Expert Systems With Applications 230 (2023) 120581

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Automatic artifact recognition and correction for electrodermal activity based on LSTM-CNN models

Jose Llanes-Jurado [a], Lucía A. Carrasco-Ribelles [a,b], Mariano Alcañiz [a], Emilio Soria-Olivas [c], Javier Marín-Morales [a,*]

[a] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, Ciudad Politécnica de la Innovación, Camino de Vera, s/n, Edif. 8B, 46022, València, Spain

[b] Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina, Gran Via de les Cortes Catalanes, 587, àtic, 08007, Barcelona, Spain

[c] IDAL, Intelligent Data Analysis Laboratory, University of Valencia, Av. Universitat s/n, 46100 Burjassot, Valencia, Spain

ARTICLE INFO

Dataset link: https://github.com/ASAPLableni/EDABE_LSTM_1DCNN, https://data.mendeley.com/datasets/w8fxrg4pv5

Keywords:
Artifact recognition
Electrodermal activity
Deep learning
Machine learning
Statistical learning
Galvanic skin response

ABSTRACT

Researchers increasingly use electrodermal activity (EDA) to assess emotional states, developing novel applications that include disorder recognition, adaptive therapy, and mental health monitoring systems. However, movement can produce major artifacts that affect EDA signals, especially in uncontrolled environments where users can freely walk and move their hands. This work develops a fully automatic pipeline for recognizing and correcting motion EDA artifacts, exploring the suitability of long short-term memory (LSTM) and convolutional neural networks (CNN). First, we constructed the EDABE dataset, collecting 74*h* EDA signals from 43 subjects collected during an immersive virtual reality task and manually corrected by two experts to provide a ground truth. The LSTM-1D CNN model produces the best performance recognizing 72% of artifacts with 88% accuracy, outperforming two state-of-the-art methods in sensitivity, AUC and kappa, in the test set. Subsequently, we developed a polynomial regression model to correct the detected artifacts automatically. Evaluation of the complete pipeline demonstrates that the automatically and manually corrected signals do not present differences in the phasic components, supporting their use in place of expert manual correction. In addition, the EDABE dataset represents the first public benchmark to compare the performance of EDA correction models. This work provides a pipeline to automatically correct EDA artifacts that can be used in uncontrolled conditions. This tool will allow to development of intelligent devices that recognize human emotional states without human intervention.

1. Introduction

Electrodermal activity (EDA) is a non-stationary signal that indicates electrical potential via the sweat glands on the surface of the skin (Boucsein, 2012). EDA represents a quantitative functional measure of sudomotor activity and, therefore, an objective assessment of emotional arousal (Ellaway, Kuppuswamy, Nicotra, & Mathias, 2010). An EDA signal can be decomposed into two different and non-redundant components: a phasic and tonic component (Benedek & Kaernbach, 2010). The phasic component is the decomposition of the rapid movements of the signal, known as the skin conductance response (SCR), which commonly provides the features used in EDA-based studies to provide valuable information for many scientific research fields (Posada-Quintero & Chon, 2020). Special attention has been given to the approach by psychology and health-related studies (Dawson, Schell, & Filion, 2000). In clinical analysis, SCR is used to assess pain, stress, schizophrenia, and peripheral neuropathy (Benedek & Kaernbach, 2010; Ellaway et al., 2010). In neuroscience and psychology, it is used to assess the subject's arousal levels (Greco, Valenza and Scilingo, 2016). For example, Anusha, Jose, Preejith, Jayaraj, and Mohanasankar (2018) used EDA signals to assess the stress of subjects in emulated real-life job scenarios, and Zangróniz, Martínez Rodrigo, Pastor García, López Bonal, and Fernández-Caballero (2017) studied EDA to distinguish between stressful and calm conditions. Liu and Du (2018) also analyzed the stress levels of a subject using the signal. Elsewhere, studies related to mental illness have utilized EDA signals,

* Correspondence to: Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, Universitat Politècnica de València, Ciudad Politécnica de la Innovación, Access N – Building 8B – 3rd Floor Camino de Vera s/n 46022, Valencia, Spain.

E-mail addresses: jllajur@upvnet.upv.es (J. Llanes-Jurado), lcarrasco@idiapjgol.info (L.A. Carrasco-Ribelles), malcaniz@i3b.upv.es (M. Alcañiz), emilio.soria@uv.es (E. Soria-Olivas), jamarmo@i3b.upv.es (J. Marín-Morales).

https://doi.org/10.1016/j.eswa.2023.120581
Received 26 March 2023; Received in revised form 10 May 2023; Accepted 27 May 2023
Available online 1 June 2023

with Greco, Valenza, Lanatà, Rota, and Scilingo (2014) finding statistical evidence concerning the relationship between healthy patients and patients with bipolar disorder using features of EDA signals and Perugia et al. (2017) discovering significant correlations between EDA signals and engagement in dementia patients.

Most previous research has collected EDA signals in laboratory environments (Shukla, Barreda-Ángeles, Oliver, & Puig, 2018), where subjects are usually seated and often cautioned not to move the hand to which the electrodes are attached. However, recent applications have recorded EDA in environments where the users can walk freely and move their hands, such as daily-life settings and virtual reality (VR) environments. Notably, many wearable devices have been developed to enable the possibility of acquiring EDA signals in a daily-life scenario, leading (Malathi, Jayaseeli, Madhuri, & Senthilkumar, 2018) to propose a wearable EDA sensor for detecting drowsiness in drivers, Leite, Henriques, Martinho, and Paiva (2013) to analyze the affective state of children in everyday situations when interacting with robots, and Kim and Fesenmaier (2015) to measure traveler emotions in real time during a four-day visit. Meanwhile, VR has been used to simulate environments where subjects can freely move and interact, which creates the sensation of being in the real world (Chicchi Giglioli, Pravettoni, Sutil, Parra, & Alcañiz Raya, 2017). VR can display different scenarios to evoke emotions or provoke cognitive processes in the subject (Marín-Morales et al., 2019; Raya, Baños, Botella, & Rey, 2003) and has been used in case studies of, for example, social adaptation in social phobia contexts, the reduction of anxiety and pain, rehabilitation, and neurological diagnosis (Bekele, Bian, Peterman, Park, & Sarkar, 2017; Li, Montaño, Chen, & Gold, 2011; Maskeliunas, Šalkevicius, Damaševičius, Maskeliunas, & Laukienė, 2019; Matijević et al., 2013; Tarrant, Viczko, & Cope, 2018). EDA has been used in VR experiments to examine sudomotor activity and arousal levels to assess anxiety and stress (Kritikos, Tzannetos, Zoitaki, Poulopoulou, & Koutsouris, 2019), conduct emotional assessments (Salgado et al., 2018), and diagnose autism (Alcañiz Raya et al., 2020).

However, among the most significant issues concerning the use of EDA signals in daily-life and VR environments is the subject's movement during data collection. Although these technologies can offer an accurate environment for recording subject responses, the absence of control over the environments can impact EDA records. Most movements can cause interferences in the contact between the skin and the recording electrodes, producing major artifacts in EDA recordings (Taylor et al., 2015). Shukla et al. (2018) suggested that artifacts in EDA signals may conceal the existence of important correlations between the signal and the subject's arousal levels due to their heavy influence on the phasic component. Therefore, ensuring the quality of the signal in uncontrolled environments represents a critical challenge.

Most EDA-based experiments manually remove major artifacts using a human expert, because there is no robust and established methodology for automatically recognizing and correcting EDA signals. Artifacts can be manually corrected using various software, including Ledalab (www.ledalab.de) and SCRalyze (Bach, 2014). However, manual correction has several disadvantages. First, it is a time-consuming and tedious task. Second, manual correction can introduce subjective human bias, with different experts correcting different signals. However, most critically, it cannot be applied in real-time or for short time periods without human intervention, as there is a demand for intelligent wearable devices that need to integrate a fully automated pipeline into the sensors. Examples of such systems include automatic systems for disorder recognition (Alcañiz Raya et al., 2020) adaptive therapies (Maskeliunas et al., 2019), mental health monitoring systems at home (Zangróniz et al., 2017), driver drowsiness detection (Malathi et al., 2018), and aesthetic evaluations (Marín-Morales et al., 2019).

Therefore, algorithms that can quickly detect and correct artifacts, ensuring data quality, appear essential for future applications of intelligent EDA-recording devices. However, works that develop automatic methods for removing artifacts remain limited (Chen et al., 2015; Hossain, Posada-Quintero, Kong, McNaboe and Chon, 2022; Shukla et al., 2018; Taylor et al., 2015; Zhang, Haghdan, & Xu, 2017) and present several limitations (see Section 2 for further details): (i) The works that recognize artifacts detect whether a segment of a signal did or did not contain an artifact, but did not provide a continuous clean signal, which is needed to compute the phasic component and assess arousal, (ii) the works that corrected signals did not compare their results with signals manually cleaned by experts, the most common method for removing artifacts, (iii) previous works did not assess the impact of the correction on the phasic component, which is related to the emotional arousal dimension and represents the most important feature in the state-of-the-art approach, and (iv) the performances of the different methods are not comparable because there is no public data benchmark. That is, no extant research has considered the development of a model that removes major EDA artifacts to provide a clean signal that does not have differences in terms of the phasic component with the signal that was cleaned manually by an expert.

