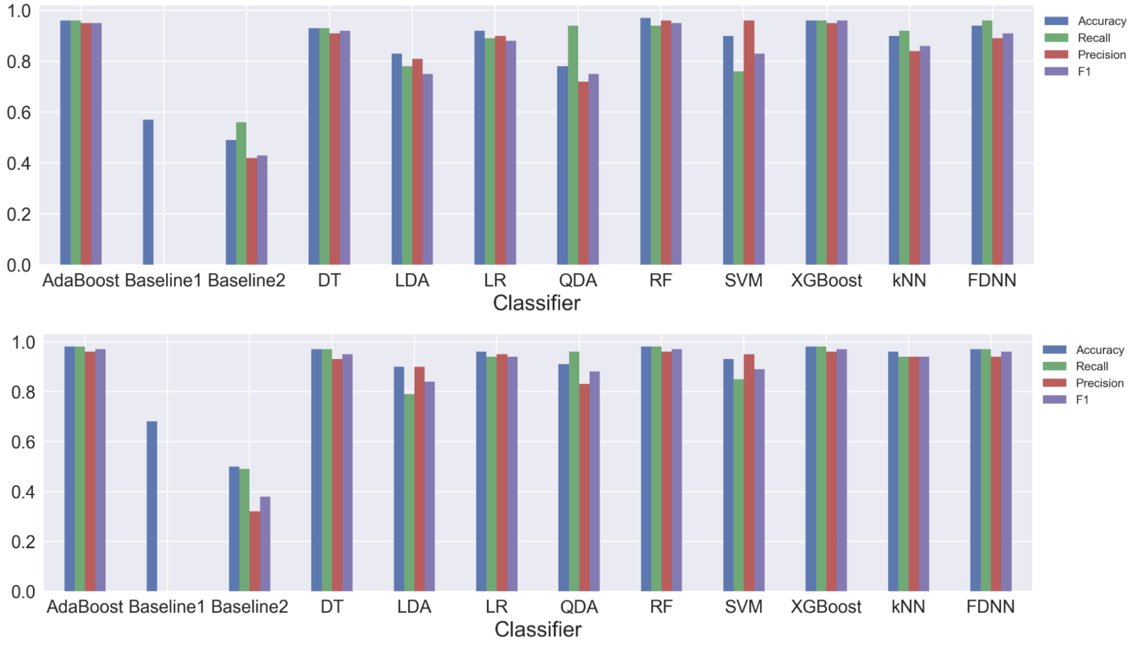

Fig. 1. Accuracy, recall, precision and F1 for all classifiers considered in this work using as input the statistical, SCR and wavelets features. The reported values refer to the mean of the metrics obtained using the **LOSO** (top) and **LODO** (bottom) validation procedures. `Baseline1` and `Baseline2` are the baseline classifiers with different strategies. To obtain the exact values of all the metrics we also include Table 9 and Table 10 in Appendix B.

of these results. Overall the classification results using LODO are slightly higher than when using the LOSO validation procedure. In particular, the XGBoost classifier has an accuracy of 98%, F1 of 97%, precision of 96% and recall of 98%, which is 1 or 2 percentage points increment from the classification results using the LOSO validation procedure. The increment when using LODO validation procedure is expected because the data from the same participant is present in both train and test sets, which simplifies the problem for the classifiers. In practice, this protocol may not scale because it requires annotated information for each new participant [28]. However in the presence of partially annotated data from a user, these results show that our approach generalizes well to the data from a new day, confirming the robustness of our approach over time.

Given that XGBoost and FDNN have a comparable performance in the validation set, in terms of recall, we further present the results using only XGBoost classifier because not only it identifies shape artifacts in the validation set well (recall 98%), but also miss-classifies only 2% of clean segments as artifacts (precision 98%). We present the rest of the results using the XGBoost with optimal parameters (`gamma=1`, `learning_rate=0.1`, `maximum_depth=4`, `number_of_estimators=100`).

5.2 Performance of Our Approach in Comparison to Taylor et al. [61]

In this section we compare our approach with Taylor et al.'s [61] approach, to which we refer as EDAExplorer.

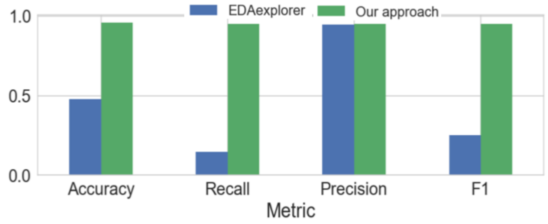

Fig. 2. Comparison between the performance of our approach (using XGBoost) and the EDAExplorer$_{Model}$ presented in [61] when classifying our data set. The results show that Taylor et al.'s approach does not generalize well to data collected in ambulatory settings.

Table 3. Comparison between the EDAExplorer$_{Pipeline}$ [61], EDAQA$_{Pipeline}$ [36] and our approach. Accuracy (A), F1, recall (R) and Cohen κ (κ) are statistically significant according to t-test with $*p<0.01$.

	EDAQA[36]	**EDAExplorer[61]**	**Our**
A	46%	95% (0.03)	**97% (0.04)***
F1	28%	93% (0.04)	**97% (0.02)***
P	76%	96% (0.05)	96% (0.05)
R	17%	90% (0.07)	**98% (0.02)***
κ	0.07	0.87% (0.09)	**0.93 (0.10)***

5.2.1 EDAExplorer$_{Model}$. We first test the EDAExplorer model robustness and generalization to other data sets. To do this, we use the publicly available EDAExplorer source code[9] to classify the segments of our Subset 1. We do not use the web-based tool[10] first because we prefer to avoid uploading sensitive data to other servers. Also EDAExplorer does not support uploading data above a certain size. Figure 2 presents the classification results for both the EDAexplorer and our approach using the XGBoost classifier. Classification results for EDAExplorer are accuracy 47%, recall 14%, precision 94%, F1 25%. The performance of XGBoost instead is accuracy 97%, recall 98%, precision 96% and F1 97%. This shows that the XGBoost achieves a significantly higher performance in comparison to the EDAexplorer for all the metrics. This is rather expected considering that EDAExplorer model has been trained with a smaller data set in comparison to ours and also the data is collected in the lab, which is difficult to generalize to the data collected in-the-wild. Indeed, Zhang et al. [70] show that the performance of artifacts detection classifiers trained with lab data decreases by 9% when tested with data collected in-the-wild. This is a tough classification task for the EDAExplorer because the model is being tested in two data sets with different activities, collected in very different settings and with different test subjects. For this reason we perform and report other comparisons in the following section.

5.2.2 EDAExplorer$_{Paper}$. We compare the results of EDAExplorer approach being trained and tested in their data set collected in laboratory settings, with our approach trained and tested with data collected in ambulatory settings. While EDAExplorer train the SVM with data collected in a controlled setting and achieve an accuracy of 95.67%, which is 19.67 percentage points higher than their baseline, we train the XGBoost classifier using data collected in-the-wild and achieve an accuracy of 96%, which is 39 percentage points higher than our baseline (Baseline2). We compare the performance only with the accuracy metric because is the only reported metric in [61]. Our approach achieves higher performance using data collected in unconstrained settings, which might be more difficult to classify given the confounding nature of EDA collected in uncontrolled settings. Further the validation procedure of EDAexplorer randomly splits the time-series data into train, validation and test set, meaning that similar data from the same user might be present both in train and test set, making the classification task easier and the approach less generizable to the data of unseen users, as discussed in [25, 51].

5.2.3 EDAExplorer$_{Pipeline}$. We then investigate the EDAExplorer pipeline-level robustness by re-implementing the whole approach described in [61] and train and test it using our data set. To do this, we first extract the features from the raw, filtered EDA signals and wavelet coefficients. We then use their best features as input to the SVM classifier to classify our data set. Since several details are missing from their approach, we make

[9] https://github.com/MITMediaLabAffectiveComputing/eda-explorer/blob/master/EDA-Artifact-Detection-Script.py
[10] https://eda-explorer.media.mit.edu/

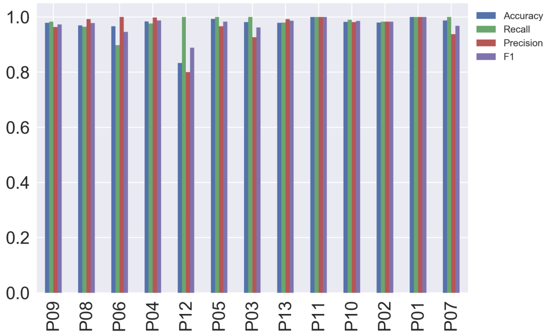

Fig. 3. Accuracy, recall, precision and F1 for XGBoost classifier using statistical, SCR and wavelets features as input and LOSO evaluation procedure.

