Detecting Deception and Suspicion in Dyadic Game Interactions

Jan Ondras
University of Cambridge
Cambridge, United Kingdom
jankondras@gmail.com

Hatice Gunes
University of Cambridge
Cambridge, United Kingdom
hatice.gunes@cl.cam.ac.uk

ABSTRACT

In this paper we focus on detection of deception and suspicion from electrodermal activity (EDA) measured on left and right wrists during a dyadic game interaction. We aim to answer three research questions: (i) Is it possible to reliably distinguish deception from truth based on EDA measurements during a dyadic game interaction? (ii) Is it possible to reliably distinguish the state of suspicion from trust based on EDA measurements during a card game? (iii) What is the relative importance of EDA measured on left and right wrists? To answer our research questions we conducted a study in which 20 participants were playing the game *Cheat* in pairs with one EDA sensor placed on each of their wrists. Our experimental results show that EDA measures from left and right wrists provide more information for suspicion detection than for deception detection and that the person-dependent detection is more reliable than the person-independent detection. In particular, classifying the EDA signal with Support Vector Machine (SVM) yields accuracies of 52% and 57% for person-independent prediction of deception and suspicion respectively, and 63% and 76% for person-dependent prediction of deception and suspicion respectively. Also, we found that: (i) the optimal interval of informative EDA signal for deception detection is about 1 s while it is around 3.5 s for suspicion detection; (ii) the EDA signal relevant for deception/suspicion detection can be captured after around 3.0 seconds after a stimulus occurrence regardless of the stimulus type (deception/truthfulness/suspicion/trust); and that (iii) features extracted from EDA from both wrists are important for classification of both deception and suspicion. To the best of our knowledge, this is the first work that uses EDA data to automatically detect both deception and suspicion in a dyadic game interaction setting.

KEYWORDS

Affective computing; Dyadic game interactions; Electrodermal activity; Deception detection; Suspicion detection

ACM Reference Format:
Jan Ondras and Hatice Gunes. 2018. Detecting Deception and Suspicion in Dyadic Game Interactions. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3242969.3242993

1 INTRODUCTION

Electrodermal activity (EDA) is a widely used indicator of sympathetic nervous system (SNS) activity and is often used to describe the degree of a person's excitement, stress, anxiety, as well as changes in arousal related to pain and anticipation [11]. EDA is also known as skin conductance or galvanic skin response and these terms will be further used interchangeably. Traditionally, EDA measurements involved attaching wired and gelled electrodes to the skin [15]. Recently, unobtrusive wearable devices such as the wireless Affectiva Q Sensor [2] used in our study have attained popularity among researchers in various fields.

EDA can be also used as an indicator of deceit since lying costs more mental effort than telling the truth [58] and the cognitive load activates SNS [13, 38]. Previous research [36, 54] found increased EDA for lying as compared to truth-telling. Moreover, EDA is an autonomic-based physiological response which makes it hard to control and therefore less susceptible to strategic manipulations [18], and is a good indicator of deceit.

Suspicion (or trust) detection is a more recent and less researched field and to the best of our knowledge there was no attempt at detecting suspicion or trust from EDA. However, the state of suspicion is often associated with an increased cognitive load and stress as compared to the state of trust [10, 50] which along with the aforementioned findings about EDA suggests that it should be possible to detect suspicion based on skin conductance measurements.

Many studies have measured the presence of EDA asymmetry on the left and right palms [25, 27, 34] and attempted to relate bilateral EDA measures to verbal/spatial, positive/negative, emotional/non-emotional specialisation, with conflicting findings. The classical understanding assumed that EDA represents one homogeneous change in arousal across the whole body, but recent works [5, 41] show that multiple brain structures contribute to elicitation of EDA, namely, two regions were identified: a limbic-hypothalamic source (EDA1) and a premotor-basal ganglia source (EDA2). The EDA1 system includes structures (such as amygdala, cingulate gyrus, anterior thalamus, fornix, hippocampus, and hypothalamus) that play an important role in emotions and it is believed to be ipsilateral[1] [51] which means that activating key emotion regions on the right side of the brain (e.g., right amygdala) produces right palmar EDA activation and analogously for the left side. In contrast, the EDA2 system (including the basal ganglia and premotor cortex) is contralateral[2]. Together with findings that amygdala is the emotional centre of the brain and that the left hemisphere primarily processes positive emotions while the right hemisphere processes primarily negative emotions [48], it might be tempting to conclude that higher skin conductance found on right hand reflects negative emotional state and analogously higher skin conductance on left hand reflects positive emotional state. Such conclusions may be flawed since the measured EDA might be also influenced by other sources of arousal (EDA2), for example, by hand movements. Based on current research findings, the plausible conclusion is that underlying negative emotions (such as fear or anxiety) only *contribute*

ICMI '18, October 16–20, 2018, Boulder, CO, USA
2018. ACM ISBN 978-1-4503-5692-3/18/10.
https://doi.org/10.1145/3242969.3242993

[1] Occurring on the same side of the body.
[2] Occurring on the opposite side of the body.

to greater right amygdala activation and thus to the EDA on right hand. Since EDA measures on only one side may lead to misjudgment of arousal [41], in this study we investigate EDA signals from both hands.

In this work we hypothesise that EDA data obtained from left and right wrists can be effectively used to detect deception and suspicion during a dyadic game interaction. Specifically, we aim to answer the following research questions: (i) Is it possible to reliably distinguish deception from truth based on EDA measurements during a dyadic game interaction? (ii) Is it possible to reliably distinguish the state of suspicion from trust based on EDA measurements during a dyadic game interaction? (iii) What is the relative importance of EDA measured on left and right wrists? We investigate these questions by undertaking a controlled study where 20 participants were asked to play a two-player variant of the card game *Cheat* and answer a post-study questionnaire. In this context, we define deception to be the action when a player discards a card different from what he/she claims and suspicion as a state when a player does not trust the opponent that the card discarded by the opponent was the same as he/she claimed. To summarise, our work has the following contributions:

- First of its kind dataset[3] for automatic deception and suspicion detection based on measurements from wearable EDA sensors, collected from 20 participants along with their personality traits.
- A prototype system for automatic detection of deception and suspicion from EDA measurements on-the-fly.

The rest of the paper is structured as follows. Section 2 details the related work in the field, Section 3 describes the details of the conducted study, and Section 4 explains the feature extraction process. Next, the classification experiments and obtained results are presented in Section 5 while Section 6 shows the importance of individual features and asymmetry in EDA. Section 7 discusses the results and Section 8 concludes the paper and highlights future research directions.