This work develops an automatic recognition and correction algorithm for EDA signals, thus providing an artifact-free corrected signal that can be used in uncontrolled environments where users can freely walk and move their hands. This involves exploring two novel approaches: a long short-term memory neural networks (LSTM) in combination with a 1D convolutional neural networks (CNN), and a 2D CNN for spectrogram analysis. We compare these approaches with two state-of-the-art methods. A total of 74.46 h of EDA signal recordings were collected in a VR environment in which the 43 participants had to perform different tasks that required hand and body movements. The signals were manually corrected by two experts, generating an artifact-free signal for use as a ground truth. The labels obtained from the manual correction procedure were used to train and test the artifact recognition models. Next, automatic correction was performed on the artifacts detected. Finally, to measure the quality of the automatic corrections, the phasic component was evaluated pairwise with the automatic correction, the manual correction, and the original raw signal using two different decomposition algorithms, namely, CDA and cvxEDA.

The rest of this paper is organized as follows. Section 2 introduces the related literature. Section 3 describes the dataset's construction and the proposed methods for recognizing and correcting the artifacts. Section 4 presents the experimental results and provides a performance analysis of the proposed model. Section 5 discusses the findings, and Section 6 concludes the research.

2. Related work

Several studies have considered EDA artifact recognition. For example, the work of Kleckner et al. (2018) explored the recognition of EDA artifacts using a model based on four rules derived from the minimum and maximum range of the EDA signal or its temporal variation. However, the research on automatic detection of artifacts on EDA signals employing ML methodologies remains limited. Adopting a sampling frequency of 8 Hz, Taylor et al. (2015) detected motion artifacts in 5 s EDA segments and extracted different features from the raw EDA signal, including statistical variables (e.g., the mean, the maximum and minimum values of the data, and wavelet coefficients) to distinguish between artifacts and non-artifacts. A dataset with a duration of 130 min is used. The method achieved 96% accuracy using a support vector machine (SVM) model. However, the proportion of artifacts was not reported, and it should be considered when interpreting the performance of the model. Gashi et al. (2020) employed the same methodologies and objectives but used a larger dataset than other experimentations, including a total of 107.56 h between 13 participants. The data collected were based on ambulatory EDA signals with a sampling frequency 32 Hz that was later resampled to 8 Hz. Validation

revealed a 98% true positive rate (TPR). However, the approach followed had certain limitations, such as recognizing artifacts using 5 s segments, a lack of evaluation of artifacts in the whole signal, and not providing a final corrected signal. In addition, the final dataset has an artifact percentage of 48.96%, which differs from the initial unbalanced percentage of artifacts (17%). Zhang et al. (2017) adopted a different approach, studying the use of unsupervised learning to identify artifacts from the raw signal, achieving competitive results compared to supervised learning. In addition, Subramanian, Tseng, Barbieri, and Brown (2022) also analyzed an unsupervised approach using synthetic data as groundtruth. Hossain, Posada-Quintero, Kong et al. (2022) presented recently a model that recognize segments of 5 s affected by artifacts with 94.7% of accuracy based on a ML model feeded by a new set of hand-crafted features. They compared the method with the methodologies of Kleckner et al. (2018) and Taylor et al. (2015), outperforming the previous results. They collected both clean and corrupted EDA signal from immobile and moving hands, respectively, and their differences were used to create the groundtruth. However, they did not perform a correction of the artifacts providing reconstructed signals, which are needed for intelligent device systems, and did not analyze the implication of the artifact recognition on the phasic component.

In contrast, several works have studied the automatic correction of EDA signals without directly recognizing the artifact. That is, these methods modify the whole signal without needing to identify the artifact. Most contributions arrive from the field of signal processing, which has proposed using low-pass filters or exponential smoothing for artifact correction such as Hernandez, Morris, and Picard (2011). However, these approaches can modify certain segments of an EDA trace, which affects genuine physiological responses, creating more artifacts (Shukla et al., 2018). Other studies have used Stationary wavelet transform models to automatically remove artifacts in EDA signals. For example, the work of Chen et al. (2015) models wavelet coefficients using a Gaussian mixture distribution. Their model required estimating three parameters using the expectation–maximization algorithm. Elsewhere, Greco, Valenza, Lanata, Scilingo and Citi (2016) made a breakthrough by studying the automatic model cvxEDA, which linearly decomposed the EDA signal into tonic components, phasic components, and a Gaussian noise term that represents the signal's white noise. Therefore, this algorithm enabled the direct decomposition of the EDA signal into two main components while simultaneously removing the noise term. This model is based on Bayesian statistics and convex optimization. Greco, Valenza, Lanata et al. (2016) showed that cvxEDA outperforms CDA in terms of finely discriminating arousal levels. Furthermore, its low computational cost and efficiency has led to its use in other experiments e.g. Can, Chalabianloo, Ekiz, and Ersoy (2019) and Ganapathy, Veeranki, and Swaminathan (2020). Meanwhile, Shukla et al. (2018) proposed a wavelet-based transformation based on the Stationary wavelet transform that used a zero-mean Laplace distribution to model the wavelet coefficients and only required estimating a single parameter. More recently, Hossain, Posada-Quintero and Chon (2022) used a deep convolutional autoencoder for automatic signal correction, which more effectively demonstrated the signal-to-noise ratio than previous methods. According to that work, "the ideal scenario would be having an extra reference clean EDA signal which then can be matched with the reconstructed signal to evaluate whether the reconstructed signal accurately recovers the underlying SCRs in the EDA signals without any distortions". However, only five subjects and 39 segments of the work include a clean EDA signal for evaluation, and the validation focused on the signal-to-noise ratio. Therefore, none of these works analyzed the implication of the correction in the phasic component of the signal to recover the underlying SCRs.

Although the correction methods used in previous works produced improvements in signal-to-noise quality, none of those studies validated their findings by using an EDA signal manually corrected by an expert as a ground truth. Having the clean signal as a reference can critically improve automated approaches by enabling not only the quantification of the existing artifacts via a comparison of raw and clean signals but also the evaluation of the correction via a comparison between the automatically corrected signal and the manually cleaned signal. Furthermore, this approach can compute the underlying phasic component of the clean signal and evaluate how the automatic correction impacts this component, the most important and common feature used in such studies. As such, emulating the manual corrections performed by experts must be the ultimate goal of ML and DL models given that most studies use manual correction for artifact correction (Posada-Quintero & Chon, 2020).

Meanwhile, no previous research has combined artifact identification followed by signal correction in the same pipeline. In addition, none has been found that presented a precise characterization of motion artifacts (e.g., total number, duration, and percentage of the signal affected), which is especially important to characterize the noise levels of the signal used in each study and understand the differences on the results between studies. This might be due to the need for a manually cleaned signal to quantify the artifacts, a reconstruction that no study has included. Finally, previous studies have not made their models available for use by the scientific community, which limits the ability to produce comparisons between models. Furthermore, there is no benchmark public data, which would enable the same test data to be used in comparisons of novel methods with the state-of-the-art. As such, there are limitations when comparing the performances reported as the performances is related to the type and number of artifacts and the methodology used.

3. Materials and methods

3.1. Participants

A group of 43 volunteers (13 females and 30 males) was recruited to participate in the experiment. The mean age of the group was 37.52 (SD = 8.38). The following inclusion criteria were applied: age between 18 and 50 years, Spanish nationality, and no previous VR experience. Before the subject's participation, they received documentary information about the study and gave their informed consent for their involvement. All methods and experimental protocols were performed according to The Code of Ethics of the World Medical Association (Declaration of Helsinki), and the experimental protocol was approved by the ethics committee of the Universitat Politècnica de València (P4_18_06_19).

3.2. Data collection: EDABE dataset

We collected and published the Electrodermal activity artifact correction benchmark (EDABE) dataset (Llanes-Jurado, Carrasco-Ribelles, Alcañiz, & Marín-Morales, 2023), which includes raw electrodermal activity signals and the signals reconstructed via manual correction for use as a ground truth. To the best of our knowledge, this is the first public dataset, enabling comparison of methods. The EDABE dataset includes a total of 74.46 h of EDA recording affected by motion artifacts from the 43 subjects. It is divided into a training set with 33 subjects (56.27 h) and a test set with 10 subjects (18.19 h). We propose the adoption of the Area Under the Curve (AUC) metric for evaluation on the test set. Given the dataset includes unbalanced classes, the AUC metric provides a more robust measure for future comparisons utilizing this dataset.

The data were collected during a VR study that had the objective of inducing stress in the subject by simulating daily situations at work in a virtual environment. The participants had to perform different tasks in the virtual scenario to achieve this objective. First, subjects were placed in an office setting, where they talked to a virtual avatar about issues related to work and personal life. Then, the subjects were moved to another scenario, a meeting with five virtual avatars in which they had to actively participate in decision making.

In all the settings that required conversations with the avatars, the subjects were able to choose between the four options displayed on the lower part of the screen. Finally, the participants played three different minigames. The first minigame involved extinguishing a fire in a virtual forest as fast as possible. The second minigame entailed reorganizing a pipe to allow water to flow through it in the minimum possible time. In the last minigame, the subjects had to complete a maze while simultaneously solving simple arithmetic equations as a parallel task. The faster the subjects solved both problems, the higher their score. In all three minigames, the participants had to move both of their hands to complete the games. As such, the EDA signal became noisier in the minigames section due to the induced stress and the subjects' rapid movements.

