Good GUIs, Bad GUIs: Affective Evaluation of Graphical User Interfaces

SYRINE HADDAD[*], National Engineering School of Tunis, University of Tunis El Manar, Tunisia
KAYHAN LATIFZADEH[*] and SARAVANAKUMAR DURAISAMY, University of Luxembourg, Luxembourg
JEAN VANDERDONCKT, Université catholique de Louvain, LouRIM, Belgium
OLFA DAASSI, National Engineering School of Carthage, University of Carthage, Tunisia
SAFYA BELGHITH, National Engineering School of Tunis, University of Tunis El Manar, Tunisia
LUIS A. LEIVA, University of Luxembourg, Luxembourg

Affective computing has potential to enrich the development lifecycle of Graphical User Interfaces (GUIs) and of intelligent user interfaces by incorporating emotion-aware responses. Yet, affect is seldom considered to determine whether a GUI design would be perceived as good or bad. We study how physiological signals can be used as an early, effective, and rapid affective assessment method for GUI design, without having to ask for explicit user feedback. We conducted a controlled experiment where 32 participants were exposed to 20 good GUI and 20 bad GUI designs while recording their eye activity through eye tracking, facial expressions through video recordings, and brain activity through electroencephalography (EEG). We observed noticeable differences in the collected data, so we trained and compared different computational models to tell good and bad designs apart. Taken together, our results suggest that each modality has its own "performance sweet spot" both in terms of model architecture and signal length. Taken together, our findings suggest that is possible to distinguish between good and bad designs using physiological signals. Ultimately, this research paves the way toward implicit evaluation methods of GUI designs through user modeling.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; User interface design; **User models**; • **Computing methodologies** → Machine learning approaches.

Additional Key Words and Phrases: Affective computing; Neurophysiological and peripheral signals; User Interface design

ACM Reference Format:
Syrine Haddad, Kayhan Latifzadeh, Saravanakumar Duraisamy, Jean Vanderdonckt, Olfa Daassi, Safya Belghith, and Luis A. Leiva. 2024. Good GUIs, Bad GUIs: Affective Evaluation of Graphical User Interfaces. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 1–4, 2024, Cagliari, Italy.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3627043.3659549

1 INTRODUCTION

Determining whether the design of a Graphical User Interface (GUI) is good or bad remains a question that needs to be answered as early as possible (the earlier a design error is detected, the less costly it will be to repair [35, 69]), as quickly as possible (the more efficient the evaluation, the faster it can improve the design, and the less costly it will prove [23]), and as little interventionist as possible (the less intrusive the evaluation is for the user, the less the user will be distracted [77] from participatory design [54]). To decide which evaluation method to apply among all methods [80],

[*] Equal contribution.

we can turn to Whitefield et al.'s framework [84], which classifies these methods according to two dimensions: with users (e.g., by user testing) or without, based on the real GUI (i.e., executable) or represented (e.g., mocked-up).

For a long time, methods *without* users attempted to automate the evaluation based on heuristics [17], guidelines [79], or user models of aesthetics [13, 60, 85], based on the assumption that what is beautiful is perceived as usable [74], even if it is not. These methods offer partial results that are subject to interpretation because their results, good or bad, remain independent of the users and their context of use [84]. In particular, prior work has focused on computational models that are expected to evaluate and predict aesthetics scores [28, 37, 60, 64]. However, these studies noted that aesthetic preferences are quite diverse, affected by the user's own taste [28], psychological traits [18], and demographic backgrounds [64]. Furthermore, aesthetic preferences collected through self-reported measures are often subject to carry-over effects and too much subjectivity [86].

On the other hand, an evaluation method *with* users is appreciated for the relevance and context-sensitivity of its results, but its implementation costs are feared [80]. Such a method is sometimes set up too late, in a costly and explicit way (the user has to explain why the design is evaluated as good or bad). We should therefore look for an evaluation method that is as early, efficient, and implicit as possible. While GUI design has been traditionally evaluated on factors such as usability (which is decomposed into effectiveness, efficiency, and satisfaction), understanding and evaluating their emotional impact [44, 45] has gained significant attention in recent years [76]. For example, Odushegun [57] emphasized that aesthetics go beyond visual beauty and encompass cognitive-affective responses, and Mastandrea et al. [49] found that positive aesthetic appreciations are beneficial to physiological well-being. Indeed, emotions can play a crucial role in shaping a better overall user experience [19, 56, 62], which goes beyond mere usability. Positive emotional experiences enhance user engagement, foster brand loyalty and engagement [25], and drive user retention [5]. Conversely, negative emotions can lead to frustration, disengagement, and abandonment of the GUI, resulting in a loss of users and potential revenue [5].

To bridge the gap between the need for objective measures and subjective user experiences towards GUI designs, this paper investigates whether an affective evaluation based on measures of affects could allow for a more holistic evaluation, providing a new angle of how GUI design might impact users. By leveraging affect-aware insights, designers can create GUIs that resonate with users, foster positive emotional experiences, and ultimately enhance the user experience. Moreover, this approach allows for implicit evaluation [31, 65], while users perceive and interact with GUIs as they would normally do. However, detecting affect in this context poses several challenges. While affective responses can also vary between individuals, we hypothesize that they can be more reliably detected than self-reported measures and lead to an overall consensus based on the principle that inter-subjectivity [30] among users becomes objective [86]. Furthermore, developing accurate affect detection models is a challenge in itself, considering that labeled affect-related datasets largely exist for images, pictures, and videos [2], but remain almost nonexistent for GUI designs.

To pursue this research, we investigate the affective responses that good and bad GUI designs produce on users. Comparing good GUI designs against bad ones seems straightforward: good designs focus on simplicity, clarity, consistency [6], and visual aesthetics [60] as opposed to bad designs which suffer from complexity, clutteredness, inconsistency, and lack of visual appeal. We hypothesize that end users may elicit some affective responses toward GUIs and their aforementioned properties, based on content and aesthetics. To test this hypothesis, we recorded neurophysiological and peripheral signals while users were exposed to various good or bad GUI designs: facial expressions [68], eye activity (e.g., fixations and pupil dilation) [3], and brain activity (electroencephalography, EEG) [32]. We also collected scores about users' subjective experiences and preferences towards GUI designs, as a means to assess the validity of our findings.

Taken together, this paper makes the following contributions:

- A controlled experiment to capture affective responses that good and bad GUI designs elicit on end users (Section 3) together with discussion about noticeable effects that we observed in the collected data (Section 4).
- Computational models for affect recognition that use different physiological signals (Section 5): eye-gaze behavior (fixations and pupil dilation), facial expressions, and brain activity (EEG).
- Contextualization of our findings with insights for designers and researchers (Section 6).

2 RELATED WORK

Our work primarily builds on two research areas: GUI aesthetics and emotion recognition.

2.1 GUI aesthetics

A GUI design can be evaluated using a large number of different methods [80], among which (perceived) aesthetics has been found to impact a variety of features, including e.g. perceived usefulness, credibility, and intention to revisit [7, 53, 74, 75]. While Gwak and Park [21] found no significant differences in user responses towards GUI designs, Bölte et al. [7] found that expert and non-expert designers can detect bad designs but experts tend to evaluate good designs as bad more often than non-experts. An automatic evaluation on websites and mobile apps [51] reinforced past findings [78] that suggested that people rely on similar visual cues [73], regardless of exposure duration.

Standardized questionnaires are typically used for GUI evaluation, including affect [83] and aesthetics [64], which is again a form of self-reported measure that requires explicit user feedback. By relying on neurophysiological measurements, recorded during natural interactions, we move away from explicit evaluation methods with users (in which users must explicitly give the result of their evaluation) and implicit methods without users (in which a user model is supposed to represent them as accurately as possible).

2.2 Emotion recognition

The research literature on emotion recognition is vast [8, 15, 20, 48, 71], therefore we refer to recent surveys [9, 38] and focus on recent approaches that used facial expressions, eye tracking, and brain signals.

2.2.1 Using facial expressions. Toisoul et al. [72] detected basic appraisals of affect[1] (valence and arousal) [66] in three datasets, achieving an average 75% accuracy on both affect dimensions. Minaee et al. [50] used an attentional CNN to mask out some face parts depending on the emotion to be detected. Their model achieved excellent accuracy (e.g., 98% on CK+ [46] and 99% on FERG [1]) to classify Eckman's six universal facial emotions, i.e., happiness, sadness, anger, disgust, fear, and surprise against a neutral status (binary classification tasks). The main limitation of these approaches is that they rely on labeled datasets collected from actors, not from real end users, thus making them unrealistic and impractical for evaluating GUI designs, as end users typically do not express their emotions in a clear way when e.g. browsing a website or looking at a GUI screenshot.

2.2.2 Using eye tracking. Eye movement-based analysis has long been used for (re)designing [4], generating [11], and evaluating GUIs [16]. However, very few studies (e.g. [29]) have explored eye tracking in the context of affect for GUI design. Since eye movements have been found to be unreliable for emotion recognition,[2] they are typically

[1] Arousal is defined as the level of autonomic activation caused by an event, and can range from calm (or low) to excited (or high). Valence is the degree of pleasantness that an event produces, and is defined along a continuum from negative to positive.