2 RELATED WORK

EDA has been widely used for various tasks such as seizure detection [44], engagement recognition during social interactions [23], analysis of EDA during sleep [56], or depression prediction based on EDA asymmetry [14].

The detection of deception (or lie detection) is a long-standing binary classification problem addressed by many studies using various sources of information. Neurophysiological signals such as Functional Magnetic Resonance Imaging (fMRI) [28, 30] and Event Related Potentials (ERP) [47] were investigated for this task. For instance, the work of [1] used electroencephalography (EEG) features extracted through wavelet transformation and they achieved a correct detection rate of 86%. Another brain-imaging technique, functional near-infrared spectroscopy (fNIRS), that measures brain activity through hemodynamic responses associated with neuron behaviour was also examined [24]. They achieved the average classification accuracy of 83.44% with subject-specific Support Vector Machine classifiers.

[3]The collected DESDEDA dataset is available to the research community at: https://www.cl.cam.ac.uk/research/rainbow/projects/desdeda/

Other approaches used cues from videos, for example, Meservy et al. [33] built an automated system that can infer deception or truthfulness from a set of features extracted from head and hands movements captured in a video, yielding 71% classification accuracy using both Support Vector Machines and a neural network.

In the majority of the lie detection settings several physiological signals including respiration, skin conductance, blood pressure, and pulse rate were employed [59]. The study of [4] reports 86.5% accuracy when fNIRS measurements were combined with physiological measures.

Several works tried to detect deception during games. For example, the study of [32] based the deceit detection on EEG measures during a poker-like card game and showed that the Wavelet analysis revealed significant differences between deceptive and truthful responses. However, they did not develop a classification model. Sung et al. [55] used a combination of physiological features (namely, skin conductance peaks (from both hands), voice pitch variation and heart rate variability) to detect stress and lying during the Poker game. They developed simple linear classifiers and identified high stress situations with the accuracy of 82% and detected deception with about 71% accuracy. They further reported that the skin conductance peaks were the most correlated features in the cases of All-In play and stressful situations in general. It is important to note that the players in their study played real live-money games of no-limit. Others [12, 61] investigated EDA in gaming but not for deception detection. For instance, Drachen et al. [12] researched correlations between heart rate, EDA and player experience in first-person shooter games. However, to the best of our knowledge, there was no previous work that attempted at detecting deception solely from EDA in a dyadic game interaction setting.

Suspicion or trust detection is a more recent and less researched field. Previous work focused mostly on videos and analysed non-verbal behaviours achieving above human detection accuracy [29]. Several studies have investigated the relationship between EDA and other dyadic/group related outcomes such as group satisfaction or engagement. For example, it has been found that team members' synchrony in EDA is associated with group satisfaction [9], periods of stress, excitement, or high levels of engagement [3, 37, 40] and it is also related to tension and negative affect [35]. Such dyadic/group related outcomes may have an effect on trust which also suggests that it might be possible to use EDA to predict suspicion in a dyadic interaction setting. The ability to accurately detect suspicion from EDA would allow longer-term analysis of trust-related behaviours in multiple contexts (e.g., human-robot interaction). However, to the best of our knowledge, to date there has been no attempt to detect suspicion or trust from EDA.

3 THE STUDY

3.1 Motivation

Our motivation to choose a card game for detection of deception and suspicion was two-fold. Firstly, games present a very common and realistic scenario where people lie and are suspicious without negative consequences. Secondly, a game with real participants and their direct interactions allows more natural reactions of players as compared to computer-based games and thus is expected to provide

more realistic results. With this motivation we designed a study to answer the three research questions described in Sec. 1.

3.2 Card game *Cheat*

We chose the card game *Cheat* (also called *Bluff* or *I Doubt It*) as it has very simple rules which allows players to better focus on their actions of deception and suspicion. To simplify the experiment we focused on a two-player variant of this game with the following rules.

All cards[4] are evenly dealt out (by the experimenter) to the two players and they can see their own cards. The goal of the game is to get rid of all cards at hand. First player calls out the suit (diamonds/clubs/spades/hearts) and discards one card face down on the discard pile. The suit called out by the first player is the *true suit* for the current discard pile. Players then take alternate turns discarding one card each time and calling out the same suit as the player in the first turn. Since the cards are discarded face down, players can cheat to finish the game faster by discarding a card of different suit than required. If one player suspects the other player, he/she can challenge the play by calling "Cheat!". Then the card played by the challenged player is exposed and one of two things happens: (i) if the exposed card is of the suit that was called, the challenger must pick up the whole discard pile; or (ii) if the card is different from the called suit, the person who played the card must pick up the whole discard pile. The player who did not pick up the pile begins the new round by discarding one card and calling out a suit that becomes the true suit for the new discard pile. The game ends when one of the players gets rid of all his/her cards at hand.

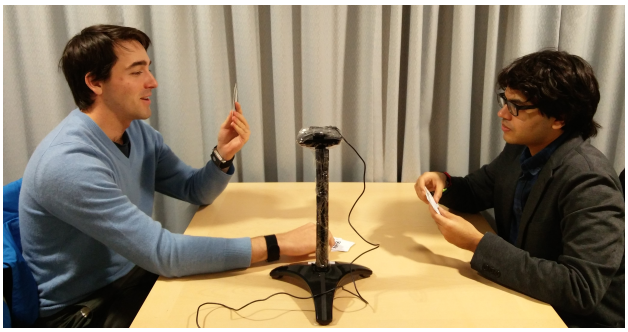

Figure 1: Experimental setup: participants playing the card game *Cheat*. Each player is wearing two Affectiva Q EDA sensors (one on the left and one on the right wrist). The discard pile is recorded by a web-camera.

3.3 Sensors

As can be seen in Figure 1, each player was wearing two Affectiva Q EDA sensors [2] (one on the left and one on the right wrist) during the game. The Q sensor measures electrical conductance (in units of μS) across the skin by passing a minuscule amount of current between two electrodes that are in contact with the skin. Each Q sensor provides the following data: EDA, skin temperature, and 3-axis of acceleration of the wrist over time. All these measures were sampled at 32 Hz (the maximum possible rate of the Affectiva Q sensors). We decided to use only EDA measurements for deception and suspicion detection, as we did not want to constrain the developed detection system to the specific use case of a card game which would be the case if acceleration measurements were also used.