The subjects performed the VR scenario with a HTC Vive Pro-eye head mounted display working at 90 Hz refresh rate with 1440×1600 pixels per eye and a field of view of 110°. EDA data were recorded at a sampling frequency of 128 Hz using a Shimmer3 together with the Consensys software. A total, 43 EDA signals were collected. The average experiment duration was 104 ± 8 min, producing a total of 74.46 h of signals. The virtual environment is developed in Unity3D platform.

3.3. Methodology overview

The proposed methodology is summarized in Fig. 1. First, two experts corrected the EDA signals to provide the ground truth. Next, two state-of-the-art and two new models fitted over the training set were developed: (i) Taylor et al. (2015), (ii) Hossain, Posada-Quintero, Kong et al. (2022), (iii) an LSTM with a 1D CNN, and (iv) a 2D CNN that analyze the signal's spectrogram. Following training and validation, the models were evaluated using the test set, with different classification metrics evaluated over each test signal. The algorithm that achieved the highest Kappa and TPR was selected as the best model.

Second, a fully automatic signal correction pipeline was developed. Artifacts were identified among the EDA signals using the best model. Then, a regression model was used to correct the detected artifacts to provide a final clean signal. Finally, the phasic component was calculated using the CDA and cvxEDA algorithms.

Validation of the complete pipeline involved comparing the phasic component of the three signals, namely, the raw signal, the automatic correction, and the expert manual correction (i.e., the ground truth). The similarity between the three signals was analyzed over the results of different regression metrics applied to each signal, namely, root mean square error (RMSE), mean absolute error (MAE), and cross-correlation. An ANOVA with a post-hoc analysis evaluated the differences between the phasic component of the signals.

3.4. Expert artifact correction

The following procedure was used to obtain the manual correction of the signal. The expert cleaned the signal using Ledalab software, which allowed them to visualize the complete EDA signal and indicate, in the signal itself, in which sample the artifact started and ended. Ledalab allows the manual correction through different interpolations as linear or spline, allowing the expert to choose between the one that best suits the segment signal affected. The expert then performed an automatic interpolation on the signal, correcting the parts of it that were determined to be artifacts according to their own criteria. Ledalab recorded the corrected samples, thereby collecting the artifact samples. These data were subsequently used as labels to perform a binary classification that divided the samples into "artifact" and "non-artifact" samples.

One expert corrected 21 signals and the other corrected 22 signals, of which 33 were randomly assigned to the training set and 10 to the test set, representing 56.27 h and 18.19 h of EDA signal. The labels for each corrected signal were used to produce a descriptive-artifact analysis table.

3.5. Artifact recognition models

This work proposes four ML and DL classification algorithms. The first two methods replicates the methodology described by Hossain, Posada-Quintero, Kong et al. (2022) and Taylor et al. (2015). The four methodologies share the same target processing, assigning artifact or non-artifact label according with the percentage of artifacts in a 0.5 s segment. If more than 50% of the segment was labeled as an artifact, the sample of 0.5 s was labeled as an artifact; otherwise, it was labeled as non-artifact. All the models were fitted using the training set. As a filter, signals with an artifact percentage below 1% were removed, leaving 51.35 h of EDA signal to train the three models.

Upon training all four models, we conducted a test evaluation of the models that collected the mean values of different metrics, including accuracy, Kappa, TPR, and true negative ratio (TNR). Due to the considerable imbalance between the proportion of artifacts and non-artifacts, the Kappa score and TPR were selected to evaluate artifact detection performance. Once the best model was selected, we applied post-processing to the labeling provided by the model. This involved re-labeling the signal segment between two artifacts as an artifact if they were separated by less than a certain time threshold, with the aim of merging nearby artifacts. The time threshold used was fixed at 2 s. Subsequently, an additional metric was implemented, namely, the percentage of artifacts detected. This metric was used because artifacts are not single points but sets of samples with a time duration. As such, this metric measures the percentage of artifact detection. To consider a detection valid, we analyzed the percentage of the duration of the artifact that the model labels an artifact. If this percentage exceeded a threshold value, the corresponding detection was considered correct.

3.5.1. Taylor et al. model

The first method (Taylor et al., 2015) is based on the extraction of several hand-crafted features from the raw EDA. The segments of 0.5 s are processed obtaining several types of features. The first is statistical features such as the minimum, maximum, mean, median, standard deviation and range. These statistical features were also computed over the first and second derivative of the segment. The same process is repeated for a low-pass filter of the signal with a frequency threshold of 16 Hz and to its first and second derivative. The last set of features was achieved from the computation of wavelet decomposition using Harr window of level three. From each level, the mean, median, maximum, standard deviation and number of coefficients above zero is computed. A total set of 62 features were obtained.

A backward feature selection (BFS) method based on SVC was used to select the best 40 features. Afterwards three different models were used, Logistic regression (LogR), Random Forest Classifier (RFC) and SVC. A parameter tunning was performed over each model to obtain the best hyperparameters, validating it through a group cross-validation of 5 folds. This type of cross-validation method was selected to ensure that the samples that belong to the same subject were not simultaneously present in train and validation split. The parameters used in the grid were 0.01, 0.1, 1, 10 and 100 for C in LogR; 200, 400 and 600 estimators, 10, 30 and 50 max. depth for RFC model; 1, 10, 100 and 1000 for C and 0.001, 0.01, 0.1, and 1 for Gamma in SVM model. The model with highest accuracy was selected as the best model.

3.5.2. Hossain et al. model

The second model reproduces (Hossain, Posada-Quintero, Kong et al., 2022) methodology. In our case, the model extracted the features and recognized whether or not an artifact was present in a signal segment of 0.5 s instead of 5 s to produce comparable results. The computed features can be divided into three groups. First, statistical features such as the mean, median, standard deviation, minimum, maximum, range and shannon entropy from the raw signal and its first and second derivatives. These characteristics are also computed from the phasic component of the EDA signal. Second, autoregresive features

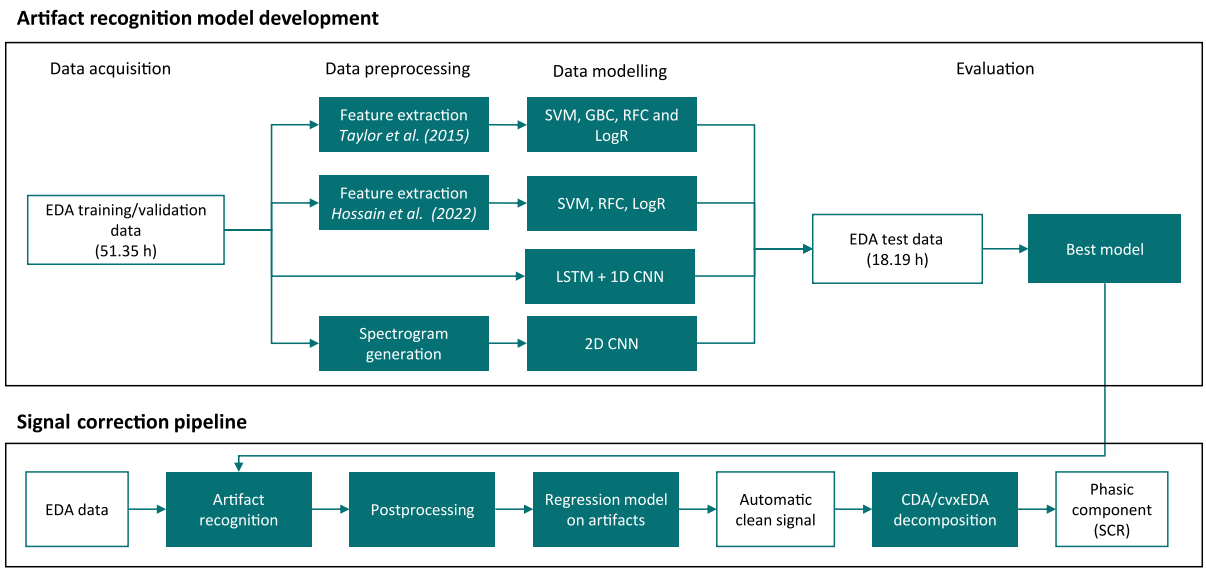


Fig. 1. Schematic representing the artifact recognition and correction pipeline.

were obtained from the coefficients of an autoregressive model over the 0.5 s signal segments, excluding the interception coefficient but adding the error variance. These type of features had also been used in other works related with time signal analysis (Hajj-Ahmad, Garg, & Wu, 2015; Moon, Hossain, & Chon, 2021). Finally, time–frequency features that were based in two different time–frequency transformations: variable frequency complex demodulation (VFCDM) (Wang, Siu, Ju, & ki, 2006) and wavelet. VFCDM was applied to the signal segment using four different frequencies: 64 Hz, 48 Hz, 32 Hz and 16 Hz. Standard deviation and mean were computed from this decomposition. From the wavelet decomposition, a three-level wavelet decomposition using Haar window is used. Mean, median, standard deviation and range of each level is obtained for each level. A total of 50 characteristics were obtained.

Following the original work, a BFS based on RFC was used to select the best 40 features. The input data were processed using standard scaler and min–max normalization. Parameter tuning was implemented using group cross-validation of 5 folds. The studied models were SVM, Gradient Boosting classifier (GBC), RFC and LogR. The parameters used in the grid were 0.01, 0.1, 1, 10 and 100 for C in LogR; 200, 400 and 600 estimators, 0.01 and 0.1 learning rate and 3, 5 and 10 max depth for GBC; 200, 400 and 600 estimators, 10, 30 and 50 max. depth for RFC model; 1, 10, 100 and 1000 for C and 0.001, 0.01, 0.1, and 1 for Gamma in SVM model. Highest accuracy defined the best model.