[2] See e.g. https://www.scientificamerican.com/article/darwin-was-wrong-your-facial-expressions-do-not-reveal-your-emotions/

complemented with other physiological methods, such as EEG [41]. Partala and Surakka [61] analyzed the valence and arousal of auditory emotional stimuli and found that the pupil size was significantly larger for both emotionally negative and positive stimuli than for neutral ones. Oliva and Anikin [58] showed that changes in pupil diameter are correlated with task efficiency and Holmes and Zanker [26] found that fixations are a good proxy of GUI's aesthetic quality. In sum, little is known about eye movement behavior for affective responses to GUI designs.

2.2.3 *Using electroencephalography.*

Affect recognition using EEG mainly involves the analysis of various frequency bands, each linked to specific mental states and activities [70]: Delta (δ: 0.5 to 4 Hz or less than 4 Hz), linked to deep sleep and unconsciousness, Theta (θ: 4 to 8 Hz), associated with light sleep and relaxation, Alpha (α: 8 to 12 Hz), connected to wakeful relaxation and tranquility, Beta (β: 13 to 30 Hz), correlated with active thinking and concentration, and Gamma (γ: 30 to 100 Hz), involved in sensory processing and attention [55].

On the DEAP dataset [33], one of the most popular ones for affect recognition, Liang et al. [39] used and unsupervised learning method that achieved 54.45% (valence) and 62.34% (arousal) accuracy. Mokatren et al. [52] used wavelet packet decomposition (WPD) to divide the EEG signal into the five sub-bands (δ, θ, α, β, and γ) and achieved a classification accuracy around 91%. Cheng et al. [10] achieved over 95% classification accuracy of valence and arousal with a 2D frame sequence that captured the spatial position relationships across EEG channels.

2.2.4 *Combining multiple inputs.*

Affect can be detected using more than one modality, for example by combining face with voice expressions [12, 22] or EEG with eye movements [47, 88]. Luo et al. [47] analyzed EEG signals and eye movements and found that positive and negative emotions are often confused in the eye domain but not so in the EEG domain. This multimodal system improved the unimodal solution by 10 percentual points. Zheng et al. [88] also combined EEG with eye tracking, and used a Support Vector Machine (SVM) that achieved better accuracy by combining both modalities than when using any modality alone.

3 METHOD

We conduct a controlled within-subjects experiment to capture the affective responses that GUI designs elicit from end users. We use three physiological signals to evaluate whether a GUI design is perceived as good or bad.

3.1 Stimuli

We selected 40 GUI designs from the LabintheWild dataset [64], that includes ground-truth judgments on a 9-point Likert scale [40] of first-impression aesthetic appeal of 418 websites from about 32k people around the world. For our experiment, we analyzed the distribution of the provided user ratings on the Labinthewild dataset and selected the top-20 rated ones as good designs and the bottom-20 rated ones as bad designs (see Figure 1). The choice of top and bottom GUIs reflects the real-world variability of user judgements and can help in building a more holistic understanding of UI design principles.

3.2 Participants

Thirty-two participants were recruited through our organization's mailing lists and flyer advertising. Two outliers were removed due to low data quality and one for data recording failure, which resulted in a final user sample of 29 participants (9 females, 20 males) aged between 18 and 46 years (M =29.28, SD=6.27, Mdn=26.5). Participants provided their written consent for the experiment and were paid 25 EUR. All of them had normal or corrected-to-normal vision.

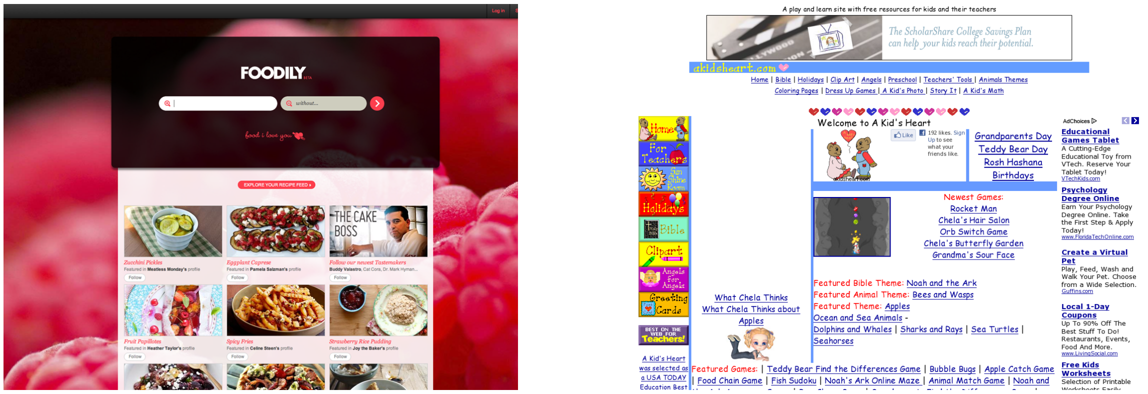

Log in
FOODILY BETA
without...
food i love you
EXPLORE YOUR RECIPE FEED »
Zucchini Pickles
Featured in Meatless Monday's profile
Follow
Eggplant Caprese
Featured in Pamela Salzman's profile
Follow
THE CAKE BOSS
Follow our newest Tastemakers
Buddy Valastro, Cat Cora, Dr. Mark Hyman...
Follow
Fruit Papillotes
Featured in Heather Taylor's profile
Follow
Spicy Fries
Featured in Celine Steen's profile
Follow
Strawberry Rice Pudding
Featured in Joy the Baker's profile
Follow
A play and learn site with free resources for kids and their teachers
The ScholarShare College Savings Plan can help your kids reach their potential.
akidsheart.com
Home | Bible | Holidays | Clip Art | Angels | Preschool | Teachers' Tools | Animals Themes
Coloring Pages | Dress Up Games | A Kid's Photo | Story It | A Kid's Math
Home
For Teachers
Sun Shine Room
Holidays
Bible
Clipart
Angels for Angels
Greeting Cards
BEST ON THE WEB FOR Teachers!
A Kid's Heart was selected as a USA TODAY Education Best
Welcome to A Kid's Heart
Like
192 likes. Sign Up to see what your friends like.
akidsheart.com
Grandparents Day
Teddy Bear Day
Rosh Hashana
Birthdays
Newest Games:
Rocket Man
Chela's Hair Salon
Orb Switch Game
Chela's Butterfly Garden
Grandma's Sour Face
What Chela Thinks
What Chela Thinks about Apples
Featured Bible Theme: Noah and the Ark
Featured Animal Theme: Bees and Wasps
Featured Theme: Apples
Ocean and Sea Animals -
Dolphins and Whales | Sharks and Rays | Sea Turtles |
Seahorses
Featured Games: | Teddy Bear Find the Differences Game | Bubble Bugs | Apple Catch Game
| Food Chain Game | Fish Sudoku | Noah's Ark Online Maze | Animal Match Game | Noah and
AdChoices
Educational Games Tablet
A Cutting-Edge Educational Toy from VTech. Reserve Your Tablet Today!
VTechKids.com
Psychology Degree Online
Earn Your Psychology Degree Online. Take the First Step & Apply Today!
www.FloridaTechOnline.com
Create a Virtual Pet
Play, Feed, Wash and Walk Your Pet. Choose from a Wide Selection.
Guffins.com
Local 1-Day Coupons
Up To 90% Off The Best Stuff To Do! Restaurants, Events, Food And More.
www.LivingSocial.com
Free Kids Worksheets
Selection of Printable

Fig. 1. Examples of a good GUI (left) and a bad GUI (right) design. These examples received the closest ratings to those of each group's centroid.

Most participants (78%) reported that they never attended any course on GUI design. This study was approved by the Ethics Review Panel of the University of Luxembourg with ID 22-071.

