IADIS International Journal on WWW/Internet
Vol. 23, No. 2, pp. 50-65
ISSN: 1645-7641

PREDICTING VISUAL AESTHETICS DISTRIBUTIONS OF MOBILE GUIS USING DEEP LEARNING

Adriano Luiz de Souza Lima[1] and Christiane Gresse von Wangenheim[2]
[1]Instituto Federal de Santa Catarina, São José - SC, Brazil
[2]Universidade Federal de Santa Catarina, Florianópolis - SC, Brazil

ABSTRACT

Visual aesthetics is a fundamental yet complex attribute of graphical user interfaces (GUIs), often represented by a single score. However, such simplification may overlook the variety of responses that GUIs elicit from different users, even when they share similar backgrounds. To address this limitation, we propose representing visual aesthetics as a distribution of ratings obtained from individual raters. From this distribution, the median can be interpreted as the degree of visual aesthetics, while the average absolute deviation reflects the level of disagreement among raters. Additionally, the histogram provides a more comprehensive visual analysis. In this work, we present a supervised deep learning approach to predict these distributions and compare different model architectures when assessing GUIs created with App Inventor. Our best results were achieved with a ResNet50 model, which obtained a mean squared error (MSE) of .021. While the correlation between individual score predictions and labels was moderate, the correlation between medians was strong (ρ = .85). Furthermore, the predicted medians showed excellent agreement with a model trained to predict single-valued aesthetics (ρ = .97), and Bland–Altman analysis revealed 95% agreement with ground-truth labels. These findings demonstrate that it is feasible to automate the prediction of visual aesthetics in mobile GUIs not only as single values but also as full distributions of human ratings, providing richer and more reliable insights into user perception.

KEYWORDS

Visual Aesthetics, Mobile Application, Score Distribution, Deep Learning, Automatic Assessment

1. INTRODUCTION

The importance of visual aesthetics for graphical user interfaces (GUIs) and its beneficial effects on users have gained increasing attention (Thielsch et al., 2019). This recognition has made visual aesthetics a key construct in human-computer interactions. It has been shown to improve perceived usability (Tractinsky et al., 2006; Tuch et al., 2012) and credibility (Oyibo et al., 2018;

Robins & Holmes, 2008) of GUIs, and acts as a differentiating factor from other systems with similar features (Bhandari et al., 2016). Furthermore, because beautiful GUIs can attract and hold users' attention (Takimoto et al., 2021), visual aesthetics should also be a priority in mobile interface designs (Guo et al., 2020). Given its importance, the visual aesthetics of mobile GUIs must be adequately assessed (Moshagen & Thielsch, 2010). However, this is a challenging task, as it is highly subjective, and agreement regarding the aesthetic quality of an image is not always easily achievable among a group of people, even when they share the same background (Gresse von Wangenheim et al., 2018).

Several approaches for assessing the visual aesthetics of GUIs have been presented (Lima & Gresse von Wangenheim, 2021). Owing to its highly subjective component, typical visual aesthetics assessments rely on target users' responses to understand their reactions upon observing the assessed GUIs. However, this approach of manually assessing individual GUIs may not be accessible to small companies or self-employed professionals because of the considerable resources demanded (Miniukovich & De Angeli, 2015). Additionally, having a reasonably large group of people for every assessment makes it time-consuming and prone to errors (Soui et al., 2020).

To overcome these challenges, alternatives aim to establish a relationship between the objective features of GUIs and subjective aesthetic judgments, allowing the automation of the assessments of visual aesthetics to reduce effort and minimize errors (Lima & Gresse von Wangenheim, 2021; Miniukovich & De Angeli, 2014a). Here, objective approaches refer to using measurable and quantifiable properties of GUIs, such as simply counting and measuring visual elements or extracting handcrafted features like symmetry (balance and evenness of elements), complexity (the amount of visual detail or variety), or colorfulness (the richness of color usage) (Altaboli & Lin, 2011). Nonetheless, the definition and design of these features require a broad knowledge of aesthetics and visual design (Bao et al., 2019). Additionally, while features like symmetry and color are clearly related to aesthetics, other visual properties might not be explicit or recognizable (Takimoto et al., 2021).

Recently, deep learning techniques have achieved superior results when applied to assess photographs (Deng et al., 2017; Lu et al., 2014; Malu et al., 2017) and GUIs (Dou et al., 2019; Khani et al., 2016; Lima, Martins, et al., 2022; Xing et al., 2021). These results indicate that it is possible to automatically extract visual features related to visual aesthetics from images without manual feature engineering (Bao et al., 2019).

Assessment techniques express visual aesthetics in the form of a categorical tag or a single numerical score representing the average aesthetic perception of an entire group of people (Dou et al., 2019; Kang et al., 2019). It is a direct indicator of the visual aesthetics of GUIs, facilitating the comprehension of which ones are beautiful and which are not (Bao et al., 2019). However, they may not represent the subjectivity of visual aesthetics because they do not truly represent the individual perceptions from which the assessment results derive (Jin et al., 2016; Murray & Gordo, 2017). For example, equivalent visual aesthetics scores might hide a high degree of consensus among human raters or disagreement about the beauty of a GUI (Figure 1) (Kang et al., 2019). Therefore, it is inaccurate to assume they are of equal quality (Takimoto et al., 2021). When a group of people assesses a GUI as attractive, they seem to reach an intersubjective agreement about its aesthetics (Zen & Vanderdonckt, 2016). However, a score representing the mean or median value when all ratings differ cannot reflect agreement.