The discard pile of cards was recorded by a web-camera (along with the audio) and the back side of every card was labelled with a QR code corresponding to the suit on the other side of the card. This allowed later reconstruction of events (whether a player lied or told truth while discarding a card) and their localisation in time which was necessary for correct annotation of the EDA data.

Prior to the experiment, all four Q sensors were time-synchronised with the system time used by the web-camera, ensuring that identical global time was used by all Q sensors and camera timestamps.

3.4 Data collection

We built an in-house dataset named Deception and Suspicion Detection from EDA (DESDEDA) by recruiting 20 participants (5 female and 15 male) to play the game *Cheat* and to answer a post-study questionnaire.

The participants were aged 19–32 and came from various cultural and educational backgrounds. All but three participants said they were right-handed (participants with IDs 09, 11, 20 were left-handed). The participants were arranged into 10 pairs so that each participant played the game only once. Before the experiment started they were informed about its procedure, game rules, and their rights by verbal introduction and through a signed consent form. However, the true goal of this study was not disclosed to them until after the experiment in order to avoid artificial behaviour (e.g., reluctance to bluff and blame the other player) and to allow more natural reactions during the game. Prior to the data collection phase the participants were given time to freely play the game to familiarise themselves with its rules. In order to settle down the measured EDA values, before the game started, participants rested for a 2-minute baseline period while listening to relaxing music. Then, one or multiple games up to a maximum duration of 30 minutes were recorded (4 streams of EDA and video recording of the discard pile).

In the post-study questionnaire the participants were asked to provide information about their gender, age, handedness[5], and take a short personality test[6]. Also, they answered 3 self-report questions: *How many times did you play this game before? On average out of 10 opportunities to lie, how many times do you think you really lied? On average out of 10 opportunities to say "Cheat", how many times do you think you really said "Cheat"?* These questions were used to collect data for future research investigations - e.g., how the player's game experience affects the detection accuracy, how well the counts of actual deception/suspicion actions match the self-reported counts for different personality types, etc.

[4] A stripped deck (US) or shortened pack (UK) of 32 cards.

[5] Tendency to use either right or left hand more naturally than the other.

[6] 20-item measure of BIG5 personality (in terms of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), available at: https://discovermyprofile.com/minIPIP/introduction.html

3.5 Data segmentation and annotation

Firstly, we defined three event types: *deception event* (player discarded the card lying), *truth event* (player discarded the card telling truth), and *suspicion event* (player called "Cheat"). Using the audiovisual recordings of the discard pile we determined the times of these three events for all players and manually annotated EDA measurements as shown in Figure 2.

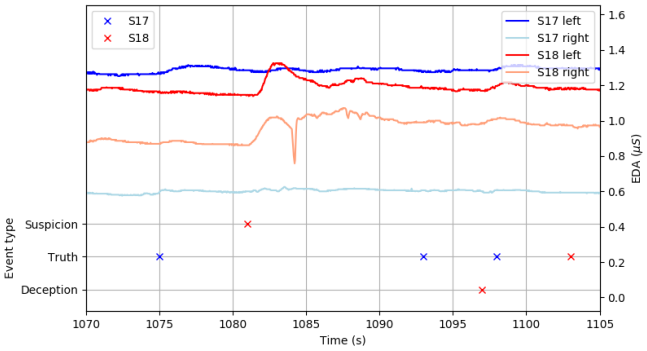

Figure 2: Annotated EDA signals from left and right hands of participants with IDs 17 and 18. EDA is annotated with three event types (Suspicion, Truth, and Deception event). The suspicion event is followed by a strong response in EDA from both hands of the corresponding player while the deception event is followed by a much weaker response.

Next, the time intervals (epochs) from which the EDA signal was used to extract features for associated events were determined. This process was motivated by the work of [54] that demonstrates that the cues in EDA are present before the event of interest. The segmentation procedure for various types of epochs follows.

3.5.1 Deception and truth epochs. The *deception epoch* D_e associated with the deception event k (or the *truth epoch* in case of the truth event) was defined as the time interval

$$D_e = [\max(t_{k-1}, t_k - \tau_{MDEL}); t_k] \tag{1}$$

where t_k is the time of the associated deception/truth event and t_{k-1} is the time of the previous event. If the previous event was not deception/truth event, then this epoch along with the event k was ignored since it was the first one in the game or the first one after the pile was picked up and it contained noise as the players were not focused yet. τ_{MDEL} is the *maximum deception epoch length* and it prevents epochs from being too long. This thresholding was necessary because in most cases the long duration between two consecutive events $k-1$ and k meant a player's distraction in the earlier stage of the time interval between these two events. Since it was not clear how to set the parameter τ_{MDEL}, we used cross-validation to find its optimal value in the discrete range $\tau_{MDEL} \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5\}s$. This range was estimated by looking at the distribution of lengths of time intervals between two consecutive deception/truth events over all players.

3.5.2 Suspicion and trust epochs. The *suspicion epoch* S_e associated to the suspicion event k was defined as the time interval

$$S_e = [\max(t_{k-1}, t_k - \tau_{MSEL}); t_k] \tag{2}$$

where t_k is the time of the associated suspicion event and t_{k-1} is the time of the previous deception/truth event. τ_{MSEL} is the *maximum suspicion epoch length* and it serves the same purpose as the threshold τ_{MDEL} and its value was also chosen by cross-validation over the discrete range $\tau_{MSEL} \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5\}s$. Analogously, this range was estimated by looking at the distribution of lengths of time intervals between the deception/truth event and the consecutive suspicion event over all players.

We further defined an imaginary *trust epoch* T_e that starts at a deception/truth event $k-1$. However, there is no observable event that would mark the end of such a trust epoch (when the player did not call "Cheat" and trusted the other player). To tackle this issue we made an assumption that the length of the trust epoch is approximately τ_{MSEL} if the next deception/truth event k is far enough (namely, further than $2 \times \tau_{MSEL}$ from the start of the trust epoch) and it is the half of the distance between the start of the trust epoch and the next deception/truth event otherwise. In other words, the *trust epoch* T_e following the deception/truth event $k-1$ at time t_{k-1} was defined as the time interval

$$T_e = \left[t_{k-1}; \min\left(t_{k-1} + \tau_{MSEL}, \frac{t_k + t_{k-1}}{2}\right)\right] \tag{3}$$

where t_k is the time of the next deception/truth event. The end of the trust epoch can be thought of as an imaginary trust event.

3.5.3 Epoch delay. Both endpoints of each epoch (of all four types) were further delayed by δ because there is a delay between a stimulus and skin conductance response [46]. This is also confirmed by Figure 2 where we can observe that the response in EDA is delayed after event occurrence. The value of the delay δ was observed to be 1–4 s, and so we determined its value by cross-validation over the discrete range $\delta \in \{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}s$.