3.5.3. LSTM-1D CNN

This section proposes a novel model that implemented artifact detection in the last 0.5 s of a 5 s signal segment. This model's main purpose is to learn from the signal's temporal evolution. The architecture of this model was inspired by the work of Antczak (2018) and Bento, Belo, and Gamboa (2020), who both used CNN and LSTM to extract features from a raw ECG signal. Our work uses a set of LSTM layers in combination with 1D CNN layers.

Fig. 2 details the model architecture. Its first two layers were LSTM layers of 16 neurons that returned the hidden state output for each input time step. Subsequently, the network included four convolutional levels, each of which featured three convolutional layers with a batch-normalization operation performed after each convolution. Finally, each level included a dropout value of 0.05 and a max-pooling operation of size 2. The numbers of filters in each level were 32, 64, 128, and 256; kernel size was 5. Finally, the model featured two fully connected layers of 256 and 16 neurons and a final fully connected layer comprising a single perceptron with a sigmoid activation function. The model was trained with the rmsprop optimizer at a learning rate of 5×10^{-5} and a batch size of 16. Due to the imbalance, the cost function used to train the model was the Dice-Sørensen coefficient (DSC). The model had an early stopping threshold of 30 epochs. The percentage of artifacts in the training set was 12.60%. No filter was applied to the raw signal. For each 5 s segment, min–max scaling was applied.

3.5.4. Spectrogram and 2D CNN

The last proposed approach involved studying the recognition of artifacts via spectrogram artifact classification and segmentation. First, a spectrogram of each segment of 32 s of signal was created using Fast Fourier Transform (FFT) with size 4096. Then, two consecutive models were used for the temporal segmentation of artifacts. The first model was an image classification model that classified a spectrogram as having an artifact or not. The second model was an image segmentation model that created a temporal segmentation inside the spectrogram to find the artifacts. This second model only studied the spectrograms classified as containing an artifact by the first spectrogram classification model. This model combination was based on the work of Kyathanahally, Döring, and Kreis (2018), and both models were based in 2D CNN layers. An overview of the pipeline appears in Fig. 3.

To obtain the spectrogram of a signal segment, the FFT algorithm was used. Using an FFT of size 4096, a resolution of 64 samples was achieved. To obtain the squared matrix, the time segments of each signal were divided into 32 s segments. A matrix representation with the dimensions 64 × 64 was obtained. In these representations, the vertical axis represents the frequencies in Hz, and the horizontal axis shows the temporal information in seconds. The spectrograms were obtained with a 50% overlap.

The classification model was a set of CNN layers used to perform image artifact recognition. The spectrogram was classified as containing an artifact if this percentage exceeded 0% based on a comparison with the ground truth. Otherwise, the spectrogram would be classified as clean. This binarization was used for labeling by the spectrogram classification model. The model architecture comprised four convolutional levels featuring between 16 and 128 filters, as Fig. 4(a) shows. The fully connected layers in the last two levels of the model had a dropout rate with a value of 0.5. The model's cost function was binary cross-entropy.

In contrast, the segmentation model followed a U-Net architecture, as Fig. 4(b) shows. The target image was a binary image in which the label 1 indicated an artifact. Therefore, the artifact was represented as a vertical segment in the spectrogram, with the width being the temporal

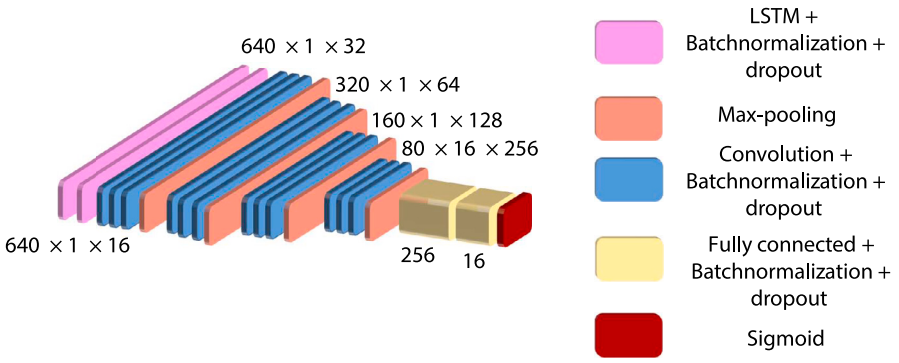


Fig. 2. Schematic representation of the architecture used for raw signal classification and LSTM-1D CNN model.

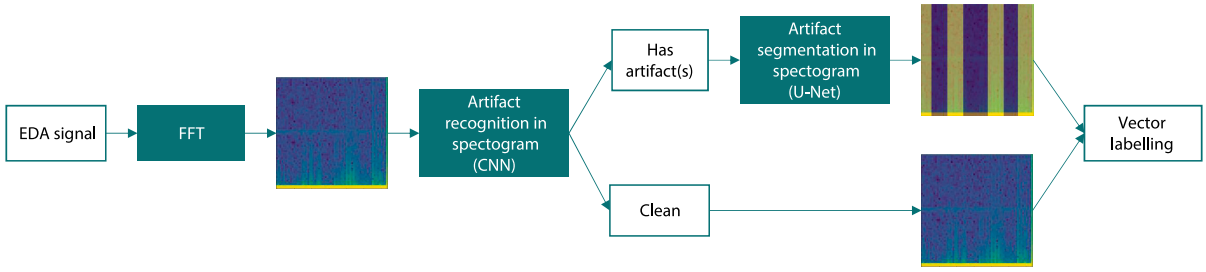
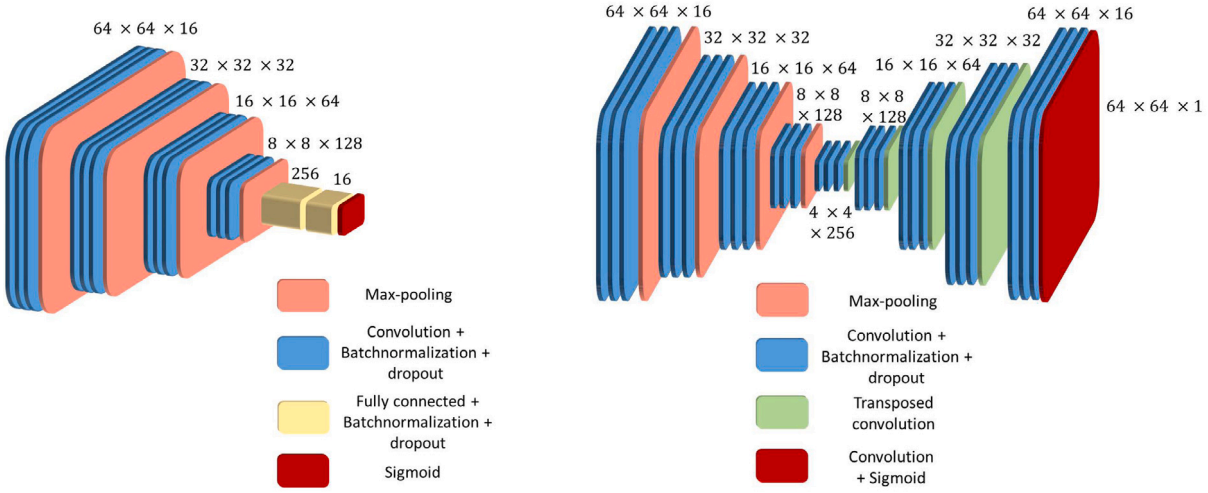


Fig. 3. Scheme of the followed methodology for the detection and segmentation of EDA artifacts in the spectogram.



(a) Spectogram artifact classification architecture

(b) Spectogram artifact segmentation architecture

Fig. 4. Architecture of the two models included in the spectrograms and 2D CNNs. Image (a) shows the artifact classification model, and image (b) shows the model that achieved the segmentation of the artifacts in the spectogram.

segmentation of the artifact demonstrated by Fig. 3. This pre-processing procedure produced the binary artifact mask image that was model's target. A maximum of 256 filters was used by the segmentation model. The kernel size for all CNNs was set to 5 × 5, and the dropout rate of the convolutional levels was set to 0.05. The model's cost function was calculated as the mean of DSC and binary cross-entropy. Using Adam optimizer with a batch size of 4, the learning rate for both models was 1×10^{-4}, and both models had an early stopping threshold of 30 epochs.

The total percentage of spectrograms with artifacts in the training set of the classification spectrogram model was 45.38%. Considering the spectrograms that contained an artifact, the total number of pixels identified as belonging to an artifact produced a total percentage of artifact pixels of 39.80%. The data introduced in the two models was a set of min–max normalized spectrograms with the dimensions 64 × 64. To increase the size of the training dataset and achieve a higher degree of model generalizability and robustness, the two models were trained using data augmentation technique (Ghosh, Das, Das, & Maulik, 2019). For this, we implemented two different types of transformation. The first involved defining random vertical or horizontal lines equal to zero that hide – at random – certain pixels in the spectrogram. The minimum and maximum threshold numbers of hidden pixels were 256 and 1024. The second transformation was the translation of the spectrogram image via a random vertical and horizontal pixel distance. The minimum and maximum threshold distances defined were 4 and 16 pixels. All the images in the dataset suffered both types of transformation, increasing the size of the dataset three times.