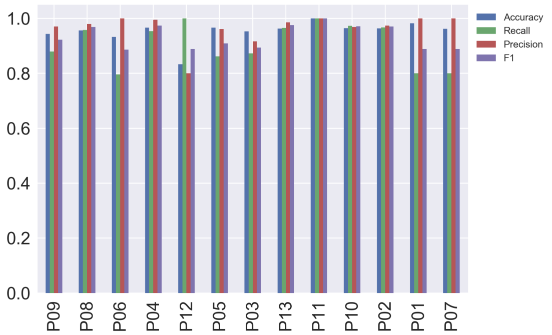

Fig. 4. Accuracy, recall, precision and F1 for SVM classifier using the same features as in Taylor et al. [61] and LOSO evaluation procedure.

the following assumptions: we filter the signals using a first order Butterworth low-pass filter with a cut-off frequency of 0.6 Hz, as we do in our approach. We then extract the wavelet coefficients from participant's raw signal since it is not clear whether this is done in the raw or filtered EDA signal. For both the approaches we use the classifier with the best performing parameters. In particular, for $\texttt{EDAExplorer}_{Pipeline}$, we use SVM with a radial basis function (RBF) kernel, β=0.1, and C=1000 and for our approach we use the XGBoost with the following parameters: γ=1, learning rate=0.1, maximum depth=4 and number of estimators=100. We use LOSO validation procedure to evaluate the generizability of their pipeline to new users and compare it to our pipeline. We perform this analysis using Subset 1 because it contains data from multiple users. Table 3 presents the classification results for the two pipelines. The increment for the $\texttt{EDAExplorer}_{Pipeline}$ performance shows that it is crucial to train the model using in-the-wild data, as we do in our approach, to ensure robustness and generizability to different settings and users. We can further see that our approach can identify artifacts with 8 percentage points higher than the $\texttt{EDAExplorer}_{Pipeline}$. The performance of our approach is higher also for the accuracy, F1 and Cohen κ. The results are statistically significant for recall, accuracy, F1 and Cohen κ according to t-test with $\texttt{p < 0.01}$. Figures 3 and 4 show the performance of our approach (left) and $\texttt{EDAExplorer}_{Pipeline}$ (right) by user. The standard deviation of the performance obtained across different LOSO folds, in terms of recall and F1, is slightly higher for $\texttt{EDAExplorer}_{Pipeline}$ than the corresponding standard deviation obtained using our approach. The instability of the performance of $\texttt{EDAExplorer}_{Pipeline}$ can also be observed in Figures 3 and 4. This implies that our approach is more robust to identify shape artifacts in EDA signals of new users. We believe that the higher performance for our approach in comparison to $\texttt{EDAExplorer}_{Pipeline}$ is first because we normalize the features to the same range of [0, 1], as suggested in [7, 17, 21, 44], to avoid the differences among participants, which might increase classifier's speed of learning [7]. We then use the XGBoost classifier, which is among the most powerful machine learning algorithms [7]. We extract the phasic component from the EDA signal, as a common practice in EDA analysis [3, 12, 16], which allows us to obtain further information from the signal.

5.3 Performance of Our Approach in Comparison to Kleckner et al. [36]

We compare the performance of our approach with the approach presented in [36], to which we refer as EDAQA.

5.3.1 $\texttt{EDAQA}_{Paper}$. We first compare the results of EDAQA approach in their data set collected in ambulatory settings.To allow this comparison we use the common evaluation metrics for the two approaches, namely accuracy, Cohen κ, and recall (or sensitivity). The $\texttt{EDAQA}_{Paper}$ recall is 91%, which 7 percentage points less than our recall.

This implies that we can identify shape artifacts better than $\texttt{EDAQA}_{Paper}$. The classifications results by XGBoost agree with human annotators with an accuracy of 97% and Cohen κ of 0.93, which is higher than the $\texttt{EDAQA}_{Paper}$, which has an accuracy of 92% and Cohen κ of 0.73. The higher agreement between our model classifications and human annotators shows that our automatic approach performs slightly better than $\texttt{EDAQA}_{Paper}$ in replicating decisions of human annotators. This implies that we could use our approach to help researchers visually inspect the signal by highlighting only the parts that need to be inspected, or to automatically recognize them by embedding our approach in their EDA analysis pipeline. To be able to evaluate the performance of $\texttt{EDAQA}_{Paper}$ in our data set, we re-implement their approach and report the results in the following section.

5.3.2 $\texttt{EDAQA}_{Pipeline}$. We re-implement the rules presented in [36], which are publicly available[11]. In particular, for the first rule we set the range of EDA values to [0.01, 100], which is the allowed range for E4 device [18] used in our study. We then implement the second rule – *EDA changes faster than ±10 S/sec.* – by calculating the slope of EDA signal on a point-by-point basis. We do not observe any case in our data set that satisfies this rule. This is because the rule considers only the extreme cases when EDA signal drops/rises significantly, as shown in Kleckner et al.'s Figure 3 (a, b, c, and d), which are not very common in general and might be specific to their context and users. After personal communication with the authors of [36], we were advised to adjust the rule based on visual inspection of the data. We thus first look at the EDA slope values in our data set, which ranges from -0.55 to 0.35. We then manually locate the abrupt peaks/drops in EDA and set the slope threshold to ±0.1 S/sec. We consider a 5-min segment as shape artifact if any of the rules is satisfied and clean otherwise. Table 3 shows the accuracy, F1, precision, recall and Cohen κ for $\texttt{EDAQA}_{Pipeline}$. We can observe that the $\texttt{EDAQA}_{Pipeline}$ can identify only 17% of the artifacts present in our data set. This is first because the $\texttt{EDAQA}_{Pipeline}$ rules capture only artifacts related to the overall shape of the EDA signal such as e.g., EDA is out of range, or there is an abrupt drop or rise in the EDA signal. Our systematic literature review, discussed in Section 6, shows that there are also other types of artifacts related to the EDA peaks such as e.g., peak drops/rises quickly, or when peaks are too close to each other. Further, this approach needs to be adjusted based on the recording device, study design and context, by visually inspecting the signals, as we did when re-implementing the rules, which might not scale to larger studies with multiple participants where each participant has their own EDA baseline.

6 ESTABLISHING GROUND TRUTH FOR ELECTRODERMAL ACITIVITY SHAPE ARTIFACTS

To validate shape artifacts detection framework, we establish a ground-truth for what shape of EDA signal is considered as artifact and what as clean. We do this in a systematic approach: we first perform a literature review of existing guidelines to label shape artifacts in EDA signals and then modify existing or add new guidelines based on knowledge from the literature. We follow this approach because there is no standard manual labeling procedure to label artifacts and using only the existing guidelines, e.g., in [36, 61, 70] is not enough to capture all shape artifacts present in ambulatory EDA signals. We then develop a dashboard to allow human annotators visualize and annotate EDA signals. We provide the guidelines, dashboard and the two subsets of our data set to two human coders to establish ground-truth labels. In this section we explain each of these steps in detail.

6.1 Method

We conduct a systematic literature review by following the guidelines presented in [46] as well as other papers, e.g., in [47, 50, 51], to identify articles that deal with identification and removal of artifacts from EDA data. We use Google Scholar and search for articles that contain at least one of the following keywords in their title: "electrodermal activity", "artifact", "artifacts" or "quality". These search criteria resulted in the following Google Scholar search query: `(intitle:electrodermal activity AND (intitle:artifact OR intitle:artifacts`

[11] https://github.com/iankleckner/EDAQA

Table 4. Overview of the existing and our shape artifacts labeling guidelines. A visual representation of our labeling guidelines is presented in Figure 5. Note that the EDA peaks specifications are representative of healthy young adults and hence, the labeling guidelines related to peaks are extracted from this specific population. Whereas, the EDA-related guidelines, used also in [36, 61, 70] are general to the whole population.

Literature	**Our approach**
1. EDA is out of range EDA not within 0.05-60 S [36] EDA tonic level ≤ 0[61] *2. EDA changes too quickly* EDA changes faster than ±10 S/sec. [36]. A sudden drop of more than 0.1 S in SC [70].	*1.1 EDA is out of range* When EDA signal value is out of range, i.e., not within 0.01-100S[12] *1.2. EDA abrupt rise* When there is no peak in the EDA signal and EDA rises more than 0.1 S in less than 1 second. *1.3. EDA abrupt drop* When there is no peak in the EDA signal and EDA drops more than 0.1 S in less than 1 second.
3. EDA peak decay A peak that does not have an exponential decay, except when two peaks are very close to each other in a short time period so that the decay of the first peak is interrupted by the second peak [61, 70].	*2.1. Peak drops quickly* EDA peak has a sudden drop of more than 0.1 S and the half decay time is less than 2 seconds, except when there are overlapping peaks. *2.2. Peak rises quickly* EDA peak has a sudden increase of more than 1 S in less than 1 second. *2.3. Peaks too close to each other* When there are 3 or more non-overlapping peaks within a 5 second segment.

`OR intitle:quality))`. We perform the search through January 2019. We restrict our search only in the article title because including also the article body returns papers that have discussed the challenge of dealing with EDA artifacts, but do not actually suggest any approach to remove artifacts. From the results returned by this query, we select for review articles that were published in English and exclude those that were published before 1980 and were not peer reviewed such as, e.g., thesis. Using the keywords mentioned above, a total of 14 manuscripts were returned by Google Scholar. We then perform forward and reverse citation to find the articles referenced in the text of obtained articles and those that cite the obtained articles. Following this procedure, we identify and analyze 17 studies that met our eligibility criteria. Only three of the reviewed articles provided a set of instructions to human annotators to identify shape artifacts. One of the authors then extracts labeling instructions from the final list of articles and compiles guidelines in Figure 5, presented also in Table 4 and adds new guidelines based on the knowledge from the literature. The guidelines were reviewed and discussed with three other authors. We then provide a small portion of the data set to two human annotators for labeling in order to verify whether the guidelines are easy to interpret and adapt the guidelines description accordingly.