Lastly, each epoch was labelled according to its type resulting in the distribution shown in Table 1. The labelled dataset excludes the deception and truth epochs from one participant (ID 05) who confused card suits during the game which made the experimenter unable to correctly label his/her epochs (this was not a problem for suspicion detection task and so the suspicion and trust epochs of this participant were kept). As we can see from Table 1, the collected data are unbalanced with a strong bias towards trust labels for suspicion detection task. To mitigate this issue, the random oversampling technique was applied to the data (as further described in Sec. 5).

Table 1: Distribution of four types of labelled epochs before balancing.

	Deception	Truth	Suspicion	Trust
#labelled epochs	496	635	300	1180

4 FEATURE EXTRACTION

Similarly to [7], as a preprocessing step, we performed epoch normalisation to allow for comparison between different epochs and between different study participants. EDA measurements from each epoch were scaled into the range $[0, 1]$ independently. In order to

remove high-frequency noise, we applied a 5^{th} order low-pass Butterworth filter with the cut-off frequency of 3 Hz using the Python library *SciPy* [26].

Inspired by [43], we chose a set of six features including: 1) Mean; 2) Standard deviation; 3) Mean of the absolute values of the first differences of the raw signal; 4) Mean of the absolute values of the first differences of the normalised signal; 5) Mean of the absolute values of the second differences of the raw signal; and 6) Mean of the absolute values of the second differences of the normalised signal. For each epoch, these six features were extracted from EDA signal from each hand and consequently, right-hand features were appended to left-hand features resulting in a 12-dimensional feature vector per epoch. This procedure was identical for all epoch types.

5 CLASSIFICATION

We approached the detection of deception and suspicion as two separate binary classification tasks. For each task we report person-independent and person-dependent testing accuracies as well as the optimal hyperparameters for detection on-the-fly. For training, hyperparameter tuning, and testing we used Support Vector Machine (SVM) classifier with linear kernel implemented in the Python framework *scikit-learn* [39]. All results from this section are summarised in Table 2.

5.1 Deception detection

Firstly, the previously discussed biased nature of the data was addressed by random over-sampling of the minority class using the toolbox *imbalanced-learn* [31]. This resulted in 635 samples for deception epoch, 635 samples for truth epoch, 1180 samples for suspicion epoch, and 1180 samples for trust epoch. Next, a nested cross-validation[7] was used to evaluate the performance of the developed method in person-independent and person-dependent manner.

5.1.1 Person-independent detection. The data from all participants were used to develop and test a general model capable of deception detection independent of a person being investigated. In this case the outer loop of the nested cross-validation was leave-one-subject-out (LOSO) cross-validation and was used for testing while the inner loop was 5-fold cross-validation and served to tune 3 hyperparameters: δ_D, τ_{MDEL}, and C_D. The parameters δ_D (deception/truth epoch delay) and τ_{MDEL} (maximum deception epoch lenght) were optimised in ranges defined in Sec. 3.5. The SVM's regularisation parameter C_D was tuned in the discrete range $C_D \in \{2^{-30}, 2^{-29}, \ldots, 2^{15}\}$. The mean testing accuracy (and the standard deviation) over all LOSO folds for person-independent deception detection was 52 ± 7 %.

5.1.2 Person-dependent detection. In this case a person-specific model was trained, tuned, and tested on each participant separately and so the outer loop of the nested cross-validation was changed to 5-fold cross-validation. Otherwise, the procedure was the same as in the person-independent detection (same set of hyperparameters was optimised but this time for each participant separately). The testing accuracies for each participant are shown in Figure 3 (left). The mean testing accuracy (and the mean standard deviation) over all participants for person-dependent deception detection was 63 ± 10 %.

5.1.3 Optimal parameters for detection on-the-fly. Lastly, optimal parameters for detection of deception on-the-fly were determined using LOSO cross-validation on the whole dataset, and consequently, a model ready for classification on-the-fly was trained on the whole dataset. The best hyperparameters that maximise the validation accuracy were found to be $(\delta_D, \tau_{MDEL}, C_D) = (3.0s, 1.0s, 2^3)$. Figure 4 (left) illustrates the search space used to determine these optimal hyperparameters, namely, it shows the mean (over all folds) validation accuracy for various parameter settings with hyperparameter C_D already optimised for each pair (δ_D, τ_{MDEL}).

5.2 Suspicion detection

For classification of suspicion/trust epochs we followed the same procedure as for deception detection with the only difference that the hyperparameter τ_{MDEL} was replaced by τ_{MSEL} and parameters δ_D, C_D were relabelled to δ_S, C_S. The testing accuracy was 57 ± 12 % and 76 ± 8 % for person-independent and person-dependent classification respectively. Figure 3 (right) shows testing accuracies for all participants in the person-dependent case. The optimal parameters that maximise the validation accuracy for suspicion detection on-the-fly were found to be $(\delta_S, \tau_{MSEL}, C_S) = (3.0s, 3.5s, 2^{-5})$ and the mean validation accuracy for various parameter settings (with C_S already optimised for each pair (δ_S, τ_{MSEL})) is shown in Figure 4 (right).

Table 2: Summary of results from deception/truth and suspicion/trust classification tasks for person-independent (PI) and person-dependent (PD) methods. In each case, the baseline accuracy is 50%.

		Deception	Suspicion
Testing accuracy [%]	PI	52 ± 7	57 ± 12
	PD	63 ± 10	76 ± 8
Best parameters	$\delta_{\{D,S\}}$ [s]	3.0	3.0
	$\tau_{M\{D,S\}EL}$ [s]	1.0	3.5
	$C_{\{D,S\}}$	2^3	2^{-5}

6 THE IMPORTANCE OF FEATURES & ASYMMETRY IN EDA

6.1 Feature importance

To answer our third research question (Sec. 1), we assessed the relative importance of the chosen features. Most importantly, we compared the informativeness of the extracted features between left and right hands. In particular, we examined SVM weights when the person-independent model was trained on the whole dataset using the optimal parameters determined according to Sec. 5.1.3.

Since we used linear-kernel SVM, there was no kernel transformation to a higher dimensional feature space and so the trained weights could be used for feature ranking, as suggested in [52] and studied in detail by [6]. The reasoning is that the larger the magnitude $|w_i|$ of weight w_i is, the larger influence the i^{th} feature has on

[7] Nested cross-validation consists of an inner and outer loop where each training fold of the outer loop is split into training and validation folds of the inner loop.