Table 1
Descriptive features for the artifacts extracted from all signals. Metrics are shown as mean and standard deviation per participant.

	Artifact duration (s)	Number of artifacts	Signal affected (%)	First artifact (s)	Minimum artifact duration (s)	Time between artifacts (s)	Total samples with artifact[a]	Total samples[a]
Train	5.37 ± 3.59	113.48 ± 97.12	9.97 ± 11.80	86.13 ± 173.98	1.08 ± 0.7	169.35 ± 291.63	44 669	405 194
Test	5.14 ± 3.01	182.30 ± 86.71	12.81 ± 10.57	45.65 ± 22.36	0.73 ± 0.55	48.47 ± 44.01	18 246	130 962
Complete dataset	5.22 ± 3.56	129.49 ± 99.16	10.63 ± 11.59	76.72 ± 153.75	0.88 ± 0.53	89.74 ± 125.76	62 915	536 156

[a]Samples are computed considering a target each 0.5 s.

3.6. Artifact correction

Following the artifact recognition task, a regression model was developed to correct the detected artifacts via the samples of signals labeled artifacts. This automatic correction process combined two interpolation methods. The first was a linear interpolation between the beginning and the end of the artifact. The second involved obtaining a polynomial of degree 8. The first and last samples of the artifact were taken to obtain this polynomial, and six additional internal and evenly spaced samples were considered. The methods produced a set of points for each sample labeled an artifact. Finally, the techniques were averaged for each point of the artifact to combine the corrections performed using the linear and nonlinear approaches. This approach partially reproduced the methodology involving the use of the Ledalab software. The method used in this work combines the two approaches, with the linear fit capturing the tendency of the artifact segment and a 8th degree polynomial estimation to adjust the interpolation to the non-linearity of the EDA signal. Subsequently, a simple moving average of eight samples was implemented. The simple moving average was applied from 0.125 s before the beginning of the corrected artifact to 0.125 s after the end of the artifact to smoothen the joint between the corrected artifact segment and the original EDA signal.

A set of metrics was computed to evaluate the quality of the automatic correction. We analyzed differences in terms of the phasic component between (1) the raw signal, (2) the automatically corrected signal, and (3) the signal manually corrected by experts. We focused on phasic component because it assessed the sympathetic activity and the central meaning of EDA is revealed by its peaks (Benedek & Kaernbach, 2010). To probe the robustness of the proposed methodology, we obtained the phasic component using two different approaches: the CDA (using the Ledapy library) and the cvxEDA algorithms. The metrics compared the three phasic signals by pairs, and the computed metrics were the RMSE, MAE, cross-correlation, and the difference in the area under the curve (DAUC). Furthermore, the phasic components of the signals were segmented into intervals of 5 min, upon which the mean could be computed. We analyzed the distribution of the means among the three signals using a one-way ANOVA test, performing a post-hoc analysis by pairs to observe statistical differences between them. The hypothesis considered is that if the automatic correction simulates the manual correction, no differences would be observed between them, while differences would be observed between the raw signal and the two corrections

4. Results

4.1. Signal and artifact description

Table 1 shows the descriptive analysis of the artifacts identified considering the train and test sets, and the complete dataset. The mean artifact presence percentage was 10.63 ± 11.59%.

4.2. Artifact recognition

Upon training and validating the four different approaches, the models were evaluated on the test set (18.19 h of recording), with the performance calculated via a binary classification each 0.5 s. Therefore, the models were tested on 130 962 samples. The performance metrics shown in Table 2 are averaged across the test set, providing the mean and standard deviation for each metric.

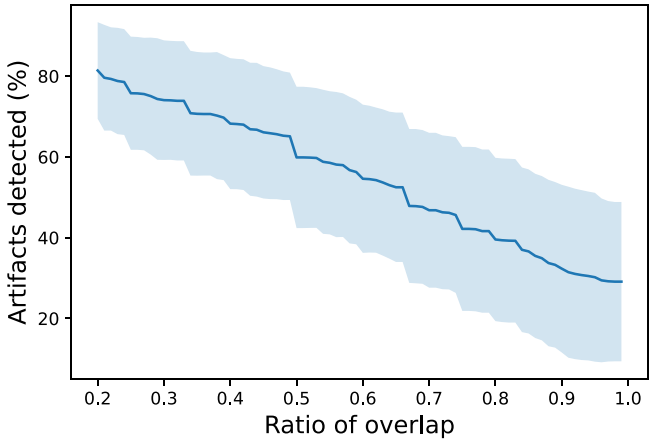


Fig. 5. Evolution of percentage of artifacts detected in terms of the overlap ratio threshold. The line represents the average in the metric; its margin area, in light blue, indicates the standard deviation above and below the mean of the metric.

Of the different ML models tested using the feature extraction and ML approach, the RFC was the best model following the features extracted from Taylor et al. (2015) whereas, the GBC outperformed the other models following the set of features of Hossain, Posada-Quintero, Kong et al. (2022). However, both performed worse than the DL approaches in terms of Kappa, TPR and AUC. The spectrogram and 2D CNN approach produced the second-best performances, achieving a TPR of 0.63 and a Kappa of 0.42. The best performance was achieved by the raw signal and LSTM-1D CNN approach, which achieved a TPR of 0.65 and a Kappa of 0.49. This performance is also corroborated by the AUC metric (0.76). This led to the selection of raw signal and LSTM-1D CNN approach as the model for recognizing artifacts to be implemented in the final pipeline.

The predictions of the raw signal and LSTM-1D CNN model were post-processed to render artifact recognition more accurate. This involved merging the artifacts separated by under 2 s. Table 3 shows an improvement in the mode's performance, producing a TPR of 0.72, a Kappa of 0.50 and an AUC of 0.79 in test set.

Next, we evaluated the percentage of artifacts detected in terms of different overlap thresholds. Fig. 5 shows a decrease in the percentage of detected artifacts according to the overlap ratio threshold. If we consider a 50% overlap threshold – that is, considering identification as valid if the model classified the artifact at least half of the time – the model detected 59.88% of the artifacts. In addition, if we considered a 20% threshold the model identification increased to 81.39%.

4.3. Artifact correction

Using the LSTM-1D CNN model with post-processing, a fully automated pipeline was implemented to the test signal data to obtain clean signals. This included a regression to interpolate the signal during the artifacts and a decomposition of the signal into phasic and tonic

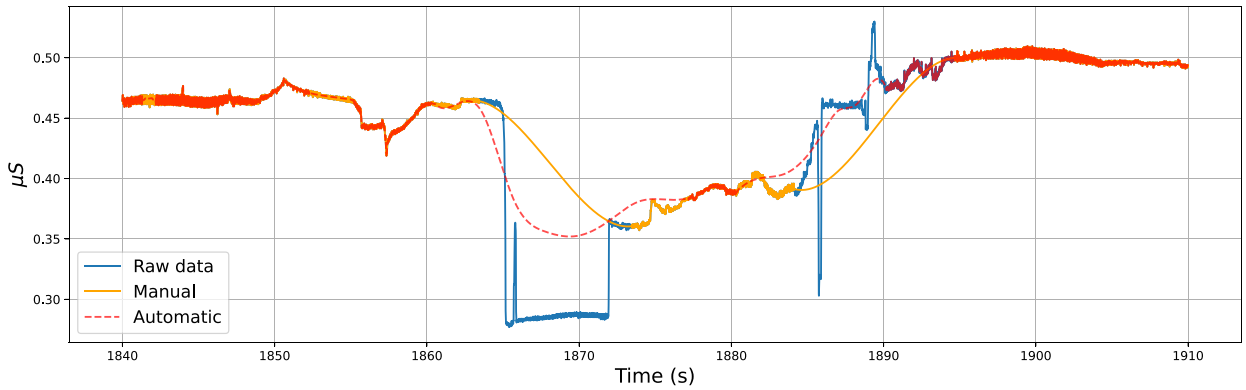


Fig. 6. Automatic correction of a certain segment of an EDA signal. The blue line is the raw signal of the segment. The orange line is the manual correction performed by an expert, and the red line is the automatic correction performed by the artifact recognition and correction algorithm.

Table 2
Evaluation of the different proposed approaches on the test set. Results appear as means and standard deviations. The model with the highest AUC, Kappa and TPR is highlighted in bold.

Model	Accuracy	TPR	TNR	Kappa	AUC	DSC
Taylor et al. (2015)	0.91 ± 0.05	0.32 ± 0.13	0.98 ± 0.04	0.39 ± 0.09	0.65 ± 0.05	0.44 ± 0.12
Hossain, Posada-Quintero, Kong et al. (2022)	0.91 ± 0.05	0.38 ± 0.18	0.96 ± 0.08	0.42 ± 0.10	0.67 ± 0.06	0.47 ± 0.14
Raw signal and LSTM-1D CNN	0.88 ± 0.09	0.65 ± 0.16	0.89 ± 0.17	0.49 ± 0.08	0.76 ± 0.06	0.57 ± 0.07
Spectrogram and 2D CNN	0.87 ± 0.10	0.63 ± 0.17	0.87 ± 0.15	0.42 ± 0.09	0.75 ± 0.06	0.50 ± 0.11

Table 3
Evaluation of the raw signal and LSTM-1D CNN model predictions on the test set after artifact recognition post-processing. Results appear as means and standard deviations.