6.2 Labeling Guidelines

Table 4 presents the list of our and existing guidelines from the literature. Each 5-second EDA segment is labeled as clean or artifact based on guidelines presented in this table. We first define two groups for the rules: (1) rules related to the overall EDA signal and (2) rules related to the EDA peaks. The rules in the second group are only valid when there are peaks in EDA signal. To identify cases when electrodes loose contact with the skin or circuit is overloaded, we add rule 1.1. *EDA is out of range*, which has been suggested in [36]. We adjust this rule to the minimum and maximum value measured by the Empatica E4 device used in this study. To identify abrupt changes in EDA signal, usually caused by electrode or body movement, we add rule 1.2 and 1.3, which are similar to the *EDA changes faster than ±10 S/sec* rule used in [36] and the *A sudden drop of more than 0.1 S/sec in EDA* rule used in [70], but Kleckner et al.'s rule accounts for only very huge jumps in the signal, e.g., above ±10 S/sec, and Zhang et al.'s rule does not consider the abrupt rise kind of artifacts. We then add rules specific to EDA peaks, to identify the corrupted peaks. To do this we consider the EDA peaks specifications of the healthy young adults [3, 12], which specify that a valid EDA peak has an amplitude in range from 0.1 to 1 S, a rise time from 1 to 3 seconds and a half recovery time from 2 to 10 seconds. Considering these specifications, whenever a peak has a drop of more than 0.1 S in less than 2 seconds (rule 2.1), or has an increase of more than 1 S in less than 1 second (rule 2.2), or there are 3 or more non-overlapping peaks within a 5-second segment (rule 2.3) is considered as a shape artifact. Authors in [61, 70] also consider the EDA peak decay among the instructions to label artifacts, however, their rule is primarily based on visual interpretation rather than the actual time and amplitude of the EDA peaks. We ask the human annotators to consider these rules whenever the EDA change within a segment is more than 0.1S because the minimum amplitude of a SCR can be 0.1S as suggested in [12, 59].

6.3 Labeling Dashboard

To allow human annotators visualize and label EDA signals, we design and implement a dashboard, called *EDArtifact*. We build the dashboard based on similar existing tools in [36, 61], and provide further functionalities to make the labeling process and the visual inspection of long EDA traces more efficient. This is because existing tools, such as e.g., EDAExplorer [61] require to traverse sequentially through all 5-second segments, which is not feasible for traces collected for several hours or days. To overcome these limitations, we first show the long EDA traces in different granularity levels, e.g., whole signal, 10 minute windows and 15 seconds, and allow human annotators to easily navigate through the signal. We also provide the possibility to upload and visualize the accelerometer sensor data but we do not use it to visually identify shape artifacts because the heuristics we developed, presented in Table 5, already capture all aspects of shape artifacts. The dashboard could be easily extended to visualize other contextual information such as, e.g., temperature, physical activities, add other labels and more. We design the dashboard by trying to avoid the common pitfalls in dashboard design, presented by Few [15]. For instance we highlight data only when necessary, i.e., to show the label type or the labeling progress, and we use only a few colors to distinguish different label types such as e.g., red for artifact, yellow for unsure, green for clean. Given the significant amount of time needed to label long EDA traces, we add features to the dashboard to make this process smoother by adding a button to mark all segments in a 10-minute window as clean or artifact and providing zoom in/out functionality to easily inspect the SCRs. We make the dashboard available [13] to other researchers for inspecting long EDA traces. We develop *EDArtifact* using the Python framework Dash [14], built on top of Flask, Plotly.js and React.js. We also provide instructions on how to install the dashboard locally.

[13] The dashboard and installation instructions are available in the following link https://github.com/shkurtagashi/EDArtifact.

[14] https://dash.plot.ly/

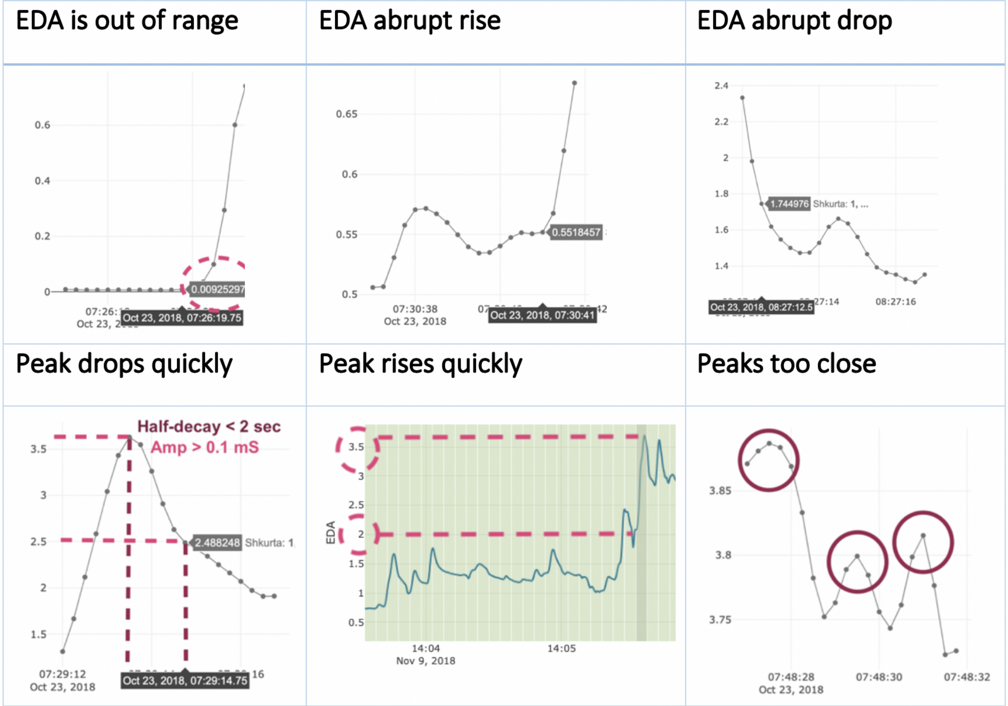

Fig. 5. Guidelines developed to aid human annotators in the data labeling process. Each of the guidelines provides an example of possible shape artifacts within a 5-second segment. The examples are related to the EDA signal in general or more specifically to the EDA peaks.

6.4 Human Annotators

We then provide the data to two human coders, who were not involved in the project, to inspect and mark shape artifacts in the data. Human annotators reviewed all the data independently and provided manual ratings of all segments. We assign two labels artifact or clean, as suggested in [61, 61, 66], to each 5-second EDA segment. We label the segment as *artifact* – when any of the rules presented in Section 6.2 is satisfied and as *clean* – if none of the rules are satisfied. We further provide the *unsure* option to allow the annotators express the cases when they are unsure of the label. We however recommended the experts to use this label only in exceptional cases. We then revise unsure labels manually and assign them to a class. We also provide nine other sub-labels that represent the shape artifact types, presented in Figure 5. In this study we focus on identification of shape artifacts from clean segments, while in future work we plan to investigate the identification of specific artifact types.

To measure the agreement between human annotators, we use the percentage agreement and the Cohen κ score as in [36, 61]. Both the metrics measure the total amount of common labels between the annotators, except that Cohen κ considers whether the agreement is due to chance [11, 36, 61]. Table 5 shows the percentage agreement and Cohen κ of the two human coders for each subset and the overall data set before and after flat responses removal. The average agreement of the two human annotators for the whole data set is 96.62% and Cohen κ 0.84, which is higher than in [61] and comparable to Kleckner et al. [36], even if the latter consider a more specific subset of artifacts. Given that the agreement between the two human annotators is high, we

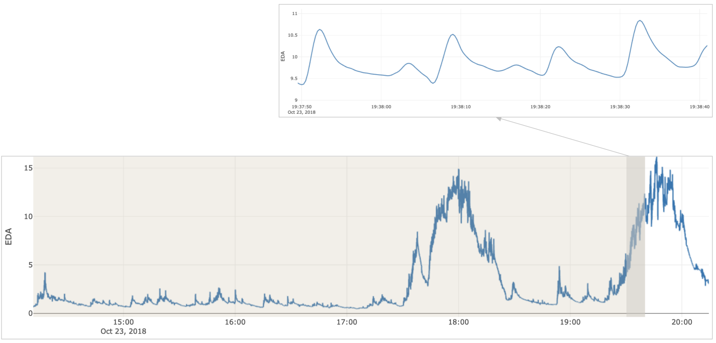

Fig. 6. Example of thermoregulation artifacts in ambulatory EDA signal. The bottom picture shows the EDA signal over the day. The top picture shows a segment of EDA signal that contains several SCRs. Based on the specifications of the SCR these peaks are valid, however looking at the acceleration sensor data we understand that the physical activity level of user is high in this window, hence increasing the possibility of having more responses due to the thermoregulation process rather than other stimuli.

Table 5. Agreement, in terms of the percentage % of cases the two human annotators agree and the Cohen's κ metric, for Subset 1, Subset 2 and the two combined. The left columns show the agreement on the whole data set, whereas, the right ones agreement after removing flat responses.

	Agreement		**Agreement**	
Data	%	κ	%	κ
Subset 1	94.92%	0.82	89.81%	0.78
Subset 2	98.01%	0.85	90.96%	0.79
All	96.62%	0.84	90.35%	0.80

consider only the segments where the two human coders agree as ground-truth to evaluate the models. We discard the segments where annotators disagree because we cannot establish a ground truth for them, thereby, we cannot assess the performance of the classifier, as also discussed in [61].

7 ELECTRODERMAL ACTIVITY QUALITY INDEX

In this section we explain our approach on first quantifying the impact of shape and thermoregulation responses, and then defining the EDA quality metric that quantifies the overall quality of the EDA signals. We perform this part of the analysis on the whole data set – both Subset 1 and Subset 2 – described in Section 3.3.