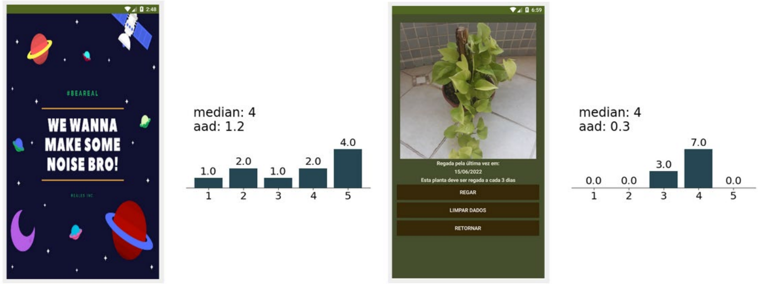

Figure 1. Mobile GUIs with different score distributions and the same median

Before computing the final score that reflects the visual aesthetics of a GUI, these methods contain the distribution of the ratings received (Murray & Gordo, 2017). The mean and median scores, variance, and average deviation can be calculated based on this information (Bao et al., 2019). These scores, together with rating histograms, can help better understand the visual aesthetics of GUIs and how much the human raters agree about them, indicating the difficulty of the assessment (Jin et al., 2016). In addition, the variance or average deviation is a way of characterizing intersubjective agreement, which can be visually represented by a histogram. In conjunction with the visual aesthetics score, this representation may better indicate the aesthetic quality of the GUIs (Bao et al., 2019).

In this article, we propose a deep learning model, the Appsthetics Histogram, to predict the score distribution of a mobile GUI developed with App Inventor, as a step further from previous models for the visual aesthetics assessment of GUIs in a single score. From the score distribution, it is possible to compute the visual aesthetics degrees of mobile GUIs and the degree of disagreement between human raters. Enabling the automated visual aesthetics assessment of mobile GUIs can contribute to improving the quality of mobile apps, tackling some usability issues early, and reducing development costs. With this research, we also expect to contribute to the understanding of how people agree about the visual aesthetics of mobile GUIs.

2. RELATED WORK

Previous research on applying deep learning techniques for visual aesthetics assessments of GUIs has focused on large screens (Dou et al., 2019; Khani et al., 2016; Xing et al., 2021), with only one exception: assessing mobile GUIs (Lima, Martins et al., 2022). Khani et al. (2016) used a pre-trained convolutional neural network (CNN) with classical machine learning techniques. They trained a model based on AlexNet that uses a support vector machine to classify web GUIs as having either "good" or "bad" visual aesthetic quality. After training with a dataset of 418 labeled website screenshots, they reported an error rate (root mean squared error) of 34.15%. Although Dou et al. (2019) also used a pre-trained CNN (CaffeNet) to train their model for assessing web GUIs, the main difference from Khani et al. is that they adopted a regression approach. Using a dataset of 398 website screenshots labeled with the mean score of all ratings received on a 9-point scale, they reported an error rate of 20.41%. Xing et al. (2021) trained a Squeeze-and-Excitation VGG model using 38,423 GUI designs labeled with the number of "likes" received on a popular website and the number of user "collections" to which they belong, as their visual aesthetics representations. They reported an error rate of 14.89% predicting "collections" and 25.38% predicting "likes", but it is unclear how they reflect visual aesthetics based on these predictions.

Despite the growing relevance of mobile GUIs, little research has addressed their visual aesthetics assessment (Lima, Gresse von Wangenheim et al., 2022). In addition to assessments based on handcrafted feature extraction (Alemerien & Magel, 2014; Miniukovich & De Angeli, 2014a, 2014b, 2015; Taba et al., 2014), one study used a pre-trained CNN to assess Android GUIs developed with App Inventor (Lima, Martins et al., 2022). They trained a model based on the ResNet-50 architecture using 481 mobile GUIs labeled according to their visual aesthetics on a 5-point scale. Using a transfer learning approach to speed up the process, an error rate of 14.85% was reported on the validation set. However, when evaluating the performance over a test set containing 20 new GUIs not used in the training, the error rate decreased to 12,88%. They justify the improvement in the model performance because the test set is composed of images with high agreement among the human raters, indicating that they present very characteristic features of each degree of visual aesthetics. To the best of our knowledge, a CNN model for predicting the visual aesthetic distribution of a GUI has not yet been presented.

Considering approaches in a similar area, for the assessment of photographs, most works use the AVA dataset containing more than 250,000 images with around 200 visual aesthetics ratings between 1 and 10 (Murray et al., 2012). As this dataset has an unbalanced distribution of scores, Jin et al. (2016) assigned weights to the images in inverse proportion to their mean score frequency. Thus, images with less frequent scores had higher weights during the training phase. They trained a VGG-16-based model by adopting a weighted chi-squared distance as the loss function. However, their model does not output a score distribution but uses it to compute the mean score and standard deviation. Other studies have presented models that predict the score distribution of photographs using different loss functions: Huber loss (Murray & Gordo, 2017), squared earth mover's distance (Talebi & Milanfar, 2018), and the cumulative Jensen-Shannon Divergence (Jin et al., 2018).

That illustrates the lack of research on machine learning to assess the visual aesthetics of mobile GUIs, particularly concerning the prediction of rating distributions that can represent the degree of agreement among users.

3. METHODOLOGY

To automate the visual aesthetics assessments of App Inventor GUIs, we developed a deep learning model following the machine learning process proposed by Ashmore et al. (2021) and Amershi et al. (2019).

Requirements analysis. Based on related work on the different types of GUIs identified through a literature review, we defined the main objective of the model and specified its target features, following Mitchell (1997). This step also entails a detailed characterization of the input and expected outputs, clearly defining the problem.

Data management. This step includes selecting available generic datasets for model pre-training and collecting GUI screenshots to build the domain-specific dataset. After the screenshot collection, we cleaned the dataset by removing duplicates. By adopting a supervised learning technique, we labeled each screenshot with a 5-dimensional vector representing the distribution of visual aesthetic ratings received from a group of volunteers. We then preprocessed the dataset and resized the collected screenshots to fit the deep learning architecture before training. It was then split into a training set to train the model and a validation set to perform an unbiased performance evaluation of the chosen model on unseen data. We also reserved 20 screenshots for testing, following the recommendations of Ripley (2007).