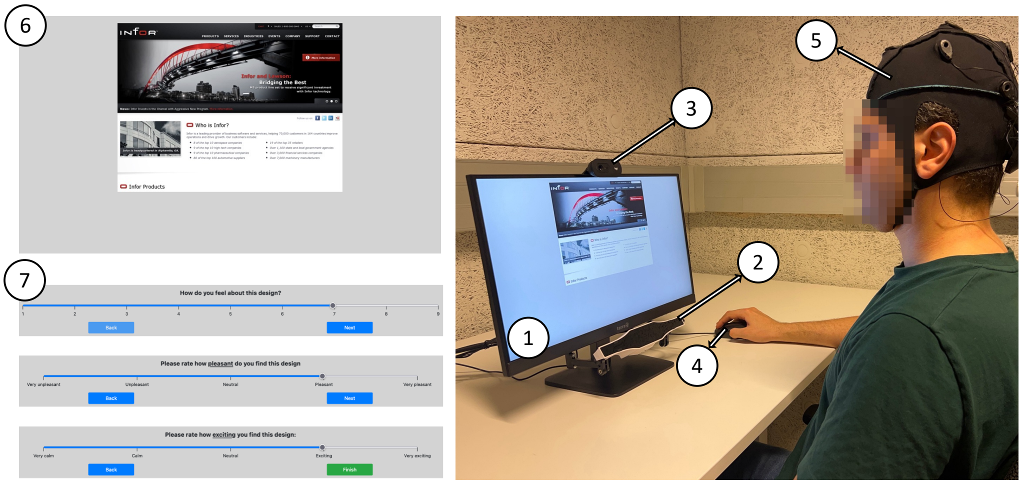
3.3 Apparatus

Figure 2 shows the experimental setup. We recorded the participants' faces with a Logitech C505e HD Business Webcam with 720p resolution at a frame rate of 30 frames per second. EEG data were collected using a Unicorn Hybrid Black device, with 8 channels (Fz, C3, Cz, C4, Pz, PO7, Oz, and PO8) at a sampling frequency of 250 Hz and the data were notch-filtered at 50 Hz to remove powerline interferences. While either dry electrodes and conductive gel can be equally used with this device, we opted for conductive gel to ensure optimal data quality collection. Eye-tracking data were recorded with a Gazepoint GP3 Eye Tracker at a sampling frequency of 150 Hz which was mounted on a 21.5" Lenovo L22e-20 monitor, offering FullHD resolution (1920 × 1080 px) and a refresh rate of 75 Hz.

3.4 Procedure

A *warm-up test* allowed participants to familiarize themselves with the setup before proceeding with the main experiment. They were presented with randomly selected GUI designs from the LabintheWild dataset. Note that these GUIs were not included in the actual experiment, only in the warm-up test.

The experiment consisted of two sessions: an *initial session* for primary evaluation and a *second session* for confirmatory evaluation, to check whether participants consistently evaluated the GUI designs (test-retest reliability). In each session, every trial started with a 5-second resting time, that served as a baseline for the physiological recordings, followed by a 10-second exposure to a randomly selected GUI design from the curated dataset, followed by a rating screen, having: a 9-point slider to provide an overall evaluation as an answer to the question "How do you feel about this design?", a 5-point slider in Likert scale [40] to answer the question "Please rate how pleasant do you find this design" ranging from 1 (very unpleasant) to 5 (very pleasant), and another 5-point slider on a Likert scale to answer the question "Please rate how exciting you find this design" ranging from 1 (very calm) to 5 (very exciting). The first question allowed us to compare the validity of the dataset, as discussed in the next section. The second and third questions were control questions related to valence and arousal (Figure 3), respectively.



How do you feel about this design?

Please rate how pleasant do you find this design

Please rate how exciting you find this design:

Fig. 2. Experimental setup. Participants wore an EEG cap (5) and sat in front of a monitor (1) equipped with an eye-tracker (2) and a webcam (3). A computer mouse (4) allowed participants to enter ratings (7) after being exposed to each GUI design (6). User ratings were used only as a quality measure.

Participants had the flexibility to switch between questions and adjust their answers accordingly before moving on to the next trial. Participants completed 20 bad GUI designs and 20 good GUI designs that were randomly shown to them. A 5 min break was allowed before starting the confirmatory session with the same stimuli, but presented in a different (also randomized) order.

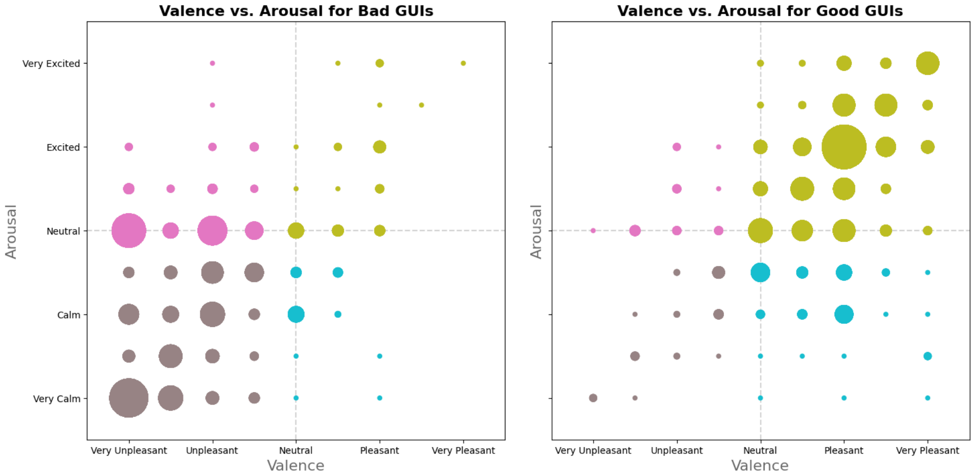


Fig. 3. Distribution of samples across the valence-arousal plane for bad (left) and good (right) GUI designs. The circle radius represents the relative sample density. Colors denote values in each quadrant.

4 RESULTS

Figure 4 shows the distribution of participants' answers to the three questions about overall rating, valence, and arousal for bad and good GUI designs. None of them followed a normal distribution (all Shapiro-Wilk tests returned a *W-stat* $\geq .05$, $p \leq .01^{***}$), therefore we computed a Wilcoxon signed-rank test for each sample to determine whether each distribution significantly departs from their respective median (*Mdn*=5 for ratings and *Mdn*=3 for valence and arousal).

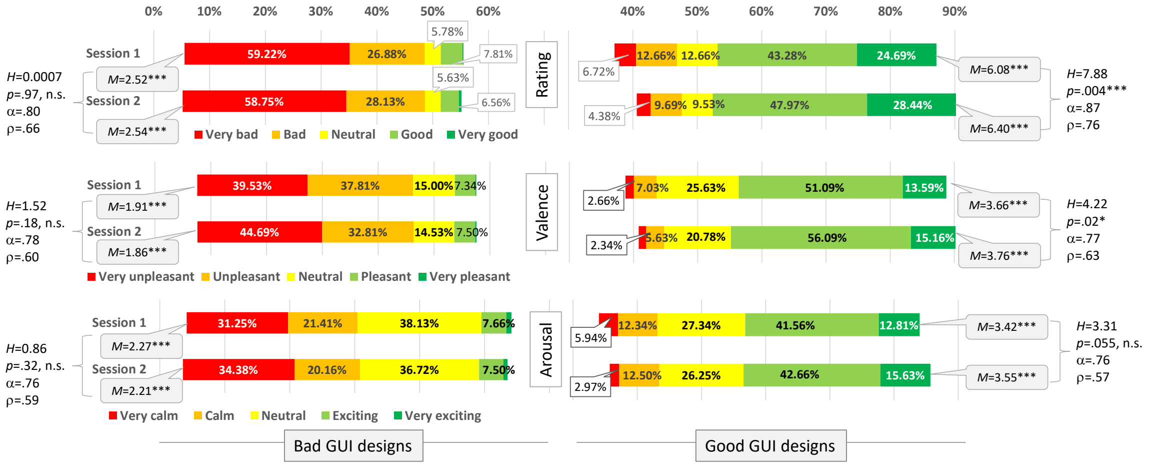
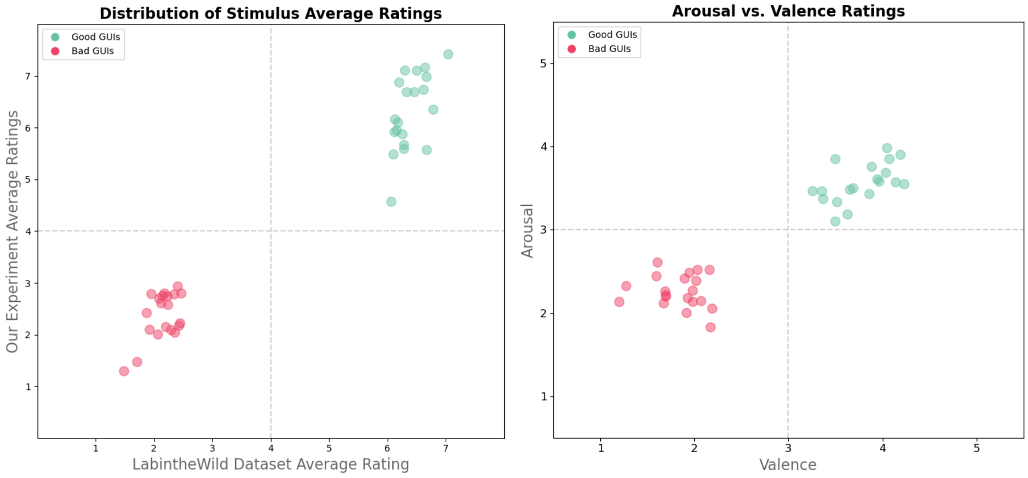


Fig. 4. Distribution of participant's answers regarding rating, valence, and arousal for bad and good GUI designs.

All values were significantly below their median for bad GUI designs, while all the respective values were significantly above their median for good GUI designs during both sessions. For example, the ratings of good GUI designs during the first session were statistically significantly above their median (n = 640, *z-score*= 11.90, $p \leq .001^{***}$ with a moderate effect size $r = .47$). All bad GUI designs received answers that were not statistically different across sessions (Kruskall-Wallis tests between sessions were not statistically significant - for example, regarding the arousal, $H = 0.86$, $p = .32$, *n.s.*). The rating for good GUI designs was different between sessions (Kruskall-Wallis test returned $H = 7.88$, $p = .004^{***}$), as well as the valence (Kruskall-Wallis test returned $H = 4.22$, $p = .02^{*}$).