7.1 Quantifying the Impact of Shape Artifacts and Thermoregulation Responses in Electrodermal Activity Signals

We perform a statistical analysis to understand the impact of the presence of shape and thermoregulation responses in EDA measurements. To this goal we first identify a set of most common features used in EDA research such as e.g., mean, standard deviation, dynamic range, number of peaks, peak's amplitude, rise time, half-recovery or decay time, width and area under the curve, which were used in e.g., [13, 14, 20, 22, 23, 26, 28, 29, 53, 56, 58, 61, 69]. We then divide the EDA traces into 5-minute segments and we compute each of the identified features in these segments. We assume a real-time system processes EDA signals every 5-minutes, as in [26], however this could easily be adapted to other window lengths. We then divide the features in two groups described in the following sections and we compare the difference between them using the Kolmogorov-Smirnov test, as in [14].

7.1.1 Shape Artifacts. To understand the impact of shape artifacts, we create two groups of the same data, one containing the shape artifacts and the other removing them based on the labels of two human annotators. We remove shape artifacts by setting the EDA level at the specific segments to a missing value. We extract the features identified from the literature in 5-minute windows before and after artifacts removal.

7.1.2 User's Movement. To understand the impact of movement in EDA, we first identify the parts with high intensity movement using the data from accelerometer (ACC) sensor. To process ACC we follow standard

Table 6. The difference between the EDA features for each factor, namely, shape artifacts, movement and temperature. Features that are statistically significant according to Kolmogorov-Smirnov test are marked with * (`*p<0.05, **p<0.01, ***p<0.001`).

	Shape Artifacts		Movement		Temperature	
Feature	**Yes**	**No**	**Low**	**High**	**Yes**	**No**
`mean`	1.05	0.32^{***}	1.05	1.05	1.02	1.81
`std`	0.21	0.03^{***}	0.23	0.21	0.19	0.68^{*}
`drange`	1.08	0.17^{***}	1.10	1.08	1.01	2.91^{***}
`peaks`	11.42	8.20^{***}	15.27	11.2	11.25	15.57^{***}
`rise`	1.89	1.96^{***}	1.76	1.90^{*}	1.89	1.85
`amp`	0.11	0.05^{***}	0.10	0.11	0.11	0.24^{***}
`decay`	2.09	2.12	1.99	2.10^{*}	2.11	1.821
`width`	2.88	2.91	2.72	2.89	2.90	2.62
`auc`	0.35	0.15^{***}	0.30	0.35	0.34	0.66

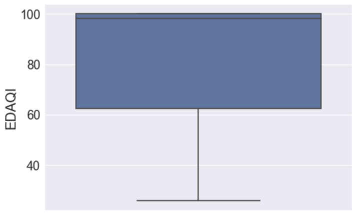

Fig. 7. The distribution of EDA quality index in our data set (`mean=83.20`, `std=20.59`, `min=25.83`, `max=100`). In this primary investigation of EDA quality index, we set the quality index weights manually as follows: $w_1 = 50$, $w_2 = 25$, and $w_3 = 25$. We plan to investigate the impact of different weights in our future work.

procedures as in [6, 49, 64]. The E4 ACC sensor is configured to measure the acceleration in range [-2g, 2g] – where g is the gravitational force – with 1/64g unit of acceleration. As a first processing step we convert the unit to g by dividing with 64, as suggested in [6, 64]. Given that the sampling frequency of the ACC sensor measured with E4 is 32Hz, we re-sample the acceleration data to the same sampling frequency as EDA sensor (4Hz), as suggested in [6]. We then calculate the vector magnitude of acceleration using the Euclidean Norm Minus One (ENMO), suggested in [64], which in formula corresponds to $max(\sqrt{x^2 + y^2 + z^2}, 0)$ with x, y and z referring to the three orthogonal acceleration axes. We compute the orientation angles of each acceleration axis as suggested in [64]. We split the acceleration magnitude and orientation to 5-minute segments, similar to EDA, and assign each 5-minute EDA segment in one of the four movement categories defined in [64], namely, sustained inactivity, inactivity, light physical activity and vigorous physical activity. Since research has shown that moderate or light physical activity has a small effect of EDA measurement [33], we group the EDA segments with vigorous physical activity into the high movement group and the other three categories in the low movement group.

7.1.3 *Temperature Change.*

To identify the parts in EDA signal measured during environment temperature change, we use the data from the skin temperature sensor. We first divide the skin temperature sensor data into 5-minute segments. We then compute the slope of temperature in each segment to understand whether the temperature is constant or not. We consider a temperature change when the skin temperature slope is not in range [-0.001, 0.001], which has been set empirically by visualising temperature sensor data. Based on the temperature slope, we divide the 5-minute EDA segments in two groups: constant and non-constant temperature and compute the most common EDA features.

7.1.4 *Impact of Shape Artifacts and Thermoregulation Responses.*

Table 6 shows the difference between the features in the two groups. We can observe that some of the features used commonly in literature are significantly different before and after shape artifacts removal, during low and high intensity movement, and during constant and non-constant temperature change. In particular the number of peaks in a 5-minute window is significantly lower after shape artifacts removal hinting at the fact that some of these peaks are corrupted. To investigate this further, we verify whether the difference in the number of peaks before and after artifacts removal is significantly related to the number of shape artifacts. Figure 8 shows that this correlation is significant based on Spearman rank correlation (`r=0.45 and p<0.001`). User's physical activity intensity movement also impacts the difference

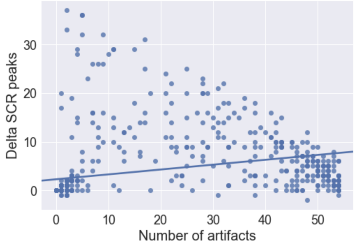

Fig. 8. Relationship between the difference in the number of peaks – before and after artifacts removal – and the number of artifacts in a 5-minute window. The relationship is significant according to Spearman rank correlation ($r=0.45$ and $p<0.001$).

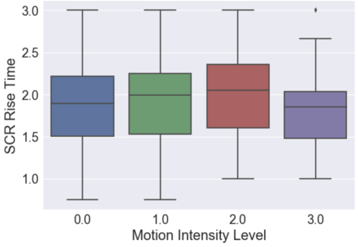

Fig. 9. Distribution of SCR rise time in 5-minute EDA segments for each movement intensity level. The difference between low (level 0, 1 and 2) and high intensity movement (level 3) is statistically significant according to Kolmogorov-Smirnov test ($p < 0.05$).

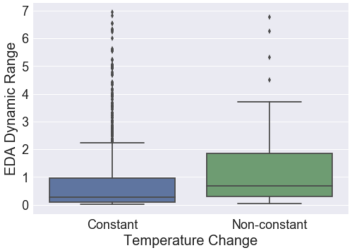

Fig. 10. Distribution of EDA dynamic range for each temperature condition (constant vs non-constant). The difference between constant and non-constant is statistically significant according to Kolmogorov-Smirnov test ($p < 0.001$).

between the features. Figure 9 for instance presents the distribution of SCR rise time for each motion intensity level of the user. The distribution of SCR rise time for motion intensity level 3.0 in comparison to the other levels is significantly different according to Kolmogorov-Smirnov test ($p < 0.01$). Figure 10 shows the distribution of the dynamic range feature in EDA segments with constant and non-constant temperature. We can observe that the dynamic range for parts of EDA signal where the temperature is constant is significantly lower – according to Kolmogorov-Smirnov test ($p < 0.001$) – than for non-constant temperature. Overall these results confirm that shape and thermoregulation responses have a significant impact in the EDA measurements and should be removed or considered before further EDA analysis.

7.2 EDA Quality Index

We then use the input from shape artifacts detection framework and thermoregulation responses to define the EDA quality index. In this section we explain the EDA quality index in more details. First the quality of EDA signal may be impacted by the amount of *shape artifacts (SA)* in the window. For instance the higher the amount of artifacts in a window the lower the quality. We formalize this as the ratio of the number of artifacts in the 5-minute window (n) and the maximum number of artifacts in the window (m). The value of (m) in a 5-minute window segmentation is 60.

$$SA = \begin{cases} \frac{n}{m} & \text{where m is the maximum amount of SA in the window and n is the number of SA in m} \\ 0 & \text{otherwise} \end{cases}$$

We then consider that a *thermoregulation responses (TR)* might happen when the user's movement intensity is high or when the temperature is not constant. We formalize this in terms of the movement level and the temperature change in the 5-minute window.

$$\text{TR}_M = \begin{cases} 1 & \text{if movement level is 3 (vigorous movement)} \\ 0 & \text{otherwise} \end{cases} \qquad \text{TR}_T = \begin{cases} 1 & \text{if temperature is not constant} \\ 0 & \text{otherwise} \end{cases}$$

We combine the three factors into a metric EDA_{QI} using Equation 1 and we propose to adjust the EDA data quality based on them.

$$EDA_{QI} = 100 - (w_1 SA + w_2 TR_M + w_3 TR_T) \tag{1}$$

The sum of the weights should be equal to 100, e.g., $w_1 + w_2 + w_3 = 100$. This means that if all the segments in a 5-minute window contain shape artifacts, if the 5-minute window of EDA is collected during vigorous movement and while temperature change the quality of the EDA signal should be very low or equal to 0.