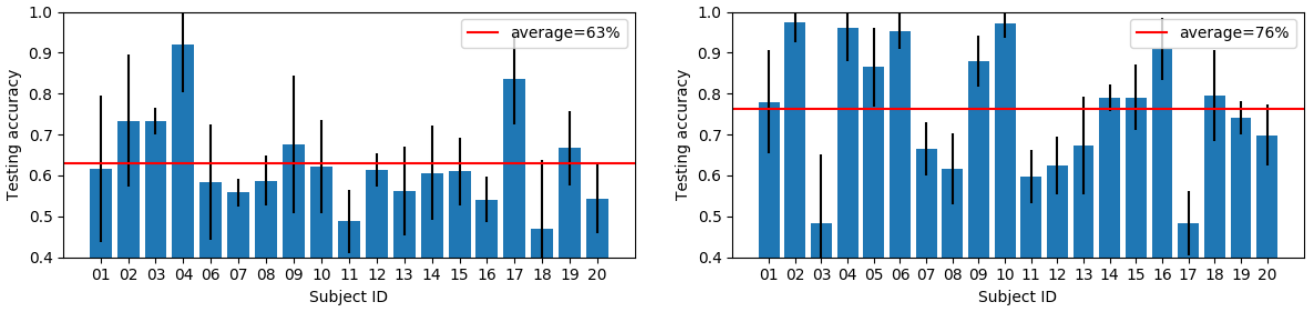

Figure 3: Testing accuracies with standard deviations for person-dependent *deception* (left) and *suspicion* (right) detection for all study participants (excluding ID 05 for deception detection, see Sec. 3.5).

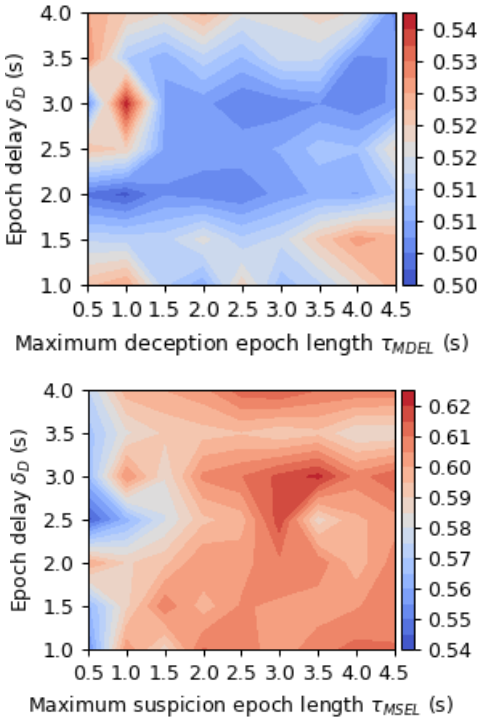

Figure 4: Mean validation accuracy from person-independent (LOSO) cross-validation on the whole dataset for *deception* (top) and *suspicion* (bottom) detection task with SVM hyperparameters C_D and C_S already optimised. The best hyperparameters were found to be $(\delta_D, \tau_{MDEL}, C_D) = (3.0s, 1.0s, 2^3)$ and $(\delta_S, \tau_{MSEL}, C_S) = (3.0s, 3.5s, 2^{-5})$.

the predictions of the classifier. The work [19] further suggests to use squares of weights as a ranking criterion to magnify relative differences between weights. Figure 5 shows squares of trained SVM weights corresponding to features extracted from EDA signals from both hands, for deception and suspicion detection tasks.

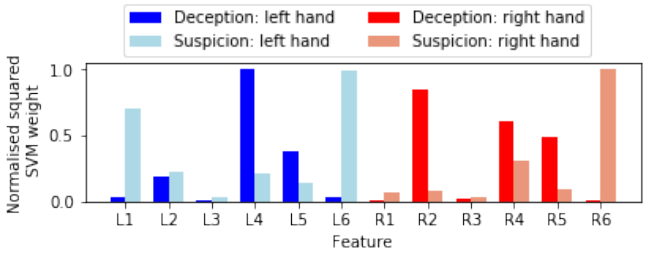

Figure 5: Feature ranking in terms of normalised squares of SVM weights trained on the whole dataset, for both *deception* (dark colours) and *suspicion* (bright colours) detection tasks. L1–L6 (blue) and R1–R6 (red) denote features extracted from EDA signals from left and right hand respectively. For definition of types of features 1–6 see Sec. 4.

6.2 Asymmetry in EDA

For 10 study participants the difference between individual EDA measurements from left and right hand never changed sign during the whole game and for the other 10 participants the difference was consistent in sign for almost the whole game (regardless of periods of deception, truth, suspicion, or trust). Therefore, there was no point in evaluating the left-right difference in an epoch-wise manner and so similarly to [45], we calculated average EDA level from each wrist for every participant (omitting the baseline period) and subtracted the left hand from the right hand mean value to obtain the mean difference Δ_{L-R}. Figure 6 shows the distribution of Δ_{L-R} comprising all participants of the study.

7 ANALYSIS AND DISCUSSION

7.1 Detection of deception and suspicion

The results in Table 2 clearly show that the detection of suspicion from EDA is more reliable than detection of deception. Detecting lie from skin conductance as a single source of information is challenging and as mentioned in Sec. 2 other methods that achieved higher detection accuracies often combined multiple sources of input. We also noticed that the differences between deception and truth events were much less visible than those between suspicion

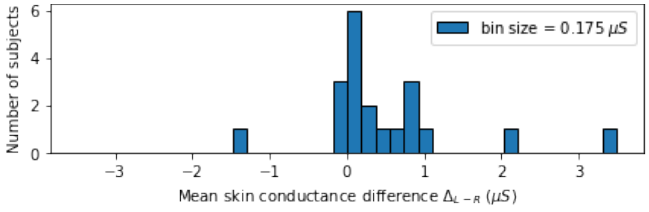

Figure 6: Distribution of mean skin conductance difference Δ_{L-R} between EDA measured on left and right hand during the whole game, omitting the baseline period. All 20 participants of the study are included.

and trust events. For example, as illustrated by Figure 2, the skin conductance response to the suspicion event has about 3-times larger magnitude than the response to the deception event. This probably also contributed to the lower deception detection performance.