Training and performance evaluation. We adopted a supervised transfer learning approach using a proven deep learning framework, starting from a pre-trained network and retraining only the input and output layers with our dataset. Once convergence was reached, all layers were unfrozen for fine-tuning with domain-specific data, dynamically optimizing hyperparameters such as momentum and learning rates. Multiple variants were trained with different hyperparameters for comparison. Model performance was evaluated using a validation set and a task-appropriate metric, with final testing conducted on previously unseen data to ensure robustness.

Model evaluation. To determine the degree to which the model predictions were equivalent to the human assessments, we performed a correlation analysis between the values resulting from the model and those attributed by the human evaluators. Additionally, correlation analysis enables the comparison of this model's results with those of other studies employing the same evaluation method. We also analyzed the Bland-Altman agreement to measure the degree of agreement between the two assessment methods (by the model and by humans) of the visual aesthetics of GUIs.

4. APPSTHETICS - HISTOGRAM: A DEEP LEARNING MODEL FOR ASSESSING THE VISUAL AESTHETICS OF MOBILE APPS

4.1 Dataset

Our dataset is adapted from a similar work developing a model for directly assessing mobile GUI visual aesthetics (Lima, Wangenheim, et al., 2024). The dataset included 820 App Inventor screenshots saved as PNG images with a resolution of 1080 x 1920 pixels. Instead of using scalars as labels representing the visual aesthetics degree of each GUI, we used score distributions representing the frequency of individual ratings received. Thus, the GUIs were labeled with a 5-dimensional vector where the i-th element was the number of ratings i received. Ten human raters assessed the visual aesthetics of each GUI on a 5-point semantic differential scale, ranging from 1 = "very ugly" to 5 = "very beautiful." In this way, a GUI with a label {1, 7, 2, 0, 0} has received one rating "1," seven ratings "2," two ratings "3," and no ratings "4" or "5." The scores were then normalized to the interval [0..1], with "0" = "no ratings received" and "1" = "all ratings received."

For comparison, we used the same test set with 20 GUIs as in the previous work. The remaining 800 GUIs were randomly split into a training set of 640 (80%) and a validation set of 160 (20%). Screenshots were downsampled from 1080 × 1920 to 448 × 448 pixels. To avoid distorting image features relevant to visual aesthetic perception, we performed no further transformations, such as cropping. The dataset is available online at https://bit.ly/app-inventor-dataset-v2.

4.2 Training

We used models pre-trained with ImageNet, one of the largest publicly available general-purpose datasets (Russakovsky et al., 2015). We used fast.ai, a robust, adaptable, and research-focused CNN framework, to build our model (Howard & Gugger, 2020). We trained three network architectures to compare and select the one with the highest performance.

VGG19. This architecture has achieved high accuracy with large-scale image recognition (Simonyan & Zisserman, 2015) and significant results with the visual aesthetics assessments of photographs (Lin, 2022; Sakaguchi et al., 2022). It uses small 3x3 convolution filters, which are the smallest possible size that still captures up/down and left/right. All hidden layers use ReLU as the activation function.

ResNet50. Residual network architectures (ResNets) employ identity connections as shortcuts to bypass several layers (He et al., 2016). ResNets provide two parallel learning paths in several network sections and avoid the typical gradient loss in very-deep networks. Although this architecture allows for much deeper networks with up to 152 layers, a network with 50 layers (ResNet50) offers the best performance compared to networks of different sizes (Lima, Martins et al., 2022).

EfficientNet B0. EfficientNet uses a fixed set of coefficients to scale up the width, depth, and image resolution uniformly (Tan & Le, 2020) and overcome the difficulty of randomly scaling up each of those dimensions by trial and error. The result is performance improvement with less use of computational resources. We selected EfficientNet B0, which has shown a performance similar to ResNet50 (Tan & Le, 2020).

The input layer was adapted to the image resolution of our dataset, which was represented by the screenshot vector and its label. The original output layers representing a categorical variable with 1,000 values used to classify the ImageNet categories (containing 1,000 neurons) were replaced by a layer with five neurons. In this way, the model output is a 5-dimensional vector, where each element corresponds to a different score on the 5-point scale. The output scores range within [0..1] and are interpreted as the percentage of ratings received, with "0" = "no ratings received (0%)" and "1" = "all ratings received (100%)."

The mean squared error (MSE) is used as the loss function. It is the average squared difference between the data labels and deep-learning model outputs. MSE values are always non-negative, and the lower the value, the better, meaning that the predictions are closer to the labels. As a quadratic function, it heavily penalizes outliers, which are common in visual aesthetic assessments.

The output layers of the networks were transfer-trained for 100 epochs, which previous work has shown to be sufficient for the validation error to stop improving (Lima, Martins et al., 2022). Each architecture was trained using two different training strategies, standard training (fit) and automated hyperparameter optimization (fit1cycle) (Smith, 2018; Smith & Topin, 2019), resulting in six models for performance comparison maintaining all default parameters.

Table 1. Best MSE for each model

Architecture	Prediction	Transfer learning		Fine-tuning	
		Strategy	MSE	LR	MSE
ResNet50	regression	*fit*	.026767	1e-04, 1e-03	**.022049**
		fi1cycle	.029282	1e-05, 1e-04	.027709
VGG19	histogram	*fit*	.02486	6.31e-07, 6.92e-05	.023841
		fi1cycle	.025081	6.31e-07, 5.75e-05	**.022129**
ResNet50	histogram	*fit*	.024029	1e-04, 1e-03	**.02095**
		fi1cycle	.025765	1e-05, 1e-04	.021359
EfficientNet B0	histogram	*fit*	.023992	1e-05, 1e-04	.023362
		fi1cycle	.024021	2e-04, 2e-05	**.022814**

After unfreezing the intermediate layers to allow all weights to adapt to our dataset, we trained each model for 20 additional epochs using the same training strategy. Learning rates were defined with the method suggested by Smith and Topin (2019), which resulted in a different range of rates for each model. All the models improved their performance after the fine-tuning phase. ResNet50 trained with the fit strategy achieved the best performance, even better than the ResNet50 model, in predicting the visual aesthetics score directly (Table 1). Therefore, we present detailed results for this sole model.

