Using users' physiological responses for the estimation of websites' aesthetic judgments

Giulio Gabrieli [1,*], Marc H. Bornstein [2,3] and Gianluca Esposito [1,4]

[1] *Psychology Program, School of Social Sciences, Nanyang Technological University, Singapore*
[2] *Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA*
[3] *Institute for Fiscal Studies, London, UK*
[4] *Department of Psychology and Cognitive Science, University of Trento, Italy*

Correspondence*:
Gianluca Esposito
gianluca.esposito@ntu.edu.sg

ABSTRACT

The aesthetic appearance of websites can influence the perception of their usability, reliability, and trustworthiness. Several studies investigated the relationship between single aesthetic features and explicit aesthetic judgments, demonstrating the existence of an attribution bias. However, only a limited amount of studies focused on the interaction between multiple visual properties and have considered not only explicit ratings, but also implicit judgments. In this work, we employ a novel approach, based on the analysis of physiological signals (implicit measures) and the application of machine learning and neural network models to predict users' perceived aesthetic pleasure from the empirical analysis of web pages' advanced visual properties (e.g. symmetry, visual complexity, colorfulness, ratio between visual and textual areas). Young adults (N=59, 33 females, Mean age = 21.52 years) assessed the aesthetic appeal of websites and emotional pictures while their physiological activity was recorded. Results using recursive partitioning and generalized linear models demonstrate the possibility of predicting the average aesthetic rating of a website using both explicit (behavioral ratings) and implicit measures (physiological activities).

Keywords: web design, aesthetics, physiology, ecg, eda, emg, pupillometry, machine learning, neural networks

1 INTRODUCTION

People interact with websites daily for work, educational purposes, and recreation. With the advent of more powerful technologies and an increasing number of active users, a new approach to the design and development of web pages was born, focusing no longer exclusively on content and functionalities, but also on pages' aesthetic appearance, which, as defined by Moshagen and Thielsch (2010) is *"an immediate pleasurable subjective experience that is directed towards an object"*.

Since the turn of the century, web design practices have evolved, encompassing a variety of disciplines, including visual design, user interface design (*UI*), user experience design (*UX*), scripting, programming, and content strategy (Robbins, 2012). User satisfaction is one of the many goals web designers aim to achieve, because satisfied users are more likely to spend more time on a page, come back to the same website in the future, and recommend a website to other possible users (Zhang and Von Dran, 2000). Users' evaluations of interactive systems are influenced by their visual appearance, and this is especially true for web pages (Karvonen, 2000; Kim et al., 2003; Zhang et al., 2001). In social psychology, the influence of aesthetic factors on other attributes is called *"halo effect"* or confirmation bias (Lindgaard et al., 2006). For example, individuals with more aesthetically pleasing faces are also perceived as more trustworthy.

Currently, the evaluation of a design artifact is an iterative process that requires time, money and external individuals who are asked to evaluate artifacts at different stages of the design process. Therefore, researchers tried to investigate possible ways to reduce the costs associated with the evaluation using heuristics or machine-based approaches.

Several attempts have been made to predict the perceived aesthetic perception of web pages, using a limited number of visual features as well as behavioral measures (Reinecke et al., 2013). Despite this, a limited number of studies have considered the role of users' exposure to different pages and their expertise in designing websites on their aesthetic evaluations. Moreover, a majority of the studies focused only on explicit behavioural measures.

In this study, we factored in multiple aesthetic features simultaneously and considered users' characteristics in different neural networks and machine learning models. Perceived aesthetic appeal of websites were estimated with both behavioral (self-reported; explicit measures) and physiological (ECG, EMG, EDA, and Pupillometry; implicit measures) measures, to overcome possible limits of self-report measures.

1.1 Visual properties and aesthetic judgment

Several features are known to affect website aesthetics judgments, including but not limited to visual complexity, perceived colorfulness, and symmetry (Cyr et al., 2010; Karvonen, 2000; Miniukovich and De Angeli, 2014). For example, Reinecke et al. (2013) assessed the impact of visual complexity and colorfulness on users' first impression of 450 websites. Results of their computational models show that visual complexity and colorfulness accounted for about half the variance in aesthetic judgments of web pages.

Visual complexity, also called visual cluttering, is a widely examined factor in aesthetic decision making, such that the more visually complex a design is, the higher the probability that it will be rated as more aesthetically pleasing (Seckler et al., 2015a; Tuch et al., 2009, 2012).

Despite its wide adoption in the literature, there is no single nor standard way to estimate a website's visual complexity. Some authors computed visual complexity by the weight of still-image (the weight of the file, expressed in kB or MB), whereas others define it by the space taken up by text and images, by calculating the number of colors, or by counting the number of images in a page (Bucy et al., 1999; Ivory et al., 2001).A promising method, proposed by Zheng et al. (2009), is based on a technique called Quadratic Tree Decomposition (QTD), often abbreviated as *quadtree decomposition*. The QTD recursively divides (horizontally and vertically) an image into areas of smaller size, if the parent area has a complexity -measured in terms of standard deviation of the area- higher than a predefined threshold. The final number of obtained squares, called *leaves*, is used as an index of visual complexity. When comparing images of the same size analyzed using the same complexity threshold value, the higher the number of leaves, the higher the visual complexity of an image. An example of Quadratic Tree Decomposition applied to websites is shown in Figure 1.

Similarly, color perception has been widely investigated in psychology, especially in relation to emotional valence and arousal (Wang and Ding, 2012). In Human-Computer Interaction, colors have an influence