To verify the inter-rater consistency, we computed Cronbach's α coefficient for each variable by category of GUI design (interpretation: 0.9>α≥0.8=good, 0.8>α≥0.7=acceptable). All α values are acceptable for the valence and the arousal and are good for the ratings ($\alpha = .80$ for bad GUI designs vs. $\alpha = .87$ for good GUI designs). We also computed Spearman's rank correlation coefficient to determine whether the variables are correlated between sessions (interpretation: $\rho \geq .70$=very strong relationship, .4≤ρ≤.69= strong relationship). Since all values range from $\rho = .57$ for the arousal of good GUI designs to $\rho = .76$ for the rating of good GUI designs, these (strongly) positive coefficients suggest that their respective values between sessions tend to occur or evolve together.

We also checked whether the participants' ratings were adequate, since the web aesthetics LabintheWild dataset we used [64] was compiled ten years ago, therefore we were unsure whether users would perceive the GUI designs in a similar way. Figure 5a shows a consistent match in the ratings (Pearson's $r(38) = 0.97, p < .0001^{***}$), indicating that the web aesthetics dataset is indeed adequate for our research. For questions 2 and 3 about pleasantness (valence) and excitement (arousal) with the GUI designs, we plotted the distribution of participants' GUI ratings across the valence-arousal plane; see Figure 5b. The bad GUI designs were rated in the '(low valence, low arousal)' coordinates, contrary to the good GUI designs which were rated in the '(high valence, high arousal)' coordinates.



(a) Comparison of user ratings.

(b) GUI designs in the valence-arousal plane.

Fig. 5. Distribution of samples across user ratings (left) and in the valence-arousal plane (right).

4.1 Investigation of eye activity

We started by analyzing the time series data for pupil size in each trial, based on three time intervals: the initial second, the first 5 seconds, and the complete 10-second duration. Missing values (e.g. due to blinks) were filtered out by interpolation. Subsequently, we smoothed the pupil size signal using a Savitzky-Golay filter [67] with a second-order polynomial [59] and a window size of 300 ms. Next, we calculated the differences between consecutive smoothed pupil size values, and then performed a min-max normalization [42] to adjust the difference values between 0 and 100. Figure 6 summarizes the results.

We observed a noticeable difference between the pupil size time series of good and bad GUIs within the first second of exposure. A t-test revealed statistically significant differences ($p < .05^*$). Upon closer examination of the data, we identified an initial increase in pupil dilation occurring approximately within the first 200 ms, followed by a subsequent decrease within the initial 500 ms. This pattern suggests that users' cognitive resources may have been fully engaged right from the outset when exposed to a poorly designed GUI. Furthermore, this initial pattern noted during the trials indicates that it may not be necessary to analyze pupil size throughout the entire duration. Instead, extracting features from the very beginning of the eye signal (up to 1 s) may suffice to tell good and bad designs apart.

We also extracted all fixations that occurred within the trial interval while users were viewing the GUIs, ignoring any fixations that fell outside these temporal boundaries. As in our initial exploration of fixation data, we examined various durations (from the first 1 second up to the tenth second) and counted the number of fixations within each interval; see Figure 7. As observed, fixation count is not a discriminative metric.

We also plotted fixation heatmaps for each design category and interval duration; see Figure 8. We observed that, overall, good GUIs tend to concentrate more symmetrical distributions of fixations, which is reasonable given that good

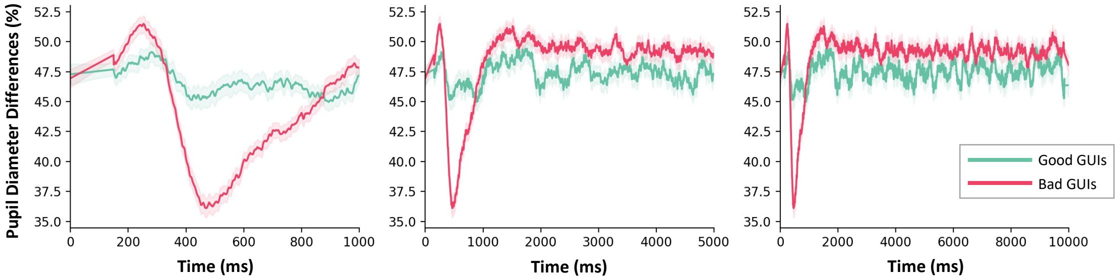


Fig. 6. Average pupil dilation for good and bad GUIs at different trial durations. Shaded areas represent 95% confidence intervals.

design guidelines recommend to ensure visual consistency by symmetry [63]. According to the paired-samples t-test, the differences between fixation distributions are statistically significant for all durations: 1s: $t(19) = 2.10, p = .04^{*}$; 5s: $t(19) = 2.98, p = .005^{***}$; 10s: $t(19) = 3.63, p = .001^{***}$.

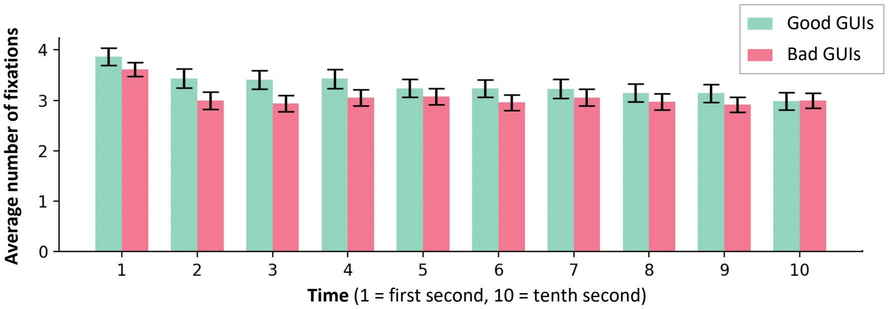


Fig. 7. Average number of fixations for good and bad GUI designs across various trial durations. Error bars represent 95% confidence intervals.

4.2 Investigation of face activity

From the camera footage captured during the experiment, we considered video fragments in which the GUI screenshots were shown to the participants (i.e., excluding pre-trials and rating times). Each participant had 40 videos of 10 seconds per session, totalling 80 video files. We extracted frames for every 0.5 seconds when we only considered the first 5 seconds, and frames for every 2 seconds when considering the whole duration trial (10 seconds). The total number of frames when processing only the first 5 seconds was 13,041 files split into 80-10-10 (80% for training, 10% for validation, and 10% for testing) whereas the total number when working on the whole trial (10 seconds) was 6,300 files and the same splitting method was applied. We did not consider the 1-second window because the number of frames was too small to train a competitive model. We ensured that each video frame contained only the participants' face, using the OpenCV library for video processing and the pre-trained Haar Cascade classifier [81] for face detection. The frames were converted to grayscale, to make our model color-invariant to color and less sensitive to lighting conditions.

In order to gain some insights about the collected data, we trained a CNN model, to be described in Section 5, and incorporated activation hooks. These hooks were attached to specific CNN layers that capture the hierarchical features of the input. After feeding the data to the trained model and generating predictions, we extracted activation maps for the

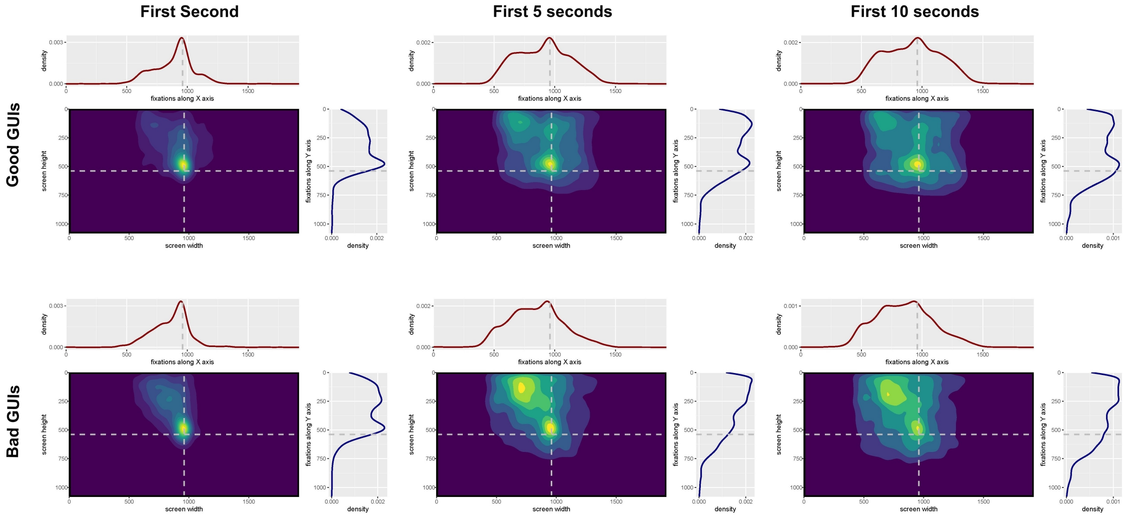
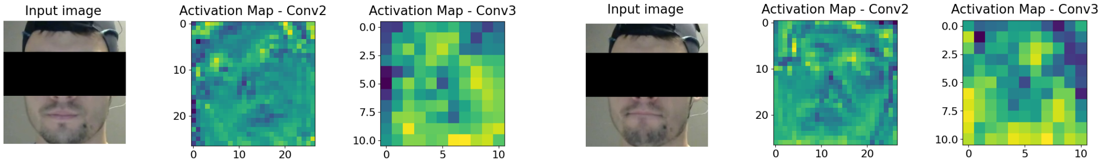


Fig. 8. Heatmaps of eye fixations for good and bad GUI designs at different trial durations.