These results demonstrate that the model performs well in predicting the visual aesthetic distributions. The medians from almost half of the predicted distributions (79 of 160) were the same as those computed from the labels. For the other 74 GUIs, the medians from the model outputs and labels differed by one point or less on a 5-point scale. This means that most of the time, the model can classify a "beautiful" GUI somewhere between "very beautiful" and "neither beautiful nor ugly" rather than "ugly", for example. On the other hand, the medians for the six distributions (3.2%) differed by one and a half points or more from their labels. The worst case was a GUI that was considered very ugly by most of the human raters, with seven ratings of "1", but was assigned a median "4" from the predicted distribution, which is interpreted as "beautiful." However, observing such divergences in only a very small number of cases can be considered satisfactory because even humans strongly disagree about visual aesthetics (Gresse von Wangenheim et al., 2018).

In general, the average absolute deviation (AAD) was higher on the predicted distributions than on the labels. This indicates the model's tendency to assign scores higher than zero. However, the AAD was lower for some predicted distributions (24%) than for the label distributions. For the worst predictions, the AAD was high, expressing the difficulty in assessing those GUIs.

A similar performance was observed with the test set, which contains new screenshots without labels. The median of the predicted distribution of one GUI was two points lower than its label. All other medians differed by one point or less, with half of the predictions resulting in identical medians as the labels. The worst predictions in the test set also resulted in high AADs, suggesting that it can be an indicator of the difficulty in assessing visual aesthetics.

5. MODEL EVALUATION

Studies using deep learning models typically evaluate their performance in predicting the sample labels in the validation set, using metrics that compute how close their predictions are to those labels. Nonetheless, it is also important to know how the model performs when assessing images unseen before. Therefore, we evaluate our model using only the test set with 20 GUIs separated from the dataset before training.

Other methods to compare the predictions from the model with the human ratings include the correlation analysis and the Bland-Altman plot analysis (B&A) (Bland & Altman, 1986). Although some studies for visual aesthetics assessments with deep learning adopt correlation analysis (Dou et al., 2019; Lima, Martins et al., 2022; Xing et al., 2021), results can be misleading because they only quantify their linear association, not how much they agree (Giavarina, 2015). Thus, to determine the degree of agreement between the model results and the human ratings, we also conduct an analysis using B&A. To the best of our knowledge, no model predicting the visual aesthetics distribution of an image (GUI or photograph) evaluates their work using correlation or B&A.

5.1 Correlation Analysis

We used the Spearman rank correlation (rho) to examine the linear association between the model results and the human assessments. The non-normality of the data justifies the selection of Spearman's rank correlation coefficient over the more often used Pearson's r correlation test (Bryman & Cramer, 1990). It is represented by a numerical score between -1 and +1, and the strength of the linear association increases as the coefficients get closer to these scores.

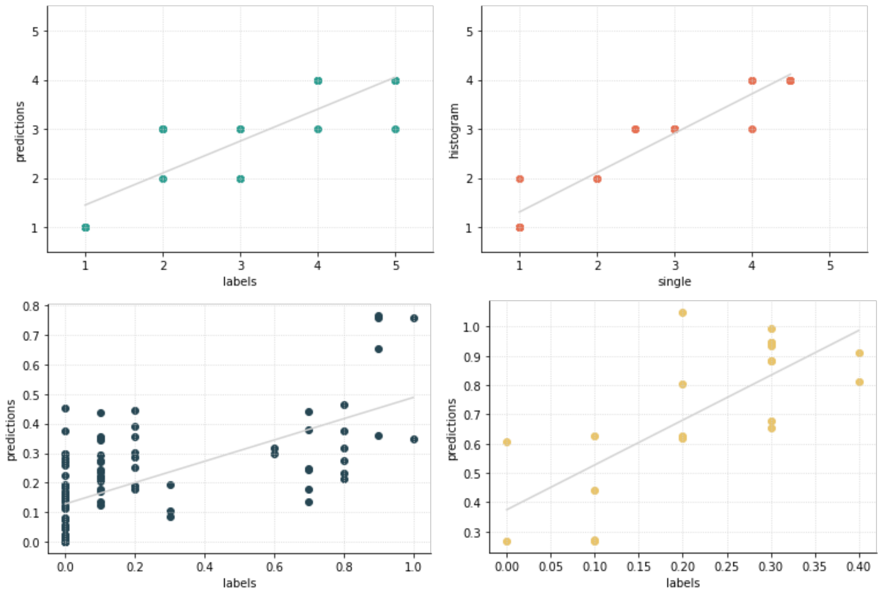

Figure 2. Correlation analysis: label distribution vs predicted distribution (top left);
label median vs predicted median (top right); direct prediction vs predicted median (bottom left);
label AAD vs predicted AAD (bottom right)

The correlation between the individual scores of the predicted distribution with those of the labels resulted in a rho = .65, which is moderate (Figure 2, top left). When analyzing the results, it is possible to notice that the model tends to distribute the scores along the scale. For example, for two of the GUIs, there was consensus among the human raters about their visual aesthetics (the right-most dots on the plot), but the model did not assign any rating with a score near 1 (Figure 3). The highest scores assigned by the model were .74 and .71.