Model	Accuracy	TPR	TNR	Kappa	AUC	DSC
Raw signal and LSTM-1D CNN with post-processing	0.87 ± 0.10	0.72 ± 0.13	0.86 ± 0.18	0.50 ± 0.10	0.79 ± 0.06	0.58 ± 0.10

Table 4
Statistical metrics for the pairwise evaluation of the phasic components of the automatic corrections, the manually cleaned signals, and the raw signals. Results appear as means and standard deviations for each participant.

Algorithm	Phasic component	RMSE	MAE	Cross correlation	DAUC	*p*-value
CDA	Automatic and manual	0.146 ± 0.096	0.054 ± 0.033	0.772 ± 0.229	0.194 ± 0.184	0.427
	Automatic and raw signal	0.171 ± 0.108	0.068 ± 0.071	0.743 ± 0.216	0.246 ± 0.247	<0.001(***)
	Manual and raw signal	0.153 ± 0.102	0.064 ± 0.055	0.795 ± 0.186	0.377 ± 0.616	0.012(*)
cvxEDA	Automatic and manual	0.339 ± 0.256	0.078 ± 0.039	0.633 ± 0.235	0.236 ± 0.168	0.246
	Automatic and raw signal	0.929 ± 0.786	0.272 ± 0.437	0.609 ± 0.207	0.478 ± 0.230	<0.001(***)
	Manual and raw signal	0.835 ± 0.809	0.255 ± 0.423	0.682 ± 0.278	0.317 ± 0.311	<0.001(***)

components. Fig. 6 shows the final interpolation result for a raw signal segment after the automatic correction process. The supplementary materials include the signals automatically corrected by the discussed algorithm.

We validated the complete pipeline by comparing the phasic component of three signals: (1) the raw signal, (2) the automatic correction, and (3) the expert manual correction (as ground truth). This involved a pairwise evaluation of the signals. Table 4 shows that automatic and manual cleaning produced lower RMSE, MAE, and DAUC values according to both decomposition algorithms (CDA and cvxEDA). The ANOVA test did not find any statistical differences (p-value > 0.05) between automatic and manual corrections. In contrast, statistical differences (p-value < 0.05) were observed between the automatic cleaning and raw signal and between the manual cleaning and the raw signal. Fig. 7 shows boxplots of the values of the phasic components for each signal and decomposition analysis. In accordance with posthoc analysis, both signals demonstrate a higher similarity in the distribution of automatic and manual clean signals compared with the raw signals.

5. Discussion

This work aimed to develop a fully automatic pipeline for recognizing and correcting artifacts in EDA signals collected in uncontrolled scenarios involving hand and body movements. The work applied two new approaches using DL algorithms: an LSTM-1D CNN applied to the raw signal and a 2D CNN applied to the spectrogram. The previous works of Hossain, Posada-Quintero, Kong et al. (2022) and Taylor et al. (2015) were used as a benchmark.

This research contributes several novelties that build upon the state-of-the-art approaches. First, some previous research on artifact recognition (Hossain, Posada-Quintero, Kong et al., 2022; Taylor et al., 2015; Zhang et al., 2017) had detected whether a segments of a signal contained an artifact. However, they did not provide a final clean signal enabling computation of the phasic component. This could be critical because, for example, Hossain, Posada-Quintero, Kong et al. (2022) analyzed segments of 5 s and, considering that many artifacts in our signals are shorter (see Table 1), this analysis could affect

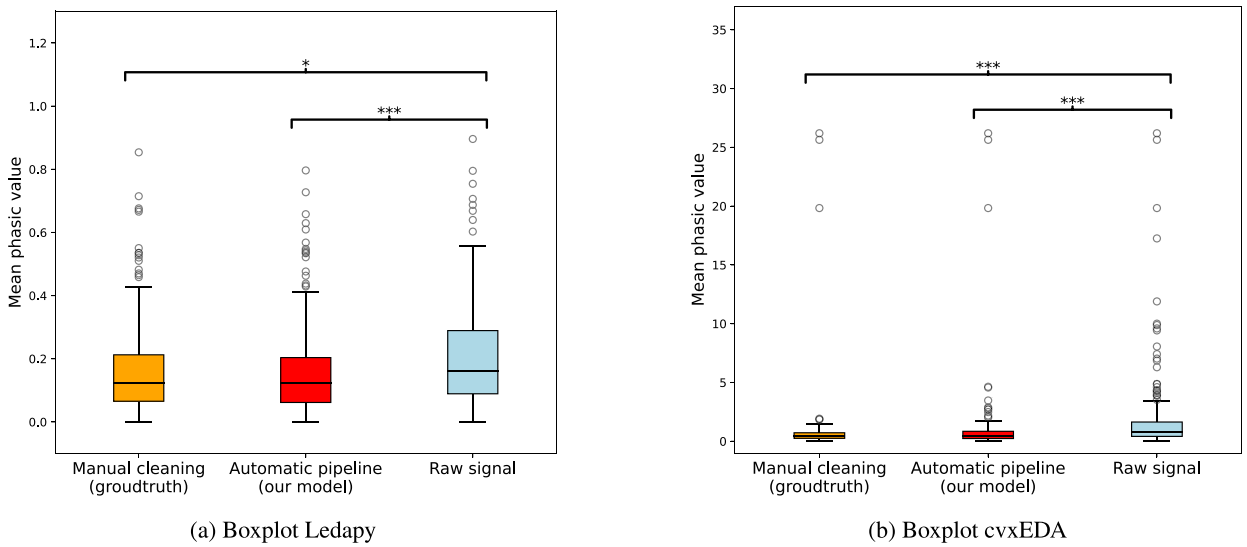


Fig. 7. Boxplot showing the distribution of different phasic values. Image (a) shows the comparison using the CDA decomposition method. Image (b) shows the results produced by the cvxEDA algorithm.

long segments of uncorrupted signal. Meanwhile, other studies had not recognized artifacts, instead aiming to directly correct signals using, for example, wavelet-based transformation (Chen et al., 2015) or convolutional autoencoders (Hossain, Posada-Quintero and Chon, 2022). However, these works did not use manually reconstructed signals as a ground truth, which represents the objective of this study, that is, to emulate the reconstruction performed by an expert by providing an artifact-free signal.

This is the first work to develop a fully automatic pipeline with three steps: (1) artifact recognition each 0.5 s, (2) post-processing of artifact recognition, and (3) correction of the signal based on artifact identification. Meanwhile, by using as a ground truth a manual reconstruction of the signal, we have been able to assess the pipeline's performance via a comparison of the automatic and manual corrections. Additionally, the dataset created contains 74.46 h of raw and manually reconstructed data, indicating labeling of more than 500,000 samples of 0.5 s. The data were collected from 43 different participants, ensuring the capacity to perform inter-subject extrapolations. The uncontrolled scenario used guaranteed the production of hand and body motion artifacts because participants needed to complete minigames causing major motion artifacts and simulating the real implementation conditions of intelligent EDA devices.

Notably, no previous work had analyzed the implications of automatic corrections for the phasic component of the signal, the most common feature used in studies because it relates to arousal (Posada-Quintero & Chon, 2020). This may be due to the need for the reconstruction of the signal to analyze the implications of the correction for the phasic component, information not contained in the majority of previous work. This work has analyzed the differences between the phasic component derived from our pipeline and the manual correction, demonstrating no differences between them. This novel result supports our pipeline as an emulation of human expert artifact correction.

Furthermore, this is the first artifact correction pipeline that is available for the use (and testing) of the scientific community and the first work that includes a dataset featuring raw data, manual reconstructions, and automated corrections. Thus, it represents a benchmark that can be used by future researchers to compare new methods and improvements using the same data, a current limitation on the state-of-the-art limitation that precludes comparison of the results because different data are used. These methodological improvements represent a breakthrough in the validation of recognition and correction algorithms for EDA signals.

We used two state-of-the-art methods as a benchmark. In the test set, both achieve the highest accuracy, but they present the lowest Kappa and AUC. It is because the TPR is relatively low, since Hossain, Posada-Quintero, Kong et al. (2022) and Taylor et al. (2015) detects the 32% and 38% of the artifacts respectively. Notably, these results were worse than those presented by previous studies. There are two potential reasons for this discrepancy. First, the type of labeling used in the present work differs from that used in other studies. That is, other studies directly assigned a complete window of 5 s the label of artifact or not artifact, while we used the comparison with the manual correction to assign this label in segments of 0.5 s, which suppose an important increasement of the precision of the correction. Second, the imbalance of the current dataset (10.63%) exceeds that of previous experiments (e.g., 48.96% in the work of Gashi et al. (2020)). This could bias the performance and the results of previous works. Note that we use a dataset collected during a VR Serious Game, which is an actual use case of the pipeline, while as an example (Hossain, Posada-Quintero, Kong et al., 2022) create a specific protocol to generate the artifacts.

