Automatic Frustration Detection Using Thermal Imaging

Youssef Mohamed
KTH: Royal Institute of Technology
Stockholm, Sweden
ymo@kth.se

Giulia Ballardini
University of Genoa
Genoa, Italy
giulia.ballardini@edu.unige.it

Maria Teresa Parreira
KTH: Royal Institute of Technology
Stockholm, Sweden
parreira@kth.se

Séverin Lemaignan
PAL Robotics
Barcelona, Spain
severin.lemaignan@pal-robotics.com

Iolanda Leite
KTH: Royal Institute of Technology
Stockholm, Sweden
iolanda@kth.se

2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) | 978-1-6654-0731-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/HRI53351.2022.9889545

***Abstract*—To achieve seamless interactions, robots have to be capable of reliably detecting affective states in real time. One of the possible states that humans go through while interacting with robots is frustration. Detecting frustration from RGB images can be challenging in some real-world situations; thus, we investigate in this work whether thermal imaging can be used to create a model that is capable of detecting frustration induced by cognitive load and failure. To train our model, we collected a data set from 18 participants experiencing both types of frustration induced by a robot. The model was tested using features from several modalities: thermal, RGB, Electrodermal Activity (EDA), and all three combined. When data from both frustration cases were combined and used as training input, the model reached an accuracy of 89% with just RGB features, 87% using only thermal features, 84% using EDA, and 86% when using all modalities. Furthermore, the highest accuracy for the thermal data was reached using three facial regions of interest: nose, forehead and lower lip.**

***Index Terms*—Human-robot interaction; Thermal imaging; Frustration; cognitive load; Action units;**

I. Introduction

In collaborative environments with robots, users are prone to feeling frustration due to the robot's behavioural errors, such as social norm violations, or technical errors, like speech recognition failure [1], [2]. This can affect acceptance of the robots [2]. Furthermore, frustration can be associated with lower levels of productivity [3], motivation [4], and trust [1], and higher levels of aggression [5], [6]. If a robot can detect frustration in a user, it could proactively employ mediation strategies or abort the interaction before that state intensifies.

Although current methods can accurately extract social signals (e.g., facial landmarks, action units and pose estimation) [7]–[9], inferring affective states and understanding those signals can be skewed, biased, and/or subjective [10], [11]. Thus, several sensors have been introduced to detect those affective states using different physiological signals, including electrocardiography, electromyography, skin conductance and body temperature [12]. However, these sensors are usually intrusive and can affect the participants' behaviour [13], making them unsuitable for real-world scenarios.

In 1997, Hirokazu Genno [14] proposed one of the first methods to evaluate stress and fatigue using thermal cameras. In spite of technical limitations in accuracy and resolution, a high correlation was observed between reported stress levels and the measured facial temperatures. This is due to the automatic reactions of the sympathetic nervous system, which are reflected in facial temperature [15]–[17]. As thermal cameras are becoming more accurate and affordable, thermal imaging has been gaining attention for detecting internal states like stress [18], cognitive load [19], [20], and deception [21].

Some researchers suggest that there are different types of frustration [22]. We focus our work in the detection of frustration in two cases that we consider relevant for human-robot interaction (HRI): failure-induced frustration and cognitive load-induced frustration. Cognitively demanding situations relate to stress and anxiety [23]; moreover, failure to resolve the situation or to change that stressful state can lead to the onset of frustration [24]. Additionally, the occurrence of a repeated failure is directly related to frustration and disappointment [25]. According to [26]–[28], frustration might be multi-faceted and can be affected by the task's length, nature, or sequence. Hence, we have reasons to believe that, by inducing frustration in different scenarios, we can create a more general prediction of frustration.

In this work, we used an infra-red thermal camera to investigate if a machine learning model can detect frustration using facial thermal data in an HRI scenario. This will be achieved by:

- Comparing the model's performance when using RGB features, i.e. Action Units (AUs), facial thermal features and ElectroDermal Activity (EDA) features;

This work was partially funded by grants from the Swedish Research Council (2017-05189), the Swedish Foundation for Strategic Research (SSF FFL18-0199), the S-FACTOR project from NordForsk, the Digital Futures research Center, the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH, and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation

- Selecting the facial thermal Regions Of Interest (ROIs) that yield the highest prediction accuracy;
- Investigating the effects of aggregating the data points into time intervals of 1, 3.5 and 7 seconds (window size).

In section II, existing approaches are discussed for thermal imaging and frustration prediction. Then, a detailed description on the data collection method and the full system architecture is explained in section III. Section IV addresses the classification methods and discusses the features extracted. Lastly, the results are stated in section V and discussed thoroughly in section VI.

II. Related Work

Understanding frustration and detecting it while people are interacting with robots is an ongoing challenge. This study will be based on the advancements made in thermal imaging, affective state detection and frustration detection.

A. Frustration Detection

Frustration has been established as one of the most important affective states to detect in HRI [29]. Hence, several approaches have been implemented to detect frustration. Taylor et al. [30] simultaneously used three wearable sensors to detect five levels of frustration with 80% accuracy using physiological data like electrodermal activity, heat flux, heart rate, skin temperature and skin conductivity. While the results were promising, the use of three different sensors is hardly applicable outside of a laboratory setting and might affect the participants' behaviour. In addition to physiological data, other non-verbal data have also been used for classification. Kapoor el al. [31] used skin conductance, pupil tracking, posture, mouse pressure and smile probability to predict frustration in a tutoring scenario with a virtual agent. The authors highlight the importance of detecting frustration in similar scenarios and compared several machine learning approaches reaching a prediction accuracy of 79%.

A data-driven approach was taken by [25] to classify frustration and disappointment caused by the same task. The authors collected the AUs, EDA and heart rate from 18 subjects within 5 seconds of the occurrence of an event. The event was based on a web form that the participants were made to believe they had to fill out to proceed to the experiment. When the participants tried to submit the form, an error would occur. The occurrence of the first error was assumed to cause disappointment, and any successive errors were assumed to cause frustration. This assumption was supported by self-reports from the participants after the experiment. The authors then created a multi-class classifier that distinguished between neutral, frustration, and disappointment states. Using different data subsets and different machine learning algorithms, they achieved a maximum accuracy of 64%. The authors used only the tonic component of the EDA without any further processing or feature extraction, which limits their results [32]. Furthermore, they used a shuffle split for cross-validation, which does not guarantee different folds, especially for small data sets.