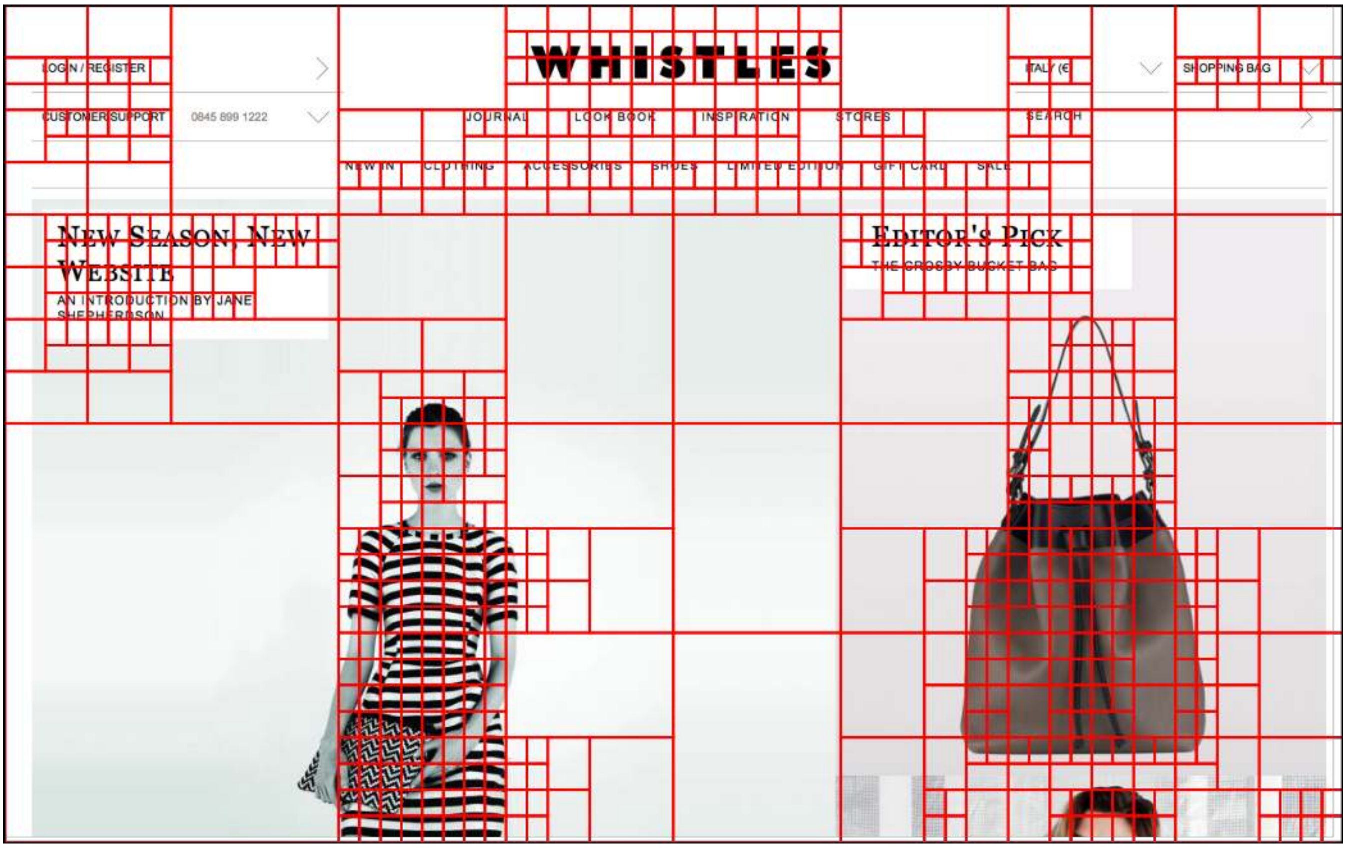

Figure 1. Visual representation of a Quadtratic ree Decomposition applied to a website of the AVI14 dataset. Visual complexity of an area is proportional to the number of leaves in that area.

on perceived trust, loyalty, and economic behavior (Cyr, 2008; Kim and Moon, 1998). The color scheme of a page can impact a user's feelings and reactions towards a page because specific colors have been demonstrated to increase —or reduce— the viewers' arousal and therefore induce excitation —or relaxation—. Cooler colors are often preferred to warmer colors because they elicit relaxed feelings (Cyr et al., 2010; Hall and Hanna, 2004; Hasler and Suesstrunk, 2003; Jacobs and Hustmyer Jr, 1974). A color is composed of a hue, a level of saturation, and a value (often defined as with brightness or luminance) (HSV model). For instance, Yendrikhovskij et al. (1998) found a correlation (r^2 = 0.91) between users' perceived colorfulness and the sum of the average saturation value of an image and its standard deviation. Additionally, a webpage's color distribution has also been proven to affect perceived brightness and perceived colorfulness (Reinecke et al., 2013). In this study, luminance is expressed as relative luminance, as per the photometric definition (Birtolo et al., 2009).
Another widely adopted feature is symmetry. First introduced by Gestalt's psychologists, symmetry has been proven to be one of the most important factors in aesthetic judgment. Symmetry indicates how well one side of an image reflects the opposite side, and it can be evaluated along a horizontal axis (top versus bottom), a vertical axis (left versus right), a radial plane around the center of the image, or using QuadTree decomposition (Miniukovich and De Angeli, 2014; Reinecke et al., 2013; Wang and Li, 2016; Zheng et al., 2009). An example of symmetry estimation using QuadTree decomposition is shown in Figure 2.

Features based not only on the appearance but also on the type of content have been proposed as well. Lin et al. (2013), for example, introduced the adoption of the ratio between graphics and text. In this work,

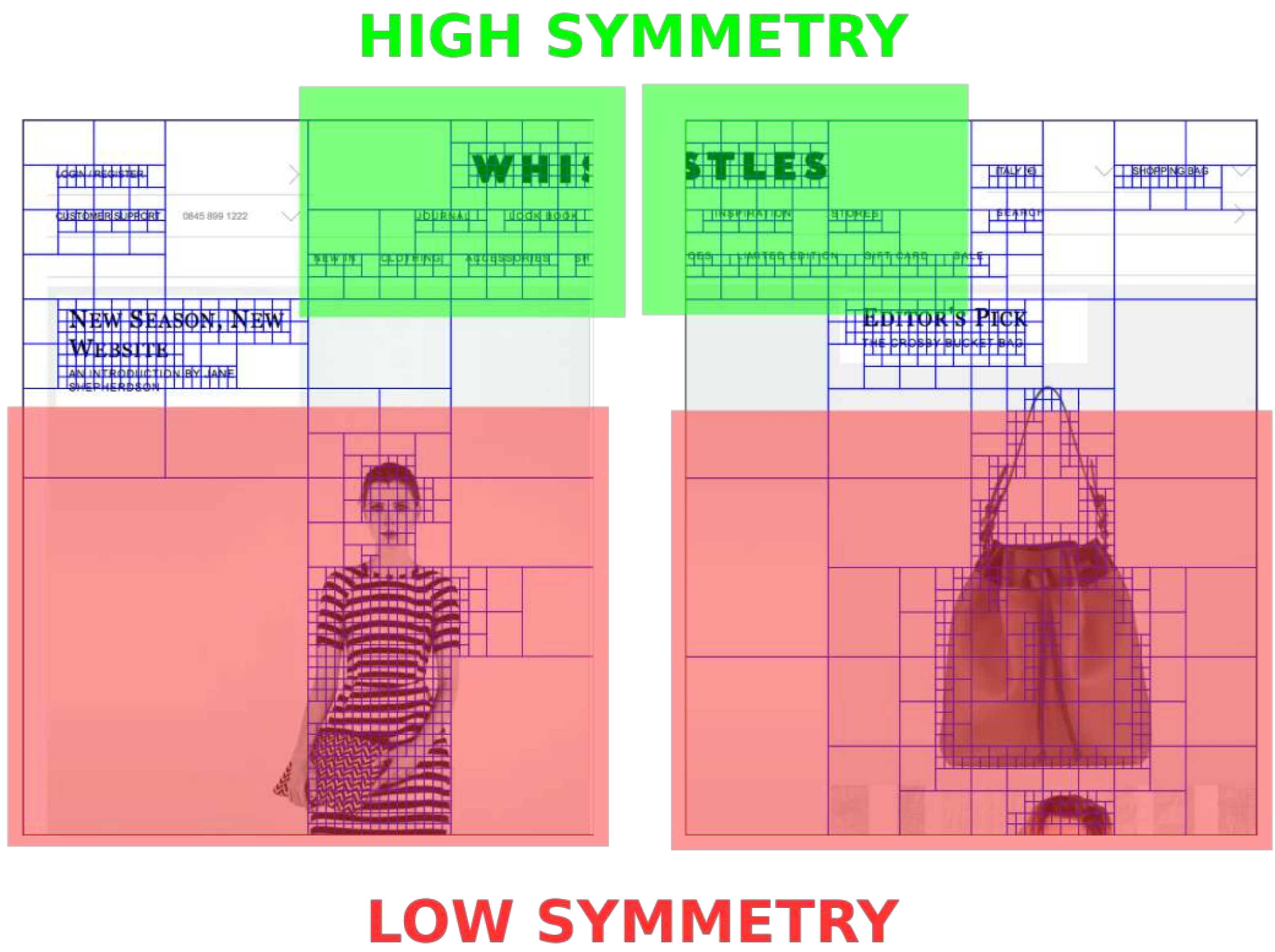

Figure 2. Example of symmetry estimation using QuadTree Decomposition (QTD) applied to a website of the AVI14 dataset. QTD is applied to the image, divided in two halves. The degree of symmetry is given by the number of overlapping rectangles between the two images. In the example above, a high degree of symmetry is present only in the upper part of the image.