CNN layers of interest (Conv2 and Conv3, see Figure 9). These activations maps are essentially visual representations of the important regions within the images that the model considers while classifying GUI screenshots.

Figure 9 shows the visualization of the activation maps alongside the original input face images. We observed clear differences between the visual representations of the last layer (Conv3) when participants were looking at good and bad GUI designs. Following these observations, we computed a t-test between the distributions of feature maps. The results revealed statistically significant differences between feature maps for both the Conv2 layer ($t(19) = 2.46, p < .01^{**}$) and the Conv3 layer ($t(19) = 2.46, p < .01^{**}$) between good and bad GUIs.



(a) Good GUI example

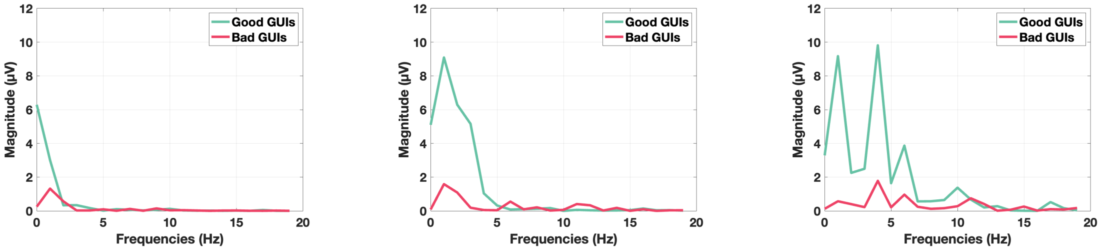
(b) Bad GUI example

Fig. 9. Visualizing the activation maps of facial expressions while looking at GUI designs.

4.3 Investigation of brain activity

EEG data were analyzed in time windows of 1, 5 and 10 seconds to find specific activities or patterns during our experiment. The data were pre-processed using the MATLAB-EEGLAB toolbox [14]. A Butterworth Infinite Impulse Response (IIR) bandpass filter was applied, with the high-pass cutoff frequency set to 0.05 Hz to remove slow drifts and baseline shifts, and the low-pass cutoff frequency set to 80 Hz to eliminate high-frequency noise. Eye movements and muscle artifacts were removed using Independent Component Analysis (ICA). In addition, a baseline correction was performed on the EEG data to mitigate potential drifts and ensure a more accurate representation of neural activity. Subsequently, the filtered and artifact-removed EEG data were categorized into bad and good GUIs to identify the respective activities and patterns.

The Welch method was employed to estimate the spectral density in different time windows, as shown in Figure 10. The magnitude of the power spectrum is significantly higher for good GUIs compared to bad GUIs, as illustrated in Figure 10, highlighting a significant difference in spectral patterns between the two types of GUI designs. The spectral activities across different brain regions were explored using brain topographical maps, as shown in Figure 11. The topographical maps of EEG were generated using 1, 5 and 10 second time windows of EEG data. In Figure 11, the activities linked to good and bad GUIs are illustrated in the first and second columns of each subplot, respectively.

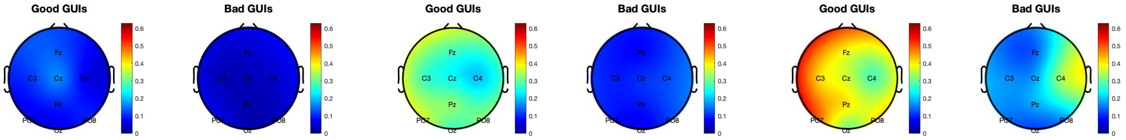


(a) First second

(b) First 5 seconds

(c) First 10 seconds

Fig. 10. EEG spectral activity at different trial durations.



(a) First second

(b) First 5 seconds

(c) First 10 seconds

Fig. 11. Brain topographic map of power distribution at different trial durations.

The observed brain activity patterns in response to good and bad GUIs reflect consistent neural responses. The reduced activity observed throughout the brain during the initial 5 seconds of perceiving bad GUIs suggests that users often need to allocate more cognitive resources to process poorly designed interfaces (Figure 11a and Figure 11b). The right hemisphere, linked to various cognitive functions such as complex spatial processing and problem solving, shows increased activity after 5 seconds while perceiving bad GUIs, indicating a heightened cognitive effort (Figure 11c). This suggests potential engagement of specific cognitive processes associated with spatial reasoning and problem solving tasks. On the other hand, the increased activity in the left hemisphere when perceiving good GUIs, particularly in areas Fz, C3, PO7, and Pz, indicates that users may experience a more focused and intuitive cognitive engagement when perceiving well-designed GUIs (Figure 11c). The left hemisphere is often associated with logical reasoning, and attention to detail.

The consistent activity observed in the Pz electrode (Figure 11b and Figure 11c) located in the parietal region, suggests that attention and cognitive processes play a significant role during GUI perception. The parietal region is known to be involved in spatial processing and attention, which aligns with the idea that users may pay more attention to well-designed interfaces. The increased activity in this area likely indicates that users are more engaged and attentive when interacting with good GUIs, possibly due to their more intuitive design.

The desynchronization of energy observed in the central and fronto-central regions of the brain's topographical maps signifies changes in cognitive processes. This desynchronization is often associated with executive functions, decision-making, temporal processing, and attention. The decreased activity in these regions when viewing good GUIs, as shown in Figure 11, suggests that users may experience reduced cognitive engagement. One interpretation of this reduced activity is that users find that good GUIs are more engaging and intuitive, requiring less mental effort to navigate and make decisions. This aligns with the idea that well-designed interfaces streamline the user experience, leading to smoother and less mentally taxing interactions. The central region, represented by electrode Cz, plays a crucial role in motor planning and execution. It also suggests that users may not need to engage in complex motor actions when dealing with poorly designed interfaces (Figure 11). From a neuroscience perspective, this observation implies that bad GUIs may demand more cognitive effort to be processed visually. Users might struggle to navigate and make sense of suboptimal interfaces, leading to altered brain activity patterns, including decreased motor-related activity.

5 COMPUTATIONAL MODELS

We investigate three computational models for each of the three input modalities we have considered: facial expressions, eye tracking, and brain activity. To preserve the rating consistency between the two sessions, we excluded the trials where the difference between the two sessions was greater than 2 points. Eventually, only 328 trials out of 3480 (less than 10% of the trials) were excluded for model training. In all cases, we considered two categories for classification: 'bad' (GUIs that received low ratings, < 4 points) and 'good' (GUIs that received high ratings, > 6 points). As a reminder, the user ratings were distributed in the $[1, 9]$ range.

5.1 GUI assessment from facial expressions

We used the CNN architecture developed by Haddad et al. [22], which was specifically designed to recognize affective responses from audio-visual inputs. We took the part that works with video inputs and fine-tuned the model hyperparameters to suit our data. We also excluded the multi-layer perceptron (MLP) component for our model.

Prior to feeding the images into the model, we resized the detected faces in each frame to a uniform size of 62x62 px. We also applied random horizontal flipping and rotation (up to 30 degrees) as data augmentation techniques, to increase the model's robustness. We considered two durations for data collection (the first 5 seconds and the whole trial of 10 seconds). We split the data into 80% of the videos for training, 10% for validation, and 10% for testing. We used cross-entropy loss and the Adam optimizer with a learning rate of 0.0001, momentums $\beta_1 = \beta_1 = 0.99$, and a weight decay of 0.0001. The model was trained for 25 epochs using early stopping with patience of 5 epochs. It achieved 72% Accuracy and 71% AUC for the first 5 seconds, and 74% Accuracy and 74% AUC for the entire trial duration (10 seconds).

5.2 GUI assessment from eye activity

We trained kNN classifiers using varying signal durations: 1 s, 5 s, and 10 s. We determined the best k value by conducting a grid search over the range of (1, 3, . . . , 9). The accuracy of each classifier is presented in Figure 12. The highest performance (71% AUC and 73% accuracy) was achieved by a 7-NN classifier using 5 s data.