B. Affective State Detection and Thermal Imaging

Using visual sensors to detect affective states is common in the literature. In [33], the authors used a Microsoft Kinect to extract action units and body movement to predict the six basic affective states: anger, fear, disgust, happiness, surprise and sadness. The authors then fed the facial expression and body movement data streams separately to a uni-modal neural network, and they applied late fusion to determine the affective state of the participant. Their model achieved an accuracy of 93% on an acted affect data set.

Image-based methods for affective state detection, however, are heavily dependant on lighting conditions, and the accuracy of their detection can be drastically affected by the self-report measures and conflicting facial expressions [34].

Alternatively, thermal cameras use far infra-red to measure the radiation emitted by warm objects, which is independent of reflected light [35]. Hence, thermal imaging can be used to overcome an RGB camera's limitations, as the thermal spectrum is not affected by light presence and it is able to record objective measures, such as changes in skin temperature [36].

Thermal imaging primarily has been used by researchers to detect the six basic affective states. For instance, the Kotani Thermal Facial Emotion data set [37] contained visual and thermal images of 26 subjects experiencing those states. Each affective state was induced by making the participants watch an emotional video clip while measuring facial thermal changes. The baseline was collected from the participants while listening to music between clips, and each affective state was labeled based on the participants' self-reports.

More complex affective states like guilt, shame, and remorse were also investigated [38]. The authors induced them by introducing the participants to storyboards with different scenarios, each designed to induce one of those affective states. They found thermal differences between the affective states, as guilt resulted in a change of at least 0.5° C higher than shame and remorse in the forehead, cheek, and mouth regions.

In addition, stress and cognitive load have been a focus for thermal imaging, as their effects on the facial temperature are established in psychology literature [39]. For example, [20] detected cognitive load induced by the Stroop effect and reading tasks, and observed a high correlation between the difficulty of the task and the facial temperature, with an increase in the nose and decrease in the forehead region. Stress detection in HRI using thermal data was discussed in [40], where a thermal camera was mounted on a Meka robot to measure facial temperature variations while playing a card-based quiz game with the robot. Several scenarios were tested with variations in setting parameters. It was observed that the closer the robot was positioned to the participant, the higher their nose temperature. Moreover, they used the RGB camera's ROI detection and overlaid it on the calibrated thermal image. This approach can accurately detect the ROI in the thermal image while eliminating the need for advanced image processing techniques (e.g., using a bilateral filter on the thermal images to preserve edges and reduce the noise [41] or generating binary images and computing their projection curves [42]).

As such, we used an RGB camera calibrated with the thermal camera to detect ROIs, which can be done using an off-the-shelf face detection model. Furthermore, other more complex facial features can be extracted from the RGB image, including action units, which can later aid in the creation of multi-modal systems with better prediction accuracy for affective states.

To the best of our knowledge, there is no study that examined frustration using thermal imaging, let alone with different types of frustration. In our study, we bridged this gap and used thermal imaging to detect frustration in two cases: cognitive load-induced frustration and failure-induced frustration.

III. Data Collection

A. Participants

A total of 25 participants (12 female, 13 male) without any known history of neurological or psychiatric disorders were recruited for the experiment. The recruitment process was through online platforms, word of mouth and flyers. Most recruits were from the surrounding area and the university campus. The age of the participants ranged from 21 to 46 years *(M = 27.80, SD = 6.18)*. The submitted work describes research with human participants and was approved by a relevant ethics committee. The data from five participants was discarded from the analysis for technical problems occurring during the experiment, as some participants did not comply with the task instructions or frequently touched their face during the data collection. The data from two participants were discarded since they self-reported (see subsubsection III-B5) that they did not get frustrated in any of the tasks. For the analysis, we used data collected from 18 participants (9 female, 9 male), with ages between 21 and 39 years old *(M = 27.28, SD = 5.67)*.

B. Task Description

Participants had to complete two tasks separated by a resting period. A NAO robot provided instructions and guided the participants through the tasks. Two cameras were mounted on a table: a thermal camera and an RGB camera, positioned high enough to ensure that the participants face was always visible, as seen in Figure 1. One task consisted of a quiz where the answers from the participant were misinterpreted by the robot, leading to frustration caused by failure; the other task involved the completion of two challenges in the laptop in front of the participant. Participants had to alternate between the two challenges when prompted by the sound of a buzzer. According to cognitive load theory, cognitive load can be reduced if the task is learned [43], hence, switching between tasks constantly is theorized to keep the participant in a constant cognitive load state. Since the participant fails to overcome the cause of the cognitive load, frustration is also expected to occur [24].

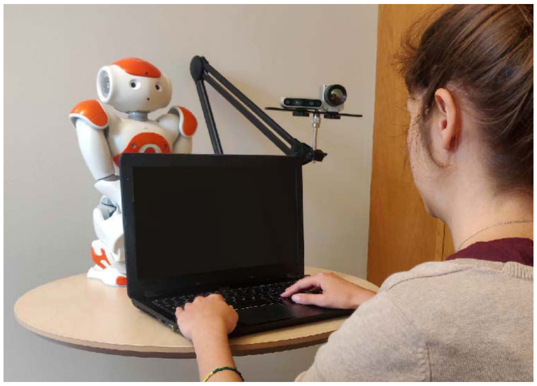

Fig. 1: Experimental setup.

As such, the experiment consisted of four stages: baseline (B), collected before the start of the first task, cognitive load-induced frustration (TCog), rest and failure-induced frustration (TFail) (Fig.2). The order of the two tasks (TFail and TCog) was balanced among participants to avoid bias due to presentation order. Before each task, the NAO robot briefly explained the instructions and during the tasks a countdown was displayed on a monitor in front of participants.

1) B: We considered as initial baseline a 1 minute time-window before the first interaction with the robot.