we used a novel method for the automatic estimation of the graphics to text ratio, based on a combination of a Space-Based Decomposition (SBD) algorithm and an Optical Character recognition system (OCR). Similar to QTD, SBD uses a recursive division of the image to identify the contours of elements within an image, namely text and graphics. Once the elements have been identified, we can apply an OCR to label each element. Finally, the ratio between the area labeled as text and graphics can be automatically computed using a machine-driven approach. This same technique allows the automatic estimation of the number of images present on a page.
In the analysis of website visual features, it is important to rely on objective measures. Visual complexity, perceived colorfulness, graphics to text ratio, number of images, and symmetry are all automatic estimable features that can be computed in an objective algorithm.

1.2 Physiology and aesthetic judgment

Until recently, researchers were only able to investigate the underlying physiological correlates of aesthetic appreciation through behavioral measures of patients suffering from neurodegenerative diseases or whose brains suffered damage (Cela-Conde et al., 2011).

Today, researchers can investigate neurophysiological signals in a more ecological way from healthy participants, using sensors applied to the surface of the body. In an electromyography (EMG) study (which investigates muscles' electrical signals), Winkielman and Cacioppo (2001) demonstrated that physiological measures reflect participants' affective responses to stimuli and implicit judgments of their beauty. The activity of the zygomaticus major correlates with positive affective responses, and activity of a region in the *corrugator supercilii* correlates with negative affective responses (Lang et al., 1993; Winkielman and Cacioppo, 2001). Electrocardiography (ECG), which measures the electrical activity of the heart, also shows relations between physiological responses and aesthetic judgments (de Jong, 1972; de Jong et al., 1973; Ray et al., 1997).

In an eye-tracking study (Yanulevskaya et al., 2012), participants focused on emotionally positive parts of pictures. Maughan, Gutnikov, and Stevens Maughan et al. (2007) found that positive aesthetic judgments of advertisements elicited sustained attention. In addition, pupil dilation in response to pleasant images, and pupil constriction in response to unpleasant images were found by Blackbourne and Schirillo Blackburn and Schirillo (2012). Similarly, both ECG and EMG signals have been proven to be suitable for the empirical analysis of websites' aesthetic features, as shown by Tuch et al. (2009).

1.3 Behaviour and aesthetic judgments

The analysis of participants' behavioral data (explicit ratings) has been widely adopted in previous studies that investigated different aspects of websites, including their complexity and aesthetic qualities. Reinecke et al. (2013), for example, employed a 9-point Likert scale to assess participants' first impressions of a website's aesthetic quality, while Seckler et al. (2015b) investigated different aesthetic facets using a 7-point Likert scale. Those results show that by collecting self-reported measures using a Likert scale, we can obtain a reliable estimate of the perceived aesthetic judgments.

1.4 Expertise, exposure, and aesthetic judgment

The mere exposure effect states that repeated exposure to a target enhances an individual's attitude towards it (Zajonc, 1968; Bornstein and D'agostino, 1992). Cox and Cox (2002) found that repeated exposure to a visually complex product design increased preference for it as compared to a simpler but novel design. Exposure effects have also been found to evoke positive affective responses, where participants who rated familiar targets as more likable than unfamiliar ones also showed more zygomatic muscle region activity when viewing familiar targets (Harmon-Jones and Allen, 2001). These results suggest that individuals' exposure to different websites needs to be considered when evaluating their aesthetic judgment. Since many websites adopt similar designs and layouts, it is possible that not only the mere exposure to a single stimuli, but also a general exposure to many different websites can play a role in shaping users' design preferences. Similarly, expertise in a field affects preferences: experts and laypersons have different preferences and make different aesthetic judgments (Müller et al., 2010; Orr and Ohlsson, 2005; Ulrich Kirk, 2009; Pihko et al., 2011). Quispel et al. (2016) found that experts preferred familiar and novel chart designs, but laypersons preferred familiar and easy-to-use designs. In addition, familiarity and perceived ease of use predict the attractiveness of designs among laypersons but not experts. Bölte et al. (2017) evaluated experts' and laypersons' event-related potentials to web pages: Experts more frequently rated aesthetic web pages as less aesthetic than laypersons. This difference was not found in ratings of unaesthetic web pages. Given the history of findings on the role of expertise in evaluating aesthetics, it is also important to consider the impact of expertise in judging the aesthetic propeprties of a web page.

The research question on which this project is built is based on the possibility of reducing the cost, both in terms of time and economical expenses, of testing the perceived aesthetic experience of a web page. Results of this project may be used, in the future, to create novel technologies that will be able to support designers by providing them with continuous evaluations of design artifacts, at a reduced cost.

1.5 Purpose of the study and Hypothesis

Despite an array of studies on website aesthetics (Reinecke et al., 2013; Seckler et al., 2015a; Miniukovich and De Angeli, 2014; Bölte et al., 2017; Tuch et al., 2009), many have focused on limited individual factors such as visual complexity or colorfulness. Few attempts have been made to study the effects of multiple different visual features together on overall perceived aesthetic. Past studies have also failed to consider how user exposure and expertise might affect website aesthetic judgment. Furthermore, even though physiological measures have been employed previously, few have employed them to predict website aesthetic judgments from multiple visual properties.
In this work, we apply a novel approach based on neural networks and machine learning models as well as recursive partitioning and generalized linear models to estimate the perceived aesthetic appeal of a website. Our proposed flow chart is illustrated in Figure 3.
We hypothesized that (1) the interaction between web pages' different visual properties (visual complexity, colorfulness, brightness, symmetry, and text ratio) can be used to predict behavioral ratings and physiologically-estimated aesthetic judgments. We also hypothesized that (2) exposure to websites moderates website aesthetic judgments. Last, we hypothesized that (3) expertise on website design, similarly, moderates website aesthetic judgments.