We implement the EDA quality index and estimate the quality of our data set based on this metric. We first count the number of shape artifacts – as identified by two human annotators – in a 5-minute window. We also calculate the movement level and temperature change and adjust the EDA quality index based on Equation 1. Considering that shape artifacts have a higher impact in EDA features, as shown in Table 2, we set the weight of the shape artifacts factor higher than for the movement level and temperature change factors, for instance, $w_1 = 50$, $w_2 = 25$, and $w_3 = 25$,. Other researchers might set these weights based on their application of interest. For instance, if the portions of interest of EDA are during a lot of physical movements such as, e.g., seizure detection [48], the movement level factor could be ignored $w_3 = 0$ and the other two factors could be given more importance. Figure 7 shows the distribution of EDA quality index for our data set. On average the data set has an EDA quality index of 83.20, with a standard deviation of 20.59. The majority of the data seem to have a high quality index (equal or above 83.20), which is expected because of the presence of flat responses i.e., when the user does not show any physiological reaction. However the minimum EDA quality index goes up to 25.83, which is critical if that part of the EDA data is of interest. In this study we focus on investigating the need for an EDA quality index and its definition, while we will investigate the impact of the weights for different use cases and the accepted quality index threshold in our future work.

8 IMPLICATIONS

In this work we show that we can automatically detect shape artifacts in EDA signals with 98%, which is 42 percentage points increment from a baseline classifier. We evaluate our approach on a data set collected in ambulatory settings – where participants perform completely unconstrained activities – and labeled by two human annotators. We present a dashboard for labeling long EDA traces, which we make publicly available. We propose a novel metric that considers the impact of not only shape artifacts, but also environmental temperature and user's physical activity in the overall EDA quality.

Our shape artifacts detection framework might be integrated in offline and online systems that monitor EDA. In particular offline detection of shape artifacts could first eliminate the need for researchers to visually inspect long traces of EDA signals collected in-the-wild, by for instance embedding our framework in their EDA processing pipeline to automatically recognize shape artifacts. This would first significantly reduce the amount of time required to analyze the vast amount of EDA data being collected in-the-wild studies. This is because manual labeling of data collected over days, weeks or even months is very time consuming for human annotators. Further manual human annotation is not feasible in real-time systems.

Researchers would then be able to improve their detection tasks by using our EDA quality index to identify parts of the EDA signal with significant physiological responses. This would make the signal more descriptive of the actual experience of the user and prevent unreliable or spurious conclusions from the analysis. Researchers will further be able to automatically assess the EDA quality in a timely manner and spot issues early on when conducting long term field studies. We anticipate that our approach can also be smoothly integrated in current or future consumer wearable devices that measure EDA. This would allow to identify unreliable parts of the signal and to prevent the need for transferring low quality data to the main server used for processing EDA. Further it would inform the EDA-based intervention systems to not act in inopportune moments or with inappropriate content. Overall, our approach will help to advance studies that utilize ambulatory EDA measures, including, but not limited to studies assessing stress [22], mood [62], engagement [13, 19, 20], or emotions [26].

9 LIMITATIONS & FUTURE WORK

While this work shows promising results in the evaluation of EDA quality, further research is needed to overcome the limitations of our work. In this section we discuss the limitations related to the EDA shape artifacts detection and the overall EDA quality, which leave room for future work.

9.1 Electrodermal Activity Shape Artifacts Detection

The first limitation stems from the fixed window segmentation of signals, which may miss the presence of artifacts that start in one window and end in the other. To avoid this limitation we add new label options, e.g., invalid left/right, in the dashboard to allow human annotators to identify these cases. We however did not analyze this data because we do not expect the types of artifacts in those segments to be significantly different from artifacts already present in our data set and they do not present a large group of cases since the total amount of invalid left/right is 180 from the total of 77443 in both Subset 1 and Subset 2. Further a segment may contain both a clean SCR and an artifact and with our approach is not possible to detect them individually. In our future work we plan to investigate customized segmentation windows and allow the human annotators to select only the segments of the signal – with variable length – that need a label.

We model the problem of distinguishing between shape artifacts and clean segments as a binary classification problem, neglecting the type of shape artifact present in signal. In future work we plan to investigate the use of ensemble methods to train models per type of shape artifact and use a voting mechanism to provide a decision about the state of the segment.

Our shape artifacts detection framework does not capture the temporal and sequential nature of the data. In future work we plan to model the temporal aspect of EDA signal by investigating the use of Long Short-term Memory (LSTM) neural networks, which might more effectively capture representations of the EDA signal.

9.2 Electrodermal Activity Quality Quantification

The first limitation on our EDA quality estimation approach stems from considering several sensors, which might not be available in all real world scenarios. The latest wearable devices however seem to have this desirable property by including multiple sensors in one device. Indeed when measuring EDA we should measure both user's physical activity and environment conditions in order to have the sound physiological reactions [16].

To estimate the EDA quality we consider only the shape and the thermoregulation responses. Literature shows that artifacts in EDA signals might be caused also by other environmental conditions such as e.g., air humidity, or respiration pattern changes e.g., sneezing, deep sights or coughing [3, 4, 16, 36, 57]. Future work should consider the data from other modalities – i.e., respiration sensors to understand when such respiration pattern changes happen or environmental sensors that measure the humidity – to quantify the presence of these artifacts. We however expect that these artifacts are less common than the ones caused by user's physical activity and environment temperature changes, as also suggested in [16, 36, 61, 70].

Another limitation of our approach is that we do not show the impact of the EDA quality index. In our future work we plan to evaluate its impact in a broader EDA-based system such as e.g., stress detection system, by understanding the relative importance of keeping or discarding the data based on the quality index. We will for instance consider the incorrect predictions of the system and understand whether they are due to the quality of the collected data. The ultimate goal of EDA quality index will be to define an appropriate threshold to consider EDA data for further analysis, as for instance the BioPatch device provides confidence levels for heart rate measurements [27].

10 RELATED WORK

In this section we review and summarize the related literature on detection of shape artifacts in EDA signals and other techniques for overall EDA quality quantification.

10.1 Electrodermal Activity Shape Artifacts Detection

The standard textbook for EDA data collection and analysis suggest to identify and remove artifacts through visual inspection [3]. This technique cannot scale to the long-term analysis of EDA signals, as e.g., in [25, 35, 36, 47], which may involve data from several participants over days and months. Additionally, it may be challenging to identify artifacts due to their high similarity to the genuine physiological responses and it may also be significantly biased by the individual who performs it. Other standard procedures in the literature to deal with artifacts in EDA include exponential smoothing (e.g., in [21]) and low-pass filtering (e.g., in [29]). These approaches only smooth out low amplitude artifacts that are more common in laboratory settings, but are less effective on higher magnitude responses which are more common in natural settings. They pass the entire EDA trace through the filter, which may distort the signal and result on removal of physiological responses [52] or modify artifacts resemble to physiological shapes [27].

Several researchers have addressed the problem of recognizing shape artifacts in EDA signals and have investigated automatic methods to identify them [35, 36, 59, 61, 66, 70]. Taylor et al. [61], for instance, propose a supervised learning approach to distinguish between shape artifacts and clean parts of the signal. This approach has been tested using data of users when performing a set of predefined tasks in a controlled environment. Further they claim that if shape artifacts remain in the signal when it is analyzed they can be misinterpreted and skew the analysis. In line with their research we provide statistical evidence that presence of shape artifacts impacts the EDA features. Kleckner et al. [36] propose a rule-based approach to identify artifacts in EDA. To test their approach they collect data at home, from a population of children with autism spectrum disorder, and for only 1.5 hours per day. Further they define four rules which might not effectively distinguish valid SCRs and invalid SCR-like artifacts. With respect to our work, these approaches might be difficult to generalize to new contexts or even new users.

Only a few researchers considered testing their approach using data from real world settings [35, 70]. Kesley et al. [35] suggest using curve fitting and sparse recovery methods to identify and remove artifacts in EDA data. In contrast to our work, they test their approach on a very small data set with 113 corrupted samples and 264 clean SCRs and use only the phasic component of the EDA signal, which might not capture shape artifacts in the mixed EDA signal, such as e.g., EDA rises quickly. Zhang et al. [70], on the other hand, investigate the use of unsupervised learning algorithms for detection of motion artifacts in EDA signals. Their results show that unsupervised learning algorithms perform competitively to the supervised algorithms presented in [61]. They test their approach on a data set of 10 hours collected from one person only. They focus specifically on motion artifacts, and do not consider other types of artifacts such as e.g., EDA out of range. In comparison to these approaches, we test our approach on the data collected in-the-wild from 13 users and a total of 107.56 hours. Our final data set contains 67318 clean and 8267 artifacts samples. Besides motion artifacts we also consider other types of artifacts related either to the overall EDA signal such as e.g., EDA out of range, or specific to SCRs such as e.g., SCR drops quickly.

In contrast to all the work presented above, where authors investigated the use of simple rule-based algorithms [36], supervised learning [61], semi-supervised learning [66] or unsupervised learning [70] techniques, we build upon their work and investigate the use of ensemble and deep learning techniques, the latter have shown to be efficient in other tasks [49], but have not previously been explored for this problem. Only Zhang et al. [70] investigate multilayer perceptron (MLP) with two hidden layers, however as discussed in [17] this approach can be still thought as a *shallow learning*. We build upon their work and further investigate more complex deep neural networks with five hidden layers and show that they can effectively detect shape artifacts with 98% recall, which is 42 percentage points higher than the baseline classifier. Further some of the existing approaches, e.g., in [61, 66], randomly split the data into train, validation and test sets and do not consider the dependence and similarity between adjacent segments in time-series, as discussed in [25, 51], which might hamper the generalizability of their approach. In contrast we train and test our approach using leave-one-day-out and leave-one-subject-out evaluation procedures to ensure the robustness of our approach over time and to new users.