2) TFail: A simple game of trivia was played between the participants and the NAO robot. The robot was teleoperated by a human wizard. The participant was instructed that they must provide 10 correct answers in less than 5 minutes in order to increase their reward by 20 SEK, from the 80 SEK they were promised. We note that the participant would receive the full compensation of 100 SEK regardless of the performance. During the interaction, NAO asked 14 obvious general questions, e.g. *'how many hours are there in a day?'*, or referred to pictures shown on the laptop in front of the participant. The order of the questions and the robot responses were predetermined. The answers to the first three questions were correctly identified by the robot, but from the fourth question onwards the robot intentionally declared an answer to be incorrect or it took time while *'processing the answer'* in order to induce frustration. This behaviour was repeated until the time was up or the participant answered all the questions. Out of the 14 questions, 8 answers were considered correct, and in 4 instances NAO took longer to process (2 ending with correct responses).

3) Rest: The participant was prompted to wait and listen to classical music for two minutes, in order to isolate the physiological responses from each task.

4) TCog: Cognitive load would be induced by a dual-task composed of a challenging coding task[1] and a mental rotation task[2] for 8 minutes. In the coding task, participant had to program (using a visual programming language interface) an animated robot to move from one place to the other and its

[1] https://oscared.github.io/level_4/

[2] https://vample.com/tools/mental-rotation/

level of difficulty was based on the participant programming background. When a loud buzzer sound was played, the participant had to solve one question in the mental rotation task and after that go back to the coding task. The timing and the number of the buzzer occurrences were adapted to the performance. In general, the closer the participant would get to solving the coding task, the smaller the intervals were between buzzer rings.

5) Self-assessment: Four different types of self-assessment questionnaires were given to the participants. They had to digitally fill out three of them before the start of the experiment:

- demographic data,
- technical affinity,
- personality traits [44].

The technical affinity questionnaire included questions about current and previous experience with robots (*'have you ever seen a robot in real life?'*). Furthermore, after each task the participants filled out the NASA-TLX [45] questionnaire, stating the amount of cognitive load and frustration felt during the previous task. We used the NASA-TLX self-reports as a manipulation check of our tasks.

C. System Implementation

The system architecture (Fig. 3) was composed of both hardware and software components, two cameras mutually calibrated (thermal IR camera: `Optris PI 640`[3] and RGB-D camera: `RealSense D435`[4]), `NAO`[5] robot and an `EDA sensor` (embedded in the Empatica E4 wristband[6]). All of the mentioned components were synchronized in real-time using Robotic Operating System (ROS), except for the EDA sensor, which was synchronized in data post-processing. In addition, OpenCV was used for image processing and camera calibration.

The frames from the thermal and RGB cameras were published to ROS (both cameras acquired 15 frames per second). Then, the RGB frames were sent into OpenFace to detect the position of the facial landmarks and the presence and intensity of 18 action units. After that, applying the calibration matrix, the landmark positions were transposed into the thermal frames (Fig. 4) by OpenCV in order to extract the thermal ROIs, i.e., a rectangle on the thermal image based on the relevant landmark positions. Finally, an average of the thermal values within the ROIs was computed. Four facial ROIs were extracted from the thermal image: nose, forehead, cheek and lower lip, as shown in Fig. 4.

Furthermore, NAO robot's SDK (NAOqi) was used to control the robot's responses. Key responses which were considered to be important events in the interaction, e.g. instances where the robot responded with *'incorrect'*,*'correct'* and *'processing'*, were published in ROS to be synchronized with the thermal and the RGB data streams.

During the experiment, participants wore the Empatica E4 wristband on the right arm. It captures skin electrical conductance by passing a minimal alternating current between two electrodes in contact with the skin. EDA samples are measured at 4 Hz rate, with a resolution of 900 pS in a measurable range of 0.01-100 μS [46].

IV. Frustration Prediction

For frustration prediction, our goal is to (1) inspect the effectiveness of thermal imaging in detecting frustration when compared to RGB and EDA data, as well as (2) find out optimal features to identify frustration from thermal, RGB and EDA features. The frustration classifier is based on a K-Nearest Neighbors (KNN) algorithm, which is widely used in affective computing with weighted distance tasks [47], [48]. To evaluate the model, we used cross validation of leaving one participant out. The models were trained based on the best features selected by the Sequential Forward Floating Selection (SFFS) algorithm. The SFFS is a wrapper method that uses several greedy search methods to select the features that would yield the highest accuracy in the model. The method was adopted due to its wide use in the affective computing literature [49]–[51], over its more simple counterpart, sequential forward selection, which does not exclude the features once they are selected.

A. Labeling

Input data for the classifier corresponded the participants that self-reported frustration in the NASA-TLX questionnaire in TCog *(M = 3.75, SD = 1)* and in TFail *(M = 3.3, SD = 0.9)*. For TCog, we assumed a constant state of cognitive load-induced frustration onset 30 s after the beginning of the task. As such, as we do with B for*'non-frustration'* instances, that period was subdivided into non-overlapping 'windows' of three possible lengths: 1, 3.5 or 7 s. The same windows were applied to all the modalities and used to extract the different features. The features from each window were used as instances to train the models (Table II). For TFail, frustration was not assumed as constant state, but inducted by failure, i.e., when the robot replied that the answer was *'incorrect'*, which occurred 6 times during the task. For that reason, in TFail frustration instances corresponded to 7-second periods after those event. The length of the period was determined as the maximum duration allowed to isolate physiological responses to frustration-inducing events, i.e., the minimum amount of time between *'incorrect'* events. Each 7-second period was then subdivided into non-overlapping windows of 1, 3.5 and 7 s. An illustration can be seen in Fig. 5.

All three data subsets (TCog, TFail and TCog+TFail) used the same baseline (B). Classification with task separation allows for a more fine grained analysis of frustration, while the combination of both types allows for a more general prediction of frustration. The number of instances for each of the data subsets is shown in Table I.

Furthermore, each subset was trained on four feature types: thermal, RGB, EDA and all features combined. Window sizes

[3] https://www.optris.global/thermal-imager-optris-pi-640

[4] https://www.intelrealsense.com/depth-camera-d435/

[5] https://www.softbankrobotics.com/emea/en/nao

[6] https://www.empatica.com/research/e4/

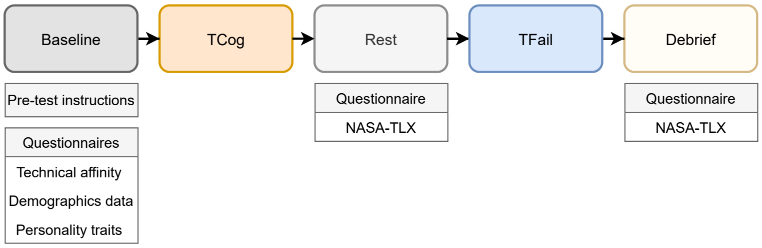

Fig. 2: Tasks procedure, the order of TCog and TFail was randomized to avoid bias.