2 METHODS

2.1 Analytic Plan

This work is structured as follows. First, an experimental procedure was conducted to collect behavioral ratings and physiological activity of participants in an image-rating task (5-point Likert scale), where participants rated screenshots of website and emotional images. Then a Multi-Layer Perceptron Neural Network (MLP NN) was trained on features estimated from the physiological activity of the participants, using the standardized valence values of the emotional images as training labels. The model was then applied to features estimated from the physiological activity recorded while participants were exposed to website images, resulting in the estimation of a valence value for the website images (implicit measures). Having both the behavioral ratings and the ratings estimated from the physiological activity, we proceeded with extracting a set of relevant features from the websites' images.
Since we know from previous studies that male and female participants may rate the aesthetic appeal of a website differently, we first excluded from the analysis all the websites that received significantly different ratings by male and female participants.
Finally, two different machine learning models —GLM and Decision Tree— were applied using websites' visual features as input and the ratings —behavioral or physiological—, as labels. To reduce the influence of a single participant on the overall accuracy of the model, bootstrapping is employed.
Performances of the models were tested not only against the 5-point Likert scale values, but also on a binomial rating, obtained by clustering ratings into two groups (High ratings: 4-5, Low ratings: 1-3). This was done to verify whether the models can be employed to obtain binary classifications (good/bad) of the aesthetic of a webpage.
Then, to compare the possible differences between experts and non-expert designers and between highly

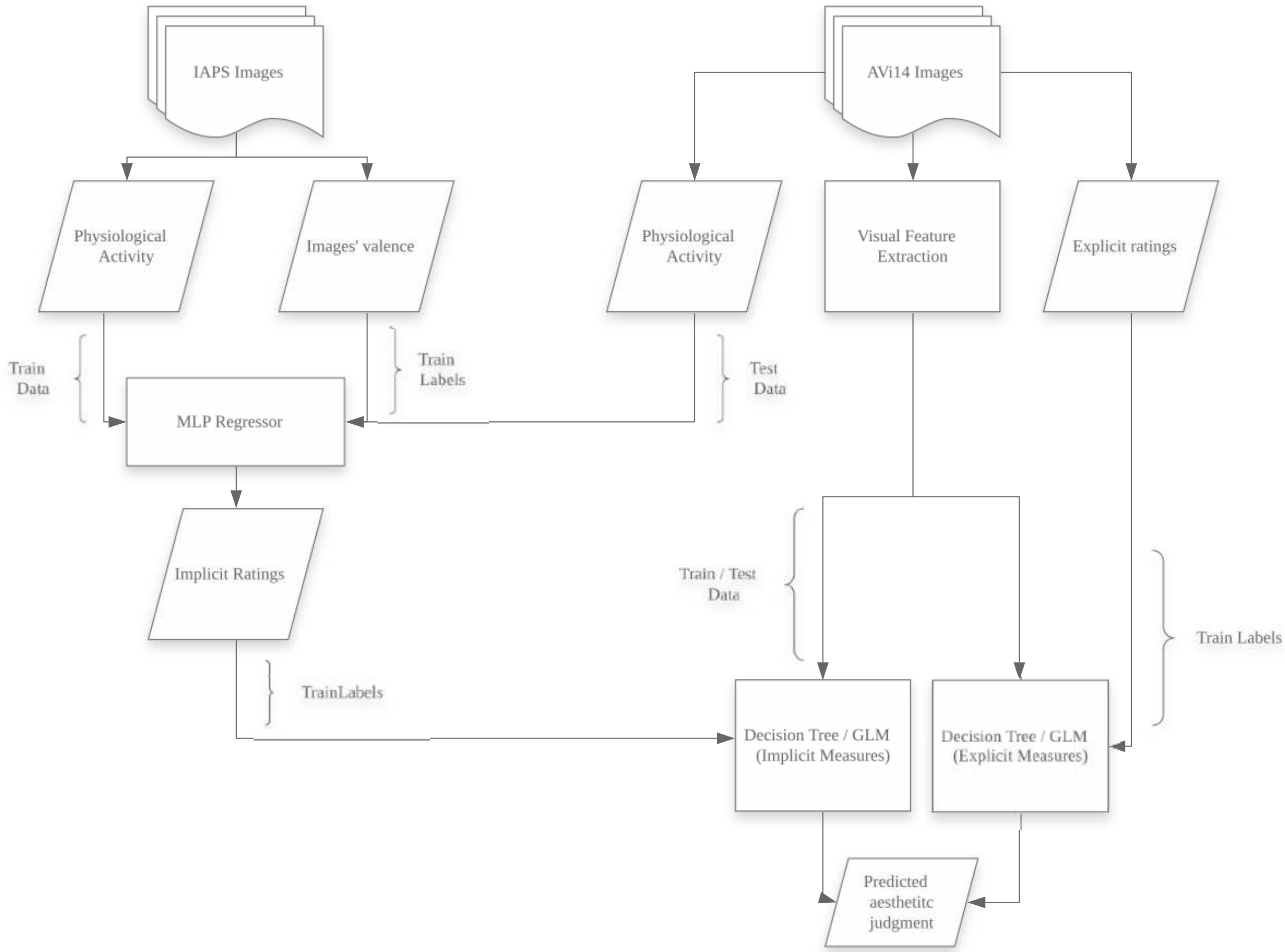

Figure 3. Flowchart of the method employed in this project.

exposed and lowly exposed users, we assigned our participants into two groups of about the same size and used these covariates as factors in our models.

2.2 Participants

59 university students (33 females, *Mean age* (in years) = 21.5 ± 3.0) voluntarily enrolled for participation. Informed consent was obtained from all the participants prior to the experimental session. The study was conducted in accordance with the declaration of Helsinki.

2.3 Stimuli

Stimuli were selected from two different datasets: the International Affective Pictures System (IAPS) (Lang and Bradley, 2007) and the AVI14 (Miniukovich and De Angeli, 2014).

2.3.1 International Affective Picture System

The International Affective Picture System (IAPS) (Lang and Bradley, 2007) is a dataset of emotionally evocative pictures, developed by the NIMH Center for the Study of Emotion and Attention (University of Florida). From the 1180 pictures included in the IAPS dataset, 50 were selected for presentation in the experimental procedure, 25 per block, balancing the mean valence value of each block[1]. The dataset is available upon request from their original authors[2].

2.3.2 AVI14

The AVI14 dataset is composed of images of 140 websites (Miniukovich and De Angeli, 2014). The dataset is available online[3]. All the websites are in English, and the original pages have no dynamic effects. Majority of the websites (*N* = 115) were selected from a public showcase of beautiful websites, and another 25 were selected to balance the overall aesthetic of the dataset. Used pages belong to four categories: a) coffee, b) chocolate bars and shops, c) online retailers, and d) design agencies. For the purposes of this work, 100 pictures were selected for presentation from the AVI14 dataset, according to their mean perceived aesthetic pleasure value (Miniukovich and De Angeli, 2014) and divided semi-randomly into two sets, one set per block.

2.4 Instrumentation

Stimuli were presented on a DELL 29" Ultrasharp Screen (U2719WM) with a fixed resolution of 1920x1080, (refresh rate = 60.00Hz). Pupil dilation signals were recorded using a Tobii X3-120 (sampling rate: 120Hz, Tobii Technology) mounted on a tripod and placed just below the screen. ECG, EDA and EMG signals were collected using a Bitalino Revolution BT board (sampling rate: 1000Hz, Wireless Biosignals S.A) (Guerreiro et al., 2013; Batista et al., 2017), using disposable 36-40mm snap connector foam electrodes (F9089/100, FIAB, Florence, Italy). The experimental paradigm and registration of physiological measurements were implemented in Python 2.7[4] (Oliphant, 2007; Van Rossum and Drake, 2011; Oliphant, 2006).