10.2 Electrodermal Activity Quality Quantification

Publication Recommendations for Electrodermal Measurements [16] suggest to account for movement and environmental factors when measuring EDA in order to keep only data about significant physiological events. Healey et al. [26] for instance remove the data portions when user's physical activity level exceeded strolling to prevent confounding factors in the data analysis. Very few approaches consider these factors to evaluate the quality of EDA signals [36, 70]. Zhang et al. [70] consider the data from accelerometer sensor for identification of shape artifacts. Their results show no significant improvement on detection of shape artifacts using features from accelerometer sensor. We believe this is because movement not only causes shape artifacts, but also generates physiological peaks that resemble genuine physiological responses but are not caused by the electrodermal system itself. Indeed shape artifacts can be effectively recognized using the EDA singal alone, as we have also shown in our work. Kleckner et al. [36] use data from temperature sensor to account for times when EDA sensor is not being worn or has not been worn long enough by looking at the temperature range. We also investigate accelerometer and temperature sensor data to evaluate the quality of EDA signal, however, in comparison to all the work above we suggest to consider not only shape artifacts but also thermoregulation responses that may cause SCRs that resemble physiological responses cause by other systems. In this manner we suggest a novel metric – EDA_{QI} – that considers both the movement and environmental conditions. Indeed Xu et al. [68] show that physiological data collected when participants are at rest achieve the highest accuracy on emotion recognition. They show that user's movement influences the physiological signals, among them EDA, and therefore it influences the results of the emotion recognition system.

11 CONCLUSION

In this paper we present an automatic approach to detect *shape artifacts* and to quantify *thermoregulation responses* in the electrodermal activity signals collected in natural settings over the day. To detect shape artifacts we first propose a systematic approach to manually label the physiological signals and to create gold-standard labels. We then suggest to extract not only features used in the literature, e.g., statistical and wavelets, but also features related to the shape of EDA signal. We use these features as input to deep neural networks, ensemble, linear and non-linear classifiers. We train the classifiers using data from 13 users, with 7729 samples in the artifacts class and 8057 in the clean class. To train the classifiers we create ground-truth labels based on the majority voting of three human annotators. We achieve a recall of 98% using the XGBoost classifiers, which corresponds to an increment of 42 percentage points from the baseline classifier. We test the generizability of our results for instance to a new user and show similar performance. We further suggest to consider also *thermoregulation responses* that may look like actual physiological responses but are caused by other systems, i.e., user's motor or thermorregulation system. In this manner we suggest a novel metric to quantify both the presence of shape and thermoregulation responses for an overall estimation of the EDA signal quality. Overall our findings suggest that we can replace human annotators or significantly reduce their effort to visually inspect the data. Our approach may further be embedded in an online manner for automatic data quality assessment.

A DATA SET

To aid reproducibility, as discussed in [41], we report details about our collected and annotated dataset. Figure 11 shows the amount of data collected by each participant per day. Figure 12 shows the amount of data collected by hour of the day. Table 7 and Table 8 present the annotated sessions by user, day and time of the day for Subset 1 and number of hours per day for Subset 2.

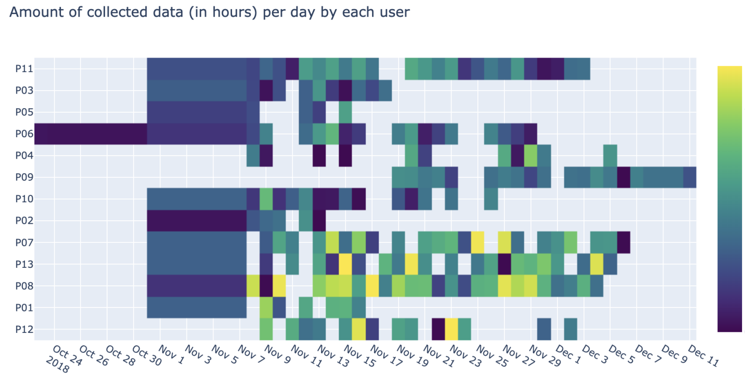

Fig. 11. Overview of the amount of collected data per participant during the period of data collection.

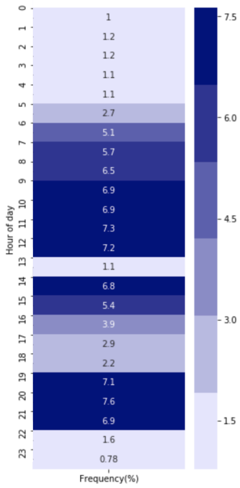

Fig. 12. Overview of the overall amount of collected data by hour.

Table 7. List of annotated sessions in Subset 1.

User	Day	Time of day
P01	8	morning
P02	8	evening
P03	14	evening
P04	4	evening
P05	11	evening
P06	32	evening
P07	11	morning
P08	8	evening
P09	16	morning
P10	11	evening
P11	12	evening
P12	9	morning
P13	10	morning
Total		**52 hours**

Table 8. List of the annotated sessions in Subset 2.

User	Day	Hours
P03	8	8
	12	7
P06	1	15
	16	13
	24	4
P09	32	11
Total	**6 days**	**55.56 hours**

B RESULTS

In this section we report numerical representations of the results presented in Section 5 and additional experiments performed in our data set. We provide details about each experiment in the table description.

Table 9. Accuracy, recall, precision and F1 for all classifiers considered in this work using as input the statistical, SCR and wavelets features. The reported values refer to the mean of the metrics obtained using the **LOSO** validation procedures and **Subset 1**.

Classifier	**Accuracy**$_V$	**F1**$_V$	**Precision**$_V$	**Recall**$_V$	**Accuracy**	**F1**	**Precision**	**Recall**	**Cohen's** κ
AdaBoost	0.97	0.97	0.97	0.97	0.96	0.95	0.95	**0.96**	0.90
Baseline1	0.50	0.00	0.00	0.00	0.57	0.00	0.00	0.00	0.00
Baseline2	0.51	0.51	0.50	0.51	0.49	0.43	0.42	0.56	-0.02
DT	0.96	0.96	0.96	0.97	0.93	0.92	0.91	0.93	0.84
LDA	0.85	0.82	0.96	0.72	0.83	0.75	0.81	0.78	0.61
LR	0.94	0.94	0.99	0.89	0.92	0.88	0.90	0.89	0.81
QDA	0.94	0.93	0.96	0.91	0.78	0.75	0.72	0.94	0.58
RF	0.97	0.97	0.97	0.97	**0.97**	0.95	**0.96**	0.94	**0.92**
SVM	0.89	0.87	0.99	0.78	0.90	0.83	**0.96**	0.76	0.73
XGBoost	0.98	0.98	0.98	0.98	0.96	**0.96**	0.95	**0.96**	0.91
kNN	0.96	0.96	0.97	0.94	0.90	0.86	0.84	0.92	0.74
FDNN	0.97	0.97	0.97	0.98	0.94	0.91	0.89	**0.96**	0.85

Table 10. Accuracy, recall, precision and F1 for all classifiers considered in this work using as input the statistical, SCR and wavelets features. The reported values refer to the mean of the metrics obtained using the **LODO** validation procedures and **Subset 2**.

Classifier	**Accuracy**$_V$	**F1**$_V$	**Precision**$_V$	**Recall**$_V$	**Accuracy**	**F1**	**Precision**	**Recall**	**Cohen's** κ
AdaBoost	0.98	0.98	0.99	0.98	**0.98**	**0.97**	**0.96**	**0.98**	0.95
Baseline1	0.49	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.00
Baseline2	0.51	0.51	0.52	0.50	0.50	0.38	0.32	0.49	-0.01
DT	0.97	0.98	0.98	0.97	0.97	0.95	0.93	0.97	0.92
LDA	0.87	0.86	0.96	0.78	0.90	0.84	0.90	0.79	0.77
LR	0.96	0.96	0.99	0.94	0.96	0.94	0.95	0.94	0.91
QDA	0.95	0.95	0.94	0.96	0.91	0.88	0.83	0.96	0.81
RF	0.98	0.98	0.98	0.98	**0.98**	**0.97**	**0.96**	**0.98**	0.95
SVM	0.92	0.91	0.97	0.86	0.93	0.89	0.95	0.85	0.84
XGBoost	0.98	0.98	0.98	0.98	**0.98**	**0.97**	**0.96**	**0.98**	**0.96**
kNN	0.96	0.96	0.97	0.94	0.96	0.94	0.94	0.94	0.91
FDNN	0.98	0.98	0.98	0.98	0.97	0.96	0.94	0.97	0.94

Table 11. Accuracy, recall, precision and F1 for all classifiers considered in this work using as input the statistical, SCR and wavelets features. The reported values refer to the mean of the metrics obtained using the **LOSO** validation procedures and **Subset 1 and Subset 2**.