Comparing the person-independent (PI) and person-dependent (PD) detections, we can conclude that the PI classification is a more challenging task than PD which is reflected in lower testing accuracies for both deception and suspicion detection tasks. Next, as shown by Figure 3, in the PD case the testing accuracies vary a lot between participants (47%–92% and 48%–98% for deception and suspicion respectively) which suggests that it is much more difficult to develop a reliable model for some people than for others. In other words, the nature of EDA signals is highly person-specific. We can also see that for some participants the testing accuracy was even below the baseline – in this case the chance level[8] of 50% for both detection tasks. This might be caused by the fact that internal factors such as hydration and medications can affect EDA measurements. Moreover, there are people who have essentially no measurable EDA [42].

As can be seen from Table 2, all 4 mean testing accuracies are above the baseline. However, the accuracy in the PI case, and especially, for deception detection is very close to the baseline. Also, the standard deviations are relatively large. One possible reason might be the fact that the game environment is very challenging for automatic detection of deception and suspicion as it is very dynamic with a wide range of response times. Moreover, players were not constrained not to talk which often caused considerable distractions. Thus, for future studies it may be appropriate to reconsider the study design.

The obtained optimal values of parameters τ_{MDEL} and τ_{MSEL} for detection on-the-fly suggest that the most informative EDA signal for deception detection is captured within 1 second before the deception/truth event and the most informative EDA signal for suspicion detection is captured within 3.5 seconds before the suspicion or imaginary trust event occurs. This is also supported by Figure 2 that illustrates that the response to the suspicion event is longer than the response to the deception event. The optimisation of epoch delays δ_D and δ_S resulted in the same value of 3.0 seconds. This indicates that the skin conductance response relevant for deception detection is delayed by the same amount of time as the skin conductance response relevant for suspicion detection. This further suggests that, in this particular dyadic game context, the informative EDA response can be captured after the same amount of time for any type of triggering stimulus (deception/truthfulness/suspicion/trust).

7.2 Feature importance

As can be seen from Figure 5, features extracted from EDA signals from both hands are important and this is the case for both deception and suspicion detection tasks. Also, it can be observed that for some feature types[9] there is a symmetry between the importance of left-hand and right-hand features. For example, features of types 4 and 5 from both hands seem to be most informative for deception detection and the feature of type 6 for suspicion detection. In other words, the mean time-changes in the normalised EDA signal over the epoch and the mean acceleration of the raw EDA signal over the epoch are most important for deception detection while the mean acceleration of the normalised EDA signal over the epoch is most informative for suspicion detection. However, such a left-right symmetry in feature importance does not hold for all feature types as illustrated by feature type 2 (variance in EDA over epoch) for deception detection and feature type 1 (mean EDA over epoch) for suspicion detection whose importances significantly differ between left and right hand. All these results confirm the conclusions of [41] that measurements from multiple points of EDA arousal are more informative than the traditional measurements taken *only* from the single non-dominant hand.

7.3 Asymmetry in EDA

Our results confirm the existence of asymmetry in EDA measured on the left and the right wrists. As can be seen from Figure 6 the magnitude of the left-hand EDA dominates over the right-hand EDA for the majority of participants. Specifically, 16 out of 20 study participants had positive mean skin conductance difference Δ_{L-R} between left and right hand. According to [41], this could be interpreted as follows: participants with dominant right-hand EDA perceived the whole game more anxiously and stressfully than participants with dominance in left-hand EDA. However, as described in Sec. 1, making such general conclusions is difficult and may be flawed.

Looking at the obtained results it seems that the dominance in skin conductance did not depend on the participant's handedness as there was 1 left-handed participant with negative Δ_{L-R} and 2 left-handed participants with positive Δ_{L-R} and also there were 3 right-handed participants with negative Δ_{L-R}. However, data from more participants would need to be collected to make more informative conclusions.

It is important to note that the above-described findings were observed irrespective of participant's position (left/right side of the table) which means that it is unlikely that they were caused by some systematic differences between Q sensors.

[8] Expected accuracy if classes are assigned by random guessing.

[9] As defined in Sec. 4.

8 CONCLUSION AND FUTURE WORK

8.1 Conclusion

This work presented a novel dataset for automatic detection of deception and suspicion from EDA measurements. Using SVM classifiers we developed models to automatically detect deception and suspicion on-the-fly. Our experimental results show that the detection of suspicion is more reliable than the detection of deception and that person-dependent models perform better than person-independent ones. Next, we found that the optimal interval of informative EDA signal is about 3.5-times shorter for deception detection than for the suspicion detection task and that the EDA signal relevant for deception/suspicion detection can be captured after approximately 3.0 seconds once a stimulus event has occurred, and regardless of the stimulus type (deception/truthfulness/suspicion/trust). Results from feature ranking show that features extracted from EDA from both hands are important for deception and suspicion classification tasks and that the importance of some feature types is symmetric between left-hand and right-hand features while it is asymmetric for other types of features. We also verified that there is an asymmetry in EDA measured on left and right wrist.

8.2 Limitations

Despite the promising results and interesting findings, there are some limitations to the methodologies we employed. Firstly, EDA is a very non-specific measure that has been related to stress, anxiety, or any kind of arousal. Therefore, the developed models might not be suitable for detection in the wild where a person may experience many different kinds of emotions and physiological responses. This work is thus bound to the specific task at hand, namely, a dyadic game interaction involving deception and suspicion.

The aim of this work was to investigate the reliability of deception/suspicion detection based purely on EDA. As described in Section 2, studies that used multiple modalities achieved higher deception detection accuracies. We therefore recommend the usage of a combination of sensors aggregating multiple modalities in order to develop more reliable deception/suspicion detectors. Another consequence of relying solely on the EDA measurements when detecting deception/suspicion is that the hand movements can confound the EDA measurements. Future work should therefore investigate to what extent the hand movements affect the detection accuracies.

It is also important to note that, in this initial study, we decided not to decompose the measured EDA signal into phasic (rapidly changing) and tonic (slowly changing) components, since the previous works [9, 21–23, 49, 57] were not conclusive on this. Future works should also investigate these aspects in more detail.

8.3 Future work

This work opens up several directions for further research. Firstly, using the method described in [17], the measured EDA could be decomposed into phasic and tonic components in order to investigate which of them is more informative for detection of deception and suspicion. Consequently, this could improve our understanding of the variability of EDA over time and over subjects.

Also, it would be interesting to compare the results obtained when RBF kernel SVMs are used and when other over-sampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) [8] or Adaptive Synthetic (ADASYN) sampling method [20] are employed.

Next, the developed models could be tested on-the-fly by conducting another dyadic game interaction study.