2.5 Experimental procedure

Participants sat approximately 50 to 70 cm away from a computer screen, in an silent and dark environment.
Before the experimental sessions, participants were instructed on the tasks they had to perform and on the physiological measures that were to be collected. The experiment consisted of two blocks, presented one after the other in a semi-randomized order, with a brief pause between blocks. Each picture was presented for 6 seconds, with an 8 second interval between consecutive images. A graphical representation of the used procedure is shown in Figure 4A.

To record participants' physiological activity, two disposable electrodes were used to record the electrodermal activity (EDA) from the left wrist, three were used to record the heart activity (ECG) —one below each clavicle and one below the last rib— and three were used to record the electromyographic activity (EMG) of the *corrugator supercilii* —one above the nose, one above the left eye and one on the left cheek—. A graphical representation of the electrodes position is reported in Figure 4B
Immediately after each picture, participants rated their aesthetic appeal on a 5-point Likert scale, by clicking on one of five buttons presented on the screen, with no time constraint.

[1] Pictures belonging to the following categories were removed prior to stimuli selection: "BurnVictim", "Mutilation", "DeadBody", "DeadMan", "headleessBody", "BabyTumour", "Tumor", "Accident", "SlicedHand", "Vomit", "BatteredFem". Remaining pictures were sorted by mean valence and the first and last 25 were semi-randomly selected and distributed in two sets, one per experimental block

[2] https://csea.phhp.ufl.edu/Media.html

[3] https://github.com/aliko-str/avi14dataset

[4] v. 2.7.12

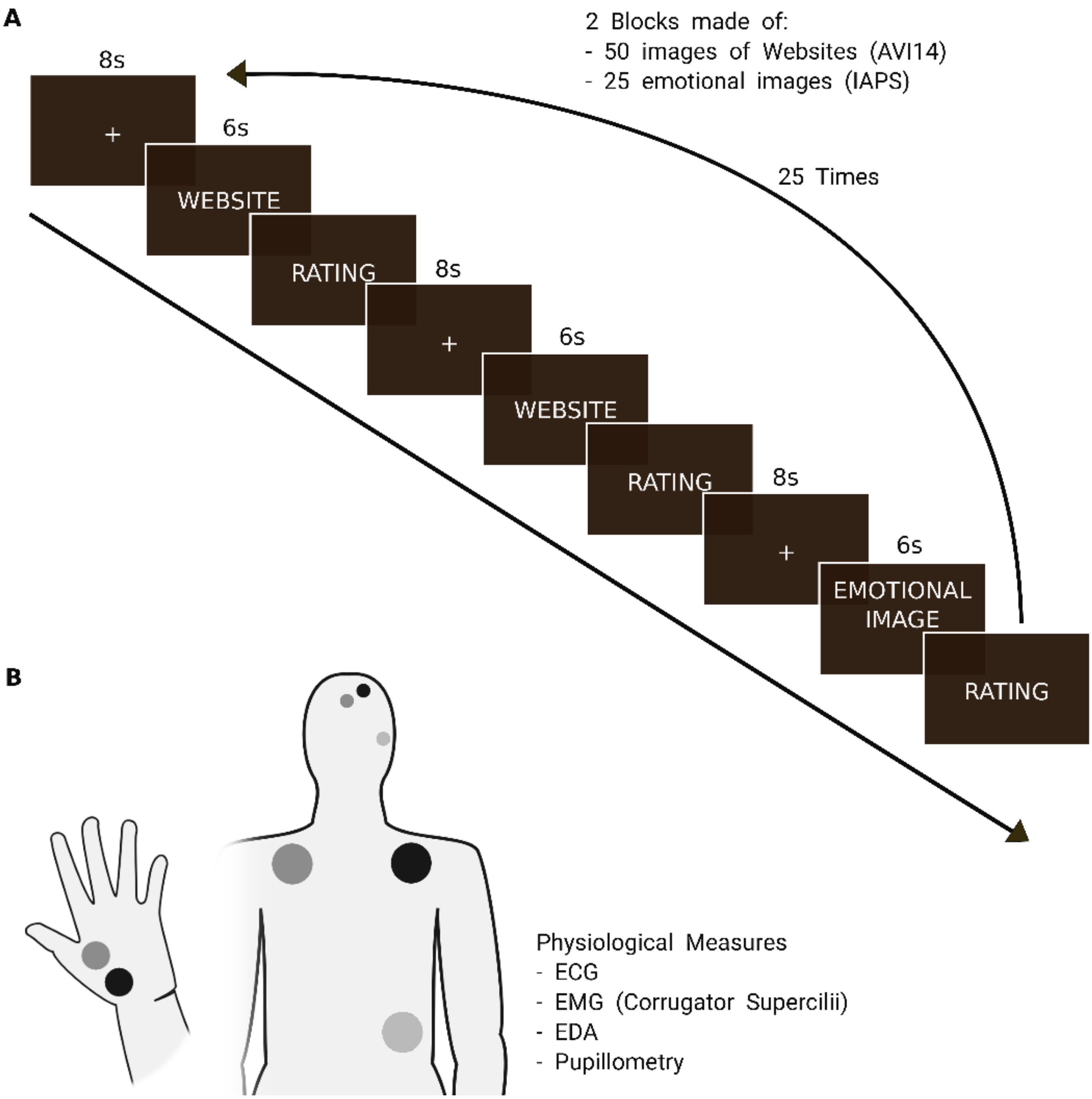

Figure 4. Graphical representation of the (**A**) experimental paradigm and (**B**) electrodes position for physiological signals recording

At the end of the experimental procedure, participants completed a 7-item survey on their browsing habits (exposure) and expertise in design, development, and management of websites. Finally, participants were debriefed.

2.6 Data modeling and analysis

2.6.1 Website Feature Extraction

Website features were extracted using a self-developed tool released under the name "*PrettyWebsite*" (Gabrieli, 2019a). The package is available through the Python Package manager (Pypi) and the project repository [5].

2.6.2 Physiological Feature Extraction

Physiological features were extracted from collected signals using *"Pysiology* (Gabrieli et al., 2020; Gabrieli, 2019b), a Python package designed for physiological signal processing.
For each stimulus, physiological measures were computed in epochs of 8 seconds. For ECG signals, the 20 seconds of recording before the first stimulus of each block was used as a baseline. For pupil diameter, signals 6 seconds preceding each stimulus served as a baseline. Detailed information about the parameters used to clean the signals and estimate features are reported in Supplementary Materials.

2.6.3 Estimation of implicit ratings

Participants' physiological activity was used to estimate the perceived valence of websites' images.
Features extracted from the epochs in which participants were engaged in viewing images from the IAPS dataset were used as training data of an MLP Regressor Solver = *"sgd"*, $\alpha = 0.0001$, number of hidden layers = 100), with the standardized valence values of the images, provided within the IAPS dataset, used as training labels. To reduce the number of input features, the best six physiological components were identified through Principal Component Analysis, standard scaled and fed to the model.
Once trained, the average accuracy of the model was tested, using bootstrapping (N=100) against real IAPS' valence value.
Finally, the model was fed with features extracted from the portions of signals where participants were rating website pictures, in order to obtain an estimated implicit valence value (implicit rating) for each website.