Classifier	Accuracy$_V$	F1$_V$	Precision$_V$	Recall$_V$	Accuracy	F1	Precision	Recall	Cohen's κ
AdaBoost	0.98	0.98	0.98	0.98	0.97	0.97	0.96	0.98	0.93
Baseline1	0.50	0.00	0.00	0.00	0.57	0.00	0.00	0.00	0.00
Baseline2	0.50	0.50	0.50	0.49	0.51	0.42	0.43	0.49	0.00
DT	0.97	0.97	0.97	0.97	0.98	0.97	0.97	0.98	0.94
LDA	0.84	0.82	0.95	0.72	0.85	0.77	0.81	0.82	0.64
LR	0.95	0.95	0.98	0.91	0.92	0.88	0.86	0.94	0.80
QDA	0.94	0.94	0.95	0.92	0.77	0.75	0.71	0.96	0.59
RF	0.98	0.98	0.98	0.98	0.98	0.98	**0.98**	0.97	0.96
SVM	0.90	0.89	0.99	0.81	0.90	0.86	0.89	0.88	0.76
XGBoost	0.98	0.98	0.98	0.98	**0.99**	**0.99**	**0.98**	**0.99**	**0.97**
kNN	0.96	0.96	0.98	0.95	0.91	0.86	0.83	0.96	0.78
FDNN	0.98	0.98	0.98	0.98	**0.99**	**0.99**	**0.98**	**0.99**	**0.97**

Table 12. Accuracy, recall, precision and F1 for all classifiers considered in this work using as input the statistical, SCR and wavelets features. The reported values refer to the mean of the metrics obtained using the **LOSO** validation procedures and **Subset 1 and Subset 2, including flat responses**.

Classifier	Accuracy$_V$	F1$_V$	Precision$_V$	Recall$_V$	Accuracy	F1	Precision	Recall	Cohen's κ
AdaBoost	0.99	0.99	0.99	0.99	0.96	0.81	0.78	**0.99**	0.80
Baseline1	0.50	0.00	0.00	0.00	0.84	0.00	0.00	0.00	0.00
Baseline2	0.50	0.50	0.50	0.50	0.50	0.19	0.16	0.51	0.00
DT	0.98	0.98	0.98	0.98	0.97	0.79	0.75	0.98	0.78
LDA	0.87	0.86	0.95	0.79	0.94	0.60	0.59	0.85	0.56
LR	0.97	0.97	0.98	0.96	**0.98**	0.75	0.70	0.97	0.74
QDA	0.96	0.97	0.94	0.99	0.89	0.56	0.49	**0.99**	0.53
RF	0.99	0.99	0.99	0.99	**0.98**	**0.82**	0.78	**0.99**	**0.81**
SVM	0.96	0.96	0.98	0.93	**0.98**	0.76	0.72	0.95	0.75
XGBoost	0.99	0.99	0.99	0.99	**0.98**	**0.82**	**0.79**	**0.99**	**0.81**
kNN	0.98	0.98	0.98	0.98	**0.98**	0.70	0.64	0.98	0.69
FDNN	0.99	0.99	0.99	0.99	**0.98**	**0.82**	**0.79**	**0.99**	**0.81**

Table 13. Accuracy, recall, precision and F1 for all classifiers considered in this work using as input the statistical, SCR and wavelets features. The reported values refer to the mean of the metrics obtained using the **LOSO** validation procedures and **Subset 1**, but with different **ground-truth** labels by considering a 5-second segment as artifact when at least one human annotator said is an artifact and clean otherwise.

Classifier	**Accuracy$_V$**	**F1$_V$**	**Precision$_V$**	**Recall$_V$**	**Accuracy**	**F1**	**Precision**	**Recall**	**Cohen's κ**
AdaBoost	0.95	0.95	0.95	0.94	0.90	0.87	0.92	**0.89**	0.78
Baseline1	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
Baseline2	0.49	0.49	0.49	0.49	0.51	0.49	0.51	0.52	0.02
DT	0.93	0.93	0.93	0.93	0.90	0.86	0.89	0.86	0.77
LDA	0.82	0.79	0.95	0.67	0.79	0.72	0.82	0.69	0.53
LR	0.91	0.90	0.98	0.84	0.85	0.80	0.87	0.81	0.67
QDA	0.90	0.89	0.94	0.84	0.80	0.79	0.78	0.88	0.56
RF	0.94	0.94	0.96	0.93	0.90	0.86	0.93	0.86	0.77
SVM	0.84	0.82	0.98	0.71	0.81	0.72	0.93	0.64	0.57
XGBoost	0.95	0.95	0.96	0.95	**0.91**	**0.88**	**0.95**	0.88	**0.81**
kNN	0.93	0.92	0.95	0.90	0.86	0.80	0.83	0.83	0.67
FDNN	0.95	0.95	0.95	0.94	0.85	0.80	0.83	0.88	0.71

Table 14. Accuracy, recall, precision and F1 for all classifiers considered in this work using as input the statistical, SCR and wavelets features. The reported values refer to the mean of the metrics obtained using the **LODO** validation procedure and **Subset 2**, but with different **ground-truth** labels by considering a 5-second segment as artifact when at least one human annotator said is an artifact and clean otherwise.

Classifier	**Accuracy$_V$**	**F1$_V$**	**Precision$_V$**	**Recall$_V$**	**Accuracy**	**F1**	**Precision**	**Recall**	**Cohen's κ**
AdaBoost	0.95	0.95	0.95	0.95	0.95	0.93	0.92	**0.95**	0.89
Baseline1	0.50	0.00	0.00	0.00	0.62	0.00	0.00	0.00	0.00
Baseline2	0.50	0.50	0.50	0.51	0.51	0.44	0.39	0.51	0.02
DT	0.94	0.94	0.94	0.94	0.93	0.91	0.89	0.93	0.84
LDA	0.84	0.82	0.95	0.72	0.87	0.81	0.90	0.74	0.71
LR	0.94	0.93	0.97	0.90	0.94	0.91	**0.94**	0.90	0.86
QDA	0.91	0.91	0.94	0.88	0.89	0.86	0.87	0.88	0.77
RF	0.96	0.95	0.96	0.95	0.95	**0.94**	**0.94**	0.94	0.90
SVM	0.89	0.87	0.96	0.80	0.90	0.85	**0.94**	0.78	0.77
XGBoost	0.96	0.96	0.95	0.96	**0.96**	**0.94**	**0.94**	**0.95**	**0.91**
kNN	0.93	0.93	0.95	0.91	0.92	0.90	0.90	0.90	0.84
FDNN	0.95	0.95	0.95	0.95	0.95	0.93	0.92	**0.95**	0.89

REFERENCES

[1] James Bergstra and Yoshua Bengio. 2012. Random Search for Hyper-parameter Optimization. *Journal of Machine Learning Research* 13, Feb (2012), 281–305.

[2] Christopher M Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer Science & Business Media.

[3] Wolfram Boucsein. 2012. *Electrodermal Activity*. Springer Science & Business Media.

[4] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. 2013. A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments. *Psychophysiology* 49, 1, 1017–1034.

[5] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

[6] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 33.

[7] Andriy Burkov. 2019. *The Hundred-page Machine Learning Book.* Andriy Burkov Quebec City, Can.

[8] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, and et al. 2019. Developing Measures of Cognitive Impairment in the Real World from Consumer-Grade Multimodal Sensor Streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2145–2155.

[9] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*. ACM, 785–794.

[10] Weixuan Chen, Natasha Jaques, Sara Taylor, Akane Sano, Szymon Fedor, and Rosalind W Picard. 2015. Wavelet-based Motion Artifact Removal for Electrodermal Activity. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE.

[11] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[12] ME Dawson, Anne M Schell, DL Filion, John T Cacioppo, Louis G Tassinary, and GG Berntson. 2000. Handbook of Psychophysiology. *Handbook of Psychophysiology, Cambridge University Press, Cambridge* (2000).

[13] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement During Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT 2018)* 2, 3 (2018).

[14] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2019. Laughter Recognition Using Non-invasive Wearable Devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2019)*. ACM, 262–271.

[15] Stephen Few. 2006. Information Dashboard Design. (2006).

[16] Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, Wolfram Boucsein, Don C Fowles, Sverre Grimnes, Gershon Ben-Shakhar, Walton T Roth, Michael E Dawson, and Diane L Filion. 2012. Publication Recommendations for Electrodermal Measurements. *Psychophysiology* 49, 8, 1017–1034.

[17] Chollet Francois. 2017. Deep Learning with Python.

[18] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. 2014. Empatica E3–A Wearable Wireless Multi-sensor Device for Real-time Computerized Biofeedback and Data Acquisition. In *Proceedings of the International Conference on Wireless Mobile Communication and Healthcare (MobiHealth 2014)*.

[19] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2018. Using Students' Physiological Synchrony to Quantify the Classroom Emotional Climate. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp 2018)*. ACM.

[20] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2019. Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT 2019)* 3, 1 (2019).

[21] Aurélien Géron. 2019. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

[22] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 2016. 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 (UbiComp 2016)*. ACM, 1185–1193.

[23] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. 2017. Monitoring Stress With a Wrist Device Using Context. *Journal of Biomedical Informatics* 73 (2017), 159–170.

[24] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2015. cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. *IEEE Transactions on Biomedical Engineering* 63, 4 (2015), 797–804.

[25] Nils Y Hammerla and Thomas Plötz. 2015. Let's (Not) Stick Together: Pairwise Similarity Biases Cross-validation in Activity Recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*. ACM.

[26] Jennifer Healey, Lama Nachman, Sushmita Subramanian, Junaith Shahabdeen, and Margaret Morris. 2010. Out of the Lab and Into the Fray: Towards Modeling Emotion in Everyday Life. In *International Conference on Pervasive Computing (Pervasive 2010)*. Springer, 156–173.

[27] Javier Hernandez, Daniel McDuff, Karen S Quigley, Pattie Maes, and Rosalind W Picard. 2018. Wearable Motion-based Heart-rate at Rest: A Workplace Evaluation. *IEEE Journal of Biomedical and Health Informatics* (2018).