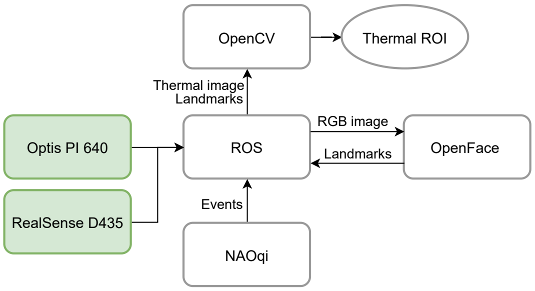

Fig. 3: System architecture.

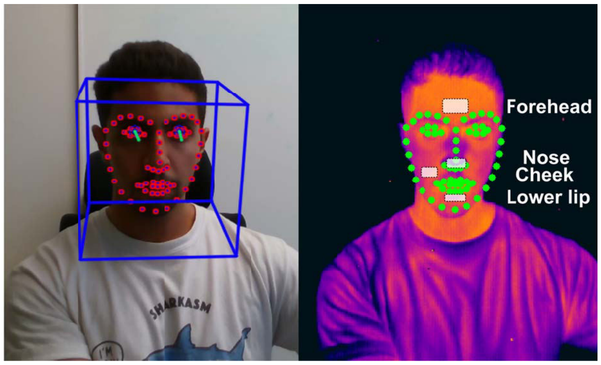

Fig. 4: (left) Landmarks positions detected in the RGB image; (right) thermal image overlaid landmarks and the ROIs, which include forehead, nose, cheek and lower lip.

TABLE I: The number of instances used for training in the cases of cognitive load-induced frustration (TCog), failure induced frustration (TFail) and baseline (B).

	No. of Instances		
Window (s)	**TCog**	**TFail**	**B**
1	7692	1127	1010
3.5	2198	322	303
7	1094	161	151

can be an indication of the duration of the affective state. Typically, a facial expression lasts between 0.5-4 seconds but a physiological effect lasts 5-15 seconds [52]. Accordingly, the window sizes of 1, 3.5 and 7 seconds were inspected for model training.

B. Pre-processing

Thermal ROIs were standardized using `RobustScaler`, a standardization method which removes the median and scales data based on the quartile range, in order to accommodate for the presence of outliers.

Similarly, the standardization based on median and quartile was applied to the EDA data. After that, the EDA was divided into tonic and phasic components [53]. The sensor sampling rate of 4 Hz only allowed window sizes of 3.5 and 7 seconds to be included, otherwise there would not be enough peaks for meaningful results.

All the data was then labelled and split into B, TCog and TFail (as described above, see Fig. 6 for reference).

All data processing was performed using Python's `scikit-learn`[7], `NeuroKit2`[8] and `MLxtend`[9] libraries.

C. Feature Extraction

For each modality, several features were extracted as seen in Table II.

TABLE II: Extracted features.

Modality	**Features**
Thermal	ROIs temperature average ROIs temperature change ROIs temperature maximum
RGB	AU Intensity average AU intensity change AU maximum intensity
EDA	Mean skin conductance level (SCL) Frequency of the peak occurrence Mean peak amplitude Peak rise time Mean peak duration Mean of inter-peak interval (IPI)

The features for the thermal data were computed for all the four ROIs: nose, forehead, cheek and lower lip. As for the

[7] https://scikit-learn.org
[8] https://github.com/neuropsychology/NeuroKit
[9] http://rasbt.github.io/mlxtend/

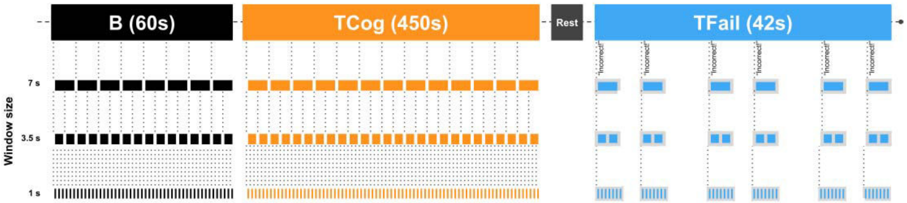

Fig. 5: The length of data of 60, 450 and 42 seconds considered in Baseline, TCog and TFail, with the window sizes to be 1, 3.5 and 7 seconds.

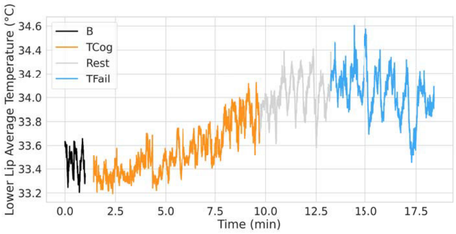

Fig. 6: An example of thermal data from lower lip from one participant, after labelling. B consists of a 60-second period.

action units extracted, they corresponded to the Facial Action Coding System (FACS): 1 (inner brow raiser), 2 (outer brow raiser), 4 (brow lowerer), 5 (Upper lid raiser), 6 (cheek raiser), 7 (lid tightener), 9 (nose wrinkler), 10 (upper lip raiser), 12 (lip corner puller), 14 (dimpler), 15 (lip corner depressor), 17 (chin raiser), 20 (lip stretcher), 23 (lip tightener), 25 (lips part), 26 (jaw drop), 28 (lip suck), and 45 (blink).

For both the temperature and the AU intensity, we computed the average, the change and the maximum within each window.

In addition, the tonic component in the EDA data included the mean Skin Conductance Level (SCL), while from the peak detection analysis we extracted standard peak features, such as time interval between consecutive peaks (IPI), frequency of peak occurrence, mean peak amplitude, mean peak rise time and mean peak duration, in accordance to [54].

V. Results

To evaluate KNN models performance, both the accuracy and the weighted F1-score were computed for each modality and window size. Considering the imbalance of the data, the accuracy alone might be unreliable [55], therefore, the F1-score can be a better metric [56]. The metrics were calculated based on the average result of the cross-validation of leaving one participant out for each test-train split.