Another research avenue could use the metadata from post-study questionnaires and explore relationships between participants' personality traits and their deceitful/truthful and suspicious/trustful behaviours as well as the deception/suspicion detection accuracies. As a starting point, one can look at the relationship between certain personality traits and the frequency of deception and suspicion events. As shown in Figure 7, the deception rate of participants is negatively correlated with the conscientiousness trait (Pearson correlation coefficient $\rho = -0.44$) while it is positively correlated with the agreeableness trait ($\rho = 0.36$). This is in accordance with other research findings reporting that conscientious individuals are less likely to lie [60] as the conscientiousness trait was found to be associated with higher levels of honesty in general [16]. However, the positive correlation between deception and the agreeableness trait contradicst previous research findings [16, 53, 60] reporting that highly agreeable individuals are less likely to lie. The prototype system and the data obtained during the dyadic game interactions presented in this paper can therefore be used to further investigate such research hypotheses and questions.

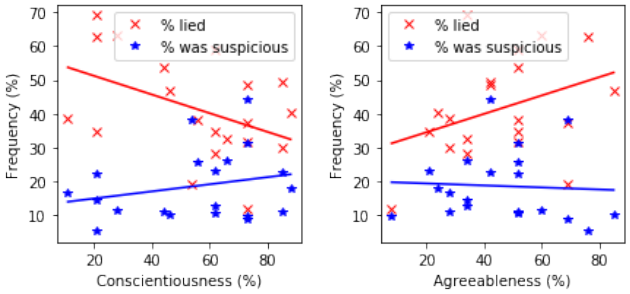

Figure 7: Frequency of deception (red) and suspicion (blue) events against personality traits: Conscientiousness (left) and Agreeableness (right). Each datapoint corresponds to one study participant.

ACKNOWLEDGMENTS

We would like to thank 1) all the participants for volunteering to take part in this study, and 2) Jan Ondras' mother Iveta Ondrasova for helping us with the manual annotation of the EDA measurements.

REFERENCES

[1] Vahid Abootalebi, Mohammad Hassan Moradi, and Mohammad Ali Khalilzadeh. 2009. A new approach for EEG feature extraction in P300-based lie detection. *Computer methods and programs in biomedicine* 94, 1 (2009), 48–57.

[2] Q Affectiva. [n. d.]. Sensor 2.0 Datasheet, 2012.

[3] Modupe Akinola. 2010. Measuring the pulse of an organization: Integrating physiological measures into the organizational scholar's toolbox. *Research in Organizational Behavior* 30 (2010), 203–223.

[4] M Raheel Bhutta, Melissa J Hong, Yun-Hee Kim, and Keum-Shik Hong. 2015. Single-trial lie detection using a combined fNIRS-polygraph system. *Frontiers in psychology* 6 (2015).

[5] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer Science & Business Media.

[6] Janez Brank, Marko Grobelnik, Natasa Milic-Frayling, and Dunja Mladenic. 2002. Feature selection using support vector machines. *WIT Transactions on Information and Communication Technologies* 28 (2002).

[7] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2017. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. *IEEE Transactions on Affective Computing* (2017).

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[9] Prerna Chikersal, Maria Tomprou, Young Ji Kim, Anita Williams Woolley, and Laura Dabbish. 2017. Deep Structures of Collaboration: Physiological Correlates of Collective Intelligence and Group Satisfaction.. In *CSCW*. 873–888.

[10] Ana Cristina Costa, Robert A Roe, and Tharsi Taillieu. 2001. Trust within teams: The relation with performance effectiveness. *European journal of work and organizational psychology* 10, 3 (2001), 225–244.

[11] Hugo D Critchley. 2002. Electrodermal responses: what happens in the brain. *The Neuroscientist* 8, 2 (2002), 132–142.

[12] Anders Drachen, Lennart E Nacke, Georgios Yannakakis, and Anja Lee Pedersen. 2010. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*. ACM, 49–54.

[13] Johan Engström, Emma Johansson, and Joakim Östlund. 2005. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 8, 2 (2005), 97–120.

[14] Szymon Fedor, Peggy Chau, Nicolina Bruno, Rosalind W Picard, Joan Camprodon, and T Hale. 2016. Can We Predict Depression From the Asymmetry of Electrodermal Activity? *Journal of Medical Internet Research* 18, 12 (2016).

[15] Don C Fowles, Margaret J Christie, Robert Edelberg, William W Grings, David T Lykken, and Peter H Venables. 1981. Publication recommendations for electrodermal measurements. *Psychophysiology* 18, 3 (1981), 232–239.

[16] Omri Gillath, Amanda K Sesko, Phillip R Shaver, and David S Chun. 2010. Attachment, authenticity, and honesty: dispositional and experimentally induced security can reduce self-and other-deception. *Journal of personality and social psychology* 98, 5 (2010), 841.

[17] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2016. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering* 63, 4 (2016), 797–804.

[18] Nurit Gronau, Gershon Ben-Shakhar, and Asher Cohen. 2005. Behavioral and physiological measures in the detection of concealed information. *Journal of Applied Psychology* 90, 1 (2005), 147.

[19] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46, 1 (2002), 389–422.

[20] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 1322–1328.

[21] 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*. Springer, 125–134.

[22] Javier Hernandez, Ivan Riobo, Agata Rozga, Gregory D Abowd, and Rosalind W Picard. [n. d.]. How Easy Are Children to Engage during Child-Adult Play? Using Electrodermal Activity as a Marker. ([n. d.]).

[23] 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 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 307–317.

[24] Xiao-Su Hu, Keum-Shik Hong, and Shuzhi Sam Ge. 2012. fNIRS-based online deception decoding. *Journal of Neural Engineering* 9, 2 (2012), 026012.

[25] Kenneth Hugdahl. 1984. Hemispheric asymmetry and bilateral electrodermal recordings: A review of the evidence. *Psychophysiology* 21, 4 (1984), 371–393.

[26] Eric Jones, Travis Oliphant, and Pearu Peterson. 2014. SciPy: open source scientific tools for Python. (2014). http://www.scipy.org/ [Online; accessed 10/01/2018].

[27] J Kayser and G Erdmann. 1991. Bilateral electrodermal activity: Effects of lateralized visual input of emotional stimuli. *Journal of Psychophysiology* 5 (1991), 110–111.

[28] Daniel D Langleben, James W Loughead, Warren B Bilker, Kosha Ruparel, Anna Rose Childress, Samantha I Busch, and Ruben C Gur. 2005. Telling truth from lie in individual subjects with fast event-related fMRI. *Human brain mapping* 26, 4 (2005), 262–272.