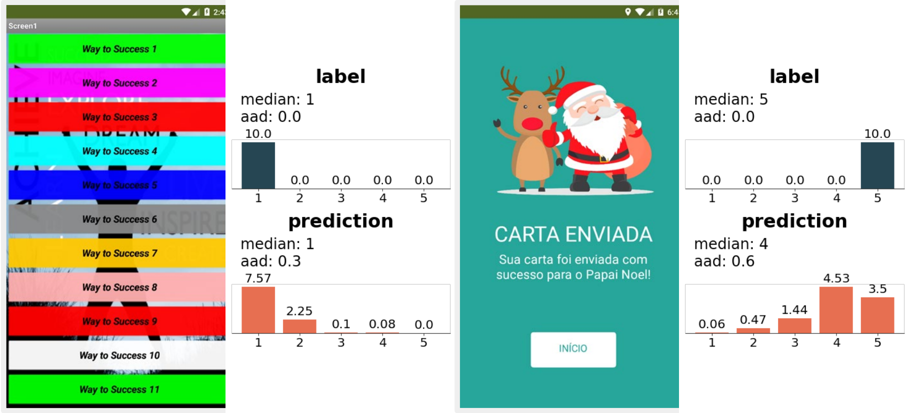

Figure 3. GUIs that generated consensus among raters

We also compared the medians of predictions with the medians of labels (Figure 2, top right). The median is preferred over the mean score as a measure of central tendency for ordinal scales (Nunnally & Bernstein, 1994). It is also less sensitive to outliers, which are very common in subjective tasks such as visual aesthetics assessments. For that reason, the correlation was much stronger, with rho=.85. In this case, the median of the predictions was at most one point from the median of the labels. For example, the predictions were all correct for the GUIs with a visual aesthetics degree of 1 = "very ugly". On the other hand, the predictions were all wrong for the GUIs with visual aesthetics of 5 = "very beautiful," although the model was consistent here, as it always predicted a median of 4. This again is an indication that the model presents a bias towards the lowest rates. This result is very similar to the model predicting the visual aesthetics directly (Lima, Gresse von Wangenheim et al., 2024) (Figure 2, bottom left). When comparing the median from our results with the visual aesthetics degree directly predicted, the correlation was rho = .97.

Comparing the AADs of labels and predictions (Figure 2, bottom right), only a moderate correlation (rho = .67) was obtained. Although this set contains GUIs with low AADs (.4 or lower), reflecting a high level of agreement between the human raters, the predicted distributions had high AADs (> .27). As an example, one label had an AAD = .2 but a predicted distribution above 1. This shows the model's tendency to spread the distribution along the 5-point scale instead of concentrating it around one single rating.

5.2 Bland-Altman Analysis

We used Bland-Altman (B&A) plot analysis to assess whether two variables could be compared, as the correlation analysis only reveals their linear association, not their differences (Giavarina, 2015). The B&A plot analysis exhibits the mean difference between two quantitative measurements and builds limits to characterize their agreement. It enables us to identify bias in the mean differences and calculate an agreement interval in which 95% of the differences between the measurements lie (Bland & Altman, 1986). The B&A results are not an indication of whether the automated assessment can adequately replace the human one or whether the agreement between the model predictions and the labels is sufficient. It merely measures the bias and a range of agreement that includes 95% of the variations between the two methods (Giavarina, 2015). The B&A recommends that 95% of the data points lie within ±2 sd of the mean difference.

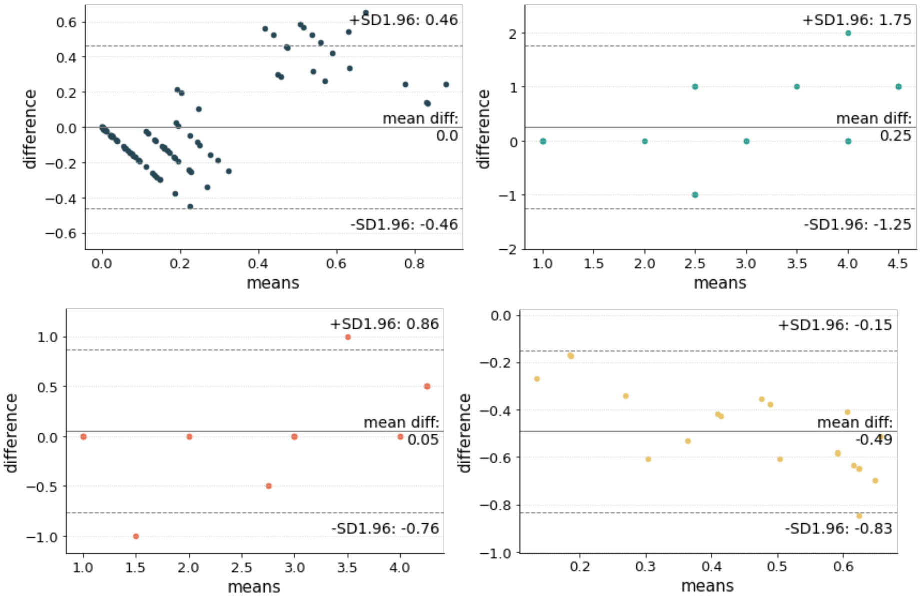

Figure 4. B&A plot analysis: label distribution vs predicted distribution (top left);
label median vs predicted median (top right); direct prediction vs predicted median (bottom left);
label AAD vs predicted AAD (bottom right)

Analyzing the B&A plot for individual scores of the predicted distribution and those of the labels (Figure 4, top left), it is possible to observe that the mean difference between them is zero. This is an indication of the absence of bias in predicting the score distribution for the whole set. Nonetheless, the confidence interval (CI) is too large (-.46 to .46), meaning that 95% of the predictions can be between half of the labels up to two times greater. It is also possible to observe that the predictions tended to be higher for the low scores and lower for the higher scores, suggesting a bias towards the middle of the scale. This may happen because many more ratings received few or no votes than ratings that received all or almost all of the votes.

The B&A plot analysis for the medians of the predicted distributions and the medians of the labels shows a different scenario (Figure 4, top right). Although the mean difference between them was .25, it is possible to see the three dots above the line pulling it up. The plot also clearly shows agreement between them because all their differences lie within the CI (-1,25 to 1,75) because the medians of the predicted distribution were never higher than one point different from the media of the labels. The analysis between the medians of the predicted distributions and the visual aesthetics degrees directly predicted displays even greater agreement (Figure 4, bottom left). The mean difference between them was half a point on a 5-point scale, and the CI was (-.76 to .86). For only two GUIs, the median for the predicted distribution was one point higher than directly predicted.

The AAD is biased on the B&A plot (Figure 5, bottom right). All differences are below zero, and the mean difference is -.49, confirming that the model outputs tend to have AADs higher than the labels.