Figs. 7 and 8 describe the performance of each modality over three window sizes.

1) Thermal: For the TFail model, increasing the window size slightly increases the accuracy, as in the 1 second window it is 59% then increases by 5% in the 7 second window. F1-score follows accordingly, as it is the highest at 64% in the 7 second window. Similarly, in TCog accuracy is the highest in the 7 second window to 83% . In TCog+TFail, the accuracy peaks in the 7 second window at 87%.

2) RGB: For TFail, maximum accuracy is achieved in the 3.5 second window (81%) and the lowest in the 1 second window (69%), while in 7 second window it goes back to 71%. In TCog+TFail, the accuracy in the 1 and 3.5 second windows is constant at 89%. The accuracy in TCog gets to 89% in the 1 and 3.5 second windows. F1-score follows the same trend, with a decline to 83% in the 7 second window.

3) EDA: The EDA data was only inspected in the 3.5 and 7 second windows, due to the low sampling rate of the wristband. In TFail, accuracy goes to 55% in the 7 second window, while the accuracy in TCog and TCog+TFail is 78% and 84% respectively and does not vary across window sizes. A similar trend can also be seen in the F1-score metric, which in TFail is 53%, while for TCog and TCog+TFail it is 73% and 78% respectively.

4) All modalities: When using all the modalities (thermal, RGB and EDA) to train the model, for TFail the accuracy is steady at 74% across both window sizes. On the other hand, TCog accuracy increases to 90% in the 7 second window. Similarly, when using TCog+TFail, accuracy increases in the 7 seconds to 86%. F1-scores follow the same trends overall for all three data subsets.

A. Feature Selection

In Table III, the best features of the SFFS are shown for each task and each modality separately. To show all three modalities, with the highest model granularity possible, the 3.5 second window was picked to illustrate the feature selection results.

1) Thermal: When using only the thermal data, for the TFail classifier it can be seen that the cheek region was discarded by the feature selector, selecting only the nose, forehead and lower lip for both temperature average and change. Similarly for the TCog and the TCog+TFail, the maximum temperatures of the nose and lower lip were also selected, in addition to nose temperature average and change for forehead and lower lip.

2) RGB: In TFail, the most relevant average intensity were AU20, AU23, AU28, AU10 and AU12, which correspond to movements in the lips, in addition to 7 and 4 which correspond to movements in the eye and brow respectively. Similarly, when using TCog the most relevant features for classification were the ones that correspond to lip and eye movements (AU07 and AU05). For the TCog+TFail subset, out of the 8 features selected, 4 were related to lip movements.

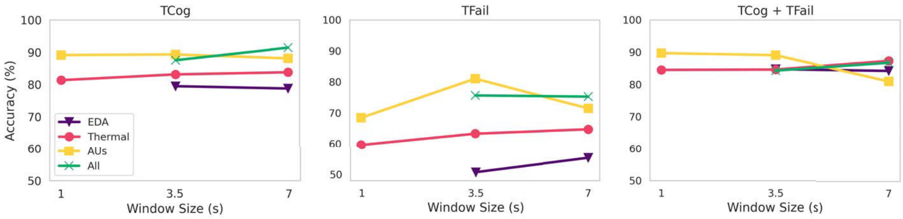

Fig. 7: Accuracy measures for cognitive load-induced frustration, failure-induced frustration and both data subsets concatenated.

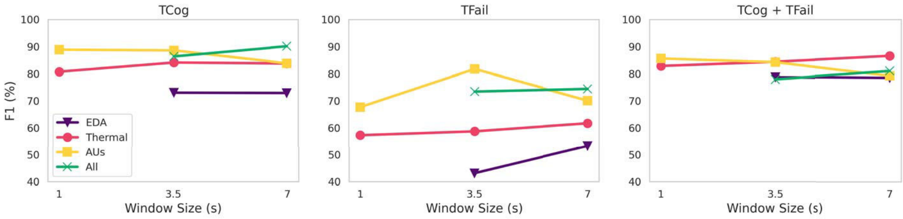

Fig. 8: F1-score for cognitive load-induced frustration, failure-induced frustration and both data subsets concatenated.

TABLE III: Results for feature selection in the 3.5 second window.

Task	Modality	Best Features	
TFail	**Thermal**	**Temp. average**	Nose, Forehead, Lower lip
		Temp. change	Nose, Forehead, Lower lip
	RGB	**Intensity average**	AU07, AU20, AU23, AU28, AU10, AU04, AU12
		Intensity change	AU02, AU09, AU04
		Intensity maximum	AU28
	EDA	**Tonic**	SCL
		Phasic	IPI, Peak amplitude, Peak duration
	All	**Temp. average**	Lower lip
		Intensity average	AU28, AU10, AU23, AU17, AU06
		Temp. change	Lower lip
		Intensity change	AU28, AU02, AU26
TCog	**Thermal**	**Temp. average**	Nose
		Temp. change	Forehead, Lower lip
		Temp. maximum	Nose, Lower lip
	RGB	**Intensity average**	AU06, AU07, AU02, AU10, AU05, AU12, AU26, AU28, AU20, AU14
		Intensity change	AU28, AU20, AU25, AU26
		Intensity maximum	AU28
	EDA	**Phasic**	IPI, Peak duration
	All	**Temp. average**	Nose
		Intensity average	AU45, AU06, AU20, AU10, AU23, AU28, AU07, AU25, AU12
		Temp. change	Lower lip
		Intensity change	AU28, AU04, AU12
		Intensity maximum	AU01, AU02
TCog+TFail	**Thermal**	**Temp. average**	Nose
		Temp. change	Forehead, Lower lip
		Temp. maximum	Nose, lower lip
	RGB	**Intensity average**	AU06, AU02, AU09, AU25, AU26, AU28
		Intensity change	AU28
		Intensity maximum	AU28
	EDA	**Phasic**	Peak duration, Peak rise
	All	**Intensity average**	AU28, AU06, AU02, AU23, AU25
		Intensity change	AU28
		Intensity maximum	AU28

3) EDA: For the TFail classifier, 4 out of the 6 features were selected: mean conductance level, inter-peak interval, peak amplitude and peak duration. For the TCog, only inter-peak interval and peak duration were selected. Finally, when both tasks are combined, peak rise time and peak duration were the most relevant features.