The two models using DL architectures outperformed the feature extraction and classical ML approach, achieving higher TPR, Kappa, AUC, and DSC values in the test set. Inspired by prior research on ECG denoising (Antczak, 2018; Bento et al., 2020), we investigated the use of a LSTM-1D CNN. Concurrently, the adoption of a 2D CNN was explored, drawing motivation from previous studies on Magnetic Resonance Spectroscopy denoising (Kyathanahally et al., 2018). The best model was the raw signal and LSTM-1D CNN model, which achieved a final accuracy of 0.88, a Kappa value of 0.49, and a TPR value of 0.65. This represents a large increase in artifact detection performance relative to previous methodologies. Meanwhile, the spectrogram and 2D CNN model achieved a Kappa value of 0.42, a TPR value of 0.63, and total accuracy of 0.87. This model's performance was inferior to that of the raw signal and LSTM-1D CNN model, likely because 2D CNN was not optimized for the study of spectrogram images due to the non-local information that a spectrogram provides, with CNNs basing their knowledge on the local information contained in the data. More specialized models, such as spectral-CNN, could be implemented in future research to study the artifact detection problem in the EDA context.

To improve artifact recognition, a post-processing method was applied to the predictions of the raw signal and LSTM-1D CNN model.

This post-processing improved artifact detection, as demonstrated the increased Kappa and TPR values (0.50 and 0.72, respectively). We also analyzed the percentage of artifacts included in the test set that were correctly identified by the model. Considering an identification valid if the model correctly labeled 50% of the artifact, the pipeline recognized 59.88% of the artifacts, with detection increasing to 81.39% with the use of a 20% threshold. Therefore, most artifacts were identified at least partially correctly, potentially because the model identifies the most aggressive segments of artifacts but not the entirety of the correction made by human experts.

Finally, the EDA signal was corrected using linear and polynomial regressions on the segments identified as artifacts. The automatic correction algorithm used in this work was designed to be similar to the type of manual correction enabled by Ledalab software. Although the results obtained fulfilled the initial objective, the type of automatic correction could be complemented or replaced by other correction methodologies. For example, the methodologies suggested by Chen et al. (2015) or Shukla et al. (2018), who implemented wavelet transformation, lowpass filters (Hernandez et al., 2011), and the cvxEDA algorithm (Greco, Valenza, Lanata et al., 2016) could enrich the corrections made by the proposed algorithm.

The complete pipeline was evaluated based on the implications of the corrections for the phasic component. This involved using a one-way ANOVA with a post-hoc test to compare the three signals: (1) the raw signal, (2) the automatic correction, and (3) the manual correction by a human expert (ground truth). According to Table 4, there was no statistical difference (p-value > 0.05) between the phasic component produced by the automatic correction and the manual correction for either the CDA or cvxEDA algorithm. Furthermore, the type of correction performed was robust against the type of signal decomposition applied, showing similar results for the two algorithms. Meanwhile, statistically significant differences (p-value < 0.05) were observed between the phasic component of the raw signal and the manual correction, as well as between the raw signal and the automatic correction. This indicates that the automatic correction features less artifact noise than the raw signal (see Fig. 7). Other metrics, namely, RMSE, MAE, and DAUC, also showed that the phasic component of the automatic correction was closer to the phasic component of the manual correction than to the phasic component of the raw signal. Therefore, the results suggest that the automatic correction accurately simulates manual correction, independently of the decomposition algorithm used. These results support this paper's main objective of providing an artifact-free corrected signal that emulates manual correction by a human expert.

However, the study does have some limitations that must be addressed in future research. First, model results can be improved by including more experts for manual correction to reduce human bias. This would enrich the signal target and, therefore, the generalizability of the models. Second, the visual inspection and manual reconstruction can create an unrealistic morphology in the EDA signal, even if it is the standard practice in experiments. The manual cleaning aims to reduce the negative impact of the artifact on the signal and, in particular, on the phasic component, but it is not capable of reconstructing the real affected EDA. The alternative approach to obtain the artifact-free signal is to create a protocol that forces one hand to generate movements while the other is stationary, collecting data from both different locations simultaneously, as performed by Hossain, Posada-Quintero, Kong et al. (2022). Even if these protocols may have a low degree of ecological validity since it is an artificial task, and EDA signal can change depending on the location (van Dooren, Janssen, et al., 2012), the model must be tested considering this alternative groundtruth. Third, future research should evaluate the model in other types of environments and tasks because the specific movements performed can modify the form of the artifacts. Validating the methodology for EDA signals collected during other types of tasks would strengthen the model and demonstrate its applicability to other contexts as real-world experimentations. Moreover, the procedure has not been tested for signals from different EDA devices or those with frequencies below 128 Hz. The methods established here could be studied at different sampling frequencies to review their performance and generalizability. Finally, future experiments should consider researching the development of fine-tuned architectures for different models, which could improve their classification metrics. For example, generative-adversarial networks and reinforcement learning represent promising alternatives to the models demonstrated in this work.

6. Conclusion

We have developed a fully automatic pipeline for recognizing and correcting EDA motion artifacts, achieving a corrected signal that does not differ from manual correction by human experts in terms of phasic component. The recognition of the artifacts outperforms two previous state-of-the-art methods. These results show that EDA signal correction in scenarios that require body movements can be achieved automatically, findings that can enhance the use of EDA signals in future experiments conducted in uncontrolled environments, including immersive VR and real-world settings. These findings also provide encouragement for the development of intelligent devices for recognizing human emotional states for healthcare services without human intervention, including implementations in the contexts of disorder recognition, adaptive therapy, remote mental health monitoring systems, and driver drowsiness detection.

CRediT authorship contribution statement

Jose Llanes-Jurado: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Lucía A. Carrasco-Ribelles:** Conceptualization, Writing – review & editing. **Mariano Alcañiz:** Funding acquisition, Resources, Writing – review & editing. **Emilio Soria-Olivas:** Writing – review & editing. **Javier Marín-Morales:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and model availability

The complete pipeline is available in https://github.com/ASAPLableni/EDABE_LSTM_1DCNN. Furthermore, the EDABE dataset is publicly available in Mendeley Data for use as a benchmark in comparisons of the performance of future models and pipelines https://data.mendeley.com/datasets/w8fxrg4pv5 (Llanes-Jurado et al., 2023).

Funding

This work was supported by the European Commission [RHUMBO H2020-MSCA-ITN-2018-813234]; the Generalitat Valenciana, Spain [REBRAND PROMETEU/2019/105]; the MCIN/AEI, Spain [PID2021-127946OB-I00]; and the Universitat Politècnica de València, Spain [PAID-10-20].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2023.120581. The following supplementary material includes the plots of the raw signals, manual corrections, and automatic corrections of the test set.

References

Alcañiz Raya, M., Chicchi Giglioli, I. A., Marín-Morales, J., Higuera-Trujillo, J. L., Olmos, E., Minissi, M. E., et al. (2020). Application of supervised machine learning for behavioral biomarkers of autism spectrum disorder based on electrodermal activity and virtual reality. *Frontiers in Human Neuroscience*, *14*, 90. http://dx.doi.org/10.3389/fnhum.2020.00090, URL: https://www.frontiersin.org/article/10.3389/fnhum.2020.00090.

Antczak, K. (2018). Deep recurrent neural networks for ECG signal denoising. CoRR, abs/1807.11551. URL: http://arxiv.org/abs/1807.11551. arXiv:1807.11551.

Anusha, A. S., Jose, J., Preejith, S. P., Jayaraj, J., & Mohanasankar, S. (2018). Physiological signal based work stress detection using unobtrusive sensors. *Biomedical Physics & Engineering Express*, *4*(6), Article 065001. http://dx.doi.org/10.1088/2057-1976/aadbd4.

Bach, D. R. (2014). A head-to-head comparison of SCRalyze and Ledalab, two model-based methods for skin conductance analysis. *Biological Psychology*, *103*, 63–68. http://dx.doi.org/10.1016/j.biopsycho.2014.08.006, URL: http://www.sciencedirect.com/science/article/pii/S0301051114001847.

Bekele, E., Bian, D., Peterman, J., Park, S., & Sarkar, N. (2017). Design of a virtual reality system for affect analysis in facial expressions (VR-SAAFE); application to schizophrenia. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(6), 739–749. http://dx.doi.org/10.1109/TNSRE.2016.2591556.

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*(1), 80–91. http://dx.doi.org/10.1016/j.jneumeth.2010.04.028, URL: http://www.sciencedirect.com/science/article/pii/S0165027010002335.

Bento, N., Belo, D., & Gamboa, H. (2020). ECG biometrics using spectrograms and deep neural networks. http://dx.doi.org/10.18178/ijmlc.2020.10.2.929.

Boucsein, W. (2012). *Electrodermal activity*. New York: Springer Science+Business Media, LLC.

Can, Y., Chalabianloo, N., Ekiz, D., & Ersoy, C. (2019). Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, *19*, http://dx.doi.org/10.3390/s19081849.

Chen, W., Jaques, N., Taylor, S., Sano, A., Fedor, S., & Picard, R. W. (2015). Wavelet-based motion artifact removal for electrodermal activity. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society* EMBC, (pp. 6223–6226). http://dx.doi.org/10.1109/EMBC.2015.7319814.

Chicchi Giglioli, I., Pravettoni, G., Sutil, L., Parra, E., & Alcañiz Raya, M. (2017). A novel integrating virtual reality approach for the assessment of the attachment behavioral system. *Frontiers in Psychology*, *8*, http://dx.doi.org/10.3389/fpsyg.2017.00959.

Dawson, M., Schell, A., & Filion, D. (2000). The electrodermal system. In *Handbook of psychophysiology*. http://dx.doi.org/10.1017/CBO9780511546396.007.

van Dooren, M., Janssen, J. H., et al. (2012). Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & Behavior*, *106*(2), 298–304.