6. DISCUSSION

Part of the challenge of assessing visual aesthetics lies in the subjective response of this type of judgment. Because people are different, even when they share a similar background, they tend to disagree about visual aesthetics (Gresse von Wangenheim et al., 2018). It is not uncommon to find people offering opposed aesthetic judgments ("beautiful" vs. "ugly") for the same objects, not to mention judgments with varying degrees ("beautiful" vs. "very beautiful"). Therefore, a single score to represent visual aesthetics hides how much disagreement a GUI can provoke before it is computed from all the individual assessments about it. On the other hand, using a distribution of ratings received allows not only to compute that same single score but also some other numerical representation for the degree of disagreement, like AAD. It can also be presented in the form of a histogram to give a visual overview of all individual assessments. This article presents the first model to assess the visual aesthetics of GUIs (desktop, web, or mobile) that outputs the distribution of individual visual aesthetic ratings for mobile GUIs created with App Inventor. From the rating distribution of each GUI, we compute the median as its visual aesthetics degree and the AAD as its disagreement degree.

Our model obtained excellent performance, with an MSE = .0209, which exceeds other models for the direct assessment of visual aesthetics of web GUIs (MSE = .042) (Dou et al., 2019) or GUI designs (MSE = .0222) (Xing et al., 2021). It also resulted in a better performance than directly assessing mobile GUIs over the same dataset (MSE = .022), with the difference that we kept those samples with high AAD.

On the other hand, the correlation between the predicted distributions and the labels is not very strong (rho = .65). This may be because the dataset is unbalanced regarding score distributions, with many more ratings receiving low scores than high ones. It also seems unavoidable when the dataset is balanced regarding visual aesthetics scores, which are the distribution medians in this case, because scores on the extremes of the scale (1 = "very ugly" or 5 = "very beautiful") need to have strongly biased distributions to the left or the right. Nonetheless, that lower correlation between the predicted distributions and the labels does not lead to a low performance in assigning the visual aesthetics degree of mobile GUIs. The correlation between the predicted medians and the label medians is below that of the model that directly assesses visual aesthetics (Lima, Gresse von Wangenheim et al., 2024). Nonetheless, the results from both models show a near-perfect correlation over the same test set (rho = .97).

Despite the good performance, the model did not get the visual aesthetics scores right for any of the four "very beautiful" GUIs in the test set. That might happen because there are very few "very beautiful" GUIs in the dataset. Creating a more balanced dataset has been complicated because a large majority of the apps available in the App Inventor Gallery have rather ugly interface designs, making it difficult to encounter beautiful designs in larger quantities. Moreover, as the final visual aesthetics score of these GUIs is the median of all individual ratings received, a GUI with a "1" or a "5" needs to have more than half of the human raters assigning those scores. Regardless of this bias, the B&A plot analysis reveals that the difference is at most one point on the five-point scale for 95% of screenshots. That means that no GUI labeled as 4 ("beautiful") received a 2 ("ugly") from the model or vice versa. The only exception was a "very beautiful" GUI that received a visual aesthetics score of 3 ("not ugly nor beautiful").

Comparing the predicted distributions with the label distributions, we can observe that our model tends to increase low scores and reduce high ones. This leads to distributions with higher AADs than labels. One possible reason for that bias may be the unbalanced dataset concerning individual ratings and, as a consequence, concerning AADs. Although it was primarily built to be as balanced as possible in terms of visual aesthetics degree (distribution median), this is not reflected in the individual ratings, as each median score may be the result of different distributions, and consequently different AADs. Maybe a larger dataset, with a balanced number of individual ratings, can contribute to mitigating the bias of the distribution and AAD prediction.

Threats to validity. A potential threat to our study relates to using a dataset that does not represent the full spectrum of possible outcomes. To minimize this threat, we tried to balance the dataset concerning the aesthetic ratings. Nonetheless, a complete balance was not achieved due to the small number of App Inventor apps with more beautiful interfaces. A further threat concerns labeling many GUIs at once, which can be affected by tired raters. For that reason, we instructed raters to interrupt labeling whenever they felt fatigued to mitigate this threat. For evaluation, we selected appropriate methods following related work and theory to evaluate correlation and agreement. Additionally, based on related work, we selected well-tested CNN architectures that have been utilized for similar tasks. Concerning external validity, we used a considerable sample size for evaluation, with a large variety of application types that allow the generalization of the results. The performance of the deep learning model was analyzed separately over a test set containing GUIs not previously used for training or validation. We chose the same GUIs as in Lima, Gresse von Wangenheim et al. (2024), which were randomly chosen from the original dataset to enable comparison.

7. CONCLUSION

This article presents an original model for the automatic assessment of the distribution of visual aesthetics in mobile GUIs of Android applications developed with App Inventor. The model predicts a distribution of ratings on a 5-point scale, from which statistical measures can be derived. Specifically, the median is interpreted as the degree of visual aesthetics, while the average absolute deviation (AAD) captures the level of disagreement among individual raters. The model was trained on a dataset of 800 GUIs labeled with 5-dimensional vectors representing the distribution of human ratings. We trained and compared three different CNN architectures, and the ResNet50 model achieved the best performance when predicting the rating distribution (MSE = .021), surpassing even a model trained to directly predict single-valued aesthetics. Evaluation on a test set of unseen GUIs indicates that the model supports visual aesthetics assessments with high correlation and agreement with human perception, although with a slight bias toward lower aesthetics scores. This approach enables effective and efficient assessment of visual aesthetics during the design of mobile application GUIs and can also provide valuable feedback to students in computing education. Future work will focus on improving the model's ability to estimate the distribution of AADs, thereby better capturing the degree of disagreement among raters. We also plan to expand the dataset and deploy the model as an online automatic assessment tool.

ACKNOWLEDGEMENT

This study was financed in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq) - Grant No.: 303674/2019-9.

REFERENCES

Alemerien, K. & Magel, K., 2014. GUIEvaluator: A Metric-tool for Evaluating the Complexity of Graphical User Interfaces. *Proceedings of the Twenty-Sixth International Conference on Software Engineering & Knowledge Engineering*, pp. 13-18.