4) All Modalities: The most relevant thermal region across all modalities for the TFail classifier was the lower lip, as the temperature average and change. In addition, 7 action units were also selected, mostly corresponding to lip movements. In TCog, two thermal regions were relevant: the nose temperature average and the lower lip temperature change. Also, 12 action units were selected, 6 of which correspond to lip movements, 3 with brows, 2 with eyes and 1 with cheek movements. When the two tasks are combined, in TCog+TFail, only action units were selected, 3 corresponding to lip movements and 2 to brow and cheek movements.

VI. Discussion

The window sizes comparison across all data subsets shows that using AUs as input data results in the highest accuracy in the 3.5-second window, while thermal and EDA data achieve the highest accuracies in the 7-second window. In other words, increasing the window size decreases the performance of the classifier which uses RGB features as input, while for thermal data as input the performance slightly increases, which coincides with Ekman's findings [52]. Ekman hypothesised that facial expressions last between 0.5 to 4 seconds after a stimuli, but the physiological reaction might take 5 to 15 seconds to completely deteriorate.

Feature selection on the thermal data shows that, among the ROIs provided, the nose, forehead and lower lip are the most relevant to detect frustration. In previous works, the cheek region has been related only to the startle affective state [39], while the other three regions were associated mainly with negative affective states like stress, fear and anxiety [39]. This could explain the correlation that we found with frustration, which is considered a negative state. Furthermore, the classifiers based on only one task (TFail or TCog) selected different features out of the thermal data. In fact, the detectors for TCog and TCog+TFail also used the maximum temperature for the nose and lower lip. This could imply that the thermal facial reaction can be dependant on the type of frustration, since TCog instances are assumed to be more related to cognitive load-induced frustration and the TFail instances to failure-induced frustration. For the classifier which uses instances from both tasks, the features selected are more similar to those of the TCog detector, which could imply these are more evidently distinguishable from a baseline state, when compared to the failure-induced frustration. However, this effect may be also related to our experimental design and processing, future investigations are needed to better address this point.

The selected features for AUs for both tasks are mainly focused on lip movements, and the common AUs across the three data subsets are AU28 (lip suck) and AU02 (outer brow raiser). As stated by [57], AU28 is associated with fidgeting and can be directly related to negative affective states. The occurrence of AU02 is explained in [58], which states that it is mainly associated with focus. Furthermore, the presence of AU04 (brow lowerer) and AU07 (lid tightner) in both the TCog and TFail detectors can indicate the occurrence of confusion [59]. Nonetheless, for the combined (TCog+TFail) classifier neither of these AUs are selected by the SFFS. According to the literature [60]–[62], there is no common consensus on which AUs relate to frustration, as it is task-dependent. However, some of the AUs mentioned were AU09, AU10, AU12, AU14, AU23 and AU24, which can be seen among the features selected for each task. Nonetheless, AU28 was repeatedly selected as one of the most relevant action units in our work.

Peak duration is the common selected feature across all data subsets in the EDA modality, in addition to peak rise time in the TCog+TFail detector. According to [63], [64], the tonic component (SCL) might be less useful to detect affective states, while the phasic component is more reactive to external stimuli. This coincides with the fact that only phasic components were selected in the TCog+TFail and TCog detectors.

Combining all the modalities can yield higher accuracies, as is the case across all detectors in the 7 second window size. However, the feature selector discarded EDA data from all the subsets, which indicates that the best combination of modalities would be AUs and thermal data.

The trained model can be extracted and used in real-time, the limiting factor for each window size is the amount of data that would be needed to give one prediction. The time window for the highest accuracy using thermal imaging is 7 seconds, in contrast to the RGB models which reach the highest accuracy in 3.5 seconds. Although, as model reactivity is a priority while running it in real-time, the increase of 3% in accuracy might be a valid compromise for faster reactivity. Furthermore, that does not hinder the capabilities of thermal imaging models as it still can detect frustration in scenarios where it is not visible to RGB cameras. Furthermore, a rolling window can be used to mitigate that effect and increase reactivity in real time retaining the accuracy.

Overall, using the thermal modality yields the highest accuracy when using larger window sizes in TCog+TFail, which is the more general model. Using RGB features for models with shorter window sizes will lead to better performance, as the KNN model had the highest accuracy at the 3.5 second window.

VII. Limitations

Considering that phasic features were extracted from EDA data, the sampling rate of 4 Hz of the used wristband is not sufficient for small window sizes [65].

Furthermore, allowing for a longer period of rest between tasks (TCog and TFail) could have provided some insight on how much time is necessary to return to a neutral state after becoming frustrated. In TFail, we have assumed that frustration would occur within 7 seconds of each frustrating event; nonetheless, the use of external annotators could have provided a more reliable ground truth for frustration.

Although we have tried to make the interactions as natural as possible while collecting the data, a more robust model would be trained on data collected outside of a lab setting.

In future work, we would like to collect a larger data set with more types of frustration and more participants, which would result in a more general and reliable model. In addition, insuring that the data in granular enough to test smaller and larger window sizes, as the thermal data might perform even better on window sizes larger than 7 seconds.

VIII. Conclusion

In this work, we investigated several models capable of predicting both cognitive load-induced and failure-induced frustration, both separately and combined. Furthermore, we investigated the effects of aggregating the data into intervals of sizes 1, 3.5 and 7 seconds (window size) on the model's performance. Several variations of the model were created for each window size, depending on the input data used: thermal, AUs, EDA and all modalities. The use of a sequential floating feature selector allowed for some insight on relevance of each feature for detection of frustration.

Thermal data can be used to detect frustration, as the model had an average accuracy of 85% in TCog+TFail across the three window sizes. However, using AUs could yield better results in shorter window sizes, as it performed better than thermal data, in TCog and TFail separately.

Using just AUs as input proved to yield the highest accuracy in both the TFail and TCog+TFail detectors, while the TCog detector using all modalities had the overall highest accuracy of 91%. Window sizes also proved to have a role in the model's performance depending on the modality, as the AUs have the best accuracy accros data subsets in the 3.5 second window, but other physiological signals needing 7 seconds or more, which is consistent with previous findings [52].