Ellaway, P., Kuppuswamy, A., Nicotra, A., & Mathias, C. (2010). Sweat production and the sympathetic skin response: improving the clinical assessment of autonomic function. *Autonomic Neuroscience*, *155*(1–2), 109—114. http://dx.doi.org/10.1016/j.autneu.2010.01.008.

Ganapathy, N., Veeranki, Y. R., & Swaminathan, R. (2020). Convolutional neural network based emotion classification using electrodermal activity signals and time-frequency features. *Expert Systems with Applications*, *159*, Article 113571. http://dx.doi.org/10.1016/j.eswa.2020.113571.

Gashi, S., Di Lascio, E., Stancu, B., Swain, V. D., Mishra, V., Gjoreski, M., et al. (2020). Detection of artifacts in ambulatory electrodermal activity data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(2), http://dx.doi.org/10.1145/3397316.

Ghosh, S., Das, N., Das, I., & Maulik, U. (2019). Understanding deep learning techniques for image segmentation. *ACM Computing Surveys*, *52*(4), http://dx.doi.org/10.1145/3329784.

Greco, A., Valenza, G., Lanatà, A., Rota, G., & Scilingo, E. (2014). Electrodermal activity in bipolar patients during affective elicitation. *IEEE Journal of Biomedical and Health Informatics*, *18*, 1865–1873. http://dx.doi.org/10.1109/JBHI.2014.2300940.

Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., & Citi, L. (2016). cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, *63*(4), 797–804, URL: http://dblp.uni-trier.de/db/journals/tbe/tbe63.html#GrecoVLSC16.

Greco, A., Valenza, G., & Scilingo, E. (2016). *Advances in electrodermal activity processing with applications for mental health*. http://dx.doi.org/10.1007/978-3-319-46705-4.

Hajj-Ahmad, A., Garg, R., & Wu, M. (2015). ENF-based region-of-recording identification for media signals. *IEEE Transactions on Information Forensics and Security*, *10*(6), 1125–1136. http://dx.doi.org/10.1109/TIFS.2015.2398367.

Hernandez, J., Morris, R. R., & Picard, R. W. (2011). Call center stress recognition with person-specific models. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (pp. 125–134). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hossain, M. B., Posada-Quintero, H., & Chon, K. (2022). A deep convolutional autoencoder for automatic motion artifact removal in electrodermal activity. *IEEE Transactions on Biomedical Engineering*.

Hossain, M.-B., Posada-Quintero, H. F., Kong, Y., McNaboe, R., & Chon, K. H. (2022). Automatic motion artifact detection in electrodermal activity data using machine learning. *Biomedical Signal Processing and Control*, *74*, Article 103483.

Kim, J. J., & Fesenmaier, D. R. (2015). Measuring emotions in real time: Implications for tourism experience design. *Journal of Travel Research*, *54*(4), 419–429. http://dx.doi.org/10.1177/0047287514550100, arXiv:https://doi.org/10.1177/0047287514550100.

Kleckner, I. R., Jones, R. M., Wilder-Smith, O., Wormwood, J. B., Akcakaya, M., Quigley, K. S., et al. (2018). Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. *IEEE Transactions on Bio-Medical Engineering*, *65*(7), 1460—1467. http://dx.doi.org/10.1109/tbme.2017.2758643, URL: https://europepmc.org/articles/PMC5880745.

Kritikos, J., Tzannetos, G., Zoitaki, C., Poulopoulou, S., & Koutsouris, D. (2019). Anxiety detection from electrodermal activity sensor with movement interaction during virtual reality simulation. In *2019 9th international IEEE/EMBS conference on neural engineering* NER, (pp. 571–576). http://dx.doi.org/10.1109/NER.2019.8717170.

Kyathanahally, S. P., Döring, A., & Kreis, R. (2018). Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy. *Magnetic Resonance in Medicine*, *80*(3), 851–863. http://dx.doi.org/10.1002/mrm.27096, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.27096.

Leite, I., Henriques, R., Martinho, C., & Paiva, A. (2013). Sensors in the wild: Exploring electrodermal activity in child-robot interaction. In *Proceedings of the 8th ACM/IEEE international conference on human-robot interaction* HRI '13, (pp. 41–48). IEEE Press.

Li, A., Montaño, Z., Chen, V. J., & Gold, J. (2011). Virtual reality and pain management: current trends and future directions. *Pain Management*, *1 2*, 147–157.

Liu, Y., & Du, S. (2018). Psychological stress level detection based on electrodermal activity. *Behavioural Brain Research*, *341*, 50–53.

Llanes-Jurado, J., Carrasco-Ribelles, L., Alcañiz, M., & Marín-Morales, J. (2023). Electrodermal Activity artifact correction BEnchmark (EDABE). http://dx.doi.org/10.17632/w8fxrg4pv5.1, URL: https://data.mendeley.com/datasets/w8fxrg4pv5.

Malathi, D., Jayaseeli, J. D., Madhuri, S., & Senthilkumar, K. (2018). Electrodermal activity based wearable device for drowsy drivers. *Journal of Physics: Conference Series*, *1000*, Article 012048. http://dx.doi.org/10.1088/1742-6596/1000/1/012048.

Marín-Morales, J., Higuera-Trujillo, J., Greco, A., Guixeres, J., Llinares, C., Gentili, C., et al. (2019). Real vs. immersive-virtual emotional experience: Analysis of psycho-physiological patterns in a free exploration of an art museum. *PLOS ONE*, *14*(10), 1–24. http://dx.doi.org/10.1371/journal.pone.0223881.

Maskeliunas, R., Šalkevicius, J., Damaševičius, R., Maskeliunas, R., & Laukienė, I. (2019). Anxiety level recognition for virtual reality therapy system using physiological signals. *Electronics*, *8*(9), 1039.

Matijević, V., Šečić, A., Mašić, V., Sunić, M., Kolak, Z., & Znika, M. (2013). Virtual reality in rehabilitation and therapy. *Acta Clinica Croatica*, *52*, 453–457.

Moon, J., Hossain, M. B., & Chon, K. H. (2021). AR and ARMA model order selection for time-series modeling with ImageNet classification. *Signal Processing*, *183*, Article 108026. http://dx.doi.org/10.1016/j.sigpro.2021.108026, URL: https://www.sciencedirect.com/science/article/pii/S0165168421000657.

Perugia, G., Rodriguez-Martin, D., Boladeras, M., Mallofre, A., Barakova, E., & Rauterberg, M. (2017). *RO-MAN 2017 - 26th IEEE international symposium on robot and human interactive communication* (pp. 1248–1254). United States: Institute of Electrical and Electronics Engineers, http://dx.doi.org/10.1109/ROMAN.2017.8172464, URL: http://www.ro-man2017.org/site/. 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2017), RO-MAN 2017 ; Conference date: 28-08-2017 Through 01-09-2017.

Posada-Quintero, H., & Chon, K. (2020). Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors*, *20*, 479. http://dx.doi.org/10.3390/s20020479.

Raya, M. A., Baños, R., Botella, C., & Rey, B. (2003). The EMMA project: Emotions as a determinant of presence. *PsychNology Journal*, *1*, 141–150.

Salgado, D., Martins, F., Braga Rodrigues, T., Keighrey, C., Flynn, R., Naves, E., et al. (2018). A QoE assessment method based on EDA, heart rate and EEG of a virtual reality assistive technology system. (pp. 517–520). http://dx.doi.org/10.1145/3204949.3208118.

Shukla, J., Barreda-Ángeles, M., Oliver, J., & Puig, D. (2018). Efficient wavelet-based artifact removal for electrodermal activity in real-world applications. *Biomedical Signal Processing and Control*, *42*, 45–52. http://dx.doi.org/10.1016/j.bspc.2018.01.009, URL: http://www.sciencedirect.com/science/article/pii/S1746809418300090.

Subramanian, S., Tseng, B., Barbieri, R., & Brown, E. N. (2022). An unsupervised automated paradigm for artifact removal from electrodermal activity in an uncontrolled clinical setting. *Physiological Measurement*, *43*(11), Article 115005.

Tarrant, J., Viczko, J., & Cope, H. (2018). Virtual reality for anxiety reduction demonstrated by quantitative EEG: A pilot study. *Frontiers in Psychology*, *9*, 1280. http://dx.doi.org/10.3389/fpsyg.2018.01280, URL: https://www.frontiersin.org/article/10.3389/fpsyg.2018.01280.

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society* EMBC, (pp. 1934–1937).

Wang, H., Siu, K., Ju, K., & ki, h. (2006). A high resolution approach to estimating time-frequency spectra and their amplitudes. *Annals of Biomedical Engineering*, *34*, 326–338. http://dx.doi.org/10.1007/s10439-005-9035-y.

Zangróniz, R., Martínez Rodrigo, A., Pastor García, J. M., López Bonal, M., & Fernández-Caballero, A. (2017). Electrodermal activity sensor for classification of calm/distress condition. *Sensors*, *17*, 2324. http://dx.doi.org/10.3390/s17102324.

Zhang, Y., Haghdan, M., & Xu, K. S. (2017). Unsupervised motion artifact detection in wrist-measured electrodermal activity data. In *Proceedings of the 2017 ACM international symposium on wearable computers* ISWC '17, (pp. 54–57). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3123021.3123054.