Altaboli, A. & Lin, Y., 2011. Objective and Subjective Measures of Visual Aesthetics of Website Interface Design: The Two Sides of the Coin. In J. A. Jacko (ed.) *Human-Computer Interaction. Design and Development Approaches.* Springer, pp. 35-44. https://doi.org/10.1007/978-3-642-21602-2_4

Amershi, S. et al., 2019. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pp. 291-300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Ashmore, R., Calinescu, R. and Paterson, C., 2021. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ACM Computing Surveys*, Vol. 54, No. 5, pp. 111:1-111:39. https://doi.org/10.1145/3453444

Bao, X., Shi, P., Pan, D. and You, J., 2019. Image Aesthetic Prediction Model Based on Deep Convolutional Neural Network. In *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp. 122-125. https://doi.org/10.1109/ICISCAE48440.2019.221601

Bhandari, U., Chua, W. Y., Neben, T. and Chang, K. (2016). Cognitive Load and Attention for Mobile Applications: A Design Perspective. In M. Kurosu (ed.) *Human-Computer Interaction. Interaction Platforms and Techniques.* Springer International Publishing, pp. 278-284. https://doi.org/10.1007/978-3-319-39516-6_26

Bland, J. M. and Altman, D. G., 1986. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet*, Vol. 327, No. 8476, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bryman, A. and Cramer, D., 1990. *Quantitative Data Analysis for Social Scientists.* Taylor & Francis/Routledge, pp. xiv, 290.

Deng, Y., Loy, C. C. and Tang, X., 2017. Image Aesthetic Assessment: An experimental survey. *IEEE Signal Processing Magazine*, Vol. 34, No. 4, pp. 80-106. https://doi.org/10.1109/MSP.2017.2696576

Dou, Q., Zheng, X. S., Sun, T. and Heng, P.-A., 2019. Webthetics: Quantifying Webpage Aesthetics with Deep Learning. *International Journal of Human-Computer Studies*, Vol. 124, pp. 56-66. https://doi.org/10.1016/j.ijhcs.2018.11.006

Giavarina, D., 2015. Understanding Bland Altman Analysis. *Biochemia Medica*, Vol. 25, No. 2, pp. 141-151. https://doi.org/10.11613/BM.2015.015

Gresse von Wangenheim, C., Porto, J. V. A., Hauck, J. C. R. and Borgatto, A. F., 2018. Do We Agree on User Interface Aesthetics of Android Apps? *arXiv*, pp. 1-5.

Guo, F. et al., 2020. How User's First Impression Forms on Mobile user Interface?: An ERPs Study. *International Journal of Human–Computer Interaction*, Vol. 36, No. 9, pp. 870-880. https://doi.org/10.1080/10447318.2019.1699745

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Howard, J. and Gugger, S., 2020. Fastai: A Layered API for Deep Learning. *Information*, Vol. 11, No. 2, Article 2. https://doi.org/10.3390/info11020108

Jin, B., Segovia, M. V. O. and Süsstrunk, S., 2016. Image aesthetic predictors based on weighted CNNs. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2291-2295. https://doi.org/10.1109/ICIP.2016.7532767

Jin, X. et al., 2018. Predicting Aesthetic Score Distribution Through Cumulative Jensen-Shannon Divergence. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, Article 1. https://doi.org/10.1609/aaai.v32i1.11286

Kang, C., Valenzise, G. and Dufaux, F., 2019. Predicting Subjectivity in Image Aesthetics Assessment. *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1-6. https://doi.org/10.1109/MMSP.2019.8901716

Khani, M. G., Mazinani, M. R., Fayyaz, M. and Hoseini, M., 2016. A Novel Approach for Website Aesthetic Evaluation Based on Convolutional Neural Networks. *Proceedings of the 2016 Second International Conference on Web Research (ICWR)*, pp. 48-53. https://doi.org/10.1109/ICWR.2016.7498445

Lima, A. L. de S. and Gresse von Wangenheim, C., 2021. Assessing the Visual Esthetics of User Interfaces: A Ten-Year Systematic Mapping. *International Journal of Human–Computer Interaction*, pp. 1-21. https://doi.org/10.1080/10447318.2021.1926118

Lima, A. L. de S., Gresse von Wangenheim, C. and Borgatto, A. F., 2022. Assessment of Visual Aesthetics through Human Judgments: A Systematic Mapping. *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, pp. 1-14. https://doi.org/10.1145/3554364.3560902

Lima, A. L. de S., Gresse von Wangenheim, C. et al., 2024. A Deep Learning Model for the Assessment of the Visual Aesthetics of Mobile User Interfaces. *Journal of the Brazilian Computer Society*, Vol. 30, pp. 102-115. https://doi.org/10.5753/jbcs.2024.3255

Lima, A. L. de S., Martins, O. P. H. R. et al., 2022. Automated Assessment of Visual aesthetics of Android User Interfaces with Deep Learning. *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, pp. 1-11. https://doi.org/10.1145/3554364.3559113

Lima, A. L. de S., Wangenheim, C. G. von, Martins, O. P. H. R. et al., 2024. A Deep Learning Model for the Assessment of the Visual Aesthetics of Mobile User Interfaces. *Journal of the Brazilian Computer Society*, Vol. 30, No. 1, Article 1. https://doi.org/10.5753/jbcs.2024.3255

Lin, R., 2022. Augmenting Image Aesthetic Assessment with Diverse Deep Features. In *2021 4th Artificial Intelligence and Cloud Computing Conference*, pp. 30-38. https://doi.org/10.1145/3508259.3508264

Lu, X., Lin, Z., Jin, H., Yang, J. and Wang, J. Z., 2014. RAPID: Rating Pictorial Aesthetics Using Deep Learning. *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 457-466. https://doi.org/10.1145/2647868.2654927

Malu, G., Bapi, R. S. and Indurkhya, B., 2017. Learning Photography Aesthetics with Deep CNNs. *arXiv*:1707.03981 [Cs]. http://arxiv.org/abs/1707.03981

Miniukovich, A. and De Angeli, A., 2014a. Quantification of Interface Visual Complexity. *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pp. 153-160. https://doi.org/10.1145/2598153.2598173

Miniukovich, A. and De Angeli, A., 2014b. Visual Impressions of Mobile App Interfaces. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, pp. 31-40. https://doi.org/10.1145/2639189.2641219

Miniukovich, A. and De Angeli, A., 2015. Computation of Interface Aesthetics. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp. 1163-1172. https://doi.org/10.1145/2702123.2702575

Mitchell, T. M., 1997. *Machine Learning* (1st ed.). McGraw-Hill Education.