The results of the feature selectors for thermal data showed that the nose, lower lip and forehead are the most relevant regions for frustration detection, while from action units, the lip and brow movements appeared to be good indicators of frustration. As for the EDA data, peak duration was the common feature selected across all three data subsets.

References

[1] Moaed A Abd, Iker Gonzalez, Mehrdad Nojoumian, and Erik D Engeberg. Trust, satisfaction and frustration measurements during human-robot interaction. In *Proceedings of the 30th Florida Conference on Recent Advances in RoboticsMay 11-12*, volume 2107, 2017.

[2] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in Psychology*, 6:931, jul 2015.

[3] Jonathan Lazar, Adam Jones, and Ben Shneiderman. Workplace user frustration with computers: An exploratory investigation of the causes and severity. *Behaviour and Information Technology*, 25(3):239–251, may 2006.

[4] Bernard Weiner. An Attributional Theory of Achievement Motivation and Emotion. *Psychological Review*, 92(4):548–573, oct 1985.

[5] Suzy Fox and Paul E. Spector. A model of work frustration-aggression. *Journal of Organizational Behavior*, 20(6):915–931, 1999.

[6] John Dollard, Neal E. Miller, Leonard W. Doob, O. H. Mowrer, and Robert R. Sears. *Frustration and aggression.* Yale University Press, oct 1939.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, jan 2021.

[8] Tadas Baltrusaitis, Peter Robinson, and Louis Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. Institute of Electrical and Electronics Engineers Inc., may 2016.

[9] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. The Social Signal Interpretation (SSI) Framework Multimodal Signal Processing and Recognition in Real-Time. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, New York, New York, USA, 2013. ACM Press.

[10] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. Affective computing for large-scale heterogeneous multimedia data: A survey, dec 2019.

[11] Rosalind W. Picard. Affective computing: Challenges. *International Journal of Human Computer Studies*, 59(1-2):55–64, jul 2003.

[12] Austin Kothig, John Munoz, Sami Alperen Akgun, Alexander M Aroyo, and Kerstin Dautenhahn. Connecting Humans and Robots Using Physiological Signals – Closing-the-Loop in HRI. pages 735–742, 2021.

[13] Paulo Novais and Davide Carneiro. The role of non-intrusive approaches in the development of people-aware systems. *Progress in Artificial Intelligence*, 5(3):215–220, 2016.

[14] Hirokazu Genno, Keiko Ishikawa, Osamu Kanbara, Makoto Kikumoto, Yoshihisa Fujiwara, Ryuuzi Suzuki, and Masato Osumi. Using facial skin temperature to objectively evaluate sensations. *International Journal of Industrial Ergonomics*, 19(2):161–171, feb 1997.

[15] Dvijesh Shastri, Arcangelo Merla, Panagiotis Tsiamyrtzis, and Ioannis Pavlidis. Imaging facial signs of neurophysiological responses. *IEEE Transactions on Biomedical Engineering*, 56(2):477–484, feb 2009.

[16] R. Sinha, W. R. Lovallo, and O. A. Parsons. Cardiovascular differentiation of emotions. *Psychosomatic Medicine*, 54(4):422–435, 1992.

[17] Christian Collet, Evelyne Vernet-Maury, Georges Delhomme, and André Dittmar. Autonomic nervous system response patterns specificity to basic emotions. *Journal of the Autonomic Nervous System*, 62(1-2):45–57, jan 1997.

[18] Carl B. Cross, Julie A. Skipper, and Douglas Petkie. Thermal imaging to detect physiological indicators of stress in humans. In *Thermosense: Thermal Infrared Applications XXXV*, volume 8705, page 87050I. SPIE, may 2013.

[19] John Stemberger, Robert S. Allison, and Thomas Schnell. Thermal imaging as a way to classify cognitive workload. In *CRV 2010 - 7th Canadian Conference on Computer and Robot Vision*, pages 231–238, 2010.

[20] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. Cognitive Heat. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, sep 2017.

[21] Bashar A. Rajoub and Reyer Zwiggelaar. Thermal Facial Analysis for Deception Detection. *IEEE Transactions on Information Forensics and Security*, 9(6):1015–1023, jun 2014.

[22] James Paul Gee. *Good video games + good learning: collected essays on video games, learning ...*, volume 49. Lang,, New York :, 2007.

[23] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z Arshad, Ahmad Khawaji, and Dan Conway. *Robust multimodal cognitive load measurement*. Springer, 2016.

[24] Barry Kort, Rob Reilly, and Rosalind W Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings IEEE international conference on advanced learning technologies*, pages 43–46. IEEE, 2001.

[25] Suhaib Aslam, Kim Gouweleeuw, Gijs Verhoeven, and Nynke Zwart. Classification of Disappointment and Frustration Elicited by Human-Computer Interaction: Towards Affective HCI. Number August, 2019.

[26] Zhongxiu Liu, Visit Pataranutaporn, Jaclyn Ocumpaugh, and Ryan Baker. Sequences of frustration and confusion, and learning. In *Educational data mining 2013*. Citeseer, 2013.

[27] Diane Marie C. Lee, Ma Mercedes T. Rodrigo, Ryan S.J.D. Baker, Jessica O. Sugay, and Andrei Coronel. Exploring the relationship between novice programmer confusion and achievement. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6974 LNCS, pages 175–184, 2011.

[28] Jennifer Sabourin, Jonathan P. Rowe, Bradford W. Mott, and James C. Lester. When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. In *AIED*, 2011.

[29] Alexandra Weidemann and Nele Rußwinkel. The Role of Frustration in Human–Robot Interaction – What Is Needed for a Successful Collaboration? *Frontiers in Psychology*, 12:707, mar 2021.

[30] Brandon Taylor, Anind Dey, Daniel Siewiorek, and Asim Smailagic. Using physiological sensors to detect levels of user frustration induced by system delays. In *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 517–528. Association for Computing Machinery, Inc, sep 2015.

[31] Ashish Kapoor, Winslow Burleson, and Rosalind W. Picard. Automatic prediction of frustration. *International Journal of Human Computer Studies*, 65(8):724–736, aug 2007.

[32] Moaed A Abd, Iker Gonzalez, Mehrdad Nojoumian, and Erik D Engeberg. Trust , Satisfaction and Frustration Measurements During Human-Robot Interaction Trust , Satisfaction and Frustration Measurements During Human-Robot Interaction. *30th Florida Conference on Recent Advances in Robotics (FCRAR)*, (May):89–93, 2017.