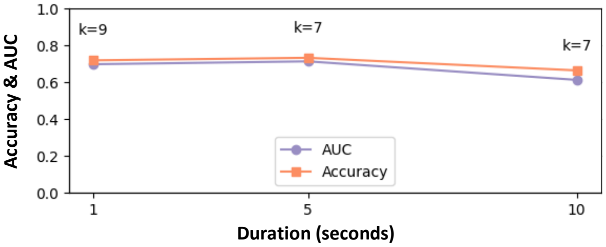


Fig. 12. Accuracy and AUC of kNN models trained on pupil dilation data at different trial durations.

In order to factor in temporal dependencies, in order to capture patterns over time, we employed three common types of recurrent neural networks (RNNs) to classify good and bad GUIs: Vanilla RNN, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). Each sample in our classification task comprised a sequence of fixations, represented by the timestamp of the fixation and its (x,y) coordinates. For each type of classifier, we trained two versions: non-bidirectional and bidirectional. The best classification results were achieved with a non-bidirectional GRU trained with 5 s of data, which achieved an Accuracy of 54.1% and an AUC of 53.4%. Results suggest that neither the number of fixations or their temporal evolution is discriminative enough for these recurrent models to tell bad and good GUIs apart.

5.3 GUI assessment from brain activity

We extracted four frequency bands (θ, α, β, and γ) by applying the fast Fourier transform (FFT) over a temporal window of 5 s. Subsequently, we computed five widely used EEG features: Three Hjorth parameters (activity, mobility, complexity) [24], spectral entropy [87], and signal energy. To create samples for EEG signal classification, we experimented with two different signal durations: the first 5 seconds and the entire trial duration (10 seconds). In the latter case, we divided each trial's signal into two samples. We concatenated features derived from all extracted frequency bands of each channel into a single feature vector. Subsequently, we divided the resulting samples into test (20%) and train (80%) sets and trained SVMs and kNN classifiers. Both SVMs and kNNs are widely used in EEG classifications since the data is so sparse that deep learning models tend to overfit [27, 36, 82]. Finally, aiming at identifing the optimal model configuration, we considered factors such as signal duration (the initial 5 seconds or the entire trial duration) and the choice between individual frequency bands or their combination. The best classification performance (67% Accuracy and AUC) was obtained by a 1-NN classifier using the Beta frequency band.

6 DISCUSSION, LIMITATIONS, AND FUTURE WORK

Evaluating GUI designs traditionally involves self-reported measures that are prone to bias and carry-over effects (e.g., subjective user feedback). We have introduced an innovative approach that considers three types of physiological data:

facial expressions, eye activity (including fixations and pupil dilation), and brain activity via EEG. In a nutshell, we have conducted a comprehensive experiment aimed at gathering affective responses that can help us distinguish between good and bad GUI designs in a free-viewing task, without having to ask for explicit user feedback.

Despite previous research suggesting that facial expressions might not be a reliable source, due to users often maintaining a neutral expression during interaction [43, 45], our computational model achieved an accuracy rate of 72% after 5 seconds of exposure and 74% after the trial duration (10 seconds). This approach can be beneficial for designers in quickly evaluating GUI quality based on facial expressions.

We anticipated observing increased pupil dilation when users encountered poorly designed GUIs, as it typically correlates with a higher cognitive load. Surprisingly, within the first second of encountering poorly designed GUIs, we noticed a visible decrease in pupil size. This contrasts with the conventional pattern observed in previous studies, where pupillary responses tend to increase with increasing task demands before stabilizing when cognitive resources are overwhelmed [34]. Furthermore, our fixation heatmaps revealed that, on bad GUI designs, fixations tend to be distributed more evenly. This suggests that users shifted their gaze to different GUI areas, possibly due to an excessive number of GUI elements (visual clutter) or poorly arranged components.

When it comes to EEG responses, feature engineering remains a critical aspect of utilizing EEG signals in affective computing. It is worth noting that collecting this type of data is a time-consuming and costly process, therefore the use of sophisticated Machine Learning models, such as deep neural networks, is challenging. Additionally, EEG data augmentation is complicated due to its non-stationary and noisy characteristics. We explored different frequency bands and signal durations using simple but effective classifiers. Again, we observed promising results in using neurophysiological signals for differentiating between good and bad GUI designs.

One of the limitations of our work is that every input modality is processed with a dedicated computational model. This was so because each modality has a "preferred" model architecture; for example, for eye pupil activity we can consider an RNN, whereas for brain activity we could not consider a neural network, given the aforementioned problem of data scarcity. Also, each signal has a "preferred" length; for example, for eye pupil activity, after 1 s of exposure, the differences between groups are not significantly different. For future work we will try to combine these different inputs in an end-to-end architecture. Another limitation of our study lies in categorizing GUIs only as good or bad, neglecting for example neutral designs or extremely bad (e.g., cluttered) and moderately bad (e.g., visually confusing) designs. We also did not consider individual preferences in our computational models, as they all are user-independent models.

To conclude, our research offers a promising avenue for advancing software interfaces and provides invaluable insights to researchers and designers. Our findings can pave the way for a new method for affective evaluation of GUIs, which hold promise for an objective assessment of GUIs in the software design industry and suggest opportunities for advancing progress thanks to more informed computational models.

ACKNOWLEDGMENTS

We thank Nima Gozalpour for helping us with the training of EEG classifiers. This work is supported by the Horizon 2020 FET program of the European Union (BANANA, grant CHIST-ERA-20-BCI-001) and Horizon Europe's European Innovation Council through the Pathfinder program (SYMBIOTIK, grant 101071147).

REFERENCES

[1] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones. 2017. Modeling Stylized Character Expressions via Deep Learning. In *Computer Vision – ACCV 2016*, Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (Eds.). Springer International Publishing, Cham, 136–153.

[2] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang. 2016. Emotion Recognition in the Wild from Videos Using Images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 433–436. https://doi.org/10.1145/2993148.2997627

[3] J. R. Bergstrom and A. Schall. 2014. *Eye Tracking in User Experience Design* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[4] A. Bojko. 2006. Using Eye Tracking to Compare Web Page Designs: A Case Study. *Journal of Usability Studies* 1, 3 (May 2006), 112–120. https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Bojko_May2006.pdf

[5] J. Bowden, J. Conduit, L. Hollebeek, V. Luoma-aho, and B. Solem. 2017. Engagement valence duality and spillover effects in online brand communities. *Journal of Service Theory and Practice* 27 (06 2017), 877–897. https://doi.org/10.1108/JSTP-04-2016-0072

[6] N. Burny and J. Vanderdonckt. 2022. (Semi-)Automatic Computation of User Interface Consistency. In *EICS '22: ACM SIGCHI Symposium on Engineering Interactive Computing Systems, Sophia Antipolis, France, June 21 - 24, 2022, Companion Volume*, Marco Winckler and Aaron Quigley (Eds.). ACM, 5–13. https://doi.org/10.1145/3531706.3536448

[7] J. Bölte, T. Hösker, G. Hirschfeld, and M. Thielsch. 2017. Electrophysiological correlates of aesthetic processing of webpages: A comparison of experts and laypersons. *PeerJ* 5 (06 2017), e3440. https://doi.org/10.7717/peerj.3440

[8] Y. Cai, X. Li, and J. Li. 2023. Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors* 23, 5 (2023). https://doi.org/10.3390/s23052455

[9] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences* 582 (2022), 593–617. https://doi.org/10.1016/j.ins.2021.10.005

[10] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen. 2021. Emotion Recognition From Multi-Channel EEG via Deep Forest. *IEEE Journal of Biomedical and Health Informatics* 25, 2 (2021), 453–464. https://doi.org/10.1109/JBHI.2020.2995767

[11] S. Cheng and A. K. Dey. 2019. I see, you design: user interface intelligent design system with eye tracking and interactive genetic algorithm. *CCF Trans. Perv. Comput. Int.* 1, 3 (2019), 224–236.

[12] K. Chengeta. 2018. Comparative Analysis of Emotion Detection from Facial Expressions and Voice Using Local Binary Patterns and Markov Models: Computer Vision and Facial Recognition. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing* (Las Vegas, NV, USA) *(ICVISP 2018)*. Association for Computing Machinery, New York, NY, USA, Article 27, 6 pages. https://doi.org/10.1145/3271553.3271574

[13] N. Chettaoui and M. S. Bouhlel. 2017. I2Evaluator: An Aesthetic Metric-Tool for Evaluating the Usability of Adaptive User Interfaces. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017, AISI 2017, Cairo, Egypt, September 9-11, 2017 (Advances in Intelligent Systems and Computing, Vol. 639)*, Aboul Ella Hassanien, Khaled Shaalan, Tarek Gaber, and Mohamed F. Tolba (Eds.). Springer, 374–383. https://doi.org/10.1007/978-3-319-64861-3_35

[14] A. Delorme and S. Makeig. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods* 134, 1 (2004), 9–21.