2.6.4 Preliminary analysis

2.6.4.1 Expertise and exposure

To compare differences between high and low expertise, and between high- and low-exposure users, each participant was assigned to one of the two groups for each classification. Assignation was done by defining a threshold that allowed the authors to obtain groups of similar sizes.

2.6.4.2 Gender differences

Previous studies highlighted the fact that males and females rate some websites with significantly different scores. To omit gender of participants from the model, we conducted a preliminary analysis to identify if, within our dataset, some of the websites received significantly different behavioral ratings by males and females and subsequently removed those websites from our analysis.

2.7 Predicting perceived visual aesthetic

Prediction of the perceived visual aesthetic from estimated website features has been performed using two different machine learning models, a generalized linear model, as implemented in *statsmodel* (Seabold

[5] **Pypi**: https://pypi.org/project/prettywebsite/
Github: https://github.com/Gabrock94/PrettyWebsite

and Perktold, 2010) (GLM) and a recursive partitioning (Decision Tree, min samples per leaf=100, max depth=5, max features=5), as implemented in Scikit-learn (Pedregosa et al., 2011).
In total, 24 different machine learning models were trained and tested using bootstrapping (N = 1000). For each rating and physiological measure of aesthetic judgments, the two classifiers were used (GLM and Decision Tree). Each classifier was tested 6 times, 3 times predicting values on a 5-point scale, and 3 times predicting values on a binomial scale (1-3, 4-5). Finally, each of these 3 models was tested three times: one with no reference to participants' expertise or exposure, one with expertise (high/low) as a factor of the model and one with exposure (high/low) as a factor of the model.

3 RESULTS

Out of the data recorded from 59 participants, data of one participant (N = 1) was removed because of technical issues in collected physiological samples. Therefore all the data described below are based on 58 participants (N = 58, F = 33, M = 25, *Mean age*: 21.4 ± 2.2).

3.1 Expertise and Exposure

With regard to expertise, ten participants (N=10) reported to have developed, and twelve (N = 12) to have managed at least one website. Of the above, five (N = 5) reported having both developed and owned at least one website. More than half of the participants (N = 35) reported having at least basic knowledge of one or more programming languages. Participants who had at least basic knowledge of two or more programming languages and have developed or managed a website were assigned to the ”expert group” (N = 29). Three websites (AVI 78, AVI 42 and AVI 128) received significantly different ratings by experts and non-experts (Table 1). Thunbnails of those websites are reported in Supplementary Material (Figure S1). Results from the t-tests showed that the power of the test was medium and only ratings given to AVI 128 were statistically significant using the KS-test. Of the three pages, the first two were rated on average higher by expert users while the last was rated higher by the non-expert users.

Table 1. p-value (and power) of Student's t, F, Kolmogorov-Smirnov tests and means of the ratings for the Image that have been rated significantly different by the expert and non expert groups.

Image	p-val. (t)	Power (t)	p-val. (F)	p-val. (K-S)	Avg. E.	Avg. Non-E.
78.png	0.0118	0.682	0.281	0.0706	3.8	3.4
42.png	0.0366	0.536	0.713	0.1951	2.6	2.3
128.png	0.0136	0.689	0.489	0.074	2.6	3.0

With respect to exposure, 4 participants browsed the web using only either a laptop or desktop, and 11 browsed the web using only mobile devices. Almost half of the participants reported browsing up to 5 different websites per day, and 30 reported browsing more than 5 websites per day. More than half of the participants spent less than 3 hours browsing websites (N = 39). Half of the participants (N = 28) reported spending the majority of their time on a single website, such as Facebook or Twitter. Participants who reported browsing 10 or more websites per day and who indicated browsing the web for more than 2 hours per day were assigned to the ”high exposure” group (N = 27). Two websites received significantly different ratings by the ”high exposure” and ”low exposure” groups (Table 2). A previes of those websites is reported in Supplementary Material (Figure S2).

Table 2. p-value (and power) of Student's t, F, Kolmogorov-Smirnov's tests and means of the ratings for the Image that have been rated significantly different by the high exposure and low exposure groups.

Image	**p-val. (t)**	**Power (t)**	p-val. (F)	**p-val. (K-S)**	**Avg. E.**	**Avg. Non-E.**
98.png	0.0439	0.621	0.239	0.338	2.7	3.1
20.png	0.0077	0.826	0.194	0.0453	3.7	3.3

3.2 Gender

Five websites received significantly different ratings from male and female participants. Results are reported in Table 3, while thumbnails of the images are reported in Supplementary Material (Figure S3) . These websites were therefore removed from subsequent analysis.

Table 3. Results of t-test, Kolmogorov-Smirnov test and Fisher's test of websites with statistically significant differences between males and females.

Image	**p-value (t-test)**	**Power (t-test)**	**p-value (F-test)**	**p-value (KS-test)**
36.png	0.0017	0.764	0.339	0.041
101.png	0.0074	0.618	0.198	0.087
132.png	0.0236	0.506	0.423	0.224
66.png	0.0133	0.576	0.381	0.256
76.png	0.004	0.695	0.261	0.194

3.3 Website features

A visual feature — e.g. Visual complexity — can be estimated using different methods. In this work, where more than one algorithm was available, we adopted the most prominent. Therefore, the index of visual complexity used was based on the QDT (as opposed to the images' weight, $R^2 = 0.5$ between the two indexes), brightness was estimated from the *BT.709*[6] index (as opposed to the *BT.601*[7], ($R^2 = 1.0$)), and colorfulness was extracted from the HSV colorscheme (as opposed to the RGB colorscheme, ($R^2 = 0.58$)), as done by Yendrikhovskij et al. (1998).
For our predicted models, we used *Symmetry, Colorfulness* (HSV), *Visual Complexity* (Quadratic Tree decomposition), *brightness* (BT709) and *number of Images* - automatically evaluated applying the Space-based decomposition and OCR- as independent variables.

3.4 Estimation of image valence from viewers' physiological activity

To obtain an estimation of an image's valence from a viewer's physiological activity, we used an MLP Regressor. First, extracted physiological features were used to estimate the valence of IAPS images. Average accuracy of MLP Regressor, tested against real IAPS' valence value, is 97.9% ($\sigma = 0.004$). Implicit ratings of website stimuli were estimated for 2792 epochs from 44 different participants (Mean number of stimuli per participant = 63.5 ± 14.5).

[6] ITU-R Recommendation BT.709

[7] ITU-R Recommendation BT.601

3.4.1 Predicting perceived visual aesthetic ratings

Average predictive accuracy of GLMs and Decision Trees are reported in Table 4. For the 5-point scale, no differences in the average prediction accuracy were reported between the explicit ratings and implicit appraisals when using GLM. However, for the same scale, when using recursive partitioning, tree-based models showed that implicit appraisals predicted better (73%) as compared to explicit ratings (60%). On the other hand, for the binary rating estimation, using GLM, the prediction of explicit ratings outperformed that of implicit appraisals by almost 10 percentage points. When using a decision tree, no differences were found in the performance when applied to the two different types of ratings.

Table 4. Comparison between average accuracy of implicit appraisals and explicit ratings from website features by model, presence of expertise/exposure factors and type of prediction (1-5 points or binary).