[28] Javier Hernandez, Rob R Morris, and Rosalind W Picard. 2011. Call Center Stress Recognition with Person-specific Models. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*. Springer, 125–134.

[29] Javier Hernandez, Ivan Riobo, Agata Rozga, Gregory D Abowd, and Rosalind W Picard. 2014. Using Electrodermal Activity to Recognize Ease of Engagement in Children During Social Interactions. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2014)*.

[30] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press.

[31] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. 2017. Multimodal Autoencoder: A Deep Learning Approach to Filling in Missing Sensor Data and Enabling Better Mood Prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 202s–208.

[32] Kyriaki Kalimeri and Charalampos Saitis. 2016. Exploring Multimodal Biosignal Features for Stress Detection During Indoor Mobility. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 53–60.

[33] Cornelia Kappeler-Setz, Franz Gravenhorst, Johannes Schumm, Bert Arnrich, and Gerhard Tröster. 2013. Towards Long Term Monitoring of Electrodermal Activity in Daily Life. *Personal and Ubiquitous Computing* 17, 2 (2013), 261–271.

[34] Malia Kelsey, Ahmed Dallal, Safaa Eldeeb, Murat Akcakaya, Ian Kleckner, Christophe Gerard, Karen S Quigley, and Matthew S Goodwin. 2016. Dictionary Learning and Sparse Recovery for Electrodermal Activity Analysis. In *Compressive Sensing V: From Diverse Modalities to Big Data Analytics*, Vol. 9857. International Society for Optics and Photonics, 98570H.

[35] Malia Kelsey, Richard Vincent Palumbo, Alberto Urbaneja, Murat Akcakaya, Jeannie Huang, Ian R Kleckner, Lisa Feldman Barrett, Karen S Quigley, Ervin Sejdic, and Matthew S Goodwin. 2017. Artifact Detection in Electrodermal Activity Using Sparse Recovery. In *Compressive Sensing VI: From Diverse Modalities to Big Data Analytics*, Vol. 10211. International Society for Optics and Photonics, 102110D.

[36] Ian R Kleckner, Rebecca M Jones, Oliver Wilder-Smith, Jolie B Wormwood, Murat Akcakaya, Karen S Quigley, Catherine Lord, and Matthew S Goodwin. 2018. Simple, Transparent, and Flexible Automated Quality Assessment Procedures for Ambulatory Electrodermal Activity Data. *IEEE Transactions on Biomedical Engineering* 65, 7.

[37] Rafal Kocielnik, Natalia Sidorova, Fabrizio Maria Maggi, Martin Ouwerkerk, and Joyce HDM Westerink. 2013. Smart Technologies for Long-term Stress Monitoring at Work. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 53–58.

[38] Robert W. Levenson. 1992. Autonomic Nervous System Differences Among Emotions. *Psychological Science* 3, 1 (1992), 23–27.

[39] Robert W. Levenson. 2014. The Autonomic Nervous System and Emotion. *Emotion Review* 6, 2 (2014), 100–112.

[40] Beth Logan, Jennifer Healey, Matthai Philipose, Emmanuel Munguia Tapia, and Stephen Intille. 2007. A Long-term Evaluation of Sensing Modalities for Activity Recognition. In *International Conference on Ubiquitous computing*. Springer, 483–500.

[41] Matthew McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Marzyeh Ghassemi, and Luca Foschini. 2019. Reproducibility in Machine Learning for Health. *arXiv preprint arXiv:1907.01463* (2019).

[42] Daniel McDuff and Mary Czerwinski. 2018. Designing Emotionally Sentient Agents. *Commun. ACM* 61, 12 (Nov. 2018), 74–83.

[43] Abhinav Mehrotra and Mirco Musolesi. 2018. Using Autoencoders to Automatically Extract Mobility Features for Predicting Depressive States. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT 2018)* 2, 3 (2018).

[44] Andreas C Müller, Sarah Guido, et al. 2016. *Introduction to Machine Learning with Python: a Guide for Data Scientists*. " O'Reilly Media, Inc.".

[45] Mohsen Nabian, Yu Yin, Jolie Wormwood, Karen S Quigley, Lisa F Barrett, and Sarah Ostadabbas. 2018. An Open-Source Feature Extraction Tool for the Analysis of Peripheral Physiological Data. *IEEE Journal of Translational Engineering in Health and Medicine* 6, 1–11.

[46] Chitu Okoli. 2015. A Guide to Conducting a Standalone Systematic Literature Review. (2015).

[47] Richard V Palumbo, Marisa E Marraccini, Lisa L Weyandt, Oliver Wilder-Smith, Heather A McGee, Siwei Liu, and Matthew S Goodwin. 2017. Interpersonal Autonomic Physiology: A Systematic Review of the Literature. *Personality and Social Psychology Review* 21, 2 (2017), 99–141.

[48] Ming-Zher Poh. 2011. *Continuous Assessment of Epileptic Seizures with Wrist-worn Biosensors*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[49] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2017. Multimodal Deep Learning for Activity and Context Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)* 1, 4 (2017), 157.

[50] Darius A Rohani, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E Bardram. 2018. Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients with Affective Disorders: Systematic Review. *JMIR mHealth and uHealth* 6, 8 (2018), e165.

[51] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. 2016. Voodoo Machine Learning for Clinical Predictions. *Biorxiv* (2016).

[52] Akane Sano, Weixuan Chen, Daniel Lopez-Martinez, Sara Taylor, and Rosalind W Picard. 2018. Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks. *IEEE Journal of Biomedical and Health Informatics* (2018).

[53] Akane Sano, Andrew J Phillips, Z Yu Amy, Andrew W McHill, Sara Taylor, Natasha Jaques, Charles A Czeisler, Elizabeth B Klerman, and Rosalind W Picard. 2015. Recognizing Academic Performance, Sleep Quality, Stress Level, and Mental Health Using Personality Traits, Wearable Sensors and Mobile Phones. In *Proceedings of the IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN 2015)*. IEEE.

[54] Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the Availability of Users to Engage in Just-In-Time Intervention in the Natural Environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 909–920.

[55] Philip Schmidt, Robert Dürichen, Attila Reiss, Kristof Van Laerhoven, and Thomas Plötz. 2019. Multi-target Affect Detection in the Wild: an Exploratory Study. In *Proceedings of the 23rd International Symposium on Wearable Computers*. ACM, 211–219.

[56] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2018. Wearable Affect and Stress Recognition: A Review. *ArXiv* abs/1811.08854 (2018).

[57] Fredric Shaffer, Didier Combatalade, Erik Peper, and Zachary M Meehan. [n.d.]. A Guide to Cleaner Electrodermal Activity Measurements.

[58] Jainendra Shukla, Miguel Barreda-Angeles, Joan Oliver, GC Nandi, and Domenec Puig. 2019. Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity. *IEEE Transactions on Affective Computing* (2019).

[59] Jainendra Shukla, Miguel Barreda-Ángeles, Joan Oliver, and Domènec Puig. 2018. Efficient Wavelet-based Artifact Removal for Electrodermal Activity in Real-world Applications. *Biomedical Signal Processing and Control* 42 (2018), 45–52.

[60] Maja Stikic, Kristof Van Laerhoven, and Bernt Schiele. 2008. Exploring Semi-supervised and Active Learning for Activity Recognition. In *2008 12th IEEE International Symposium on Wearable Computers*. IEEE, 81–88.

[61] Sara Taylor, Natasha Jaques, Weixuan Chen, Szymon Fedor, Akane Sano, and Rosalind Picard. 2015. Automatic Identification of Artifacts in Electrodermal Activity Data. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE.

[62] Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* (2017).

[63] Terumi Umematsu, Akane Sano, Sara Taylor, and Rosalind Picard. 2019. Improving Students' Daily Life Stress Forecasting using LSTM Neural Networks. *IEEE-EMBS Biomedical and Health Informatics 2019* (2019).

[64] Dafne van Kuppevelt, Joe Heywood, Mark Hamer, Séverine Sabia, Emla Fitzsimons, and Vincent van Hees. 2019. Segmenting Accelerometer Data From Daily Life With Unsupervised Machine Learning. *PloS one* 14, 1 (2019), e0208692.

[65] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking Depression Dynamics in College Students using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT 2018)* 2, 1 (2018), 43.

[66] Victoria Xia, Natasha Jaques, Sara Taylor, Szymon Fedor, and Rosalind Picard. 2015. Active Learning for Electrodermal Activity Classification. In *Signal Processing in Medicine and Biology Symposium (SPMB) IEEE*. IEEE.

[67] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT 2019)* 3, 3 (2019), 116.

[68] Yaqian Xu, Isabel Hübener, Ann-Kathrin Seipp, Sandra Ohly, and Klaus David. 2017. From the Lab to the Real-world: An Investigation on the Influence of Human Movement on Emotion Recognition using Physiological Signals. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom 2017)*. IEEE, 345–350.

[69] Roberto Zangróniz, Arturo Martínez-Rodrigo, José Pastor, María López, and Antonio Fernández-Caballero. 2017. Electrodermal Activity Sensor for Classification of Calm/Distress Condition. *Sensors* 17, 10 (2017), 2324.

[70] Yuning Zhang, Maysam Haghdan, and Kevin S Xu. 2017. Unsupervised Motion Artifact Detection in Wrist-measured Electrodermal Activity Data. *arXiv preprint arXiv:1707.08287*.