[15] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas. 2020. Human Emotion Recognition: Review of Sensors and Methods. *Sensors* 20, 3 (2020). https://doi.org/10.3390/s20030592

[16] P. Emami, Y. Yiang, Z. Guo, and L. A. Leiva. 2024. Impact of Design Decisions in Scanpath Modeling. In *Proceedings of the ACM Symposium on Eye Tracking Research an Applications (ETRA).*

[17] R. A. Fernandez, J. A. Deja, and B. P. V. Samson. 2018. Automating Heuristic Evaluation of Websites Using Convolutional Neural Networks. In *Proceedings of the Asian HCI Symposium'18 on Emerging Research Collection* (Montreal, QC, Canada) *(Asian HCI Symposium'18)*. Association for Computing Machinery, New York, NY, USA, 9–12. https://doi.org/10.1145/3205851.3205854

[18] K. Z. Gajos and K. Chauncey. 2017. The Influence of Personality Traits and Cognitive Load on the Use of Adaptive User Interfaces. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI 2017, Limassol, Cyprus, March 13-16, 2017*, George A. Papadopoulos, Tsvi Kuflik, Fang Chen, Carlos Duarte, and Wai-Tat Fu (Eds.). ACM, 301–306. https://doi.org/10.1145/3025171.3025192

[19] J. A. Galindo, S. Dupuy-Chessa, N. Mandran, and E. Céret. 2018. Using user emotions to trigger UI adaptation. In *2018 12th International Conference on Research Challenges in Information Science (RCIS)*. 1–11. https://doi.org/10.1109/RCIS.2018.8406661

[20] J. M. Garcia-Garcia, V. M. R. Penichet, and M. D. Lozano. 2017. Emotion Detection: A Technology Review. In *Proceedings of the XVIII International Conference on Human Computer Interaction* (Cancun, Mexico) *(Interacción '17)*. Association for Computing Machinery, New York, NY, USA, Article 8, 8 pages. https://doi.org/10.1145/3123818.3123852

[21] S. Gwak and K. Park. 2023. Designing Effective Visual Feedback for Facial Rehabilitation Exercises: Investigating the Role of Shape, Transparency, and Age on User Experience. *Healthcare* 11 (06 2023), 1835. https://doi.org/10.3390/healthcare11131835

[22] S. Haddad, O. Daassi, and S. Belghith. 2023. Emotion Recognition from Audio-Visual Information based on Convolutional Neural Network. In *2023 International Conference on Control, Automation and Diagnosis (ICCAD)*. 1–5. https://doi.org/10.1109/ICCAD57653.2023.10152451

[23] H. B. Hassan and Q. I. Sarhan. 2020. Performance Evaluation of Graphical User Interfaces in Java and C#. In *2020 International Conference on Computer Science and Software Engineering (CSASE)*. 290–295. https://doi.org/10.1109/CSASE48920.2020.9142075

[24] B. Hjorth. 1970. EEG analysis based on time domain properties. *Electroencephalography and clinical neurophysiology* 29, 3 (1970), 306–310.

[25] L. D. Hollebeek and T. Chen. 2014. Exploring positively- versus negatively-valenced brand engagement: a conceptual model. *Journal of Product & Brand Management* 23, 1 (2014), 62–74. https://doi.org/10.1108/JPBM-06-2013-0332

[26] T. Holmes and J. M. Zanker. 2012. Using an Oculomotor Signature as an Indicator of Aesthetic Preference. *i-Perception* 3, 7 (2012), 426–439. https://doi.org/10.1068/i0448aap arXiv:https://doi.org/10.1068/i0448aap PMID: 23145294.

[27] K. M. Hossain, M. A. Islam, S. Hossain, A. Nijholt, and M. A. R. Ahad. 2023. Status of deep learning for EEG-based brain–computer interface applications. *Frontiers in computational neuroscience* 16 (2023), 1006763.

[28] W. Hoyer and N. Stokburger-Sauer. 2011. The role of aesthetic taste in consumer behavior. *Journal of the Academy of Marketing Science* 40 (01 2011), 167–180. https://doi.org/10.1007/s11747-011-0269-y

[29] Y. M. Hwang and K. C. Lee. 2022. An eye-tracking paradigm to explore the effect of online consumers' emotion on their visual behaviour between desktop screen and mobile screen. *Behaviour & Information Technology* 41, 3 (2022), 535–546.

[30] I. Kant. 1987. *The Critique of judgment.* Hackett Publishing. https://monoskop.org/images/7/77/Kant_Immanuel_Critique_of_Judgment_1987.pdf

[31] A. Kaushik and G. J. F. Jones. 2023. Comparing Conventional and Conversational Search Interaction Using Implicit Evaluation Methods. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2023, Volume 2: HUCAPP, Lisbon, Portugal, February 19-21, 2023,* Alexis Paljic, Mounia Ziat, and Kadi Bouatouch (Eds.). SCITEPRESS, 292–304. https://doi.org/10.5220/0011798500003417

[32] W. Klimesch. 1999. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews* 29, 2 (1999), 169–195. https://doi.org/10.1016/S0165-0173(98)00056-3

[33] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[34] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one* 13, 9 (2018), e0203629.

[35] T. Landauer. 1996. *The trouble with computers: Usefulness, usability, and productivity.* The MIT Press. https://mitpress.mit.edu/9780262621083/the-trouble-with-computers/

[36] K. Latifzadeh, N. Gozalppour, V. J. Traver, T. Ruotsalo, A. Kawala-Sterniuk, and L. A. Leiva. 2024. Efficient Decoding of Affective States from Video-elicited EEG Signals: An Empirical Investigation. *ACM Transactions on Multimedia Computing Communications and Applications* (2024).

[37] L. Leiva, M. Shiripour, and A. Oulasvirta. 2022. Modeling how different user groups perceive webpage aesthetics. *Universal Access in the Information Society* (08 2022). https://doi.org/10.1007/s10209-022-00910-x

[38] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen. 2022. EEG Based Emotion Recognition: A Tutorial and Review. *ACM Comput. Surv.* 55, 4, Article 79 (nov 2022), 57 pages. https://doi.org/10.1145/3524499

[39] Z. Liang, S. Oba, and S. Ishii. 2019. An Unsupervised EEG Decoding System for Human Emotion Recognition. *Neural Netw.* 116, C (aug 2019), 257–268. https://doi.org/10.1016/j.neunet.2019.04.003

[40] R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 22, 140 (1932), 55–. http://psycnet.apa.org/record/1933-01885-001

[41] J. Z. Lim, J. Mountstephens, and J. Teo. 2020. Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors* 20, 8 (2020), 2384.

[42] X. Liu, Y. P. Sanchez Perdomo, B. Zheng, X. Duan, Z. Zhang, and D. Zhang. 2022. When medical trainees encountering a performance difficulty: evidence from pupillary responses. *BMC Medical Education* 22, 1 (2022), 1–9.

[43] D. Lockner and N. Bonnardel. 2014. Emotion and Interface Design How to measure interface design emotional effect?

[44] D. Lockner and N. Bonnardel. 2015. Towards the Evaluation of Emotional Interfaces. In *Human-Computer Interaction: Design and Evaluation - 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 9169),* Masaaki Kurosu (Ed.). Springer, 500–511. https://doi.org/10.1007/978-3-319-20901-2_47

[45] D. Lockner, N. Bonnardel, C. Bouchard, and V. Rieuf. 2014. Emotion and Interface Design. In *Proceedings of the 2014 Ergonomie et Informatique Avancée Conference - Design, Ergonomie et IHM: Quelle Articulation Pour La Co-Conception de l'interaction* (Bidart-Biarritz, France) *(Ergo'IA '14).* Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/2671470.2671475

[46] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops.* 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[47] S. Luo, Y.-T. Lan, D. Peng, Z. Li, W.-L. Zheng, and B.-L. Lu. 2022. Multimodal Emotion Recognition in Response to Oil Paintings. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).* 4167–4170. https://doi.org/10.1109/EMBC48229.2022.9871630

[48] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, Prabal Datta Barua, M. Murugappan, Y. Chakole, and U. R. Acharya. 2022. Automated emotion recognition: Current trends and future perspectives. *Computer Methods and Programs in Biomedicine* 215 (2022), 106646. https://doi.org/10.1016/j.cmpb.2022.106646

[49] S. Mastandrea, S. Fagioli, and V. Biasi. 2019. Art and Psychological Well-Being: Linking the Brain to the Aesthetic Emotion. *Frontiers in Psychology* 10 (4 Apr 2019), 739. https://doi.org/10.3389/fpsyg.2019.00739

[50] S. Minaee, M. Minaei, and A. Abdolrashidi. 2021. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors* 21 (04 2021), 3046. https://doi.org/10.3390/s21093046

[51] A. Miniukovich and A. De Angeli. 2015. Computation of Interface Aesthetics. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15).* Association for Computing Machinery, New York, NY, USA, 1163–1172. https://doi.org/10.1145/2702123.2702575

[52] L. S. Mokatren, R. Ansari, A. E. Cetin, A. D. Leow, O. A. Ajilore, H. Klumpp, and F. T. Yarman Vural. 2021. EEG Classification by Factoring in Sensor Spatial Configuration. *IEEE Access* 9 (2021), 19053–19065. https://doi.org/10.1109/ACCESS.2021.3054670

[53] M. Moshagen and M. T. Thielsch. 2010. Facets of visual aesthetics. *International Journal of Human-Computer Studies* 68, 10 (2010), 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006

[54] M. Muller. 2007. *Participatory design: The third space in HCI (revised).* Erlbaum, Mahway NJ USA.