Prediction	Model	Expertise / Exposure	Average accuracy Explicit	Average accuracy Implicit
5-Points	GLM	None	67.7%	67.7%
		Exposure	62.1%	64.0%
		Expertise	62.8%	69.2%
	Decision Tree	None	60.1%	72.9%
		Exposure	60.1%	68.2%
		Expertise	61.3%	66.0%
Binary	GLM	None	97.1%	89.9%
		Exposure	97.7%	86.6%
		Expertise	96.9%	87.7%
	Decision Tree	None	87.6%	87.1%
		Exposure	88.5%	86.5%
		Expertise	88.0%	85.0%

4 DISCUSSION

4.1 Prediction of perceived visual aesthetic

Our results showed that by using automatic estimable features from still images of web-pages, regressive models can be used to predict with reasonably high accuracy if a page will be explicitly and/or implicitly perceived as aesthetically pleasant, thereby supporting our first hypothesis.

Our finding provides further support to the existing literature whereby visual properties have been found to play a role in aesthetic judgment. More importantly, this finding provides insight into the predictive capabilities of these visual properties on both explicit and implicit aesthetic judgments and how they can be utilized effectively depending on the type of scale the researcher prefers.

More specifically, depending on the type of desired data, different models can be selected to predict the different ratings. When a 5-point scale is preferred, either GLM or Decision Tree model can be used to both predict explicit ratings or implicit appraisals. However, when a binary scale is preferred, using GLM will provide a higher prediction accuracy for explicit ratings than for implicit ratings while using Decision Tree will provide similar accuracy for both implicit appraisals and explicit ratings.

4.2 Does the level of exposure to different websites influences perceived visual appeal?

With regards to our second hypothesis, we predicted that the level of exposure to different web pages moderates users' aesthetic judgments.
Despite the fact that two websites received significantly different ratings by participants of the high and low-exposure groups, the addition of the level of expertise as a factor of our regressive model resulted in no significant increase in their accuracy. We can, therefore, conclude that our second hypothesis, within given limits of the number of websites and participants, is not confirmed.

4.3 Does the level of expertise play a role in perceived visual appeal?

For our third hypothesis, we predicted that participants' expertise in the design and development of web pages affects their aesthetic judgment. Similar to the comparison between highly-exposed and low-exposed participants, three websites received significantly different ratings by participants of the two groups, but the addition of the users' expertise as a factor of the models led to no significant improvement of their accuracy, hence not supporting our third hypothesis.

4.4 Limitations

As is common to all experimental studies, limitations are inevitable and should be mentioned. With respect to the physiological measurements utilized to assess participants' implicit appraisals, we are unable to control for the participants' physiological state at the beginning of each session. Despite the fact that a baseline correction is applied during feature extraction, possible differences in pre-experimental physiological arousal and valence may still be present and should be taken into consideration. Next, it should also be noted that participants' explicit ratings and expertise/exposure responses are all self-reported measurements and the social desirability factor could affect the reliability of the reporting. Participants may feel the social pressure in not stating the truth to questions about the amount of time they spend on the Internet and about the average number of pages browsed per day. Another important consideration is that the number of existing websites are almost impossible to be determined and, as such, the usage of a limited number of websites may not be suitable if used pages are not representative of the whole dataset. Thus, it is ideal for further studies to be conducted to determine how representative the used pages are of the entire dataset.
Across different age groups and different cultures, the daily usage of websites can vary greatly and as such, future studies should also take these factors into consideration and select appropriate thresholds for their sample. More specifically, future studies can consider including participants with a broader range in design and development knowledge, time spent browsing pages and number of different pages browsed per day. Different indicators of expertise and exposure can then be considered. In addition, our sample is not a perfect representation of the actual age range of Internet users. Therefore, future studies should also involve younger and older participants in order to test the reliability of our models on a more varied sample.

5 CONCLUSION

In this work, we investigated the possibility of predicting both implicit and explicit user aesthetic judgment of websites from visual properties while considering expertise and exposure as possible predictive factors. Results showed that by investigating the visual properties of web pages, it is possible to predict, with a good degree of accuracy, if a website will be perceived -explicitly or implicitly- as aesthetically pleasing by possible users. Although differences in ratings given by experts and non-experts as well as high-exposure

and low-exposure users have been found, the accuracy of predictive models was not enhanced by the addition of expertise and exposure as factors.
Findings from this study will help designers uncover the most critical aspects that they should consider in sketching the layout of digital interfaces.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

G.G. and G.E. conceived and planned the experiments; G.G. conducted the study and analyzed the data; G.G., M.H.B. and G.E. discussed and interpreted the results; G.G. wrote the original draft; all the authors reviewed and edited the submitted version.

FUNDING

This study was supported by the NAP-SUG program of the Nanyang Technological University and by the Singapore Ministry of Education Academic Research Grants - Tier1.

ACKNOWLEDGMENTS

Thank to Chiara Iannaccone, Giulia Garbin, and Mengyu Lim for their help.

DATA AVAILABILITY STATEMENT

The raw and processed data, a copy of the python packages and of the scripts used in this work can be found in the data repository of the Nanyang Technological University `https://doi.org/10.21979/N9/YCDXNE`.

REFERENCES

Batista, D., Silva, H., and Fred, A. (2017). Experimental characterization and analysis of the bitalino platforms against a reference device. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE), 2418–2421

Birtolo, C., Pagano, P., and Troiano, L. (2009). Evolving colors in user interfaces by interactive genetic algorithm. In *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on* (IEEE), 349–355

Blackburn, K. and Schirillo, J. (2012). Emotive hemispheric differences measured in real-life portraits using pupil diameter and subjective aesthetic preferences. *Experimental Brain Research*

Bölte, J., Hösker, T. M., Hirschfeld, G., and Thielsch, M. T. (2017). Electrophysiological correlates of aesthetic processing of webpages: a comparison of experts and laypersons. *PeerJ* 5, e3440

Bornstein, R. F. and D'agostino, P. R. (1992). Stimulus recognition and the mere exposure effect. *Journal of personality and social psychology* 63, 545

Bucy, E. P., Lang, A., Potter, R. F., and Grabe, M. E. (1999). Formal features of cyberspace: Relationships between Web page complexity and site traffic. *Journal of the American Society for Information Science* 50, 1246–1256

Cela-Conde, C. J., Agnati, L., Huston, J. P., Mora, F., and Nadal, M. (2011). The neural foundations of aesthetic appreciation. *Prog. Neurobiol.* 94, 39–48

Cox, D. and Cox, A. D. (2002). Beyond first impressions: Th effects of repeated exposure on consumer liking of visually complex and simple product designs. *Journal of the Academy of Marketing Sciences*

Cyr, D. (2008). Modeling web site design across cultures: relationships to trust, satisfaction, and e-loyalty. *Journal of Management Information Systems* 24, 47–72

Cyr, D., Head, M., and Larios, H. (2010). Colour appeal in website design within and across cultures: A multi-method evaluation. *Int. J. Hum. Comput. Stud.* 68, 1–21

de Jong, M. A. (1972). A physiological approach to aesthetic preference. I. Paintings. *Psychother. Psychosom.* 20, 360–365

de Jong, M. A., van Mourik, K. R., and Schellekens, H. M. C. (1973). A Physiological Approach to Aesthetic Preference. *Psychother. Psychosom.* 22, 46–51