[29] Jin Joo Lee, W Bradley Knox, Jolie B Wormwood, Cynthia Breazeal, and David DeSteno. 2013. Computationally modeling interpersonal trust. *Frontiers in psychology* 4 (2013).

[30] Tatia Lee, Ho-Ling Liu, Li-Hai Tan, Chetwyn CH Chan, Srikanth Mahankali, Ching-Mei Feng, Jinwen Hou, Peter T Fox, and Jia-Hong Gao. 2002. Lie detection by functional magnetic resonance imaging. *Human brain mapping* 15, 3 (2002), 157–164.

[31] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. http://jmlr.org/papers/v18/16-365.html

[32] Anna Caterina Merzagora, Scott Bunce, Meltem Izzetoglu, and Banu Onaral. 2006. Wavelet analysis for EEG feature extraction in deception detection. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 2434–2437.

[33] Thomas O Meservy, Matthew L Jensen, John Kruse, Judee K Burgoon, Jay F Nunamaker, Douglas P Twitchell, Gabriel Tsechpenakis, and Dimitris N Metaxas. 2005. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems* 20, 5 (2005), 36–43.

[34] Yanik Miossec, Marie-Claude Catteau, Esteve Freixa i Baqué, and Jean-Claude Roy. 1985. Methodological problems in bilateral electrodermal research. *International journal of psychophysiology* 2, 4 (1985), 247–256.

[35] Dan Mønster, Dorthe Døjbak Håkonsson, Jacob Kjær Eskildsen, and Sebastian Wallot. 2016. Physiological evidence of interpersonal dynamics in a cooperative production task. *Physiology & behavior* 156 (2016), 24–34.

[36] Makoto Nakayama. 2002. Practical use of the concealed information test for criminal investigation in Japan. *Handbook of polygraph testing* (2002), 49–86.

[37] Reiner Nikula. 1991. Psychological correlates of nonspecific skin conductance responses. *Psychophysiology* 28, 1 (1991), 86–90.

[38] Nargess Nourbakhsh, Yang Wang, Fang Chen, and Rafael A Calvo. 2012. Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. ACM, 420–423.

[39] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[40] Suzanne J Peterson, Christopher S Reina, David A Waldman, and William J Becker. 2015. Using physiological methods to study emotions in organizations. In *New ways of studying emotions in organizations*. Emerald Group Publishing Limited, 1–27.

[41] Rosalind W Picard, Szymon Fedor, and Yadid Ayzenberg. 2016. Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Review* 8, 1 (2016), 62–75.

[42] Rosalind W Picard, Szymon Fedor, and Yadid Ayzenberg. 2016. Response to Commentaries on âĂIJMultiple Arousal Theory and Daily-Life Electrodermal Activity AsymmetryâĂİ. *Emotion Review* 8, 1 (2016), 84–86.

[43] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence* 23, 10 (2001), 1175–1191.

[44] Ming-Zher Poh, Tobias Loddenkemper, Claus Reinsberger, Nicholas C Swenson, Shubhi Goyal, Mangwe C Sabtala, Joseph R Madsen, and Rosalind W Picard. 2012. Convulsive seizure detection using a wrist-worn electrodermal activity and accelerometry biosensor. *Epilepsia* 53, 5 (2012).

[45] Ming-Zher Poh, Tobias Loddenkemper, Nicholas C Swenson, Shubhi Goyal, Joseph R Madsen, and Rosalind W Picard. 2010. Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 4415–4418.

[46] Psychlab. [n. d.]. Skin conductance explained. http://www.psychalb.com/SC_explained.html http://www.psychlab.com/SC_explained.html [Online; accessed 10/01/2018].

[47] J Peter Rosenfeld. 2002. Event-related potentials in the detection of deception, malingering, and false memories. (2002).

[48] Harold A Sackeim, Mark S Greenberg, Andrew L Weiman, Ruben C Gur, Jean Pierre Hungerbuhler, and Norman Geschwind. 1982. Hemispheric asymmetry in the expression of positive and negative emotions. *Archives of Neurology* 39 (1982), 210–218.

[49] Akane Sano, Rosalind W Picard, and Robert Stickgold. 2014. Quantitative analysis of wrist electrodermal activity during sleep. *International Journal of Psychophysiology* 94, 3 (2014), 382–389.

[50] Thomas Schill, Carmen Toves, and Nerella V Ramanaiah. 1980. Interpersonal trust and coping with stress. *Psychological Reports* (1980).

[51] Hans Schliack and Roland Schiffter. 1979. Neurophysiologie und-pathophysiologie der Schweißsekretion. In *Normale und Pathologische Physiologie der Haut II*. Springer, 349–458.

[52] Vikas Sindhwani, Pushpak Bhattacharya, and Subrata Rakshit. 2001. Information theoretic feature crediting in multiclass support vector machines. In *Proceedings of the 2001 SIAM International Conference on Data Mining*. SIAM, 1–18.

[53] Kasey Stanton, Stephanie Ellickson-Larew, and David Watson. 2016. Development and validation of a measure of online deception and intimacy. *Personality and Individual Differences* 88 (2016), 187–196.
[54] Sabine Ströfer, Matthijs L Noordzij, Elze G Ufkes, and Ellen Giebels. 2015. Deceptive intentions: can cues to deception be measured before a lie is even stated? *PloS one* 10, 5 (2015), e0125237.
[55] Michael Sung and Alex Pentland. 2005. PokerMetrics: Stress and lie detection through non-invasive physiological sensing. *Tech. Rep., MIT Media Lab* (2005).
[56] Sara Taylor, Akane Sano, and Rosalind Picard. [n. d.]. Structure of Electrodermal Responses During Sleep. *REM* 1, 3 ([n. d.]), 4.
[57] Sara Taylor, Akane Sano, and Rosalind Picard. 2017. Structure of Electrodermal Responses During Sleep. *REM* 1, 3 (2017), 4.
[58] Aldert Vrij. 2008. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.
[59] Jeffrey J Walczyk, Frank P Igou, Alexa P Dixon, and Talar Tcholakian. 2013. Advancing lie detection by inducing cognitive load on liars: a review of relevant theories and techniques guided by lessons from polygraph-based approaches. *Frontiers in psychology* 4 (2013).
[60] Jason T Weber. 2017. When deception gets personal: an exploration into personality's link to deception. (2017).
[61] J Christopher Westland. 2011. Electrodermal response in gaming. *Journal of Computer Networks and Communications* 2011 (2011).