Moshagen, M. and Thielsch, M. T., 2010. Facets of Visual Aesthetics. *International Journal of Human-Computer Studies*, Vol. 68, No. 10, pp. 689-709. https://doi.org/10.1016/j.ijhcs.2010.05.006

Murray, N. and Gordo, A., 2017. A deep architecture for unified aesthetic prediction. *arXiv*. (No. arXiv:1708.04890). https://doi.org/10.48550/arXiv.1708.04890

Murray, N., Marchesotti, L. and Perronnin, F., 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408-2415. https://doi.org/10.1109/CVPR.2012.6247954

Nunnally, J. C. and Bernstein, I. H., 1994. *Psychometric Theory* (3rd ed.). McGraw-Hill.

Oyibo, K., Adaji, I., Orji, R. and Vassileva, J., 2018. What Drives the Perceived Credibility of Mobile Websites: Classical or Expressive Aesthetics? In M. Kurosu (ed.) *Human-Computer Interaction. Interaction in Context*. Springer International Publishing, pp. 576-594. https://doi.org/10.1007/978-3-319-91244-8_45

Ripley, B. D., 2007. *Pattern Recognition and Neural Networks*. Cambridge University Press.

Robins, D. and Holmes, J., 2008. Aesthetics and credibility in web site design. *Information Processing and Management: An International Journal*, Vol. 44, No. 1, pp. 386-399. https://doi.org/10.1016/j.ipm.2007.02.003

Russakovsky, O. et al., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211-252. https://doi.org/10.1007/s11263-015-0816-y

Sakaguchi, D., Takimoto, H. and Kanagawa, A., 2022. Study on relationship between composition and prediction of photo aesthetics using CNN. *Cogent Engineering*, Vol. 9, No. 1, 2107472. https://doi.org/10.1080/23311916.2022.2107472

Simonyan, K. and Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* (No. arXiv:1409.1556). https://doi.org/10.48550/arXiv.1409.1556

Smith, L. N., 2018. A Disciplined Approach to Neural Network Hyper-parameters: Part 1 -- Learning Rate, Batch Size, Momentum, and Weight Decay. *arXiv*:1803.09820 [Cs, Stat]. http://arxiv.org/abs/1803.09820

Smith, L. N. and Topin, N., 2019. Super-convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, Vol. 11006, 1100612. https://doi.org/10.1117/12.2520589

Soui, M., Chouchane, M., Mkaouer, M. W., Kessentini, M. and Ghedira, K., 2020. Assessing the Quality of Mobile Graphical User Interfaces Using Multi-Objective Optimization. *Soft Computing*, Vol. 24, No. 10, pp. 7685-7714. https://doi.org/10.1007/s00500-019-04391-8

Taba, S. E. S., Keivanloo, I., Zou, Y., Ng, J. and Ng, T., 2014. An Exploratory Study on the Relation between User Interface Complexity and the Perceived Quality. In S. Casteleyn, G. Rossi and M. Winckler (eds) *Web Engineering*. Springer International Publishing, pp. 370-379. https://doi.org/10.1007/978-3-319-08245-5_22

Takimoto, H., Omori, F. and Kanagawa, A., 2021. Image Aesthetics Assessment Based on Multi-stream CNN Architecture and Saliency Features. *Applied Artificial Intelligence*, Vol. 35, No. 1, pp. 25-40. https://doi.org/10.1080/08839514.2020.1839197

Talebi, H. and Milanfar, P., 2018. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, Vol. 27, No. 8, pp. 3998-4011. https://doi.org/10.1109/TIP.2018.2831899

Tan, M. and Le, Q. V., 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv,* (No. arXiv:1905.11946). https://doi.org/10.48550/arXiv.1905.11946

Thielsch, M. T., Scharfen, J., Masoudi, E. and Reuter, M., 2019. Visual Aesthetics and Performance: A First Meta-Analysis. *Proceedings of Mensch Und Computer 2019*, vol. 199-210. https://doi.org/10.1145/3340764.3340794

Tractinsky, N., Cokhavi, A., Kirschenbaum, M. and Sharfi, T., 2006. Evaluating the Consistency of Immediate Aesthetic Perceptions of Web Pages. *International Journal of Human-Computer Studies*, Vol. 64, No. 11, pp. 1071-1083. https://doi.org/10.1016/j.ijhcs.2006.06.009

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K. and Bargas-Avila, J. A., 2012. Is Beautiful Really Usable? Toward Understanding the Relation Between Usability, Aesthetics, and Affect in HCI. *Computers in Human Behavior*, Vol. 28, No. 5, pp. 1596-1607. https://doi.org/10.1016/j.chb.2012.03.024

Xing, B., Si, H., Chen, J., Ye, M. and Shi, L., 2021. Computational model for predicting user aesthetic preference for GUI using DCNNs. *CCF Transactions on Pervasive Computing and Interaction*, Vol. 3, No. 2, pp. 147-169. https://doi.org/10.1007/s42486-021-00064-4

Zen, M. and Vanderdonckt, J., 2016. Assessing User Interface Aesthetics based on the Inter-subjectivity of Judgment. *Proceedings of the 30th International BCS Human Computer Interaction Conference. 30th International BCS Human Computer Interaction Conference*, Poole, UK. https://doi.org/10.14236/ewic/HCI2016.25