[Dataset] Gabrieli, G. (2019a). Gabrock94/prettywebsite: Version 0.0.3 - build 1. doi:10.5281/zenodo.3187316

[Dataset] Gabrieli, G. (2019b). Gabrock94/pysiology: Version 0.0.9 - build 3. doi:10.5281/zenodo.2622204

Gabrieli, G., Azhari, A., and Esposito, G. (2020). Pysiology: A python package for physiological feature extraction. In *Neural Approaches to Dynamics of Signal Exchanges* (Springer). 395–402

Guerreiro, J., Martins, R., Silva, H., Lourenço, A., and Fred, A. L. (2013). Bitalino-a multimodal platform for physiological computing. In *ICINCO (1)*. 500–506

Hall, R. H. and Hanna, P. (2004). The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behav. Inf. Technol.* 23, 183–195

Harmon-Jones, E. and Allen, J. J. B. (2001). The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. *Personality and social pyschology bulletin*

Hasler, D. and Suesstrunk, S. E. (2003). Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII* (International Society for Optics and Photonics), vol. 5007, 87–96

Ivory, M. Y., Sinha, R. R., and Hearst, M. A. (2001). Empirically validated web page design metrics. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (ACM), 53–60

Jacobs, K. W. and Hustmyer Jr, F. E. (1974). *Effects of four psychological primary colors on GSR, heart rate and respiration rate*, vol. 38 (SAGE Publications Sage CA: Los Angeles, CA)

Karvonen, K. (2000). The beauty of simplicity. In *Proceedings on the 2000 conference on Universal Usability* (ACM), 85–90

Kim, J., Lee, J., and Choi, D. (2003). Designing emotionally evocative homepages: an empirical study of the quantitative relations between design factors and emotional dimensions. *Int. J. Hum. Comput. Stud.* 59, 899–940

Kim, J. and Moon, J. Y. (1998). Designing towards emotional usability in customer interfaces—trustworthiness of cyber-banking system interfaces. *Interact. Comput.* 10, 1–29

Lang, P. and Bradley, M. M. (2007). The international affective picture system (iaps) in the study of emotion and attention. *Handbook of emotion elicitation and assessment* 29

Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 261–273

Lin, Y.-C., Yeh, C.-H., and Wei, C.-C. (2013). How will the use of graphics affect visual aesthetics? a user-centered approach for web page design. *International Journal of Human-Computer Studies* 71, 217–227

Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behav. Inf. Technol.* 25, 115–126

Maughan, L., Gutnikov, S., and Stevens, R. (2007). Like more, look more. look more, like more: The evidence from eye-tracking. *Journal of Brand management* 14, 335–342

Miniukovich, A. and De Angeli, A. (2014). Quantification of interface visual complexity. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces - AVI '14* (New York, New York, USA: ACM Press), 153–160. doi:10.1145/2598153.2598173

Moshagen, M. and Thielsch, M. T. (2010). Facets of visual aesthetics. *International journal of human-computer studies* 68, 689–709

Müller, M., Höfel, L., Brattico, E., and Jacobsen, T. (2010). Aesthetic judgments of music in experts and laypersons—an erp study. *International Journal of Psychophysiology* 76, 40–51

Oliphant, T. E. (2006). *A guide to NumPy*, vol. 1 (Trelgol Publishing USA)

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering* 9, 10–20

Orr, M. G. and Ohlsson, S. (2005). Relationship between complexity and liking as a function of expertise. *Music Perception: An Interdisciplinary Journal*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830

Pihko, E., Virtanen, A., Saarinen, V.-M., Pannasch, S., Hirvenkari, L., Tossavainen, T., et al. (2011). Experiencing art: the influence of expertise and painting abstraction level. *Frontiers in human neuroscience* 5, 94

Quispel, A., Maes, A., and Schilperoord, J. (2016). Graph and chart aesthetics for experts and laymen in design: The role of familiarity and perceived ease of use. *Information Visualization* 15, 238–252

Ray, G., Kaplan, A. Y., and Jovanov, E. (1997). Morphological variations in ecg during music-induced change in consciousness. In *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE* (IEEE), vol. 1, 227–230

Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., et al. (2013). Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM), 2049–2058

Robbins, J. N. (2012). *Learning Web Design: A Beginner's Guide to {HTML}, {CSS}, {JavaScript}, and Web Graphics* ("O'Reilly Media, Inc.")

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* (Scipy), vol. 57, 61

Seckler, M., Opwis, K., and Tuch, A. N. (2015a). Linking objective design factors with subjective aesthetics: An experimental study on how structure and color of websites affect the facets of users' visual aesthetic perception. *Computers in Human Behavior* 49, 375–389. doi:10.1016/j.chb.2015.02.056

Seckler, M., Opwis, K., and Tuch, A. N. (2015b). Linking objective design factors with subjective aesthetics: An experimental study on how structure and color of websites affect the facets of users' visual aesthetic perception. *Computers in Human Behavior* 49, 375–389

Tuch, A. N., Bargas-Avila, J. A., Opwis, K., and Wilhelm, F. H. (2009). Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *Int. J. Hum. Comput. Stud.* 67, 703–715

Tuch, A. N., Presslaber, E. E., Stöcklin, M., Opwis, K., and Bargas-Avila, J. A. (2012). The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human Computer Studies* 70, 794–811. doi:10.1016/j.ijhcs.2012.06.003

Ulrich Kirk, M. S. C. N. N., Martin Skov (2009). Brain correlates of aesthetic expertise: A parametric fmri study. *Brain and Cognition*

Van Rossum, G. and Drake, F. L. (2011). *The python language reference manual* (Network Theory Ltd.)

Wang, M. and Li, X. (2016). Effects of the aesthetic design of icons on app downloads: evidence from an android market. *Electr. Commerce Res.* 17, 83–102

Wang, S. and Ding, R. (2012). A qualitative and quantitative study of color emotion using valence-arousal. *Frontiers of Computer Science*

Winkielman, P. and Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *J. Pers. Soc. Psychol.* 81, 989–1000

Yanulevskaya, V., Uijlings, J., Bruni, E., Sartori, A., Zamboni, E., Bacci, F., et al. (2012). In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings. In *Proceedings of the 20th ACM international conference on Multimedia* (ACM), 349–358

Yendrikhovskij, S., Blommaert, F. J., and de Ridder, H. (1998). Optimizing color reproduction of natural images. *Color and Imaging Conference* 1998, 140–145

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*

Zhang, P. and Von Dran, G. M. (2000). Satisfiers and dissatisfiers: a two-factor model for website design and evaluation. *Journal of the American Society for Information Science and Technology* 51, 1253–1268. doi:10.1002/1097-4571(2000)9999:9999⟨::AID-ASI1039⟩3.0.CO;2-O

Zhang, P., Von Dran, G. M., Blake, P., and Pipithsuksunt, V. (2001). Important Design Features in Different Web Site Domains: An Empirical Study of User Perceptions. *e-Service Journal* 1, 77–91

Zheng, X. S., Chakraborty, I., Lin, J. J.-W., and Rauschenberger, R. (2009). Correlating low-level image statistics with users - rapid aesthetic and affective judgments of web pages. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09* (New York, New York, USA: ACM Press), 1. doi:10.1145/1518701.1518703