[55] C. S. Nayak and A. C. Anilkumar. 2021. *EEG Normal Waveforms.* StatPearls Publishing, Treasure Island (FL). https://www.ncbi.nlm.nih.gov/books/NBK539805/

[56] M. Ninaus, S. Greipl, K. Kiili, A. Lindstedt, S. Huber, E. Klein, H.-O. Karnath, and K. Moeller. 2019. Increased emotional engagement in game-based learning – A machine learning approach on facial emotion detection data. *Computers & Education* 142 (2019), 103641. https://doi.org/10.1016/j.compedu.2019.103641

[57] L. Odushegun. 2023. Aesthetic semantics: Affect rating of atomic visual web aesthetics for use in affective user experience design. *International Journal of Human-Computer Studies* 171 (2023), 102978. https://doi.org/10.1016/j.ijhcs.2022.102978

[58] M. Oliva and A. Anikin. 2018. Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports* 8 (03 2018). https://doi.org/10.1038/s41598-018-23265-x

[59] E. Ostertagova and O. Ostertag. 2016. Methodology and Application of Savitzky-Golay Moving Average Polynomial Smoother. *Global Journal of Pure and Applied Mathematics* 12, 4 (08 2016), 3201–3210. https://www.ripublication.com/gjpam16/gjpamv12n4_35.pdf

[60] A. Oulasvirta, S. D. Pascale, J. Koch, T. Langerak, J. Jokinen, K. Todi, M. Laine, M. Kristhombuge, Y. Zhu, A. Miniukovich, G. Palmas, and T. Weinkauf. 2018. Aalto Interface Metrics (AIM): A Service and Codebase for Computational GUI Evaluation. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings, UIST 2018, Berlin, Germany, October 14-17, 2018*, Patrick Baudisch, Albrecht Schmidt, and Andy Wilson (Eds.). ACM, 16–19. https://doi.org/10.1145/3266037.3266087

[61] T. Partala and V. Surakka. 2003. Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies* 59, 1 (2003), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X Applications of Affective Computing in Human-Computer Interaction.

[62] J. L. Plass, S. Heidig, E. O. Hayward, B. D. Homer, and E. Um. 2014. Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction* 29 (2014), 128–140. https://doi.org/10.1016/j.learninstruc.2013.02.006

[63] R. Reber, N. Schwarz, and P. Winkielman. 2004. Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience? *Personality and Social Psychology Review* 8, 4 (2004), 364–382. https://doi.org/10.1207/s15327957pspr0804_3 arXiv:https://doi.org/10.1207/s15327957pspr0804_3 PMID: 15582859.

[64] K. Reinecke and K. Z. Gajos. 2014. Quantifying Visual Preferences around the World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/2556288.2557052

[65] T. Ruotsalo, V. J. Traver, A. Kawala-Sterniuk, and L. A. Leiva. 2024. Affective Relevance. *IEEE Intelligent Systems* (2024).

[66] J. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39 (12 1980), 1161–1178. https://doi.org/10.1037/h0077714

[67] A. Savitzky and M. J. E. Golay. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36, 8 (1964), 1627–1639. https://doi.org/10.1021/ac60214a047 arXiv:https://doi.org/10.1021/ac60214a047

[68] K. R. Scherer, H. Ellgring, A. Dieckmann, M. Unfried, and M. Mortillaro. 2019. Dynamic Facial Expression of Emotion and Observer Inference. *Frontiers in Psychology* 10 (2019). https://doi.org/10.3389/fpsyg.2019.00508

[69] J. Tan, K. Otto, and K. Wood. 2017. A comparison of design decisions made early and late in development. In *Proceedings of the 21st International Conference on Engineering Design* (Vancouver, Canada) *(ICED '17, Vol. 2)*. 41–50. https://www.designsociety.org/publication/39558/A+comparison+of+design+decisions+made+early+and+late+in+development

[70] M. Teplan et al. 2002. Fundamentals of EEG measurement. *Measurement science review* 2, 2 (2002), 1–11.

[71] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan. 2018. Emotion in a Century: A Review of Emotion Recognition. In *Proceedings of the 10th International Conference on Advances in Information Technology* (Bangkok, Thailand) *(IAIT 2018)*. Association for Computing Machinery, New York, NY, USA, Article 17, 8 pages. https://doi.org/10.1145/3291280.3291788

[72] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* 3 (01 2021). https://doi.org/10.1038/s42256-020-00280-0

[73] N. Tractinsky, A. Cokhavi, M. Kirschenbaum, and T. Sharfi. 2006. Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies* 64, 11 (2006), 1071–1083. https://doi.org/10.1016/j.ijhcs.2006.06.009

[74] N. Tractinsky, A. Katz, and D. Ikar. 2000. What is beautiful is usable. *Interacting with Computers* 13, 2 (2000), 127–145. https://doi.org/10.1016/S0953-5438(00)00031-X

[75] A. N. Tuch, S. P. Roth, K. Hornbæk, K. Opwis, and J. A. Bargas-Avila. 2012. Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior* 28, 5 (2012), 1596–1607. https://doi.org/10.1016/j.chb.2012.03.024

[76] S. Tzvetanova, M. Tang, and L. Justice. 2007. Emotional Web Usability Evaluation. In *Human-Computer Interaction. HCI Applications and Services, 12th International Conference, HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 4553)*, Julie A. Jacko (Ed.). Springer, 1039–1046. https://doi.org/10.1007/978-3-540-73111-5_114

[77] N. van Berkel, M. J. Clarkson, G. Xiao, E. Dursun, M. Allam, B. R. Davidson, and A. Blandford. 2020. Dimensions of ecological validity for usability evaluations in clinical settings. *J. Biomed. Informatics* 110 (2020), 103553. https://doi.org/10.1016/j.jbi.2020.103553

[78] P. van Schaik and J. Ling. 2009. The role of context in perceptions of the aesthetics of web pages over time. *International Journal of Human-Computer Studies* 67, 1 (2009), 79–89. https://doi.org/10.1016/j.ijhcs.2008.09.012

[79] J. Vanderdonckt and A. Beirekdar. 2005. Automated Web Evaluation by Guideline Review. *J. Web Eng.* 4, 2 (2005), 102–117. http://www.rintonpress.com/xjwe4/jwe-4-2/102-117.pdf

[80] A. P. O. S. Vermeeren, E. L.-C. Law, V. Roto, M. Obrist, J. Hoonhout, and K. Väänänen-Vainio-Mattila. 2010. User Experience Evaluation Methods: Current State and Development Needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (Reykjavik, Iceland) *(NordiCHI '10)*. Association for Computing Machinery, New York, NY, USA, 521–530. https://doi.org/10.1145/1868914.1868973

[81] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Vol. 1. I–I. https://doi.org/10.1109/CVPR.2001.990517

[82] B. Wang, C. M. Wong, F. Wan, P. U. Mak, P. I. Mak, and M. I. Vai. 2009. Comparison of different classification methods for EEG-based brain computer interfaces: a case study. In *2009 International Conference on Information and Automation*. IEEE, 1416–1421.

[83] J. Wang, Y. Liu, Y. Wang, J. Mao, T. Yue, and F. You. 2021. SAET: The Non-Verbal Measurement Tool in User Emotional Experience. *Applied Sciences* 11, 16 (2021). https://doi.org/10.3390/app11167532

[84] A. Whitefield, F. Wilson, and J. Dowell. 1991. A framework for human factors evaluation. *Behaviour & Information Technology* 10, 1 (1991), 65–79. https://doi.org/10.1080/01449299108924272 arXiv:https://doi.org/10.1080/01449299108924272

[85] M. Zen and J. Vanderdonckt. 2014. Towards an evaluation of graphical user interfaces aesthetics based on metrics. In *IEEE 8th International Conference on Research Challenges in Information Science, RCIS 2014, Marrakech, Morocco, May 28-30, 2014*, Marko Bajec, Martine Collard, and Rébecca Deneckère (Eds.). IEEE, 1–12. https://doi.org/10.1109/RCIS.2014.6861050

[86] M. Zen and J. Vanderdonckt. 2016. Assessing User Interface Aesthetics Based on the Inter-Subjectivity of Judgment. In *HCI 2016 - Fusion! Proceedings of the 30th International BCS Human Computer Interaction Conference, BCS HCI 2016, Bournemouth University, Poole, UK, 11-15 July 2016 (Workshops in Computing)*, Shamal Faily, Nan Jiang, Huseyin Dogan, and Jacqui Taylor (Eds.). BCS. http://ewic.bcs.org/content/ConWebDoc/56903

[87] A. Zhang, B. Yang, and L. Huang. 2008. Feature extraction of EEG signals using power spectral entropy. In *2008 international conference on BioMedical engineering and informatics*, Vol. 2. IEEE, 435–439.

[88] W.-L. Zheng, B.-N. Dong, and B.-L. Lu. 2014. Multimodal emotion recognition using EEG and eye tracking data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 5040–5043. https://doi.org/10.1109/EMBC.2014.6944757