[33] Athanasios Psaltis, Kyriaki Kaza, Kiriakos Stefanidis, Spyridon Thermos, Konstantinos C. Apostolakis, Kosmas Dimitropoulos, and Petros Daras. Multimodal affective state recognition in serious games applications. In *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 435–439, 2016.

[34] Angeliki Fydanaki and Zeno Geradts. Evaluating OpenFace: an open-source automatic facial comparison algorithm for forensics. *https://doi.org/10.1080/20961790.2018.1523703*, 3(3):202–209, jul 2018.

[35] J. M. Lloyd. *Thermal Imaging Systems*. Springer US, Boston, MA, 1975.

[36] Thu Nguyen, Khang Tran, and Hung Nguyen. Towards Thermal Region of Interest for Human Emotion Estimation. In *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering, KSE 2018*, pages 152–157. Institute of Electrical and Electronics Engineers Inc., dec 2018.

[37] Hung Nguyen, Kazunori Kotani, Fan Chen, and Bac Le. A thermal facial emotion database and its analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8333 LNCS, pages 397–408. Springer Verlag, oct 2014.

[38] Braj Bhushan, Sabnam Basu, Pradipta Kumar Panigrahi, and Sourav Dutta. Exploring the Thermal Signature of Guilt, Shame, and Remorse. *Frontiers in Psychology*, 11:2874, nov 2020.

[39] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology*, 51(10):951–963, oct 2014.

[40] Mihaela Sorostinean, François Ferland, and Adriana Tapus. Reliable stress measurement using face temperature variation with a thermal camera in human-robot interaction. In *IEEE-RAS International Conference on Humanoid Robots*, volume 2015-Decem, pages 14–19. IEEE Computer Society, dec 2015.

[41] Mohd Norzali Haji Mohd, Masayuki Kashima, Kiminori Sato, and Mutsumi Watanabe. Mental Stress Recognition based on Non-invasive and Non-contact Measurement from Stereo Thermal and Visible Sensors. *International Journal of Affective Engineering*, 14(1):9–17, 2015.

[42] Shangfei Wang, Menghua He, Zhen Gao, Shan He, and Qiang Ji. Emotion recognition from thermal infrared images using deep boltzmann machine. *Frontiers of Computer Science*, 8(4):609–618, 2014.

[43] Fred G W C Paas and Jeroen J G Van Merriënboer. The Efficiency of Instructional Conditions: An Approach to Combine Mental Effort and Performance Measures. *Human Factors*, 35(4):737–743, 1993.

[44] Donald W. Fiske. Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, 44(3):329–344, jul 1949.

[45] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 52(C):139–183, jan 1988.

[46] Stefania Cecchi, Agnese Piersanti, Angelica Poli, and Susanna Spinsante. Physical Stimuli and Emotions: EDA Features Analysis from a Wrist-Worn Measurement Sensor. In *IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, CAMAD*, volume 2020-Septe. Institute of Electrical and Electronics Engineers Inc., sep 2020.

[47] Jianhua Tao and Tieniu Tan. Affective Computing: A Review. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3784 LNCS:981–995, oct 2005.

[48] R. Chinmayi, G. Jayachandran Nair, Mantha Soundarya, D. Sai Poojitha, Gayathri Venugopal, and Jishnu Vijayan. Extracting the features of emotion from EEG signals and classify using affective computing. *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2017*, 2018-January:2032–2036, feb 2018.

[49] Elias Vyzas and Rosalind W Picard. Offline and online recognition of emotion expression from physiological data. In *Workshop on Emotion-Based Agent Architectures at the Third International Conference on Autonomous Agents*, volume Technical, 1999.

[50] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.

[51] Vitaliy Kolodyazhniy, Sylvia D. Kreibig, James J. Gross, Walton T. Roth, and Frank H. Wilhelm. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. *Psychophysiology*, 48(7):908–922, jul 2011.

[52] Paul Ekman. *Emotions revealed: recognizing faces and feelings to improve communication and emotional life*. 2003.

[53] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods 2021 53:4*, 53(4):1689–1696, feb 2021.

[54] Mohsen Nabian, Yu Yin, Jolie Wormwood, Karen S. Quigley, Lisa F. Barrett, and Sarah Ostadabbas. An open-source feature extraction tool for the analysis of peripheral physiological data. *IEEE Journal of Translational Engineering in Health and Medicine*, 6, 2018.

[55] Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery 2012 28:1*, 28(1):92–122, oct 2012.

[56] László A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251, 2013.

[57] Alexandria K Vail, Joseph B Wiggins, Joseph F Grafsgaard, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. The affective impact of tutor questions: Predicting frustration and engagement. *International Educational Data Mining Society*, 2016.

[58] Ebrahim Babaei, Namrata Srivastava, Joshua Newn, Qiushi Zhou, Tilman Dingler, and Eduardo Velloso. *Faces of Focus: A Study on the Facial Cues of Attentional States*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2020.

[59] Dana Kulíc and Elizabeth Croft. Affective state estimation for human-robot interaction. In *IEEE Transactions on Robotics*, volume 23, pages 991–1000, oct 2007.

[60] Klas Ihme, Anirudh Unni, Meng Zhang, Jochem W Rieger, and Meike Jipp. Recognizing frustration of drivers from face video recordings and brain activation measurements with functional near-infrared spectroscopy. *Frontiers in human neuroscience*, 12:327, 2018.

[61] SK D'Mello, SD Craig, B Gholson, S Franklin, R Picard, and AC Graesser. Integrating affect sensors into an intelligent tutoring system. In *Affective interactions: The computer in the affective loop. Proceedings of the 2005 International Conference on Intelligent User Interfaces*, pages 7–13, 2004.

[62] Bethany McDaniel, Sidney D'Mello, Brandon King, Patrick Chipman, Kristy Tapp, and Art Graesser. Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

[63] Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. 2017.

[64] George I Christopoulos, Marilyn A Uy, and Wei Jie Yap. The Body and the Brain: Measuring Skin Conductance Responses to Understand the Emotional Experience. *Organizational Research Methods*, 22(1):394–420, 2019.

[65] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments. *Psychophysiology*, 49(1):1017–1034, 2013.